



## Article

# Statistical Modeling of RPCA-FCM in Spatiotemporal Rainfall Patterns Recognition

Siti Mariana Che Mat Nor <sup>1</sup>, Shazlyn Milleana Shaharudin <sup>1,\*</sup> , Shuhaida Ismail <sup>2</sup>, Sumayyah Aimi Mohd Najib <sup>3</sup>, Mou Leong Tan <sup>4</sup>  and Norhaiza Ahmad <sup>5</sup>

- <sup>1</sup> Department of Mathematics, Faculty of Science and Mathematics, Universiti Pendidikan Sultan Idris, Tanjong Malim 35900, Perak, Malaysia; M20181001458@siswa.upsi.edu.my
- <sup>2</sup> Department of Mathematics and Statistics, Faculty of Applied Sciences and Technology, Universiti Tun Hussein Onn Malaysia, Panchor 84600, Johor, Malaysia; shuhaida@uthm.edu.my
- <sup>3</sup> Department Geography and Environment, Faculty of Human Sciences, Universiti Pendidikan Sultan Idris, Tanjong Malim 35900, Perak, Malaysia; sumayyah@fsk.upsi.edu.my
- <sup>4</sup> Geoinformatic Unit, Geography Section, School of Humanities, Universiti Sains Malaysia, Gelugor 11800, Pulau Pinang, Malaysia; mouleong@usm.my
- <sup>5</sup> Department of Mathematical Sciences, Faculty of Science, Universiti Teknologi Malaysia, Skudai 81310, Johor, Malaysia; norhaiza@utm.my
- \* Correspondence: shazlyn@fsmt.upsi.edu.my



**Citation:** Che Mat Nor, S.M.; Shaharudin, S.M.; Ismail, S.; Mohd Najib, S.A.; Tan, M.L.; Ahmad, N. Statistical Modeling of RPCA-FCM in Spatiotemporal Rainfall Patterns Recognition. *Atmosphere* **2022**, *13*, 145. <https://doi.org/10.3390/atmos13010145>

Academic Editors: Rezzy Eko Caraka, Youngjo Lee, Toni Toharudin, Rung-Ching Chen, Heri Kuswanto and Maengseok Noh

Received: 6 December 2021

Accepted: 6 January 2022

Published: 16 January 2022

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



**Copyright:** © 2022 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/>).

**Abstract:** This study was conducted to identify the spatiotemporal torrential rainfall patterns of the East Coast of Peninsular Malaysia, as it is the region most affected by the torrential rainfall of the Northeast Monsoon season. Dimension reduction, such as the classical Principal Components Analysis (PCA) coupled with the clustering approach, is often applied to reduce the dimension of the data while simultaneously performing cluster partitions. However, the classical PCA is highly insensitive to outliers, as it assigns equal weights to each set of observations. Hence, applying the classical PCA could affect the cluster partitions of the rainfall patterns. Furthermore, traditional clustering algorithms only allow each element to exclusively belong to one cluster, thus observations within overlapping clusters of the torrential rainfall datasets might not be captured effectively. In this study, a statistical model of torrential rainfall pattern recognition was proposed to alleviate these issues. Here, a Robust PCA (RPCA) based on Tukey's biweight correlation was introduced and the optimum breakdown point to extract the number of components was identified. A breakdown point of 0.4 at 85% cumulative variance percentage efficiently extracted the number of components to avoid low-frequency variations or insignificant clusters on a spatial scale. Based on the extracted components, the rainfall patterns were further characterized based on cluster solutions attained using Fuzzy C-means clustering (FCM) to allow data elements to belong to more than one cluster, as the rainfall data structure permits this. Lastly, data generated using a Monte Carlo simulation were used to evaluate the performance of the proposed statistical modeling. It was found that the proposed RPCA-FCM performed better using RPCA-FCM compared to the classical PCA coupled with FCM in identifying the torrential rainfall patterns of Peninsular Malaysia's East Coast.

**Keywords:** principal component analysis; robust principal component analysis; rainfall patterns; Tukey's biweight correlation; spatiotemporal

## 1. Introduction

Rainfall is undoubtedly one of the most significant natural phenomena that plays a key role in the natural life and habitat of the earth. However, excessive rainfall can trigger natural disasters that put millions of lives at serious risk. In Malaysia, the rainy season, also known as the Northeast Monsoon season, runs from November to January each year and contributes 60% of the annual rainfall in East Malaysia. Along the east coastline of Malaysia, states such as Kelantan and Terengganu receive very heavy rainfall with over

3500 mm of rainfall annually. Each subsequent rainfall event over the rainy season increases the discharge of rainfall. Since Malaysia often encounters such issues, hydrologists have studied the rainfall trend in Malaysia with an emphasis on extreme rainfall events [1–3].

For instance, Malaysia experienced a massive flood in November of 2005. The flood hit various locations in Malaysia, including Kota Bharu, Kelantan. The disastrous event was described as the worst natural flood in history. Afterwards, in the years 2006, 2007, and 2008, heavy monsoonal rainfall again triggered major floods along the East Coast as well as in other parts of the country [4]. The hardest-hit areas were again along the East Coast of Peninsular Malaysia in the states of Kelantan, Terengganu, and Pahang.

In hydrological studies, these events are known as torrential rainfall. The term torrential rainfall has been used to describe extreme rainfall exceeding the threshold of 60 mm/day based on the classification of rainfall intensity by the Department of Irrigation and Drainage, Malaysia (DID) [5]. The East Coast region of Peninsular Malaysia often receives torrential rainfall during the Northeast Monsoon season. Due to this, several states located in the East Coast region face massive flooding every year. Climatologists or hydrologists could put this data to use to recommend additional measures to mitigate flood damage or to take precautions when flooding occurs [6].

Over the last few decades, there has been abundant research on spatial and temporal rainfall as well as its effect on the catchment areas [7]. The study of spatial and temporal rainfall has proved its importance in modeling and forecasting future rainfall patterns [8]. The study of spatial and temporal rainfall in hydrology is comprised of two approaches: regression-based modeling and cluster-based modeling. Previous studies [9–11] used regression-based modeling that generally aims to characterize the rainfall distribution patterns. With this approach, identifying spatial and temporal patterns of rainfall focuses on detecting trends rather than describing the regional characteristics of each pattern.

Nevertheless, studies related to cluster-based approaches in identifying spatial and temporal rainfall patterns aim to quantify the characteristics of a set of observations that fit into the same group, meaning that the patterns are similarly highly structured [12]. In addition, this approach is perceived as an efficient statistical tool in handling tasks such as grouping each region and identifying time periods based on the groups to reflect the occurrence of rainfall events.

Cluster analysis is often used to identify the rainfall patterns. Classical clustering methods, such as k-means clustering, require certain assumptions that contradict the characteristics of rainfall by dividing the database of rainfall patterns into clusters with the assumption that every weather pattern exclusively belongs to only one specific cluster. Using classical clustering, each day is respectively assigned to only one cluster [13]. These assumptions may not reflect reality, since one day could include several different types of weather at different stations.

Researchers have proven the superiority of using the Fuzzy C-means (FCM) as opposed to other clustering algorithms [14,15]. As a result, several clusters can belong to one data point when the Fuzzy C-means are used. The distance between the data point and the cluster center is used to assign the membership of each data point. It is more feasible to use a cluster located near a cluster center for particular sets of data. For instance, a study [16–18] proved that the FCM were very stable compared to other clustering algorithms, even when outliers and overlapped variables were present. The FCM algorithm incorporates noise, outliers, and non-uniform mass distributions to make the method applicable to data with unequal cluster sizes [19]. The FCM algorithm has proven its capability in handling datasets contaminated by noise and outliers. Previous researchers have applied this algorithm in various areas, including image processing [20–22].

Aside from traditional cluster algorithms, the usage of PCA has proven to be effective in identifying the number of clusters and improving cluster accuracy. Previous studies [23,24] have demonstrated the capability and applicability of PCA in identifying the number of clusters for rainfall patterns. However, the classical PCA can be insensitive

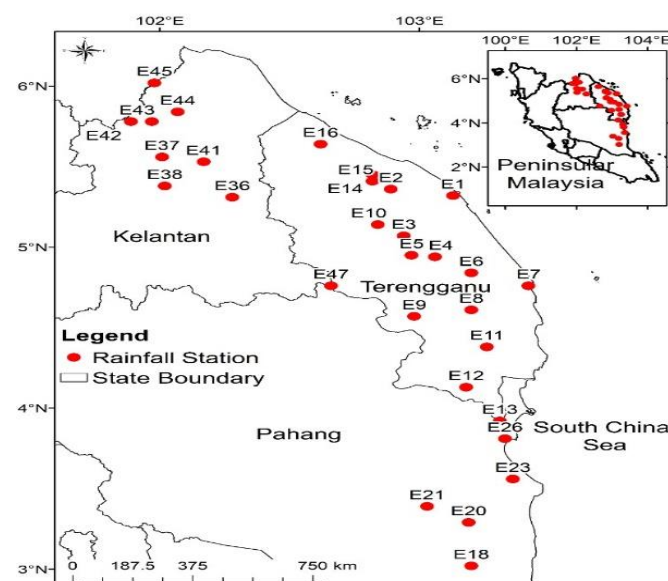
to outliers, as it assigns equal weights to each set of observations [25]. Hence, applying the classical PCA may affect overall cluster partitions of the rainfall patterns.

In clustering torrential rainfall patterns of the East Coast of Peninsular Malaysia, a long time series of observed rainfall data trends is needed. This is because the amount of rainfall for tropic regions does not vary significantly from day to day as compared to daily rainfall averages in four-season regions [12]. However, high-dimensional datasets tend to contain many irrelevant features that will affect the accuracy of the results [26]. Therefore, feature selection is considered to be an essential procedure in the processing of high-dimensional data [27]. There are few widely used dimensionality reduction methods that deal with spatial and temporal datasets. For instance, researchers [28] applied discriminant analysis (DA) in finding the relation between rainfall and landsides, while others [29] demonstrated the capability of non-linear principal component analysis in dealing with a climate dataset. On the other hand, many other researchers [30–32] have applied PCA in identifying the rainfall patterns for climate analysis.

The aims of this study are twofold: to propose statistical modeling using a combination of Robust PCA and the Fuzzy C-Means clustering method and to validate the outcome of the proposed model against the classical PCA in terms of the quality of internal and external clustering methods based on clustering validity indices. The proposed model is expected to improve the cluster partitions by using effective inputs from significant underlying patterns from the torrential dataset. It is anticipated to counter the shortcomings of the clustering approach for determining rainfall patterns on the East Coast of Peninsular Malaysia. Consequently, a robust measure of location and scale in the PCA was incorporated using Tukey's biweight correlation to downweight observations that were far from the data center and resistant to outlying observations due to their characteristics. The Fuzzy C-Means method exhibited the capability for overlapping data set in cluster partitions based on the membership level.

## 2. Study Area and Data Pre-Processing

The datasets used in this study were obtained from DID, Malaysia. This study focuses on daily rainfall data from 1 January 1987 until 31 December 2018 with a 32-year interval. In total, 48 rainfall stations (with 8.59% missing values), located at different geographical coordinates all over the East Coast of Peninsular Malaysia, were analyzed [33]. The dataset of torrential rainfall for the East Coast of Peninsular Malaysia is shown in Figure 1 and Table 1.



**Figure 1.** The locations of the observed 30 rainfall stations on the east coast of Peninsular Malaysia.

**Table 1.** Geographical coordinates of the observed 30 rainfall stations on the east coast of Peninsular Malaysia.

Stations	Code	Latitude (°)	Longitude (°)
Setor JPS KT	E1	5.32	103.13
Kg. Sg. Tong	E2	5.36	102.89
Kg. Dura	E3	5.07	102.94
Kg. Menerong	E4	4.94	103.06
Kg. Embong Sekayu	E5	4.95	102.97
Jambatan Jerangau	E6	4.84	103.2
SM Sultan Omar	E7	4.76	103.42
Bandar Al-Muktafi	E8	4.61	103.2
Rumah Pam Paya Kempian	E9	4.57	102.98
Sg. Gawi	E10	5.14	102.84
Jambatan Tebak	E11	4.38	103.26
Kg. Ban Ho	E12	4.13	103.18
Hulu Jabor	E13	3.92	103.31
Kg. Batu Hampar	E14	5.45	102.82
Klinik Chalok Barat	E15	5.41	102.82
Inst. Pertanian Besut	E16	5.64	102.62
Temeris	E18	3.02	103.2
Kg. Unchang	E20	3.29	103.19
Kg. Gong Batu	E21	3.39	103.03
Rumah Pam Pahang Tua	E23	3.56	103.36
Pejabat JPS Pahang	E26	3.81	103.33
Kg. Laloh	E36	5.31	102.28
Ulu Sekor	E37	5.56	102.01
Dabong	E38	5.38	102.02
JPS Kuala Krai	E41	5.53	102.17
Air Lanas	E42	5.78	101.89
Kg. Durian Daun	E43	5.78	101.97
Bendang Nyior	E44	5.84	102.07
Rumah Kastam	E45	6.02	101.98
Gunung Gagau	E47	4.76	102.66

To allow for a reasonable discrepancy of factors that signifies a day of torrential rainfall, it was thus important to select certain standards that would contribute to the setting of a threshold. Representing tropical climates, the greatest recorded threshold was 60 mm/day based on the categorization of rainfall intensity by the DID, Malaysia. The unit for rainfall calculation is presented as mm/day or millimeters per day, precisely, indicating the total rainwater depth (mm) in 24 h (day). After filtering the total days from the data, the days with rainfall amounts of over 60 mm/day were accepted for research purposes [23]. After the 11,680-day datasets of the more than 48 stations was filtered accordingly to the threshold standards for at least 1.5% of the overall stations, the new dataset obtained was 175 days and 30 stations, which was adequate for the representation of the major torrential rainfall centers.

### 3. Methodology

#### 3.1. Principal Component Analysis (PCA)

By retaining most of the original variability in the data, PCA is designed to reduce the dimensions of large data to a lower dimension [34]. This is achieved by converting a set of observations of possibly correlated variables into a set of linearly uncorrelated principal components. The first principal components account for much of the original data's variability. Then each successive component accounts for the remaining variation subject to the previous component being uncorrelated.

When PCA is applied, the covariance or correlation matrix, derived from the data matrix is used to calculate its eigenvalues and eigenvectors [35]. In this study, the eigenvectors and eigenvalues were calculated from the correlation matrix to find the components that best explain variance in the data.

For the extraction of the number of components, it is recommended to take at least 70% of the cumulative percentage of total variation as a benchmark to cut off the eigenvalues in the large data sets [36]. The reduced matrix is the component matrix of eigenvector “loadings” that defines the new variables consisting of the linear transformation of the original variables maximizing the variance in the new. The steps involved in the PCA algorithm are shown in Figure 2.

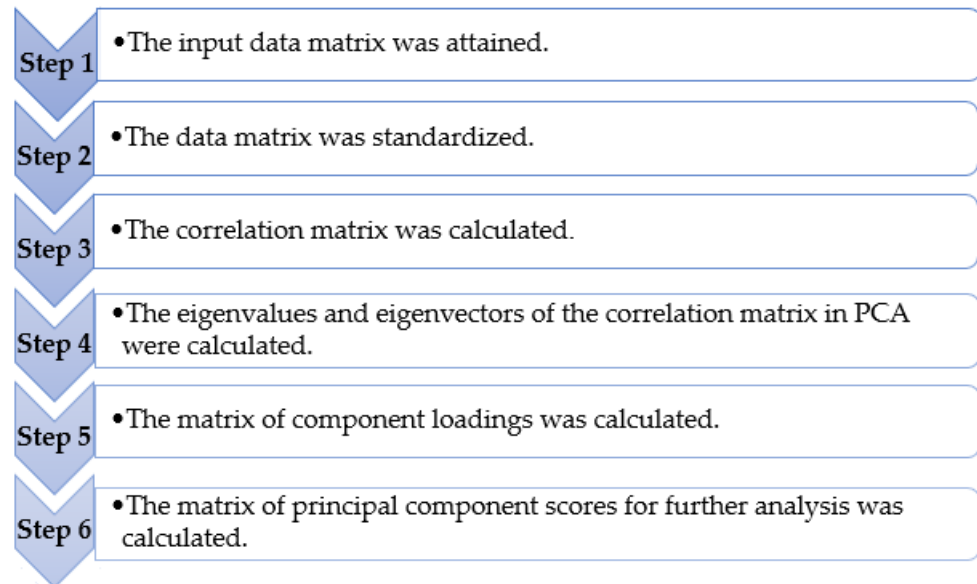


Figure 2. Procedure of classical PCA.

### 3.2. Pearson Correlation Matrix

In applications of hydrology and climatology, the Pearson correlation is typically used in the PCA for calculating the eigenvectors and eigenvalues [37,38]. The Pearson correlation coefficient between two variables is defined as the covariance of the two variables divided by the product of their standard deviations. Typically, the Pearson correlation is used to measure the distance (or similarities) before implementing a clustering algorithm. The Pearson correlation coefficient between two vectors of observations is as follows:

$$r_{ij} = \frac{\sum_{i=1}^n (X_i - \bar{X}_i)(X_j - \bar{X}_j)}{\sqrt{\sum_{i=1}^n (X_i - \bar{X}_i)^2 \sum_{i=1}^n (X_j - \bar{X}_j)^2}} \quad (1)$$

where  $X_i$  and  $X_j$  refer to the vectors of observations in matrix data  $X$  with  $n$  observations, with  $\bar{X}_i$  and  $\bar{X}_j$  referring to the mean of the vectors.

### 3.3. Tukey's Biweight Correlation

Tukey's biweight correlation is based on Tukey's biweight function that relies on the M-estimator used in robust correlation estimates. The M-estimator has a derivative function  $\psi$  which determines the weight assigned to the observations in the data set. It has the ability to down-weight observations to reflect their influence from the mean of the data [39].

Let matrix  $X$  be the rainfall data in the east coast of Peninsular Malaysia comprising  $I$  observations (i.e., rainfall days) defined by  $J$  variables (i.e., rainfall stations) denoted by the  $I \times J$  matrix  $X$ , whose generic element is  $x_{ij}$ . After that, the data was standardized using median and mean absolute deviation (MAD). This was conducted to prevent any masking or swamping effects [40]. The standardization of  $X$  is computed as follows

$$X^* = x_{ij}^* = \frac{x_{ij} - \bar{x}}{\text{median}(|x_{ij} - \text{median}(x_{ij})|)} \quad (2)$$

such that  $x_{ij}$  refer to elements in the input matrix,  $X$ .

Let  $X^* = (x_{ij}^*)$  denote an  $I \times J$  matrix  $X^*$ , in which its generic element is  $x_{ij}^*$  comprising  $I$  observations (i.e., rainfall days) described by  $J$  variables (i.e., rainfall stations). Let  $T$  be the valuation that minimizes

$$\sum_{i=1}^n \rho(X^*, t) \quad (3)$$

where  $\rho(X^*, t)$  is described as the objective function and the M-estimator of location  $T_n(x_{11}^*, x_{12}^*, \dots, x_{ij}^*)$ . Consequently,

$$\sum_{i=1}^n \psi(X^*, t) = 0 \quad (4)$$

where  $\psi(X^*, t)$  denotes the derivative of  $\rho(X^*, t)$  with respect to  $t$  which is used to simplify the calculation of the M-estimator by computing a value of  $t$ .

To calculate  $\rho$  and  $\psi$ , it is mandatory to measure the observation and choose some estimators of matrix, called  $S_n$ , and also a function of each element,  $(x_{11}^*, x_{12}^*, \dots, x_{ij}^*)$  of matrix  $X^*$ . Hence, the observations are converted to

$$u_i = (X^* - T_n)^T (c^* S_n)^{-1} (X^* - T_n) \quad (5)$$

where  $T_n$  is the location matrix,  $S_n$  is the shape matrix, and  $c^*$  is the positive constant, which is an equally tuning constant. To simplify the interpretation, the algorithm is shown in one dimension as:

$$u_i = \frac{x_{ij}^* - T_n}{c^* S_n} \quad (6)$$

Equation (6) substituted in Equations (3) and (4) to obtain

$$\sum_{i=1}^n \rho(u_i) \quad (7)$$

and

$$\sum_{i=1}^n \psi(u_i) = 0 \quad (8)$$

where  $T_n$  and  $S_n$  that solve Equation (7) are the consequential M-estimates of location and shape, correspondingly.

The derivative function is derived as follows:

$$\psi(u) = \begin{cases} u(1-u)^2 & |u| \leq 1 \\ 0 & |u| > 1 \end{cases} \quad (9)$$

It can be seen that if  $|u|$  is large enough, then  $\psi(u)$  reduces to zero. One of the important aspects of measuring the resistance to outlying data values of M-estimators is its breakdown point. According to the study, the breakdown point is used in measuring their resistance to outlying data values [41]. However, in the PCA-based Tukey's biweight correlation, the breakdown point is used to determine the ideal number of components to extract. Adjustments of the breakdown point will have an effect in determining the number of components to obtain using the PCA method [12]. This study tested the performance of the biweight correlation under various point changes in order to determine the number of components to extract from the PCA in order to identify the pattern of torrential rainfall. In this study, Tukey's biweight with breakdown points at 0.2, 0.4, 0.6, and 0.8 are compared.

The biweight estimate of correlation is produced by first calculating the location estimate  $\tilde{T}$  and then updating the shape estimate,  $\tilde{S}$ . The  $(i, j)$ th element of  $\tilde{S}$ , i.e.,  $\tilde{s}_{ij}$  acts



as a robust estimate of the covariance between the two vectors,  $X_i$  and  $X_j$ . The biweight correlation of these two vectors is calculated as follows:

$$\tilde{r}_{ij} = \frac{\tilde{s}_{ij}}{\sqrt{\tilde{s}_{ii}\tilde{s}_{jj}}} \quad (10)$$

with

$$T_n^{(k+1)} = \frac{\sum_{i=1}^n X_i w(u_{i(k)})}{\sum_{i=1}^n w(u_{i(k)})} \quad k = 0, 1, 2, \dots \quad (11)$$

$$S_n^{(k+1)} = \frac{\sum_{i=1}^n w(u_{i(k)}) (X_i - T^{(k+1)}) (X_i - T^{(k+1)})^t}{\sum_{i=1}^n w(u_{i(k)}) (u_{i(k)})} \quad (12)$$

where  $T_n^{(k+1)}$  is a location vector and  $S_n^{(k+1)}$  is a shape matrix such that  $k = 0, 1, 2, \dots$

Thus, a PCA-based Tukey's biweight correlation is more likely to produce a better cluster partition that is more resistant to outlying values than the Pearson correlation in PCA.

### 3.4. K-Means

This algorithm consists of two separate phases; the first step is to randomly select  $k$  objects, each of which initially represents a cluster mean or center. The next phase is to assign each data point  $e_L$  to the nearest cluster center. The Euclidean distance method is generally used to determine the distance  $d(e_L, c)$  between each data point  $e_L$  and centroid  $c_L$  as shown below:

$$d(e, c) = \sqrt{\sum_{i=1}^n (e_L - c_L)^2} \quad (13)$$

When all the data points are assigned to each of the clusters, the cluster centroid is recalculated. A schematic procedure of the k-means method is also presented below; the k-means clustering algorithm works as follows:

Step 1: Randomly choose  $k$  objects from a data set as an initial cluster centroid.

Step 2: Calculate the distance between each data point  $d$  and assign each item  $d$  to the cluster which has the closest centroid. Recalculate the cluster centroid for each cluster until the convergence criteria is reached.

### 3.5. Fuzzy C-Means Clustering (FCM)

Bezdek [42] introduced a clustering approach, FCM, which is widely used in solving the problems connected with feature analysis, clustering, and classifiers. Generally, fuzzy clustering refers to assigning values to membership levels and using them to assign data elements to one or more clusters. When using k-means clustering algorithms, each pattern belongs to only one cluster. However, fuzzy clustering goes a step further and assigns or associates every pattern present in an image to every cluster of image data using a membership function.

The clusters in FCM are formed according to the distance between data points and the cluster centers formed. As compared to K-means, FCM clustering is well performed in that it allows the data to belong to more than one cluster [43–45]. A dataset is grouped into  $p$  clusters with every data point in the dataset related to every cluster; each data point that lies close to the center will have a high degree of connection to that cluster while another point which lies far from the center of a cluster will have a low degree of connection to that cluster [46].

The FCM approach uses a fuzzy membership function that assigns a membership degree to each class. FCM splits the dataset  $X$  into  $c$  fuzzy clusters. This algorithm holds that every object belongs to a specific cluster with various degrees of membership, while a

cluster would be considered as a fuzzy subset on the dataset. FCM finds the partition that minimizes the objective function using the algorithm below [47]:

$$J = \sum_{i=1}^n \sum_{j=1}^p (u_{ij})^m d(X_i, C_j)^2 \quad (14)$$

where  $n$  is the number of data;  $p$  is the number of clusters;  $u_{ij}$  is the degree of relevance of the sample  $X_i$  to the  $j$ -th cluster;  $m \in [1, \infty]$  is a measure of fuzziness that determines how membership between fuzzy clusters is shared;  $d$  is the Euclidean distance between  $X_i$  and  $C_j$ ;  $X_i$  is the data vector, with  $i = 1, 2, \dots, n$ , which signifies a data attribute; and  $C_i$  is the center of fuzzy clustering.

The FCM method relies on the fuzziness parameter  $m$ , and it works well when  $m \in (1.5, 2.5)$  [48]. In this study, the fuzziness parameter applied is  $m = 2$  as a default from the package that has been used. The quadratic total of the weighed distances from the samples to the cluster centroid in every cluster is known as the objective function [49]. The objective function,  $J$  is minimized, and the membership degrees,  $u_{ij}$  are produced as the following formula:

$$u_{ij} = \left[ \sum_{i=1}^c \left( \frac{d(X_i, C_j)}{d(X_i, C_j)} \right)^{2/(m-1)} \right] \quad (15)$$

where  $C_j$  can be obtained from:

$$C_j = \frac{\sum_{i=1}^n (u_{ij})^m X_i}{\sum_{i=1}^n (u_{ij})^m} \quad (16)$$

In the interval  $(0, 1)$ , the degrees of membership,  $u_{ij}$  represent the probabilities that could be generated by a uniform distribution [50,51]. The clustering is modified at each iteration as the steps involved in the Figure 3 shown below:

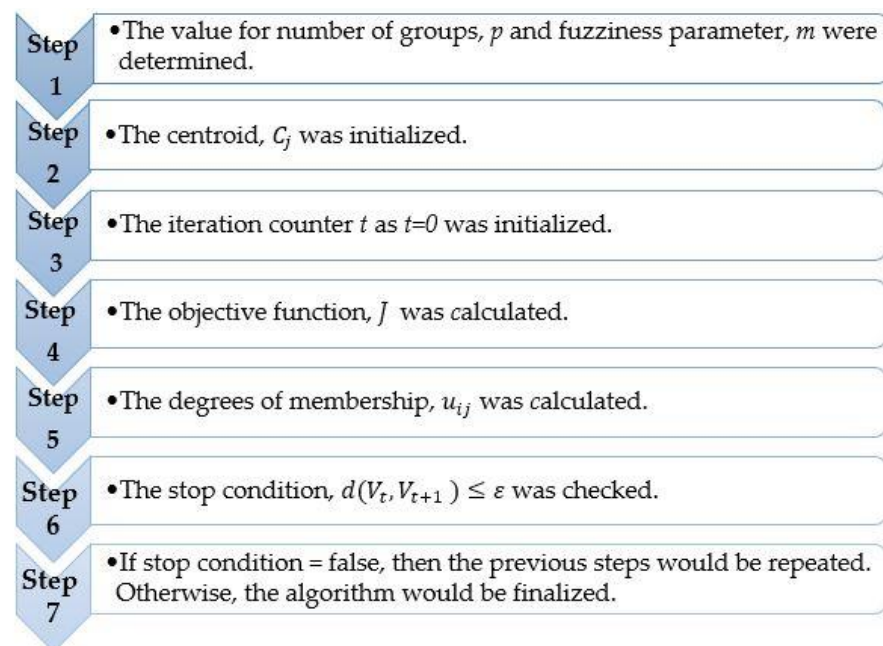


Figure 3. Procedure of FCM.

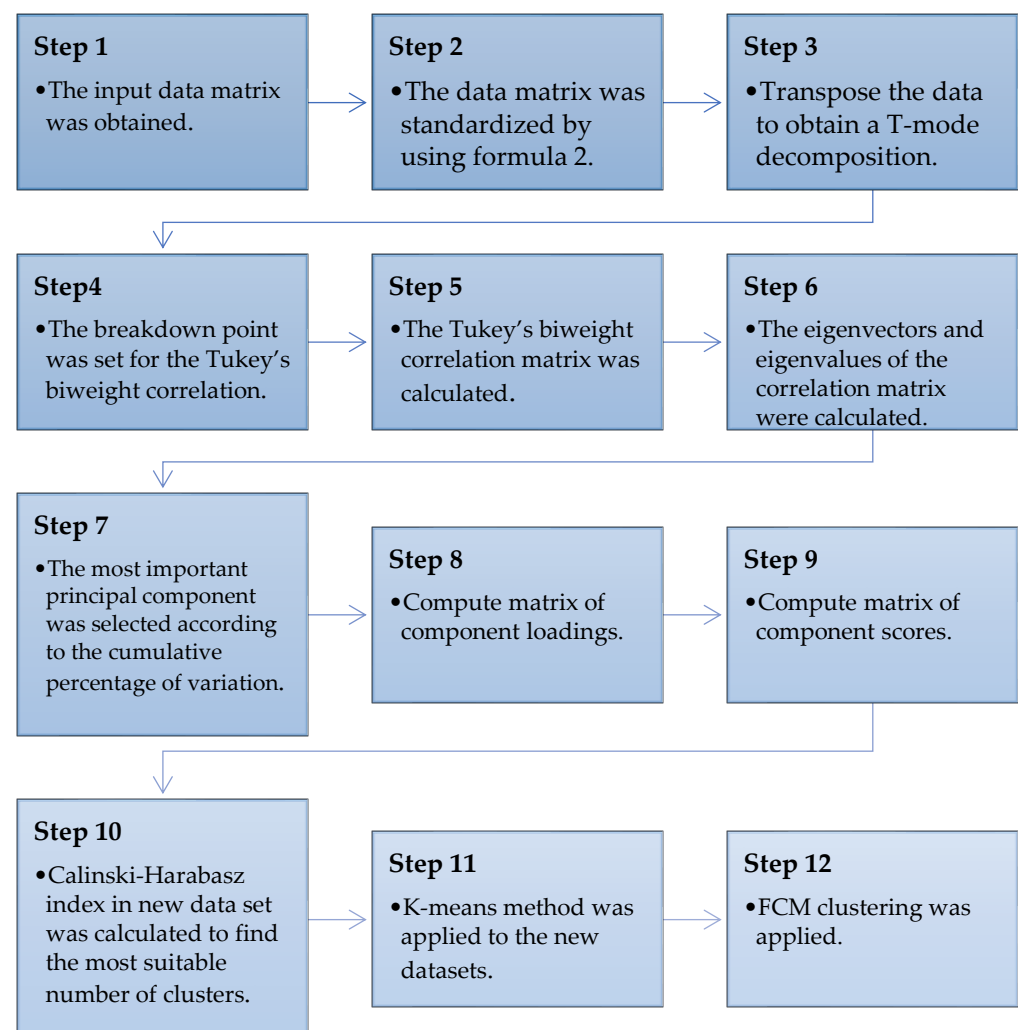
### 3.6. Proposed Statistical Modeling of RPCA-FCM

Robust PCA (RPCA) which is PCA based Tukey's biweight correlation, is used to overcome the problems that affect cluster partitions and generate extremely unbalanced

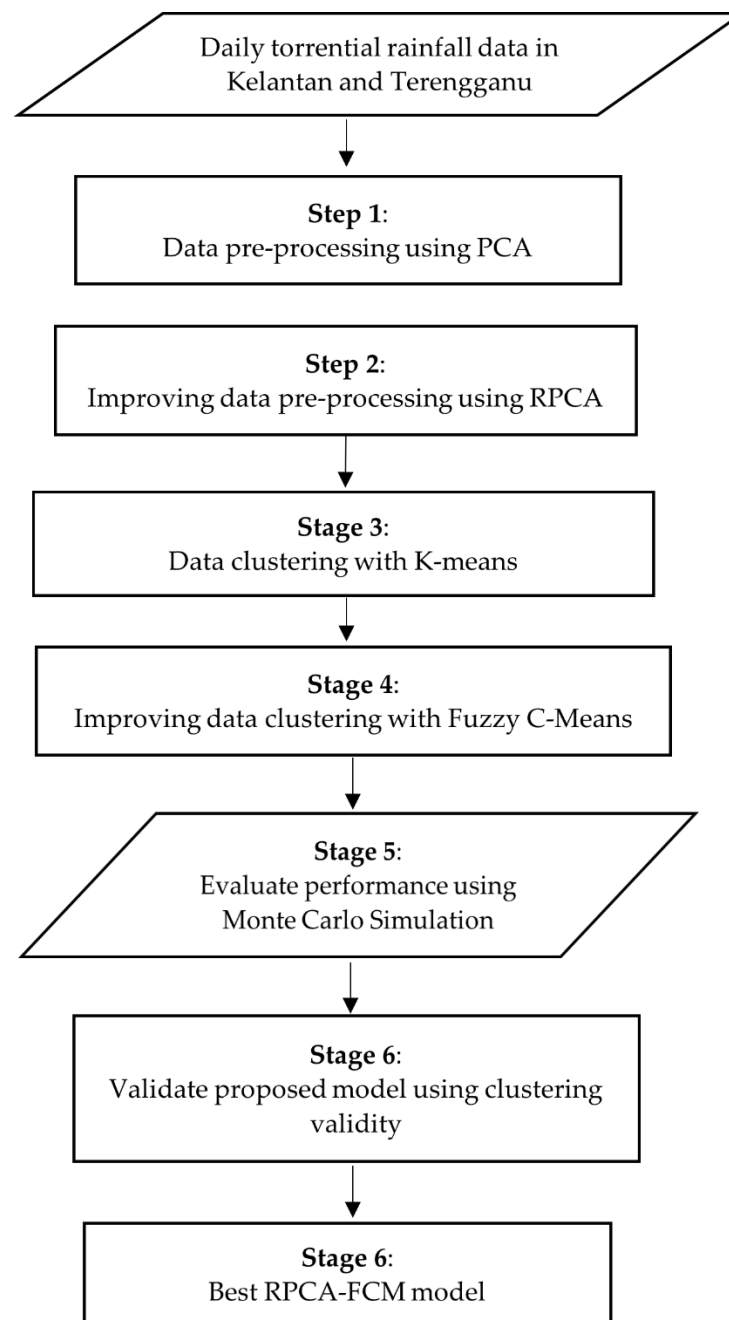


clusters in a high-dimensional space. RPCA is the robust version of classical PCA where the Pearson correlation is replaced with Tukey's biweight correlation. This solves the issue of the number of clusters obtained by using classical PCA where the Pearson correlation assigns equal weight to the outliers. Outliers in hydrological data are potentially gained from various mechanism of physical phenomenon or from errors occurring during the measuring process. Additionally, outliers could be carrying substantial information about irregular conditions of hydrologic phenomenon. While Tukey's biweight correlation is capable of assigning the weightage to the outliers in order to obtain the various number of clusters for the dataset used. To avoid any masking or swamping effects, the original data matrix is standardized using a robust location and scale estimator before proceeding. The reduced data set is then applied to the K-means cluster analysis to obtain the cluster partitions. The K-means method requires specifying the number of clusters before the algorithm is applied. To overcome this problem, the Calinski–Harabasz Index [52] is used as a measure to determine the optimal number of cluster partitions for the input data. Lastly, the steps are continued by combining the FCM to allow each element in the data to belong to more than one cluster partition. This is due to the drawback of K-means clustering where the element in the data is only allowed to belong to one and only one cluster.

The steps involved in the proposed model are presented in Figure 4 and the flowchart of the developed RPCA-FCM model is illustrated in Figure 5.



**Figure 4.** Procedure of RPCA-FCM.



**Figure 5.** Flowchart of the developed model of RPCA-FCM.

### 3.7. Clustering Validity and Model Validation

The use of validity indices is important in order to contrast the clustering algorithms and compare the performance of clusters in terms of compactness and connectedness. Hence, the validation of the clustering results is a base for the clustering process. In this study, all of the external and internal criteria were used to calculate the quality of the clustering findings using two separate approaches in the correlation matrix.

Internal clustering validation uses internal information from the clustering process to assess the decency of a clustering structure. It can also be employed for guesstimating the number of clusters and suitable clustering algorithms. In this study, the silhouette index is used to reflect the compactness, connectedness, and separation of the cluster partitions to assess the quality of the clustering findings. The silhouette width is the average of each observation's silhouette value. This technique finds the silhouette width for each single data point, the average silhouette width for every cluster, and the overall average silhouette

width for the overall data set [53]. The largest overall average silhouette width indicates the best clustering. Therefore, the number of clusters with maximized overall average silhouette width is considered the optimal number of the clusters. It can be concluded that the silhouette value measures the degree of confidence in a clustering study and lies in the interval  $(-1, 1)$ , with well-clustered observations showing values near 1 and poorly clustered observations having values near  $-1$ .

The external criteria tested whether the points of the data set were randomly structured. There were many external quality indices and this study focused on the Rand index as a means of validating the clustering results. This index was calculated by determining the fraction of properly classified data points to every data point. The derivation of the Rand index has been described in previous studies [54]. The ranges of the Rand index are between 0 and 1 where the value near 0 indicates that the pair is organized similarly under both clusterings, and the value approximate to 1 shows identical clustering.

The effectiveness of the properly identified clusters is shown by the following validity measures: partition coefficient, partition entropy, the fuzzy silhouette index, and the modified partition coefficient [55,56].

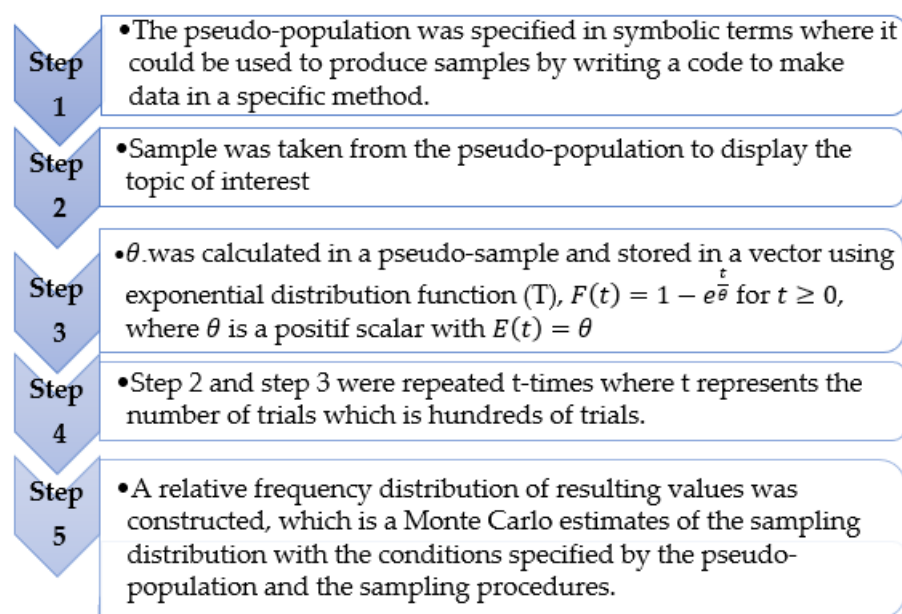
Based on the fuzzy partition matrix, the partition coefficient (PC) measures the degree of the fuzzy division of the clusters, as well as the quality of the partition; the larger its value, the better the partition results. Using the fuzzy partition matrix, the partition entropy (PE) measures the fuzzy degree of the final divided cluster, and the smaller its value, the better the partition. Meanwhile, the modified partition coefficient (MPC) decreases the monotonic tendency and the maximum value is shown at the best clustering of a given data set. The fuzzy silhouette index (FSI) is ideal for showing the maximum number of clusters so that the index takes the maximum value.

### 3.8. Monte Carlo Simulation

In order to assess the performance of the proposed model (RPCA-FCM), several sets of simulated data matrices generated by the Monte Carlo simulation were used to illustrate the methods described in Section 3.6. These simulated data matrices were generated to mimic the real data and to follow the pattern of the original torrential rainfall data from East Coast Peninsular Malaysia. The purpose of using simulation is to determine an appropriate breakdown point and to evaluate the performance of the RPCA-FCM against the classical PCA in the analysis of identifying spatial cluster torrential rainfall patterns in East Coast Peninsular Malaysia.

The Monte Carlo simulation is capable of testing the performance of the normal and non-normal distribution of data [57]. During the process, values are sampled at random from the input probability distributions. A sample is defined as a collection of samples. The outcome is reported from each sample. The Monte Carlo simulation is conducted through hundreds of iterations in order to obtain the optimal outcome.

The distribution is obtained for the creation of a  $n \times p$  matrix with  $n = 175$  and  $p = 30$  to reflect 175 torrential rainfall days and 30 rainfall stations correspondingly, from the original torrential rainfall data collected over 32 years from East Coast Peninsular Malaysia. All of the data produced shows values of about 60 and above, representing the torrential rainfall threshold of 60 mm/day. The standard algorithm of the Monte Carlo Simulation was illustrated in Figure 6 [58].



**Figure 6.** Procedure using RPCA-FCM.

Every set of the produced data was further applied to two approaches that are either classical PCA or RPCA. From both approaches, the results were then contrasted based on the cluster partition, the number of components, and the validity measure of the original datasets.

## 4. Results and Discussion

### 4.1. Descriptive Analysis

Data preparation is one of the vital steps before proceeding with the analysis; it involves screening, evaluating, selecting records, and reducing the number of variables to a manageable range. A brief overview of the torrential rainfall data on the study area is provided in Table 2 including descriptive statistics such as mean, standard deviation, and skewness.

**Table 2.** Summary statistics of average daily torrential rainfall amounts (mm) for 48 stations.

Mean	Standard Deviation	Skewness	Kurtosis	Coefficient Variation
7.78	17.80	6.83	139.47	2.29

All the stations give positive skewness with value 6.83. The results illustrated that the shape of rainfall distribution for the 48 stations is skewed because the values of the skewness are far from zero. Based on the data, it is likely that daily torrential rainfall does not originate from a normal population. Meanwhile, the kurtosis has the value of 139.47, indicating that the data has heavier tails compared to a normal distribution.

Table 2 shows that the average mean and average standard deviation of daily torrential rainfall amounts in East Coast Peninsular Malaysia are 7.78 and 17.80, respectively. Typically, the average coefficient of variation for the 48 rainfall stations is quite small, with the value of 2.29, exhibiting that the data have distributed evenly.

### 4.2. Choice of Breakdown Point for RPCA

The breakdown point is a significant element in the M-estimator for its tolerance towards the outlying data values. This study analyzed the choices of the breakdown point for  $r = 0.2$ ,  $r = 0.4$ ,  $r = 0.6$ , and  $r = 0.8$ . Table 3 visualizes how the choice of the breakdown point affected the number of extracting components, as well as the clusters, in this approach.

**Table 3.** Number of components, clusters, and validations for the chosen breakdown point.

	Cum. Percentage (%)	No. of Components	No. of Clusters	Internal Validity	External Validity
				Silhouette Index	Rand Index
$r = 0.2$	60	4	15	NA	0.6069
	65	5	7	NA	0.6137
	70	8	5	NA	0.6451
	75	11	4	0.8719	0.6438
	80	16	4	0.9043	0.6412
	85	24	3	0.9298	0.6032
	90	37	2	0.9496	0.5484
$r = 0.4$	60	7	7	NA	0.5024
	65	10	4	NA	0.5944
	70	14	4	0.9029	0.6037
	75	18	4	0.9214	0.6044
	80	23	4	0.9353	0.6045
	85	29	4	0.9442	0.6062
	90	36	2	0.9524	0.6069
$r = 0.6$	60	4	16	NA	0.6872
	65	5	13	NA	0.6883
	70	7	14	NA	0.6566
	75	8	13	NA	0.6566
	80	9	12	NA	0.6565
	85	10	7	NA	0.6459
	90	12	7	0.8943	0.6445
$r = 0.8$	60	4	18	NA	0.4946
	65	5	14	NA	0.4945
	70	7	12	NA	0.4931
	75	8	7	NA	0.4958
	80	9	7	NA	0.4962
	85	10	7	NA	0.4958
	90	11	7	0.8869	0.4956

From Table 3, choosing  $r = 0.2$  was not suitable for this dataset where the highest internal validity measured at 90% cumulative percentage of variation while the highest value of the Rand index was at 70% cumulative percentage of variation. However, 60% to 70% of the cumulative percentage of variation obtained undefined the value for the silhouette index. This shows that  $r = 0.2$  is an unstable breakdown point for this dataset. While from Table 3, it could be clearly seen that  $r = 0.6$  and  $r = 0.8$  can also be concluded as unstable breakdown points for RPCA. Based on the internal validity, the silhouette index, NA, was obtained for 60% to 85% cumulative percentage of variation used. In addition, the number of components to extract in both breakdown points was too small, affecting the validation measure of the clusters. However, it was impossible to increase the cumulative percentage of variation with the aim of increasing the number of components to extract. This is because higher cumulative percentage is defined as higher noise in the data. It is clearly proven that the cutoff for the breakdown point at  $r = 0.6$  and  $r = 0.8$  were not suitable for this dataset. Based on Table 3, 90% cumulative percentage of variation obtained well-performed internal and external validity measures. In contrast, the clusters obtained for the 90% cumulative percentage contained two clusters that were not suitable for the rainfall study. The number of components to retain for 90% was also the highest (36 components) compared to other cumulative percentages of variation. Especially in hydrological study, if excessive components were extracted, it would show the insignificant variations of low frequency or spatial scale. Due to this fact, the 85% of cumulative percentage of variation using the breakdown point,  $r = 0.4$  was focused on in this study to improve the performance of the clustering.

It can be concluded that the best selection of the breakdown point for RPCA is  $r = 0.4$ . This result is supported by previous researchers [12,41] who agreed that the breakdown point of  $r = 0.4$  offered a good balance to extract the best number of components since it retains only sufficient outstanding components.

The time consumed in calculating the validity index for the validity measure with a higher number of components takes a longer time compared to lower number of components. There is a limitation, as high-end computers with faster processing speeds are needed to acquire reliable and accurate results. The results can only be generated with the help of high-end computers and the application of machine learning methods that can improve the speed and accuracy of the test. In this study, we also found that extracting too low a number of components also affected the calculation for obtaining the validity measures of the clusters. This basically refers to a number of fewer than 10 components. In fact, the results that came out as NA for the number of components was less than 10. Hence, choosing the most accurate breakdown point is very important in using RPCA.

Based on the internal validity, the silhouette index, NA, was obtained for 60% to 85% cumulative percentage of variation used. Based on Table 3, 90% cumulative percentage of variation obtained the well-performed internal and external validity measures. However, the clusters obtained at 90% cumulative percentage were two clusters which were not a suitable cluster identification for the rainfall study since they mask the true structure of the data. The number of components to retain for 90% was also the highest, which was 36 components, compared to other accumulative percentages of variation. Especially in hydrological study, if excessive components were extracted, it would show the insignificant variations of low frequency or spatial scale. Due to that, 85% of cumulative percentage of variation in which the breakdown point,  $r = 0.4$  was focused on in this study to improve the performance of the clustering.

#### 4.3. Evaluating Performance of Classical PCA against RPCA

From Table 4, it shows that in order to obtain at least 60% of the total percentage of variance in contrast to the classical PCA, RPCA required less components to be extracted. Fourteen components were retained using RPCA as compared to the classical PCA, which required 30 components at 70% cumulative percentage of variation. This reflects the results of the number of components in each level of cumulative percentage of variation in which RPCA required a smaller number of components to extract compared to the classical PCA. However, the selection of excessively high cumulative percentages of variation was not a good cutoff for the number of principal components in identifying the rainfall patterns. This is due the fact that extracting a higher cumulative percentage of variation results in a higher number of components to retain. Inclusion of too many principal components inflates the importance of outlier, and the results become unsatisfactory in identifying the rainfall patterns [25].

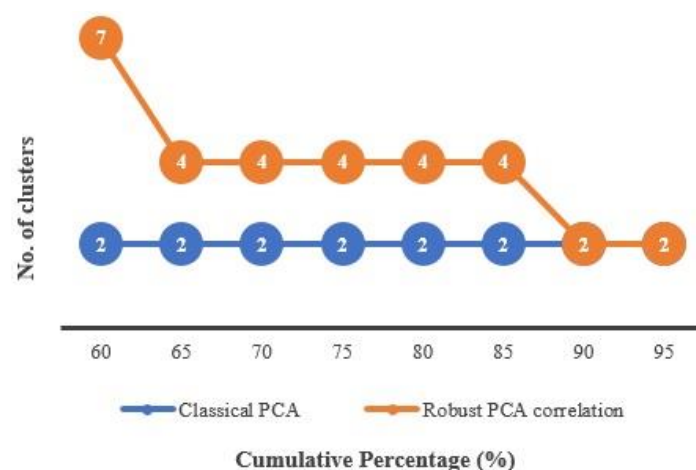
**Table 4.** Number of components obtained for classical PCA against RPCA for the original rainfall data.

Cum. Percentage (%)	No. of Components	
	Classical PCA	RPCA
60	22	7
65	26	10
70	30	14
75	36	18
80	42	23
85	49	29
90	58	36
95	62	45

In terms of cluster partitions, Figure 7 also shows that in contrast to the classical PCA, RPCA is more sensitive to the number of clusters according to the number of components



retained. At 60% of cumulative percentage, RPCA resulted in the highest number of clusters (seven clusters) compared to the classical PCA. The number of cluster partitions becomes stabilized at 65% to 85% cumulative percentage of the variation of the data (four clusters). However, the resulting number of clusters was maintained as the classical PCA when obtained at more than 85% cumulative percentage of variance. Moreover, the number of clusters using the classical PCA appeared to stabilize at just two clusters, despite the cumulative percentage of variation employed.



**Figure 7.** The number of clusters obtained based on using the classical PCA against the RPCA correlation using torrential rainfall data from East Coast Peninsular Malaysia.

This is due to the fact that the detection and treatment of the outliers in these observations is less effective when employing the Pearson correlation in the classical PCA because of its vulnerability towards the outliers. Additionally, the Pearson correlation in PCA assigns equal weight to each pair of observations. This would reduce the efficiency of the data due to the fact that the outliers undoubtedly stick out, causing information in the observation to be weakened. As a result, this correlation becomes ineffective in detecting the outliers.

In hydrological studies, particularly in identifying rainfall patterns, it is more reasonable to obtain more than two cluster partitions to explain various types of rainfall patterns. This shows that RPCA is capable of downweighting the outliers closer to the center as it helps in obtaining differentiated numbers of clusters compared to using the classical PCA.

#### 4.4. Validity Measures of Clusters

Table 5 illustrates that the RPCA shows the relatively highest silhouette index with approaching 1 (0.9442) while the classical PCA obtained the smallest silhouette index at 0.6047. Similarly, RPCA also obtained a higher value of Rand index compared to the classical PCA. Therefore, from the clustering results, it can be concluded that RPCA correlation is more efficient in clustering hydrological data, especially for clustering the rainfall data of tropic regions.

**Table 5.** Indices to measure the quality of clustering results for torrential rainfall data.

Validity Measures	Classical PCA	RPCA
Silhouette Index	0.6047	0.9442
Rand Index	0.5221	0.6062

#### 4.5. Validity Measures of RPCA-FCM

The clusters obtained from the RPCA correlation were maintained as four optimal clusters. These clusters were then validated using the fuzzy silhouette index (FSI), partition

entropy (PE), the partition coefficient (PC) and the modified partition coefficient (MPC) as shown in Table 6. As a guideline, the smallest value of PE with the largest value of FSI, PC, and MPC represents a well-clustered partition.

**Table 6.** Validity measures of the statistical models for original data.

Methods	FSI	PE	PC	MPC
Classical PCA-FCM	0.6833	0.6837	0.6363	0.5150
RPCA-FCM	0.7857	0.3471	0.7895	0.5789

From Table 6, it can be seen that RPCA-FCM obtained the largest value of FSI, PC, and MPC with the lowest value of PE, while the combination of classical PCA with FCM obtained poor results. This result shows that the proposed statistical modeling, RPCA-FCM, performed well in clustering the torrential rainfall patterns of East Coast of Peninsular Malaysia compared to the classical procedure.

#### 4.6. Evaluating the Performance of Proposed Models Based on Simulation Results

The data that mimicked the original dataset was then generated using the Monte Carlo simulation. In Table 7, similar results were shown with the original data where the classical PCA required a greater number of components to extract in order to achieve at least 60% of the cumulative percentage of variation compared to the RPCA. For example, seven components were retained in the RPCA as compared to the classical PCA which required 28 components at 60% cumulative percentage of variation. This reflected the results obtained from the original data where RPCA required a smaller number of components to extract.

**Table 7.** Number of components and clusters obtained using classical PCA and RPCA based on simulation data.

Cum. Percentage (%)	No. of Components		No. of Clusters	
	Classical PCA	RPCA	Classical PCA	RPCA
55	24	4	2	11
60	28	7	2	7
65	33	10	2	4
70	38	14	2	4
75	44	18	2	4
80	52	24	2	3
85	60	30	2	3
90	72	38	2	3

In terms of the number of clusters obtained, it can be seen that in contrast to the classical PCA, RPCA was more sensitive to the number of clusters according to the number of components retained. RPCA also showed similar results similar to the original data, whereas the classical PCA appeared to stabilize at only two clusters regardless of the cumulative percentage of variation used while the RPCA showed differentiating patterns of the number of clusters obtained.

Table 8 illustrates that RPCA showed a relatively higher silhouette index against the classical PCA. On the other hand, the classical PCA obtained a lower Rand index compared to the RPCA, proving that the RPCA is more well-clustered.

**Table 8.** Indices to measure the quality of the clustering results of the simulation data.

Validity Measures	Classical PCA	RPCA
Silhouette Index	0.5846	0.9257
Rand Index	0.4958	0.5766

The analysis was then continued by using 85% cumulative percentage of variation for the simulated data in the FCM clustering. From Table 9, the classical PCA combined with FCM obtained smaller values of FSI, PC, and MPC compared to RPCA combined with FCM. It also showed a higher value of partition entropy at 0.4419 compared to RPCA combined with FCM, which obtained a smaller value of 0.4126. This result shows the same results as implied by the original dataset where the proposed statistical modeling, RPCA-FCM, performed well compared to the classical PCA in clustering the torrential rainfall patterns of East Coast Peninsular Malaysia.

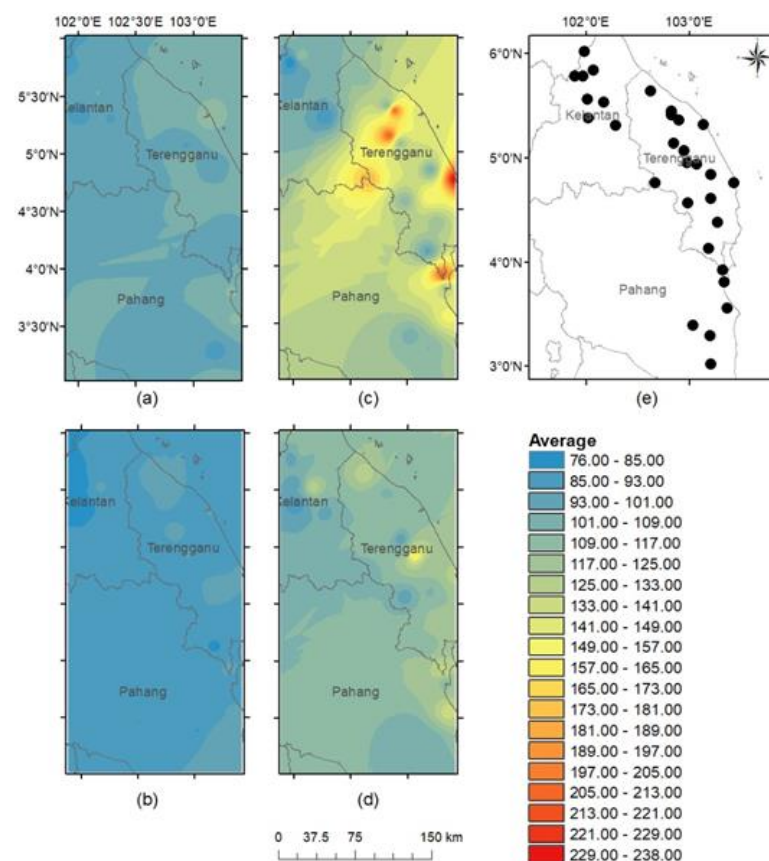
**Table 9.** Validity measures of the statistical models for the simulation data.

Methods	FSI	PE	PC	MPC
Classical PCA-FCM	0.6579	0.4419	0.7184	0.4369
RPCA-FCM	0.6906	0.4126	0.7406	0.4813

#### 4.7. Description of Clustering of Rainfall Patterns

Using the results of the proposed statistical model RPCA coupled with FCM, four clusters were obtained. The clusters were then analyzed in order complete the cluster mapping.

As shown in Figure 8, the pattern of torrential rainfall on the East Coast of Peninsular Malaysia is divided into daily torrential rainfall amounts from RP1 through RP4. As we compare the four clusters presented, it can be seen that the distinction between the torrential rainfall patterns were quite evident and that each pattern represented a different distribution of the rainfall. In each torrential rainfall patterns, there were distinct locations that were especially affected by the torrential rainfall. The highest rainfall recorded with an average amount of 229–238 mm.



**Figure 8.** Torrential rainfall composites on the East Coast of Peninsular Malaysia. (a) RP1, (b) RP2, (c) RP3, (d) RP4, (e) rainfall stations.

Figure 8e shows the main torrential rainfall stations after filtering out days with rainfall more than 60 mm for at least 1.5% of the total number of stations.

Figure 8a refers to RP1, or generally exhibited moderate torrential rainfall, in the entire regions of Kelantan, Pahang, and Terengganu with an average amount of 85–117 mm where the highest percentage of days associated in this pattern was 40% of the torrential days over 32 years. The maximum average of torrential rainfall for RP1 occurred at Setor JPS KT (E1) in Terengganu with the record of 117.71 mm/day.

Figure 8b refers to RP2, or light torrential rainfall within the range of 81–101 mm throughout the whole region. The second-highest percentage of days associated in this pattern was 35.4% of torrential days over 32 years. Kg. Unchang (E20) recorded the maximum average of torrential rainfall at 100.06 mm/day.

Figure 8c refers to RP3, characterized as extreme torrential rainfall, since it generally represents heavy rainfall with an average amount of 76–238 mm in most regions. The maximum average torrential rainfall in RP3 is 237.80 mm/day at (E7). RP3 was associated with the lowest days of torrential days at 5.7%.

RP4 as referred to in Figure 8d represents intense torrential rainfall with an average amount of torrential rainfall of 85–166 mm. This rainfall pattern was associated with 18.9% of torrential days over 32 years. It contributed to the second-highest maximum average of torrential rainfall of 165.8 mm, measured at Kg. Menerong (E4).

These results shows that the eastern areas are strongly affected by the Northeast Monsoon season that brings torrential rainfall to the East Coast regions from November until February. Therefore, the eastern part of Peninsular Malaysia is considered the wettest area during the Northeast Monsoon season. In particular, a maximum intensity is clearly distinguishable on the windward side of the mountain range. A large intensity gradient is observed across the mountain range, where greater values are recorded on the windward side of the mountain range and lesser values on its leeward side. The mountain range located in the center of the country forces the winds to rise and the rising air cools and condenses, resulting in heavy rains. The increased of intensity observed in October all across the peninsula may be attributable to the equatorward migration of another broad area of convergence. Intensity values are also greater in October than in April/May due to the fact that the convergence is smaller in the latter months than in the former. In spite of the fact that the intensity all across the peninsula increases in October, the same pattern prevails with larger values on the windward side of the mountain range and lesser values on its leeward side.

## 5. Conclusions

In this paper, a Robust PCA using Tukey's biweight correlation, combined with the FCM also known as RPCA-FCM was introduced. The experimental results showed that the proposed RPCA-FCM was able to effectively capture the clustered torrential rainfall patterns of East Coast Peninsular Malaysia. This study has demonstrated a substantial improvement in terms of the numbers of cluster partitions analyzed in order to avoid the use of inaccurate imbalanced clusters in high-dimensional datasets. The validation of the clusters obtained has been proven by the internal and external validity of the clusters. In addition, simulation studies were conducted to evaluate the performance of the RPCA-FCM compared to the classical procedure. In particular, the proposed method managed to downweight the far-from-center outliers and develop a crisper cluster partition. The results proved that several clusters can belong to one data point when the Fuzzy C-means method was employed. Four distinct clusters were identified to explain the spatiotemporal rainfall patterns that occur in the East Coast region of Malaysia. The main features between clusters are particularly characterized with regards to their significant locations and the period of torrential rainfall patterns.

**Author Contributions:** Conceptualization, S.M.S.; methodology, S.M.C.M.N. and N.A.; software applications, S.I.; validations, S.A.M.N. and M.L.T.; writing—original draft preparation, S.M.C.M.N. and S.M.S.; writing—review and editing, S.I. and M.L.T.; project administration, S.M.S. All authors have read and agreed to the published version of the manuscript.

**Funding:** This study was produced under the Fundamental Research Grants Scheme Vot No. 2019-0132-103-02 (FRGS/1/2019/STG06/UPSI/02/4) provided by the Malaysian Ministry of Education.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Not applicable.

**Acknowledgments:** The authors would like to thank the Ministry of Higher Education, Malaysia (MOHE) for supporting this research under the Fundamental Research Grant Scheme.

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

## References

- Wong, C.L.; Ismail, T.; Yusop, Z. Trend of Daily Rainfall and Temperature in Peninsular Malaysia based on Gridded Data Set. *Int. J. GEOMATE* **2018**, *14*, 65–72. [\[CrossRef\]](#)
- Wong, C.L.; Liew, J.; Yusop, Z.; Ismail, T.; Venneker, R.; Uhlenbrook, S. Rainfall Characteristics and Regionalization in Peninsular Malaysia Based on a High Resolution Gridded Data Set. *Water* **2016**, *8*, 500. [\[CrossRef\]](#)
- Khan, N.; Pour, S.H.; Shahid, S.; Ismail, T.; Ahmed, K.; Chung, E.S.; Nawaz, N.; Wang, X. Spatial distribution of secular trends in rainfall indices of Peninsular Malaysia in the presence of long-term persistence. *Meteorol. Appl.* **2019**, *26*, 655–670. [\[CrossRef\]](#)
- Alias, N.E.; Mohamad, H.; Chin, W.Y.; Yusop, Z. Rainfall Analysis of the Kelantan Big Yellow Flood. *J. Teknol.* **2014**, *78*, 83–90. [\[CrossRef\]](#)
- Shaharudin, S.M.; Ahmad, N.; Nor, S.M.C.M. A modified correlation in principal component analysis for torrential rainfall patterns identification. *IAES Int. J. Artif. Intell.* **2020**, *9*, 655–661. [\[CrossRef\]](#)
- Son, C.-H.; Baek, J.-I.; Ban, Y.-U.; Ha, S.-R. The Effects of Mitigation Measures on Flood Damage Prevention in Korea. *Sustainability* **2015**, *7*, 16866–16884. [\[CrossRef\]](#)
- Ochoa-Rodriguez, S.; Wang, L.-P.; Gires, A.; Pina, R.D.; Reinoso-Rondinel, R.; bruni, G.; Ichiba, A.; Gaitan, S.; Cristiano, E.; van Assel, J.; et al. Impact of Spatial and Temporal Resolution of Rainfall Inputs on Urban Hydrodynamic Modelling Outputs: A Multi-Catchment Investigation. *J. Hydrol.* **2015**, *531*, 389–407. [\[CrossRef\]](#)
- Norliyana, W.I.W.; Suhaila, J. Smoothing Wind and Rainfall Data through Functional Data Analysis Technique. *J. Teknol.* **2015**, *74*, 105–112. [\[CrossRef\]](#)
- Zhang, B.; Cao, P. Classification of high dimensional biomedical data based on feature selection using redundant removal. *PLoS ONE* **2019**, *14*, e0214406. [\[CrossRef\]](#) [\[PubMed\]](#)
- Rahman, A.S.; Rahman, A. Application of Principal Component Analysis and Cluster Analysis in Regional Flood Frequency Analysis: A Case Study in New South Wales, Australia. *Water* **2020**, *12*, 781. [\[CrossRef\]](#)
- Moutinho, L.; Hutcheson, G.; Moutinho, L. Exploratory or Confirmatory Factor Analysis. In *The SAGE Dictionary of Quantitative Management Research*; SAGE Publications: Thousand Oaks, CA, USA, 2014; pp. 111–116. [\[CrossRef\]](#)
- Shaharudin, S.M.; Nor, S.M.C.M.; Tan, M.L.; Samsudin, M.S.; Azid, A.; Ismail, S. Spatial Torrential Rainfall Modelling in Pattern Analysis Based on Robust PCA Approach. *Pol. J. Environ. Stud.* **2021**, *30*, 3221–3230. [\[CrossRef\]](#)
- Padilha, V.A.; Campello, R.J.G.B. A systematic comparative evaluation of biclustering techniques. *BMC Bioinform.* **2017**, *18*, 55. [\[CrossRef\]](#)
- Alam, M.S.; Paul, S. A comparative analysis of clustering algorithms to identify the homogeneous rainfall gauge stations of Bangladesh. *J. Appl. Stat.* **2019**, *47*, 1460–1481. [\[CrossRef\]](#)
- Mingoti, S.A.; Lima, J.O. Comparing SOM neural network with Fuzzy c-means, K-means and traditional hierarchical clustering algorithms. *Eur. J. Oper. Res.* **2006**, *174*, 1742–1759. [\[CrossRef\]](#)
- Suganya, R.; Shanti, R. Fuzzy C-Means Algorithm-A Review. *Int. J. Sci. Res. Publ.* **2012**, *2*, 1–3.
- Grover, N. A study of various Fuzzy Clustering Algorithms. *Int. J. Eng. Res.* **2014**, *3*, 177–181. [\[CrossRef\]](#)
- Atiyah, I.A.; Mohammadpour, A.; Taheri, S.M. KC-Means: A Fast Fuzzy Clustering. *Adv. Fuzzy Syst.* **2018**, *2018*, 2634861. [\[CrossRef\]](#)
- Askari, S. Fuzzy C-Means clustering algorithm for data with unequal cluster sizes and contaminated with noise and outliers: Review and development. *Expert Syst. Appl.* **2021**, *165*, 113856. [\[CrossRef\]](#)
- Beliakov, G.; Li, G.; Vu, H.Q.; Wilkin, T. Characterizing Compactness of Geometrical Clusters Using Fuzzy Measures. *IEEE Trans. Fuzzy Syst.* **2014**, *23*, 1030–1043. [\[CrossRef\]](#)
- Chen, L.; Chen, C.L.P.; Lu, M. A Multiple-Kernel Fuzzy C-Means Algorithm for Image Segmentation. *IEEE Trans. Syst. Man Cybern.* **2011**, *41*, 1263–1274. [\[CrossRef\]](#)



22. Askari, S.; Montazerin, N. A high-order multi-variable Fuzzy Time Series forecasting algorithm based on fuzzy clustering. *Expert Syst. Appl.* **2015**, *42*, 2121–2135. [\[CrossRef\]](#)
23. Nor, S.M.C.M.; Shaharudin, S.M.; Ismail, S.; Kismiantini, K. A RPCA-Based Tukey's Biweight for Clustering Identification on Extreme Rainfall Data. *Environ. Ecol. Res.* **2021**, *9*, 114–118. [\[CrossRef\]](#)
24. Shaharudin, S.M.; Ismail, S.; Nor, S.M.C.M.; Ahmad, N. An Efficient Method to Improve the Clustering Performance using Hybrid Robust Principal Component Analysis-Spectral biclustering in Rainfall Patterns Identification. *IAES Int. J. Artif. Intell.* **2019**, *8*, 237–243. [\[CrossRef\]](#)
25. Bolon-Canedo, V.; Sanchez-Marono, N.; Alonso-Betanzos, A. *Feature Selection for High-Dimensional Data*; Springer Nature: Basingstoke, UK, 2020.
26. Pes, B. Ensemble feature selection for high-dimensional data: A stability analysis across multiple domains. *Neural Comput. Appl.* **2020**, *32*, 5951–5973. [\[CrossRef\]](#)
27. Ramos-Cañón, A.M.; Prada-Sarmiento, L.F.; Trujillo-Vela, M.G.; Macías, J.P.; Santos-R, A.C. Linear discriminant analysis to describe the relationship between rainfall and landslides in Bogotá, Colombia. *Landslides* **2016**, *13*, 671–681. [\[CrossRef\]](#)
28. Bueso, D.; Piles, M.; Camps-Valls, G. Nonlinear PCA for Spatio-Temporal Analysis of Earth Observation Data. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 5752–5763. [\[CrossRef\]](#)
29. Jardim, A.M.D.R.F.; da Silva, M.V.; Silva, A.R.; dos Santos, A.; Pandorfi, H.; de Oliveira-Júnior, J.F.; de Lima, J.L.; de Souza, L.S.B.; Júnior, G.D.N.A.; Lopes, P.M.O.; et al. Spatiotemporal climatic analysis in Pernambuco State, Northeast Brazil. *J. Atmos. Sol.-Terr. Phys.* **2021**, *223*, 105733. [\[CrossRef\]](#)
30. Othman, M.; Ash'Aari, Z.H.; Mohamad, N.D. Long-term Daily Rainfall Pattern Recognition: Application of Principal Component Analysis. *Procedia Environ. Sci.* **2015**, *30*, 127–132. [\[CrossRef\]](#)
31. Jiang, Y.; Cooley, D.; Wehner, M.F. Principal Component Analysis for Extremes and Application to U.S. Precipitation. *J. Clim.* **2020**, *33*, 6441–6451. [\[CrossRef\]](#)
32. Nor, S.M.C.M.; Shaharudin, S.M.; Ismail, S.; Zainuddin, N.H.; Tan, M.L. A comparative study of different imputation methods for daily rainfall data in east-coast Peninsular Malaysia. *Bull. Electr. Eng. Inform.* **2020**, *9*, 635–643. [\[CrossRef\]](#)
33. Jolliffe, I.T.; Cadima, J. Principal component analysis: A review and recent developments. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* **2016**, *374*, 20150202. [\[CrossRef\]](#)
34. Kim, H.-J. Common Factor Analysis Versus Principal Component Analysis: Choice for Symptom Cluster Research. *Asian Nurs. Res. Korean. Soc. Nurs. Sci.* **2008**, *2*, 17–24. [\[CrossRef\]](#)
35. Jolliffe, I.T. Discarding Variables in a Principal Component Analysis. I: Artificial Data. *J. R. Stat. Soc. Ser. C Appl. Stat.* **1972**, *21*, 160–173. [\[CrossRef\]](#)
36. Cipolla, G.; Francipane, A.; Noto, L.V. Classification of extreme rainfall for a Mediterranean region by means of atmospheric circulation patterns and reanalysis data. *Water Resour. Manag.* **2020**, *34*, 3219–3235. [\[CrossRef\]](#)
37. Romero, R.; Ramis, C.; Guijarro, J.A. Daily rainfall patterns in the Spanish Mediterranean area: An objective classification. *Int. J. Climatol.* **1999**, *19*, 95–112. [\[CrossRef\]](#)
38. Hardin, J.; Mitani, A.; Hicks, L.; VanKoten, B. A robust measure of correlation between two genes on a microarray. *BMC Bioinform.* **2007**, *8*, 220. [\[CrossRef\]](#) [\[PubMed\]](#)
39. Choulakian, V. Robust Q-mode principal component analysis in L1. *Comput. Stat. Data Anal.* **2001**, *37*, 135–150. [\[CrossRef\]](#)
40. Owen, M. Tukey's Biweight Correlation and the Breakdown. Phd Thesis, Pomona College, Claremont, CA, USA, 2010. Bezdek, J.C. Cluster Validity with Fuzzy Sets. *J. Cybern.* **1973**, *3*, 58–73. [\[CrossRef\]](#)
41. Taufik, A.; Ahmad, S.S. A Comparative Study of Fuzzy C-Means And K-Means Clustering Techniques. *Malays. Tech. Univ. Conf. Eng. Technol. 8th MUCET* **2014**, *1*, 10–11.
42. Dubey, A.K.; Gupta, U.; Jain, S. Comparative Study of K-means and Fuzzy C-means Algorithms on The Breast Cancer Data. *Int. J. Adv. Sci. Eng. Inf. Technol.* **2018**, *8*, 18–29. [\[CrossRef\]](#)
43. Chuan, Z.L.; Wan Yusoff, W.N.S.; Senawi, A.; Akramin, M.R.M.; Fam, S.-F.; Shinyie, W.L.; Ken, T.L. A comparative effectiveness of hierarchical and nonhierarchical regionalisation algorithms in regionalising the homogeneous rainfall regions. *Pertanika J.* **2022**, *30*, 1–24. [\[CrossRef\]](#)
44. Ghosh, S.; Dubey, S.K. Comparative Analysis of K-Means and Fuzzy C-Means Algorithms. *Int. J. Adv. Comput. Sci. Appl.* **2013**, *4*, 35–39. [\[CrossRef\]](#)
45. Gomes, E.P.; Blanco, C.J.C.; Pessoa, F.C.L. Identification of homogeneous precipitation regions via Fuzzy c-means in the hydrographic region of Tocantins-Araguaia of Brazilian Amazonia. *Appl. Water Sci.* **2019**, *9*, 6. [\[CrossRef\]](#)
46. Zerouali, B.; Chettih, M.; Abda, Z.; Mesbah, M.; Santos, C.A.G.; Neto, R.M.B. A new regionalization of rainfall patterns based on wavelet transform information and hierarchical cluster analysis in northeastern Algeria. *Theor. Appl. Climatol.* **2021**. [\[CrossRef\]](#)
47. Ren, M.; Liu, P.; Wang, Z.; Yi, J. A Self-Adaptive Fuzzy c-Means Algorithm for Determining the Optimal Number of Clusters. *Comput. Intell. Neurosci.* **2016**, *2016*, 2647389. [\[CrossRef\]](#)
48. Alcantara, A.L.; Ahn, K.-H. Probability Distribution and Characterization of Daily Precipitation Related to Tropical Cyclones over the Korean Peninsula. *Water* **2020**, *12*, 1214. [\[CrossRef\]](#)
49. Ye, L.; Hanson, L.S.; Ding, P.; Wang, D.; Vogel, R.M. The probability distribution of daily precipitation at the point and catchment scales in the United States. *Hydrol. Earth Syst. Sci.* **2018**, *22*, 6519–6531. [\[CrossRef\]](#)
50. Calinski, T.; Harabasz, J. A dendrite method for cluster analysis. *Commun. Stat.* **1974**, *3*, 1–27. [\[CrossRef\]](#)



51. Ansari, Z.; Azeem, M.F.; Ahmed, W.; Babu, A.V. Quantitative Evaluation of Performance and Validity Indices for Clustering the Web Navigational Sessions. *World Comput. Sci. Inf. Technol. J. WCSIT* **2011**, *1*, 217–226.
52. Halkidi, M.; Batistakis, Y.; Vazirgiannis, M. On Clustering Validation Techniques. *J. Intell. Inf. Syst.* **2001**, *17*, 107–145. [[CrossRef](#)]
53. Campello, R.J.G.B.; Hruschka, E.R. A fuzzy extension of the silhouette width criterion for cluster analysis. *Fuzzy Sets Syst.* **2006**, *157*, 2858–2875. [[CrossRef](#)]
54. Liu, Y.; Zhang, X.; Chen, J.; Chao, H. A Validity Index for Fuzzy Clustering Based on Bipartite Modularity. *J. Electr. Comput. Eng.* **2019**, *2019*, 2719617. [[CrossRef](#)]
55. Zainuddin, N.H.; Lola, M.S.; Kamar, N.S. The Performance of BB-MCEWMA Model: Case Study on Normal & Non-Normal Data. *Soc. Sci. Res. J.* **2016**, *4*, 155–163.
56. Feldman, R.M.; Valdez-Flores, C. Basics of Monte Carlo Simulation. In *Applied Probability and Stochastic Processes*; Springer: Berlin/Heidelberg, Germany, 2010; pp. 45–72.
57. Peñarrocha, D.; Estrela, M.J.; Millán, M. Classification of daily rainfall patterns in a Mediterranean area with extreme intensity levels: The Valencia region. *Int. J. Clim.* **2002**, *22*, 677–695. [[CrossRef](#)]
58. Wang, F. Factor Analysis and Principal-Components Analysis. In *International Encyclopedia of Human Geography*; Elsevier: Warwick, UK, 2009; Volume 4, pp. 1–7.