SPATIAL AND TEMPORAL TORRENTIAL RAINFALL GUIDED CLUSTER PATTERN BASED ON DIMENSION REDUCTION METHODS

SHAZLYN MILLEANA SHAHARUDIN

A thesis submitted in fulfilment of the
requirements for the award of the degree of
Doctor of Philosophy (Mathematics)

Faculty of Science
Universiti Teknologi Malaysia

FEBRUARY 2017

To my beloved mother and father

My lovely husband and my children

My supportive supervisor

Lecturers and Friends

ACKNOWLEDGEMENT

First and foremost, I am very grateful to Allah The Almighty for giving me the strength to finish this project. I wish to acknowledge the extraordinary help in variety of ways I received from my supervisor, Dr. Norhaiza Ahmad for her guidance, comments and encouragement. I am really appreciating for her direction and for demonstrating her utmost patience in handling me. I have learned tremendously from her brilliance, high standards and strive for excellence, including how to choose and pursue interesting research directions and how to present them both in paper and talks. To Assoc. Prof. Dr. Fadhilah Yusof, thank you for introducing me and got me excited about the field of hydrology especially in rainfall.

I would like to extend my most sincere gratitude to my family especially to my beloved mother and father, Norasmah and Shaharudin, my lovely husband and children, Wan Azizi, Qaleesya and Hakimi for their support, encouragement and precise advice. Not forgotten to my sister, Awin and my family that help me a lot in this journey.

Besides, I would like to thank Malaysian Ministry of Higher Education for the financial support through MyPhD Scholarship.

Last but not least, I owe special thanks to all my lecturers who have taught me lots of knowledge. I am also grateful to all my friends wherever they are. Thanks for all your support and understanding.

ABSTRACT

This thesis identifies the spatial and temporal cluster patterns for torrential rainfall data in Peninsular Malaysia. Two dimension reduction methods are used to improve the cluster patterns of the torrential rainfall data. Firstly, a robust dimension reduction method in Principal Component Analysis (PCA) is used to rectify the issue of unbalanced clusters in rainfall patterns due to the skewed nature of rainfall data. A robust measure in PCA using Tukey's biweight correlation to downweigh observations is introduced and the optimum breakdown point to extract the number of components in PCA using this approach is proposed. The simulated data indicates a breakdown optimum point of 0.4 at 70% cumulative percentage of variance to give a good balance in extracting the number of components to avoid variations of low frequency or insignificant spatial scale in the clusters. The results show a significance improvement with the robust PCA than the PCA based Pearson correlation in terms of the average number of clusters obtained and its cluster quality. Secondly, based on the decomposing properties in Singular Spectrum Analysis (SSA), a two-way approach to identify the range of local time scale for a cluster of torrential rainfall pattern by discriminating the noise in a time series trend is introduced. Firstly, appropriate window length for the trajectory matrix and adjustments on the coinciding singular values obtained from the decomposed time series matrix based on a restricted singular value decomposition (SVD) using iterative oblique SSA (Iterative O-SSA) is proposed. In addition, a guided clustering method called Robust Sparse k-means (RSk-means) to discriminate the eigenvectors from this iterative procedure is suggested to identify the trend and noise components more objectively. The modified SSA indicates strongest separability between the reconstructed components based on a simulated skewed and short time series rainfall data to effectively identify the local time scale.

ABSTRAK

Tesis ini mengenalpasti ruang dan masa corak kelompok untuk data hujan lebat di Semenanjung Malaysia. Dua kaedah pengurangan dimensi yang digunakan untuk memperbaiki corak kelompok data hujan lebat. Pertama, kaedah pengurangan dimensi tahan lasak dalam Analisis Komponen Prinsipal yang digunakan untuk memperbetulkan isu kelompok yang tidak seimbang dalam corak hujan kerana sifat pencong terhadap data hujan. Pendekatan tahan lasak dalam Analisis Komponen Prinsipal menggunakan korelasi Tukey biweight untuk mengurangkan pemberatan pemerhatian diperkenalkan dan titik pecahan optimum untuk mengesktrak bilangan komponen dalam Analisis Komponen Prinsipal menggunakan pendekatan ini adalah dicadangkan. Data simulasi menggambarkan titik pecahan optimum 0.4 pada 70% peratusan kumulatif varians untuk memberi keseimbangan yang baik dalam mengekstrak bilangan komponen untuk mengelakkan variasi frekuensi rendah atau skala ruang yang tidak penting dalam kelompok. Keputusan menunjukkan penambahbaikan yang penting dengan Analisis Komponen Prinsipal tahan lasak berbanding Analisis Komponen Prinsipal yang berasaskan korelasi Pearson dari segi purata bilangan kelompok diperolehi dan kualiti kelompoknya. Kedua, berdasarkan sifat-sifat pengkomposan dalam Analisis Spektum Singular pendekatan dua hala untuk mengenalpasti julat skala masa tempatan untuk kelompok corak hujan lebat dengan membezakan ralat dalam trend siri masa diperkenalkan. Pertama, panjang tetingkap yang sesuai untuk matriks trajektori dan pengubahsuaian pada persamaan nilai tunggal diperolehi daripada penguraian siri masa matriks berdasarkan nilai penghuraian singular terbatas menggunakan Analisis Spektum Singular lelaran serong adalah dicadangkan. Sebagai tambahan, kaedah pengelompokan berpandu dipanggil *Robust Sparse k-means* (RSk-Means) untuk mendiskriminasi vektor eigen daripada prosedur lelaran ini dicadangkan untuk mengenalpasti trend dan komponen ralat dengan lebih objektif. Pengubahsuaian Analisis Spektum Singular menggambarkan sifat dapat dipisahkan paling kuat antara pembinaan semula komponen berdasarkan simulasi pencong dan siri masa data hujan pendek untuk mengenalpasti skala masa tempatan secara berkesan.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS

$\mathbf{X}$	-	Data matrix
x_{ij}	-	Elements in the input matrix, **X** i.e. rainfall amount
$\mathbf{X}^*$	-	Data matrix is standardized with median and mean absolute deviation
x_{ij}^*	-	Elements in the input matrix, $\mathbf{X}^*$ after standardized
C_{tt}	-	Correlation matrix
n, I	-	Rainfall days
p, J	-	Rainfall stations
$\mathbf{E}_t$	-	Matrix with the columns holding the eigenvectors
$\overrightarrow{e_t}$	-	Elements of eigenvector matrix
Z	-	Rank of $\mathbf{X}'$
λ	-	Eigenvalue
t_m	-	The cumulative percentage of variation in PCA
m	-	Number of components
l_k	-	The variance of the kth principal component
C_i	-	The i^{th} cluster
c_i	-	The centroid of cluster C_i
K	-	Number of clusters
CH	-	Calinski and Harabasz index
B	-	The sum of squares among the clusters
W	-	The sum of squares within the clusters
h	-	The number of principal component loadings
ρ	-	Objective function
T_n	-	M-estimator of location
ψ	-	Derivate function

S_n	-	Scale the observation
c^*	-	Tuning constant
$w(u)$	-	Weight function
$\tilde{S}$	-	Biweight estimate of shape
$\tilde{s}_{ij}$	-	Element of $\tilde{S}$
$\tilde{r}_{ij}$	-	The biweight correlation of $\mathbf{X}^*$
P	-	Partition of the data
SS	-	Both data points belong to the same cluster of the clustering structure C and to the same group of partition P
SD	-	Data points belong to the same cluster of C and to different groups of P
DS	-	Data points belong to different clusters of C and to the same group of P
DD	-	Both points belong to different clusters of C and to different groups of P
M	-	The maximum number of all pairs in the data set
N	-	The total number of points in the data set
R	-	Rand index
$S(i)$	-	Silhouette index
a_i	-	The average distance between data points, i and all other data points in the same cluster
b_i	-	The average distance between i and the data points in the nearest neighbouring cluster
DB_{nc}	-	Davies-Bouldin index
d_{ij}	-	Dissimilarity measure between two clusters
R_{ij}	-	Similarity measure of clusters
s_i	-	Dispersion measure of a cluster
α	-	Significance level
PCA	-	Principal component analysis
GPD	-	Generalized Pareto distribution
r	-	Breakdown point
RP	-	Rainfall pattern

SW	-	Southwest
NE	-	Northeast
SSA	-	Singular Spectrum Analysis
SVD	-	Singular Value Decomposition
L	-	Window length
$\mathbb{Y}_T$	-	A one dimensional time series
$\mathbb{X}_i$	-	A multi-dimensional time series
Σ	-	Diagonal matrix
σ_i	-	Singular values
u_i	-	Left singular vectors
v_i	-	Right singular vectors
$\mathbf{X}_I$	-	Resultant matrix
ρ^w	-	w-correlation
RSVD	-	Restricted SVD
†	-	Pseudo-inverse
RSk-means	-	Robust Sparse k-Means

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Introduction

Patterns of change in extreme rainfall events vary with region and time. This is parallel with many scientific studies relating to the physics of a warming climate. In extreme rainfall events, these changes are termed as spatial and temporal variations which exist in the observed torrential rainfall data. These variations can be used to display similar characteristics and behavior of structured spatial rainfall patterns. From the identified cluster, variations in the time series observations can be decomposed and reconstructed to locate the time period in which the extreme rainfall events occur. Thus, the identification of spatial and temporal cluster rainfall patterns is useful for hydrologist in analyzing environmental models and improves assessment on climate change.

1.2 Background of Problem

Rainfall in Malaysia is strongly influenced by three monsoon seasons classified as the Northeast (NE), Southwest (SW) and intermonsoon seasons. The SW monsoon season occurs annually from May to September, NE monsoon between November to March and intermonsoon occurs in the months of April and October.

Figure 1.1 illustrates the prevailing wind flows of monsoon seasons in Malaysia. Monsoons bring high volumes moisture to various parts of Malaysia due to her location in the equatorial zone with a tropical climate. As a result, the total amount of rainfall in Malaysia is rather high, between 2000 mm to 4000 mm, with 150 to 200 rainy days annually (Suhaila and Jemain, 2007).

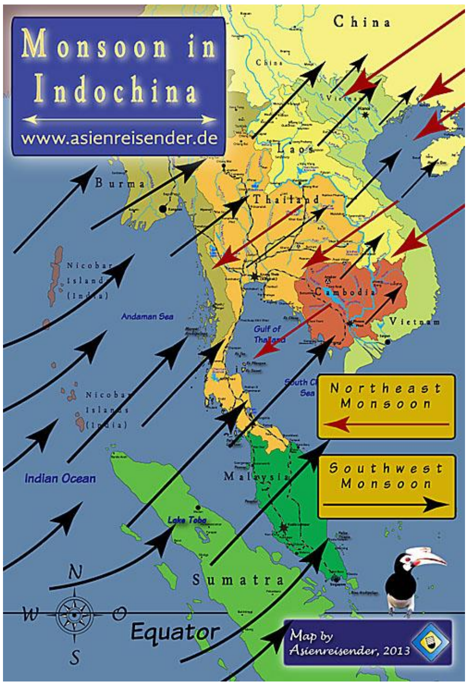

Figure 1.1 Monsoon seasons in Peninsular Malaysia (Map from http://www.asienreisender.de/)

One particularly relevant feature of the rainfall regime in Malaysia is the occurrence of episodes of rains of torrential character which has potential to become hazardous in Malaysia. For instance, consecutive waves of flash floods hitting the southern and south-western coasts of Peninsular Malaysia from 19-27 December 2006 and 13-20 January 2006 (International Federation of Red Cross and Red Crescent Societies, 2007). The more recent disaster on 28 December 2014 where massive flood hit Kelantan was the worst in the history in Peninsular Malaysia (http://www.themalaymailonline.com/). For that reason, rainfall pattern in Malaysia is analysed by meteorologists with focus on events of heavy rainfalls. This is a very significant issue and the results of such studies could be used as a guide to climatologist or hydrologists to recommend actions in mitigating the flood damages and taking necessary precautions when they happen.

In general, the study of spatial and temporal rainfall patterns in hydrology use two approaches which are regression based modelling and clustering based approach. Studies related to regression based modelling include Xu et al.(2014), Kaliraj et al.(2012), Koumare (2014), Lee (2015), Szyniszewska and Waylen (2012) and Mohd Deni et al.(2008) which generally aims to characterize the rainfall distribution patterns. Identifying spatial and temporal rainfall patterns using this approach is more focused on detecting trend compared to describe the regional characteristics of each pattern for rainfall data. It is because the results from this approach will be used to accomplish the goal of forecasting. However, studies related to clustering based approach in identifying spatial and temporal rainfall patterns is to quantify the characteristics of a set of observations that place data into same groups which imply that rainfall patterns are similarly highly structured (Romero et al., 1999). Furthermore, this approach tends to recognized as efficient statistical tool to deal with tasks for grouping regions and determine time periods based on grouping results that reflect the occurrence of rainfall events.

For instance, Kansakar et al.(2004), characterized the nature of precipitation regimes across Nepal and helps to identify the key controlling factors for spatial patterns in seasonal precipitation behavior. SaeedSoltani and Modarres (2006), used hierarchical and divisive cluster analysis to classify rainfall spatial time distribution pattern and also to evaluate annual and seasonal temporal pattern over Iran. Another study with similar objectives include Penarrocha et al.(2002) which identified the spatial distribution patterns of heavy rainfall in the Valencia region, Spain. Most of the literatures concerning these studies are related to subtropical climate. The characteristics of the rainfall pattern in these areas are typically highly varied which depends on four seasons that occurred every year in that country. Thus, such rainfall patterns can be easily distinct based on standard classification methods such as clustering.

However, standard clustering methods to find spatial and temporal rainfall patterns might not be suitable for tropical climate due to several characteristics of the rainfall data in the tropics. One, regions in the tropics generally experience rainfall throughout the year. Thus, a long time series of observed rainfall data tend to produce a highly dimensional data set. This leads to the data becoming complex to extract significant information as it may contain high degree of irrelevant and redundant information that could degrade the results of the analysis. Two, even if tropical regions have two dry and wet seasons, the amount of rainfall does not significantly vary much as compared to that in the four season regions. Thus, this makes it difficult to discern a particular cluster pattern for this type of rainfall data. Three, such large data set of recorded rainfall data are bound to contain noise in the rainfall measurements. This noise could be an intrinsic error structure due to either technical or human recording error.

In essence, identifying rainfall patterns under the circumstances outlined above can be daunting task. Therefore, it is essential that any analysis should take into account as many of these factors as possible in order to identify the spatial and temporal cluster of rainfall patterns.

In this thesis, we present two statistical strategies based on a reduction dimension approach in identifying the spatial and temporal torrential rainfall patterns in Peninsular Malaysia by considering the issues mentioned above. The general strategies are namely (i) providing guided cluster rainfall pattern associated with torrential rainfall events based on region (ii) effectively guide the identification of local time scale indicated at the peak occurrence of temporal torrential rainfall.

This chapter begin with an overview of rainfall data in Peninsular Malaysia and its descriptive statistics. Then, problem statements will be discussed in this chapter due to several issues that highlighted in this thesis. Next, aim of thesis also described here according to the problem statement that discussed before. Followed by the significance of the study and notations are included in this chapter. Finally, the synopsis of the thesis ends this chapter.

1.3 Rainfall Data in Peninsular Malaysia

Malaysia consists of two noncontiguous areas, Peninsular Malaysia (West Malaysia) and the states of Sabah and Sarawak, known together as East Malaysia, with a total area of 329,750 sq km. This study of rainfall focuses on the long narrow Peninsula Malaysia that covers from latitude $1°20'$ north to latitude $6°40'$ north, and from longitude $99°35'$ east to longitude $104°20°$ east. The region is chosen based on the length, reliability and quality of daily rainfall data covering more than 30 years.

The initial daily rainfall data from 75 stations over Peninsular Malaysia were obtained from Jabatan Pengairan dan Saliran (JPS) which measure rainfall using bucket rain gauge as shown in Figure 1.2. The rainfall data is complete without missing values from year 1975 to 2007 with a total of 903,375 daily measurements. Within those years, the data have 8 leap days, the data of which are excluded for this study. The rainfall data set considered for the purpose of this study is taken from 75 stations and 12,045 days which constitute enough data to allow for the identification of the main rainfall patterns. Rainfall stations are located at different geographical coordinates all over Peninsular Malaysia on four regions, East, Southwest, West and Northwest. Figure 1.3 shows the locations of all 75 stations indicated by the letters N,S,W,E reflecting the four region followed by a number. Detail information regarding these rainfall stations can be referred to in Appendix A.

The total annual rainfall statistics from 1975 to 2007 at 75 rainfall stations shows that the study area receives an average annual rainfall of 2327.3 mm, with a standard deviation of 442.2 mm. Figure 1.4 shows the plot of total annual rainfall from 1975 to 2007 against rainfall station. The rainfall stations are numbered as listed in Table A.1 in Appendix A. The red lined show the average annual rainfall of 2300 mm. It is observed that 37 stations from 75 stations receive more than the average rainfall yearly. These stations are found to be located in the eastern region of Peninsular Malaysia.

Figure 1.2 A tipping bucket rain gauge has a receiving funnel leading to two small metal collectors (buckets) and the maximum rainfall rate is 200 mm/hr for the funnel

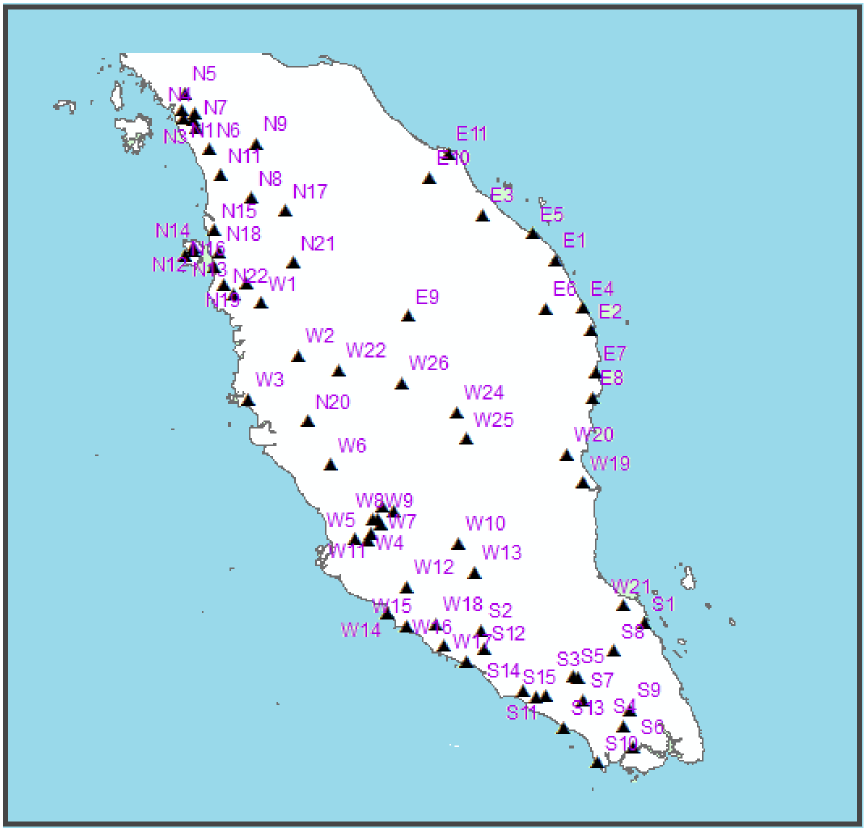

Figure 1.3 The location of 75 rainfall stations in Peninsular Malaysia

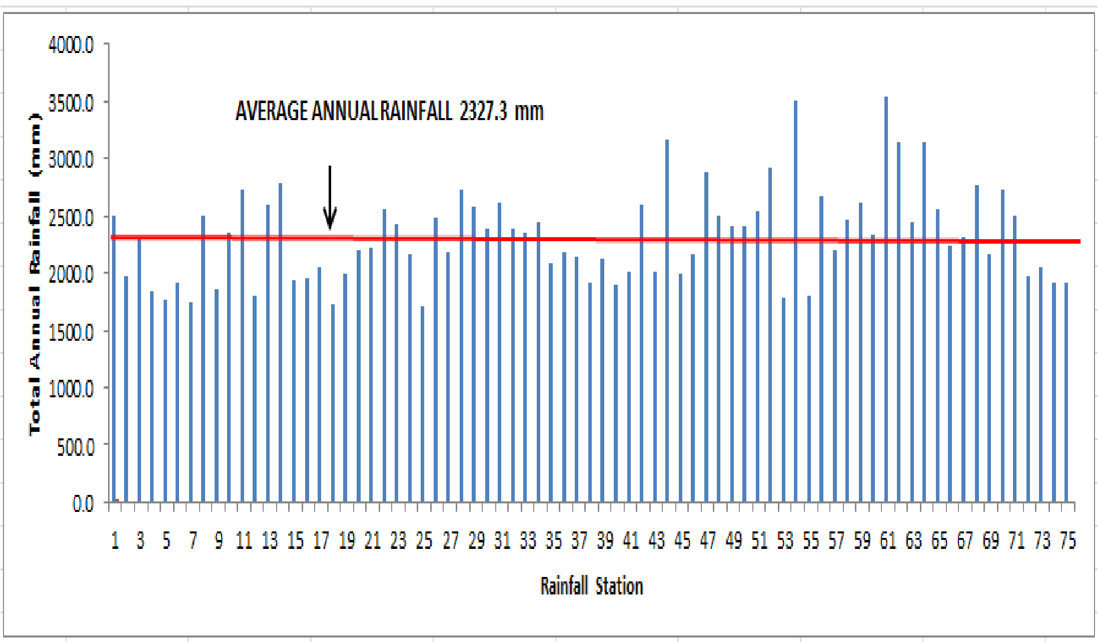

Figure 1.4 Total annual rainfall at 75 rainfall stations from 1975-2007

1.3.1 Database for Torrential Rainfall Data

In this study, we focus on the occurrence of extreme rainfall event described as torrential rainfall. It was therefore necessary to choose some criteria that would lead to the establishment of a threshold, in order to allow for a clear distinction between what constitutes a day of torrential rainfall in the Peninsular Malaysia region and what does not. The range of threshold for torrential rainfall data in Peninsular Malaysia is 60 mm/day. This threshold is chosen based on the categorization of rainfall intensity by Jabatan Pengairan dan Saliran (JPS). By filtering days with rainfall more than 60 mm for at least 2% of overall stations, we managed to obtain 250 days and 15 rainfall stations which in turn are suffice enough to represent the main torrential centers.

Figure 1.6 shows the matrix of daily torrential rainfall data after filtering from raw data based on the threshold that set to the data. The rainfall day in the first column refers to the rainfall observation and the rainfall station in the first row refers to the variable. Rainfall is often expressed in millimeters per day (mm/day) which represents the total depth of rainwater (mm), during 24 hours. It appears that the locations where these torrential rainfall occur and largely located at two regions specifically the Northern and Eastern region. As shown in Figure 1.5 and Table 1.1, we note that these locations are subsets of the rainfall stations that receive more than the average annual rainfall in Table A.1 in Appendix A. This is expected since north-easterly monsoon wind tends to bring heavy rainfall to these locations (Jamaludin et al., 2010).

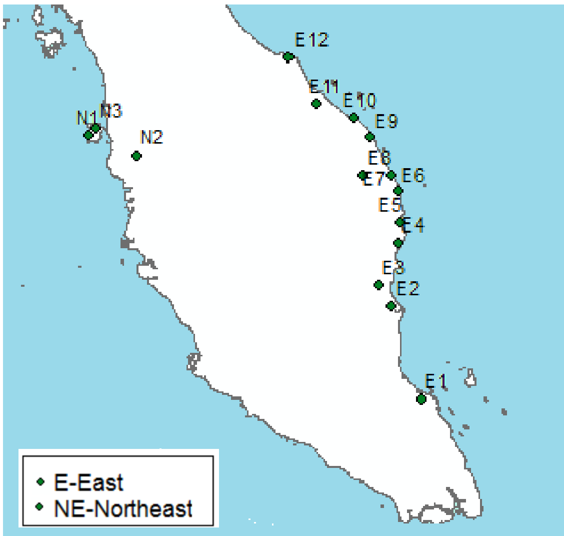

Figure 1.5 Rainfall stations that represent the main torrential centers in Peninsular Malaysia

Table 1.1: List of the rainfall stations according the monsoon occurred

Region	Station	Code
Northeast	PintuA.Bagan,AirItam	N1
	Selama	N2
	KlinikBkt. Bendera	N3
East	KlinikBidan ,JambuBongkok	E7
	Sek. Keb. Kemasek	E5
	Sek. Keb. Kg. Jabi	E11
	Kg. Merang ,Setiu	E10
	Endau	E1
	Rumah Pam Pahang Tua,Pekan	E2
	Kuantan	E3
	JPS Kemaman	E4
	Sek.Men. Sultan Omar, Dungun	E6
	Kg. Menerong	E8
	Stor JPS Kuala Trengganu	E9
	Kota Bharu	E12

Rainfall day/ Rainfall station	Pintu A.Bagan	Klinik Bidan,Jambu Bongkok	Sek.Keb.Kemasek	Sek.Keb.Kg.Jabi	Kg.Merang,Setiu	Endau	Rumah Pam Pahang Tua,Pekan	Kuantan
1	96.0	126.0	75.0	62.0	189.5	60.0	132.2	131.3
2	122.5	270.0	215.5	81.5	75.0	75.0	81.8	240.5
3	81.0	167.5	140.5	243.0	97.0	90.0	70.0	129.2
4	95.5	65.0	118.5	220.5	73.5	86.0	444.8	164.5
5	89.0	92.0	92.5	78.0	82.5	67.5	60.6	72.2
6	116.0	61.5	87.0	95.0	60.5	65.0	148.8	72.3
7	79.0	131.0	69.5	68.5	65.0	60.0	77.9	89.8
8	180.0	145.0	122.5	60.0	93.5	79.0	62.1	68.0
9	63.5	91.5	96.5	103.5	60.5	200.0	160.4	66.3
10	116.0	138.0	77.0	73.5	65.0	82.5	104.8	62.5
11	73.0	95.5	88.5	64.0	72.5	111.0	161.8	73.6
12	79.0	153.0	84.6	61.0	150.0	112.5	175.0	114.1
13	73.0	139.0	135.0	78.5	134.0	88.0	101.6	63.0
14	83.0	80.0	90.1	127.0	72.0	84.0	95.2	80.8
15	133.5	198.0	256.6	75.0	89.0	65.0	142.7	527.5
16	89.0	123.0	82.9	83.0	61.5	60.0	107.5	62.2
17	105.0	90.5	81.6	87.5	175.0	65.0	69.4	109.0
18	104.0	126.5	101.2	72.0	233.0	69.0	65.5	62.4
19	81.0	92.0	103.8	155.0	83.5	76.0	81.5	68.5
20	66.5	121.0	121.5	133.0	71.5	81.0	167.9	124.2
21	81.0	94.5	80.0	70.0	83.0	86.0	108.0	111.5
22	104.0	60.0	79.5	61.5	123.0	100.5	61.0	109.7
23	72.5	60.0	68.9	71.0	135.0	90.0	124.0	79.2

Figure 1.6 An example of a snapshot of the daily torrential rainfall data consisting of daily amount of rainfall data recorded at several locations

1.3.2 Descriptive Statistics for Torrential Rainfall Data

Data preparation is one of the important step before proceed to further analysis where this step is involving data screening, data evaluation, selection of records, operations to bring the number of variables to manageable range. Descriptive statistics of torrential rainfall data such as mean, standard deviation and skewness as shown in Table 1.2 are calculated to provide a brief overview of the torrential rainfall data on the study area.

From the Table 1.2, we can illustrate that the highest average mean and average standard deviation of daily torrential rainfall amount in Peninsular Malaysia is located in East region. We observed that, there are differences in the coefficient of variation between regions. The stations at the East region show the largest variability of torrential rainfall amounts, which range from 46% to 70%. The lowest coefficient variation is found in the Northeast station with variation less than 39%. Northeast and East region give positive skewness with starting value from 1.8 mm to 5.8 mm. The results illustrate that the shape of rainfall distribution for the stations in this two regions is skewed due to the values of the skewness are far enough from zero. We can conclude that the daily torrential rainfall data probably did not come from a normal population.

1.4 Problem Statement

Identify daily spatial and temporal of torrential rainfall patterns is not an easy task due to several issues.

- The daily torrential rainfall data are of high dimensions and are difficult to visualize and interpret.
  However, some variables are particularly significant and need to be identified in order to convey important information to an understanding of the underlying process. For this purpose, we need a statistical approach to reduce the dimensionality of the data.

Table 1.2 : Summary statistics of daily torrential rainfall amount (mm) for each station divided by regions

Region	Station	Value					
		Mean	Standard Deviation	Min	Max	Skewness	Coefficient Variation (%)
Northern	Pintu A.Bagan	85.0	26.8	60.0	245.5	2.6	32
	Klinik Bkt. Bendera	88.5	34.5	60.0	299.0	5.8	39
	Selama	78.6	18.4	60.0	176.0	1.8	23
Eastern	Sek.Keb. Kemasek	107.6	52.6	60.0	321.0	1.8	49
	Kg.Merang,Setiu	113.8	60.6	60.0	365.5	1.9	53
	Sek.Men. Sultan Omar, Dungun	104.0	53.3	60.0	572.0	3.7	51
	Kg. Menerong	109.4	76.8	60.0	676.0	4.0	70
	Sek.Keb.Kg.Jabi	102.0	46.7	60.0	329.5	1.8	46
	JPS Kemaman	109.0	58.2	60.0	434.0	2.5	53
	Stor JPS Kuala Terengganu	118.1	68.4	60.0	520.4	2.4	58
	Klinik Bidan, Jambu Bongkok	111.4	75.7	60.0	790.0	2.7	68
	Kota Bharu	115.9	73.4	60.0	591.5	3.0	63
	Endau	107.4	52.5	60.0	353.5	1.8	49
	Rumah Pam Pahang Tua,Pekan	114.0	63.3	60.4	444.8	2.4	56
	Kuantan	106.6	59.6	60.0	527.5	3.0	56

- Besides that, the amount of rainfall does not significantly vary much as compared to that in the four season regions of subtropical climate. This leads to difficulty to discern a particular cluster pattern for this type of rainfall data. In study of identifying daily spatial and temporal torrential rainfall patterns, we need more clusters which could explain the various types of rainfall patterns in region where each pattern exhibits specific characteristics. Thus, a much robust procedure is essential to improve the cluster patterns.

- Another limiting factor of identifying daily spatial and temporal torrential rainfall data is that the data contain noise. Noise in rainfall data is defined as the observations are never completely accurate due to either technical or human recording error. When there are noises in the data, statistical approach that used to analyze the data set would yield poor accuracies of the results. Thus, we need a statistical approach to separate noise from rainfall data to make them more readily observable.

1.5 Objectives

This thesis aims to deal with torrential rainfall data in Peninsular Malaysia: (i) providing guided rainfall clusters based on a robust dimension reduction method (ii) decomposing and reconstructing the time series components in the data to locate the time period of occurrence in rainfall events.

Specifically, we want to be able to:

a) identify daily spatial and temporal of torrential rainfall patterns in Peninsular Malaysia using dimension reduction methods which is Principal Component Analysis (PCA) and Singular Spectrum Analysis (SSA).

b) provide a robust procedure to improve the cluster partitions of a PCA guided clustering in identifying rainfall patterns.

c) propose modification of SSA to locate the range of time period from the extracted trend components that is free from noise, at a particular location of torrential rainfall events.

1.6 Scope of the Study

This study deals with long term tropical climate daily rainfall data focusing on torrential events (more than 60mm per day). The original data set was obtained from the Malaysian Department of Irrigation and Drainage, *Jabatan Pengairan dan Saliran (JPS)* consisting of 75 rainfall stations in Peninsular Malaysia and 30 years of daily rainfall records from 1975 until 2007 with no missing values. Two main approaches were used in this study: Principal Component Analysis (PCA) and Singular Spectrum Analysis (SSA). Programming codes and subroutine were carried out using R statistical software.

1.7 Significance of the Study

This study proposes to:

a) establish the pattern of daily torrential rainfall patterns in Peninsular Malaysia using clustering method, where each cluster is identified to be linked to certain topographic characteristics.

b) identify local time scale to determine when torrential rainfall events occurred at a particular location.

c) establish a novel approach in multivariate technique for certain hydrologic applications especially in tropical climate.

d) help in further analysis and development of appropriate models for prediction of torrential rainfall events over Peninsular Malaysia.

1.8 Notations

To ease computation process, the database is set up to take into form of a large rectangular n rows by p columns matrix **X**, with $n > p$. We denote $x_{ij} \in \mathbf{X}$ to be the rainfall amount for each i^{th} at each j^{th} rainfall station where $i = 1, \dots, n,\ j = 1, \dots, p$.

Throughout this thesis, we use interchangeably the terms ‘rows’ of the data matrix as ‘rainfall observations’ , ‘rainfall days’, ‘elements’ or ‘data points’, ‘columns’ of the data matrix as the ‘rainfall station’.

1.9 Research Methodology

Briefly, this thesis focuses on two statistical strategies in identifying spatial and torrential rainfall patterns in Peninsular Malaysia, namely

(i) providing guided cluster rainfall pattern associated with torrential rainfall events based on region

(ii) effectively guide the identification of local time scale indicated at the peak occurrence of temporal torrential rainfall

The research methodology is as in Figure 1.7. In order to achieve these two statistical strategies, Principal Component Analysis (PCA) and Singular Spectrum Analysis (SSA) are used in this study. These approaches to find spatial and temporal rainfall patterns might not be suitable for tropical climate due to several issues of the rainfall data in the tropics. To counter these issues, we proposed a robust approach in PCA and SSA that could be applied in torrential rainfall data in Peninsular Malaysia. In addition, the performance of the modification of the PCA and SSA are evaluated using simulation.

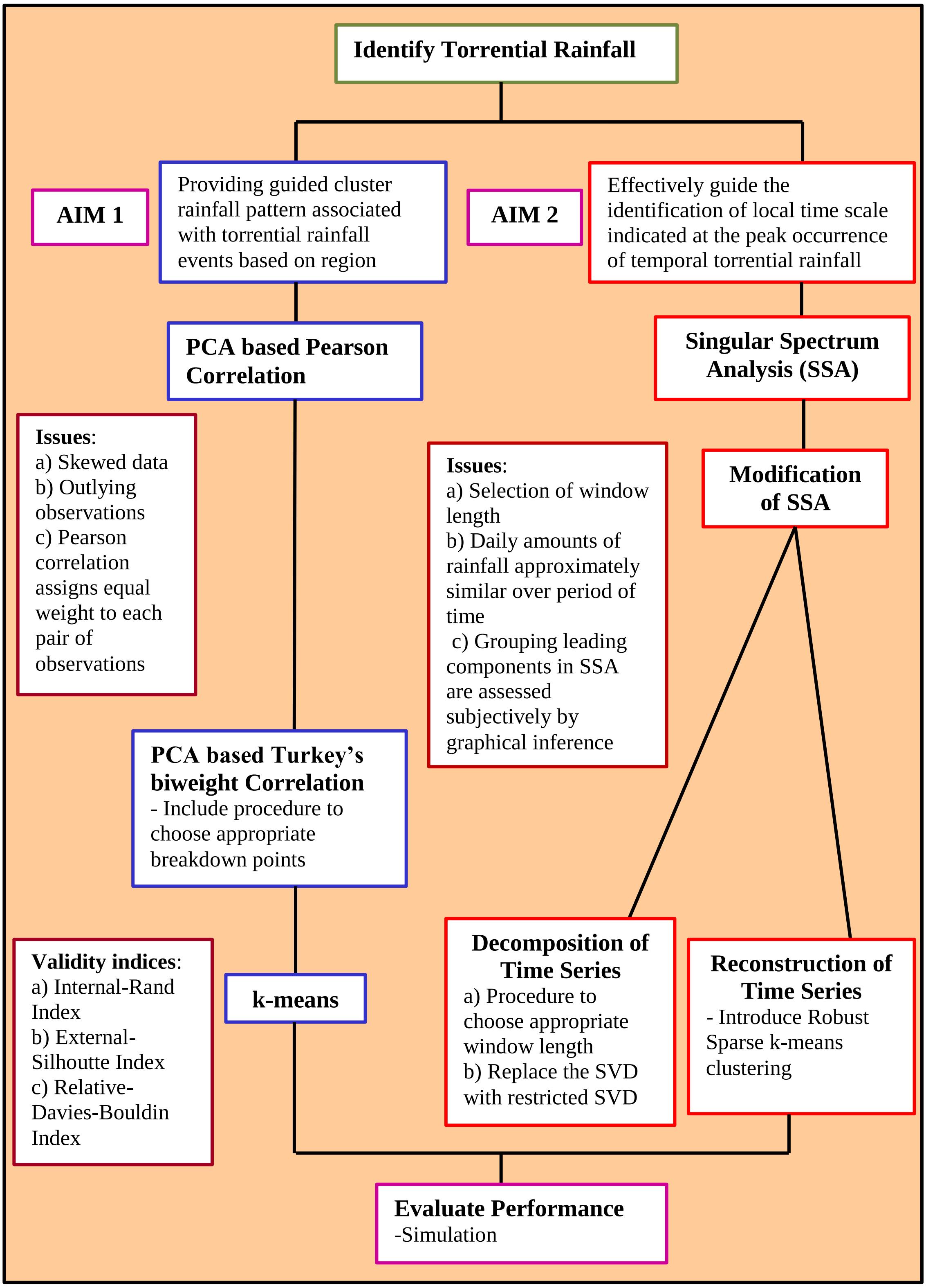

Figure 1.7 Flow Chart of Research Methodology

1.10 Thesis Organization

Chapter 1 introduces the background of the problem and the related issues to identify the spatial and temporal daily torrential rainfall patterns in Peninsular Malaysia. It also presents the description of the study area and describes the specific dataset that were used in this study. In addition, descriptive of daily torrential rainfall statistics, determination of threshold for rainfall data, scope of the study, significance of the study, notations and research methodology also presented in this chapter. In essence, this chapter provides a general overview of the thesis. Furthermore in Chapter 2, the existing works related to the area of research and the methods related to this study are reviewed. The chapter begins with identification of rainfall pattern using multivariate analysis where we focus on past literature that have similar objectives and which employ the same method to the data analysis. Then, we proceed to the next section which focuses on robust methods to determine the robust procedure suitable for use in our data set.

In Chapter 3, a novel approach to identify the spatial daily torrential rainfall patterns in Peninsular Malaysia is discussed. We clarify the issues related to the high dimensional data which explain the typical methods used in hydrology area and from there, we propose a robust method that is more appropriate to our data set. Simulation procedure is also included in this chapter to assess the performance of the robust PCA. Afterwards, Chapter 4 present and discuss the results for original torrential rainfall data in identifying the spatial torrential rainfall patterns using robust PCA. In this chapter, we also compare the results from PCA and robust PCA. We also present and discuss the results of simulated data for assessing the performance robust PCA. The results of simulated data could be used as guide to analyze the original torrential rainfall data.

Later, we provide a modification of SSA in order to make an adjustment for eigen time series torrential rainfall data obtained from SVD in SSA and to introduce a statistical measure to group the decompose eigenvector appropriately in Chapter 5.

Simulation procedure is also included in this chapter to assess the performance of the modification of SSA. Further, Chapter 6 is present and discusses the results for original torrential rainfall data in identifying the temporal torrential rainfall patterns using modified SSA. This chapter also compared the results from SSA and modified SSA. We also present and discuss the results of simulated data for assessing the performance of modified SSA. The results of simulated data could be used as guide to analyze the original time series of the torrential rainfall data.

Finally, Chapter 7 conclude the results and discussions of all the problems investigated in this thesis. This chapter is also discusses future research that may be conducted for deeper understanding of the problems considered.

APPENDIX B

From this research, some results have been published/accepted/submitted in journals and presented in the international conferences as listed in the following:

B1 Journal

Shaharudin, S.M., Ahmad, N., and Yusof, F. (2013). The Comparison of T-Mode and Pearson Correlation Matrices in Classification of Daily Rainfall Patterns in Peninsular Malaysia. *Matematika*. 29(1c), 187-194.

Shaharudin, S.M., Ahmad, N., and Yusof, F. (2013). Improved Cluster Partition in Principal Component Analysis Guided Clustering. *International Journal of Computer Applications*. 75(11), 22-25.

B2 Conference Proceedings (Scopus)

Shaharudin, S.M., Ahmad, N., and Yusof, F. (2015). Effect of Window Length with Singular Spectrum Analysis in Extracting the Trend Signal on Rainfall Data. *Proceedings of the International Statistical Conference (ISM,2014).* 12-14 August. Kuantan, Pahang, 321-326.

B3 Conference Proceedings

Shaharudin, S.M., Ahmad, N., and Yusof, F. (2013). Choice of Percentage of Cumulative in Principal Component Analysis to Define Region. *Proceedings of ICOWOBAS-RAFSS*. 3-5 September. Johor Bahru.

Shaharudin, S.M., Ahmad, N., and Yusof, F. (2015). Patterns of Daily Torrential Monsoon Rainfall in Peninsular Malaysia based on a Robust Correlation Measure. *Proceedings of the International Conference on Applied Analysis and Mathematical Modeling*. 8-12 June. Istanbul, Turki.

Shaharudin, S.M., Ahmad, N., and Yusof, F. (2016). Classification of Daily Torrential Rainfall Patterns in Peninsular Malaysia based on a Robust Correlation Measure. *Proceedings of the International Conference on Computational Physics, Mathematics and It's Application*. 7-8 November. Tokyo, Jepun.

Khaleghi, M., Zeinivand, H., and Moradipour, S. (2014). Rainfall and River Discharge Trend Analysis: A Case Study of Jajrood Watershed, Iran. International Bulletin of Water Resources and Development. 2(3).

Khan, M.A.R, and Poskitt, D.S. (2013). Moment tests for Window Length Selection in Singular Spectrum Analysis of Short- and Long-Memory Processes. *Journal of Time Series Analysis*, 34(2), 141-155.

Khan, M.A.R., and Poskitt, D.S(2013). A Note on Window Length Selection In Singular Spectrum Analysis. Aust. New Zeal. *J. Stat*. 55(2), 87-108.

Khan, M.A.R., and Poskitt, D.S. (2011). *Window Length Selection and Signal-Noise Separation and Reconstruction in Singular Spectrum Analysis* .Working Paper, Monash University.

Kondo, Y. (2009). *Robustification of the Sparse K-Means Clustering Algorithm*. Master. The University of British Columbia.

Kondo, Y., Salibian-Barrera, M., and Zamar, R. (2012). A Robust and Sparse K-means Clustering Algortihm. arXiv:1201.6082v1.

Koumare, I. (2014). Temporal/Spatial Distribution of Rainfall and the Associated Circulation Anomalies over West Africa. *Pakistan Journal of Meteorology*. 10(20), 1-11.

Kovacs, F., Legany, C., and Babos, A. (2005). Cluster Validity Measurement Techniques. *6th International Symposium of Hungarian Researchers on Computational Intelligence*.

Lee, H.S. (2015). General Rainfall Patterns in Indonesia and the Potential Impacts of Local Seas on Rainfall Intensity Water. 7, 1751-1768.

Li, G., and Chen, Z. (1985). Projection-Pursuit Approach to Robust Dispersion Matricesand Principal Components: Primary theory and Monte Carlo. *Journal of the American Statistical Association*. 80 (391), 759-766.

Liew, A.W.-C., and Yan, H. (2009). Reliable Detection of Short Periodic Gene Expression Time Series Profiles in DNA Microarray Data. *Proceedings of the 2009 IEEE International Conference on Systems, Man, and Cybernetics*. October. San Antonio, USA, 4274-4279.

Lisi, F., Nicolis, O., and Sandri, M. (1995). Combining Singular-Spectrum Analysis and Neural Networks for Time Series Forecasting. *Neural Processing Letters*. 2(4), 6-10.

Marques, C.A.F, Ferreira, J.A., Rocha, A., Castanheira, J.A., Melo-Goncalves, P., Vaz, N., and Dias, J.M. (2006). Singular Spectrum Ananlysis and Forecasting of Hydrological Time Series, *Physics and Chemistry of the Earth*. 31(18), 1172-1179.

Maslennikova, Y.S., and Bochkarev, V.V. (2012). Training Algorithm for Neuro-Fuzzy Based on Singular Spectrum Analysis. *World Conference on Information Technology*. 14-16 November. Barcelona, Spain, 606-610.

Meng, J., and Yang, Y. (2012). Symmetrical Two-Dimensional PCA with Image Measures in Face Recognition. *Int J Adv Robotic Sy*. 9.

Mierlus-Mazilu (2010).On Generalized Pareto Distributions. *Romanian Journal of Economic Forecasting*. 8, 107-117.

Mika, J., Kim, B.-J., and Park, J.-K.(2006). Classification of Daily Precipitation Patterns in South Korea using Multivariate Statistical Methods. *Journal of the Environmental Sciences*. 1125-1139.

Mills, G.F. (1995). Principal Component Analysis of Precipitation and Rainfall Regionalization in Spain.*Theoretical and Applied Climatology*. 50, 169-183.

Mimmack, G.M., Mason S.J., and Galpin, J.S. (2002). Choice of Distance Matrices in Cluster Analysis : Defining Regions. *Journal of Climate*. 14, 2790-2797.

Moller, S.F., Frese, J.V., and Bro, R. (2005). Robust Methods for Multivariate Data Analysis. *Journal of Chemometrics*. 19, 549-563.

Mondal, A., Kundu, S., and Mukhopadhyay, A. (2012). Rainfall Trend Analysis by Mann-Kendall Test: A Case Study of North-Eastern Part of Cuttack District, Orissa. *International Journal of Geology, Earth and Environmental Sciences*. 2(1), 70-78.

Montfort, M.A.J.V., and Witter, J.V. (1986). The Generalized Pareto Distribution Applied to Rainfall Dephts. *Hydrological Sciences Journal*. 31(2), 151-162.

Moron, V., Robertson, A. W., Qian, J.-H., and Ghil, M. (2015). Weather Types Across the Maritime Continent: from the Diurnal Cycle to Interannual Variations. *Frontiers in Environmental Science*. 3(44).

Munoz-Diaz, D., and Rodrigo, F.S. (2004). Spatio-Temporal Patterns of Seasonal Rainfall in Spain (1912-2000) using Cluster and Principal Component Analysis: Comparison. *Annales Geophysicae*. 22, 1435-1448.

Nathan, R., and Weinmann, E. (2013). *Australian Rainfall and Runoff Discussion Paper: Monte Carlo Simulation Techniques*. Australia: Engineering House.

Neal, R.A., and Phillips, I.D. (2009). Summer Daily Precipitation Variability over the East Anglian Region of Great Britain. *International Journal of Climatology*. 29, 1661-1679.

Nojumuddin, N.S., Yusof, F., and Yusof, F., and Yusop, Z. (2015). Identification of Rainfall Patterns in Johor. *Applied Mathematical Sciences*. 9(38), 1869-1888.

Norman, C. (1987). *Analyzing Multivariate Data*. SanDiego:Harcourt Brace Jovanovich.

Owen, M. (2010).*Tukey'sBiweight Correlation and the Breakdown*. Thesis.Pomona College.

Pelczer, I.J., and Cisneros-Iturbe, H.L. (2008). Identification of Rainfall Patterns over the Valley of Mexico.*International Conference on Urban Drainage*. Edinburgh, Scotland, 1-9.

Penarrocha, D., Estrela, M.J., and Millan, M. (2002). Classification of Daily Rainfall Patterns in a Mediterranean Area with Extreme Intensity Levels: The Valencia Region. *Internation Journal of Climatology*.22,677-695.

Priya, R.L., and Manimannan, G. (2014). Rainfall Fluctuation and Region Wise Classification in Tamilnadu: Using Geographical Information System. *IOSR Journal of Mathematics*. 10(5), 5-12.

Ramos, M.C. (2001). Divisive and Hierarchical Clustering Techniques to Analyse Variability of Rainfall Distribution Patterns in A Mediterranean Region. *Atmospheric Research*. 57(2), 123-138.

Rexstad, E.A., Miller, D.D., Flather, C.H., Anderson, E.M., Hupp, J.W., and Anderson, D.R.(1988).Questionable Multivariate Statistical Inference in Wildlife Habitat and Community Studies *Journal of Wildlife Management*. 52, 794-798.

Richman, M.B.(1986). Rotation of Principal Components. *J.Climatol*. 6, 293-335.

Rizwan, M., and Kim, T.-W. (2013). Application of a Mixed Gumbel Distribution to Construct Rainfall Depth-Duration-Frequency (DDF) Curves Considering Outlier Effect in Hydrologic Data. *IOSR Journal of Environmental Science*. 6 (2), 54-60.

Rodriguez-Aragon, L.J., Zhiglkavsky, A.(2010). Singular Spectrum Analysis for Image Processiong. *Statistics and Its Interface*. 3, 419-426.

Romero, R., Ramis, C., and Guijarro, J.A. (1999). Daily Rainfall Patterns in the Spanish Mediterranean Area: An Objective Classification. *International Journal of Climatology*.19, 95-112.

Romero, R., Summer, G., Ramis, C., and Genoves, A. (1999). A Classification of the Atmospheric Circulation Patterns Producing Significant Daily Rainfall in the Spanish Mediterranean Area. *Int. J. Climatol*. 19, 765-785.

Rousseeuw, P., and Leroy, A. (1987). *Robust Regression and Outlier Detection*. New York, USA: John Wiley and Sons, Inc.

Rousseeuw, P., and Van Driessen, K. (1999). A Fast Algorithm for the Minimum Covariance Determinant Estimator. *Technometrics*.41, 212–223.

Rousseeuw, P.J. (1985). Multivariate Estimation with High Breakdown Point. *Mathematical Statistics and Applications*.

SaeedSoltani*, and Modarres, R.(2006). Classification of Spatio -Temporal Pattern of Rainfall in Iran using A Hierarchical and Divisive Cluster Analysis. *Journal of Spatial Hydrology*. (2), 1-12.

Siva, G.S., Rao, V.S., and Babu, D.R. (2014). Cluster Analysis Approach to Study the Rainfall Pattern in Visakhapatnam District. *Weekly Science Research Journal*. 1(31).

Stathis, D., and Myronidis, D. (2009). Principal Component Analysis of Precipitation in Thessaly Region (Central Greece). *Global NEST Journal*. 11(4), 467-476.

Suhaila, J., and Jemain, A. A. (2007). Fitting Daily Rainfall Amount in Malaysia using the Normal Transform Distribution. *Journal of Applied Sciences*. 7(14), 1880-1886.

Sumner, G. (1996). Daily precipitation patterns over Wales: Towards a Detailed Precipitation Climatology. *Transactions of the Institute of British Geographer*. 21(1), 157–176.

Sumner, G., Guijarro. J.A., and Ramis, C. (1995). The Impact of Surface Circulations on the Daily Rainfall over Mallorca. *International Journal of Climatology*. 15, 673–696.

Szyniszweska, A.M., and Waylen, P.R. (2012). Determining the Daily Rainfall Charateristics from the Monthly Rainfall Totals in Central and Northeastern Thailand. *Applied Geography*. 35, 377-393.

Teodoro, P.E., O.-Junior, J.F.D., Cunha, E.R.D., Correa, C.C.G.C, Torres, F.E., Bacani, V.M., Gois, G., and Ribeiro, L.P. (2015). *MeteorolAtmos Phys*. DOI 10.1007/s00703-015-0408-y.

Turkes, M., Koc, T., and Saris, F.(2009). Spatiotemporal Variability of Precipitation Total Series over Turkey. *International Journal of Climatology*. 29(8), 1056-1074.

Vautard, R., and Ghil, M. (1989). Singular Spectrum Analysisin Nonlinear Dynamics, with Applications to Paleoclimatic time series. *Physica D.* 35, 395–424.

Wasowicz, P., Pasierbinski, A., Przedpelska-Wasowicz, E.M., and Kristinsson, H. (2014). Distribution Patterns in the Native Vascular Flora of Iceland. *Plos One*. 9(7), e102916. doi:10.1371/journal.pone.0102916.

Weare, B. C., and Nasstrom, J. S. (1982). Examples of Extended Empirical Orthogonal Functions. *Monthly Weather Review*.110, 481–485.

White, D., Richman, M., and Yarnal, B. (1991). Climate Regionalization and Rotation of Principal Components. *Int.J. Climatol*. 11, 1–25.

Wickramagamage, P. (2010). Seasonality and spatial pattern of rainfall of Sri Lanka: Exploratory factor analysis. *International Journal of Climatology*. 30, 1235-1245.

Widmann, M., and Schar, C. (1997). A Principal Component and Long-Term Trend Analysis of Daily Precipitation in Switzerland. *International Journal of Climatology*. 17, 1333-1356.

Xu, T., Croke, B., Hutchinson, M. (2014). Identification of Spatial and Temporal Patterns of Australian Daily Rainfall under a Changing Climate. *7th Int. Congress on Env.Modelling and Software*.15-19 June. San Diego, California, USA, 1-8.

Zhigljavsky, A., Hassani, H., and Heravi, S. (2009). Forecasting European Industrial Production with Singular Spectrum Analysis. *International Journal of Forecasting*. 25(1), 103-118.

Zotov, L.V.(2012). Application of Multichannel Singular Spectrum Analysis to Geophysical Fields and Astronomical Images. *Advances in Astronomy and Space Physics*. 2, 82-84.

Zotov, L.V., and Shum, C.K. (2010). Multichannel Singular Spectrum Analysis of the Gravity Field Data from GRACE Satellites. *Proceedings of the 4th Gamow*

International Conference on Astrophysics and Cosmology After Gamow and the 9th Gamow Summer School ``Astronomy and Beyond: Astrophysics, Cosmology, Radio Astronomy, High Energy Physics and Astrobiology''. Odessa, Ukraine, 473-479.