

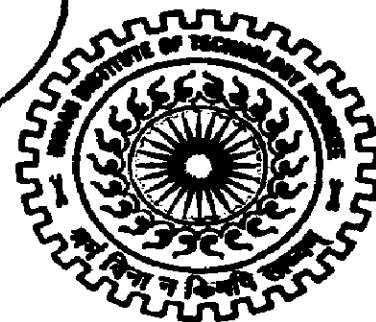
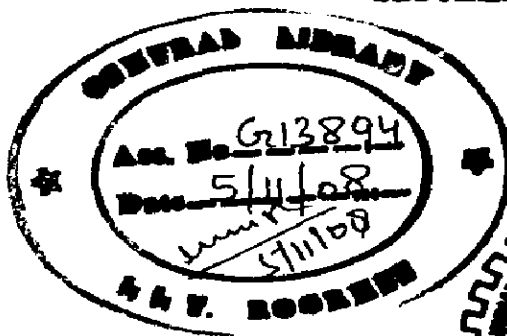
# **FLOOD ESTIMATION IN MAHANDI BASIN**

**A DISSERTATION**

*Submitted in partial fulfilment of the  
requirements for the award of the degree  
of*  
**MASTER OF TECHNOLOGY  
in  
HYDROLOGY**

By

**ANIL KUMAR KAR**



**DEPARTMENT OF HYDROLOGY  
INDIAN INSTITUTE OF TECHNOLOGY ROORKEE  
ROORKEE -247 667 (INDIA)  
JUNE, 2008**

## CANDIDATE'S DECLARATION.

I hereby certify that the work which is being presented in the Dissertation work, entitled 'FLOOD ESTIMATION IN MAHANADI BASIN' in partial fulfillment of the requirement for the award of the **degree of Master of Technology** submitted in the **Department of Hydrology, Indian Institute of Technology, Roorkee** is an authentic record of my own work carried out during the period from July, 2007 to June, 2008 under the supervision of **Dr. N. K. Goel** and **Mr. G. P. Roy**.

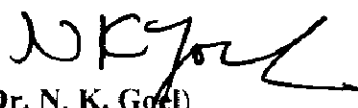
The matter embodied in this thesis has not been submitted earlier by me for the award of any other degree of this institute or any other University.

Date 30/06/2008

  
(Anil Kumar Kar)

This is to certify that the above statement made by the candidate is correct to the best of our knowledge.

  
(Mr. Gopal Prasad Roy)  
Deputy Director  
Secha Sadan  
Bhubaneswar

  
(Dr. N. K. Goel)  
Professor  
Department of Hydrology  
Indian Institute of Technology  
Roorkee

## ACKNOWLEDGEMENTS

It is my proud privilege to express my sincere gratitude to Dr. N. K. Goel, Professor, Department of Hydrology, who suggested the field for the present study and has been main source of guidance. His untiring efforts and patience to listen and suggestions to improve the work, are gratefully acknowledged. Indeed, I will remain ever indebted to him for his keen interest and whole-hearted co-operation all through the pursuance of the study.

I am also indebted to Mr.G.P.Roy Deputy Director, Secha Sadan who is also my co-guide for his valuable help, guidance and encouragement for this study.

At the beginning I wish to express my deepest gratitude to Department of Water Resources, Government of Orissa for nominating me for the 35 th International Hydrology course at the Department of Hydrology at I.I.T. Roorkee. I am also thankful to Office of the Engineer-in-Chief, Secha Sadan for providing necessary data for the study of I am also grateful to Dr. B. S. Mathur, Dr. D. K. Srivastava, Dr. R. Singh, Dr. D. C. Singhal, Dr. H. Joshi, Dr. M. Perumal, Dr. M. K. Jain and Dr. D. S. Arya of the Department of Hydrology and other staff for their valuable teaching and cooperation during tenure of my stay at I.I.T.Roorkee.

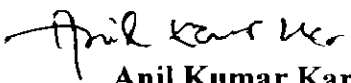
My special gratitude goes to Mr. Shibayan Sarkar, Mr. Jhajharia, Mr. C. S. Padhi, Mr. R. M. Das, Mr. H. Mohanty and all other research scholars of Department of Hydrology for their active help in my study.

I am also thankful to Mr. Ashok Ranjan, Department of AHEC, for his help in Autocad and GIS used in my study. Also I want to thank to all my colleagues, whose help and support did possible to finish my study in Roorkee.

I am thankful to the members of my family for their support and tolerance during my stay at Roorkee.

Last but no ways least, I bow with reverence to Lord Jagannath and my parents who has made all this possible.

**Roorkee**

  
**Anil Kumar Kar**  
**( Orissa, India)**

## SYNOPSIS

Mahanadi is one of the major east-flowing peninsular rivers draining into Bay of Bengal. It ranks second to the Godavari amongst the peninsular rivers. Mahanadi is known for its large water potential as well as devastations due to floods. Its 99% catchment lies on two states i.e Chhatisgarh and Orissa. The coast line of this river is highly developed as far as urbanization and industrial growth is concerned. The river steps down to flat land as it reaches Hirakud and passes through mostly populated area where flooding causes great loss to mankind and infrastructures.

The information on flood magnitudes and their frequencies are needed for the design of hydraulic structures such as dams, spillways, road, railway bridges and culverts etc. When river flow records are not available at or near the site of interest, it is difficult for hydrologists or engineers to derive reliable flood estimates directly. In such a situation the use of flood formulae developed for the regions is the only alternative method. When the available flow data of a catchment is too short to conduct frequency analysis, a regional analysis is adopted.

Our aim remains to form hydrologically homogeneous region / regions from the statistical point of view is considering flow data alongwith available catchment characteristics. The observed values of annual maximum flow data (AFS) is collected for these stations for different length of periods. Available long time data from neighboring catchments are tested for homogeneity and a group of stations satisfying the test are identified. For deciding the homogeneity, the L-moment approach of Hosking and Wallis has been used. This group of station constitutes a region and all the station data of this

region are pooled and analysed as a group to find the frequency characteristics of the region.

After testing this basin for homogeneity, entire basin is divided into two regions/clusters. Clustering has been done by using MATLAB in two methods, K-mean clustering, Fuzzy C-mean, and PAST software for doing Hierarchical clustering in Ward's method. Finally the number of optimum clusters has been decided by using Kohonen Self organizing map. The same has been done by using ANN Clust. Software.

The regional growth parameters for both the clusters has been drawn. The catchment characteristics are also considered during clustering process as well as when regressing for average flow in case of ungauged sites. The flood value for different return periods has also been calculated for different return periods for both gauged and ungauged sites of the basin. This study will be helpful for deriving flood values for longer periods, predicting flood values for ungauged sites. The present study can further be extended for better flood risk mapping and disaster warning system.

# CONTENTS

<b>Description</b>	<b>Page No.</b>
CANDIDATE'S DECLARATION	i
ACKNOWLEDGEMENTS	ii
SYNOPSIS	iv
CONTENTS	vi
LIST OF TABLES	ix
LIST OF FIGURES	xi
ABBREVIATIONS	xii
CHAPTER-I INTRODUCTION	1
1.1 GENERAL	1
1.2 OBJECTIVES OF THE STUDY	3
1.3 LAYOUT OF THESIS	4
CHAPTER-II LITERATURES REVIEW	5
2.1 GENERAL	5
2.2 DIRECT ANALYSIS FLOOD PEAKS	5
2.3 CLUSTER ANALYSIS	8
CHAPTER-III STUDY AREA, DATA AVAILABILITY AND <b>PRELIMINARY PROCESSING OF DATA</b>	13
3.1 GENERAL	13
3.2 STUDY AREA	13
3.2.1 Basin shape	18
3.2.2 Topography	18
3.2.3 Physiography	18
3.2.4 Climate	18
3.2.5 Rainfall	18
3.2.6 Temperature	19

3.2.7 Soil	19
3.3 DATA AVAILABILITY	20
3.4 PRELIMINARY PROCESSING OF DATA	24
<b>CHAPTER-IV METHODOLOGY</b>	27
4.1 GENERAL	27
4.2 SCREENING OF DATA SETS	27
4.2.1 Anderson's correlogram test	28
4.2.2 Kendall rank correlation test	28
4.2.3 Grubbs and Beck test	28
4.3 L-MOMENT AND ITS ADVANTAGES	29
4.4 STATISTICAL MEASURES OF L-MOMENT	31
4.4.1 Discordancy measure	31
4.4.2 Heterogeneity measure	32
4.4.3 Goodness of fit measure	34
4.4.3.1 L-moment ratio diagram	34
4.4.3.2 Z-statistics	34
4.5 NOTE ON SIZE AND MODIFICATION OF POOLING GROUPS	35
4.5.1 Size	35
4.5.2 Modifications	36
4.6 CLUSTERING METHODS	36
4.6.1 Selection of attributes	37
4.6.2 Standardisation of data	37
4.6.3 Ward's method(HC)	38
4.6.4 K-mean clustering method(KM)	40
4.6.5 Fuzzy C-mean	41
4.6.6 Kohonen self organization map	42
4.7 DISCUSSION ON DIFFERENT CLUSTERING METHODS	43
<b>CHAPTER-V RESULTS AND DISCUSSION</b>	47
5.1 GENERAL	47
5.2 REGIONAL HOMOGENEITY	47
5.2.1 Screening of dataset	47



5.2.2 Discordany measure	50
5.2.3 Heterogeneity measure	51
5.3 CLUSTER FORMATION	52
5.3.1 Normalisation of variables	52
5.4 APPLICATION OF CLUSTERING METHODS	53
5.4.1 Result of Ward's method	53
5.4.2 Result of K-mean method	55
5.4.3 Result of Fuzzy C-mean method	56
5.4.4 Result of Self organization map method	57
5.5 ANALYSIS OF DIFFERENT CLUSTERING METHODS	58
5.6 FLOOD ESTIMATION	63
5.7 PREDICTION OF DISCHARGE FOR DIFFERENT RETURN PERIODS	67
5.8 COMPARISON OF RESULTS	70
CHAPTER-VI CONCLUSION AND SCOPE FOR FURTHER STUDY	73
6.1 CONCLUSIONS	73
6.2 SCOPE FOR FURTHER STUDY	75
CHAPTER-VII REFERENCES	77

## LIST OF TABLES

Table. No.	Description	Page No.
3.1	Statewise catchment details of Mahanadi Basin	14
3.2	Basin details of subzone -3(d)	14
3.3	Variation of rainfall and temperature in Mahanadi basin	19
3.4	Locational details of Mahanadi basin G&D sites	20
3.5	Site characteristics of Mahanadi basin	23
3.6	Statistical Parameters of AFS (Original series)	24
3.7	Statistical Parameters of AFS (Log transformed series)	25
4.1	Critical values for Hosking and Wallis Discordancy test	32
4.2	Types of Attributes	37
5.1	Results of Anderson's correlogram test	48
5.2	Results of Kendall rank test	49
5.3	Results of Grubbs and Beck test	50
5.4	Results of L-statistics and $D_1$ values	51
5.5	Normalised values of site characteristics	52
5.6	Site allocation in K-mean clustering method	56
5.7	Site allocation in Fuzzy C-mean clustering method	57
5.8	Site allocation in SOM clustering method	58
5.9	Results of different clustering methods with heterogeneity measure	59
5.10	Interrelationship between different clustering methods	59
5.11	Final allocation of sites to both clusters	60
5.12	Final features of two clusters	63
5.13	Regional weighted parameters	64
5.14	Regional parameters of GEV distribution	65
5.15	Growth factors for both clusters	66
5.16	Extreme flood values(cumecs) for different sites of Mahanadi Basin for gauged catchments	68
5.17	Extreme flood values (cumecs) for different return periods for ungauged catchments	69

## LIST OF FIGURES

Figure. No.	Description	Page No.
3.1	Location map of Mahanadi basin	15
3.2	Location of subzone -3(d)	17
3.3	Drainage details of Mahanadi basin with G&D sites	21
5.1	Dendrogram for Hierarchical clustering ( Ward's method)	54
5.2	Representation of silhouette value of each site (K-mean)	55
5.3	Representation of cluster allocation (SOM)	57
5.4	Map of Mahanadi basin in two clusters	61
5.5	Typical L-moment ratio diagram showing robust distribution for Mahanadi basin	64

# CHAPTER-I

## INTRODUCTION

### 1.1 GENERAL

Since time immemorial water is the most precious gift of nature for growth and also for devastation of civilization. Flood hazards have become ever-increasing natural disasters resulting the highest economic damage among all kinds of natural disasters around the world. Our country is known for its much dense river network. So many big rivers like Ganga, Brahmaputra, Godavari and others are flowing in length and breadth of India. Out of all, the river Mahanadi is the sixth largest river of our country and is said to be the lifeline of two states like Orissa and Chhatisgarh. Simultaneously, it is also dangerous for the disastrous floods it carried during different periods of time.

On an average the basin receives 1088 millimeters of rainfall during the South-west monsoon (mid June to mid-October). Due to heavy rainfall the delta of the basin i.e the river below Naraj is subject to annual floods, which are aggravated by high tides and heavy rainfall directly on the delta. Floods of recent time like that of 1980, 1982, 2001 are still in the mind of the people of the basin. As it is more flat on part of Orissa (carrying 46% of the total catchment) devastation due to floods are more in Orissa. Although, some reservoirs and storage structures are now being made at Chhatisgarh and Orissa to discharge a controlled flow but that is insufficient. Even the original design flood of Hirakud dam was 42,500 m<sup>3</sup>/s. More recent calculations indicate that the maximum probable flood is 69500 m<sup>3</sup>/s or 63% greater than original calculation. (Sengupta et al.,2006 ). In this regard it is time to think over the design of the structures or to adopt some warning system for the safety of the structures and related devastation. So the story does not only confined to big structures like Hirakud but

the design of different types of hydraulic structures and flood plain zoning, the economic evaluation of flood protection projects etc. also require information on flood magnitudes and their frequencies.

The availability of gauge and discharge data (G&D) in Mahanadi basin is very poor as for as its network and utility is concerned. The main reason for the same are inaccessibility of the area, lack of funds to establish G&D sites and shortage of man power. In spite of the limited data availability the two state Chhatisgarh and Orissa are marching ahead with all kinds of developmental activities as the hydrologic engineering evaluation and decisions cannot be delayed for non-availability of systematic records or to obtain longer records.

For design flood estimation number of methods like rational formula, unit hydrograph based approach and flood frequency analysis are in vogue. The choice of method generally depends on the design criteria applicable to the structure and availability of data. As per Indian design criteria frequency based floods find their applications in estimations of design floods for almost all the types of hydraulic structures (Kumar and Chatterjee, 2005).

Regional flood frequency analysis resolves the problem of estimation of the extreme flood events for the catchments having short data records or ungauged catchments by substituting 'space for time' data from various sites for estimating floods for different return periods, particularly for small to medium catchments.

The basic aim of regionalization is to form the homogeneous regions. In this regard the non-homogeneous regions are converted to homogeneous region by different clusterisation methods. The robust distribution of the cluster depends upon the characteristics of the stations inside the cluster.

## 1.2 OBJECTIVES OF THE STUDY

In literatures, there are two broad approaches for ensuring regional homogeneity. These approaches are (i) flexible boundary approach and (ii) fixed boundary approach. The flexible boundary approach is based on clustering. The application of this approach is slightly difficult for field engineers but more number of station years data can be included in this approach. On the other hand the application of fixed boundary approach (geographically contiguous regions) is simple and straight forward. However, only limited number of station years data can be included in this approach.

The broad objective of the work is (i) to develop a method of subdivision of basin, which is having the advantages of both the approaches and (ii) to develop regional flood formulae for the Mahanadi basin after applying the newly developed approach for sub division of the basin(s).

Keeping the broad objective in view, the analysis consisted of the following major steps:

- (a) Checking the homogeneity of entire basin.
- (b) Division of the basin into suitable number of parts (Clusters).
- (c) Checking the homogeneity of clusters formed and select the robust distribution.
- (d) Development of  $Q_r/Q_m$  relationship for gauged catchments and
- (e) Extension of  $Q_r/Q_m$  relationship for ungauged catchments by regressing catchments characteristics with  $Q_m$ .
- (f) Flood magnitudes for different return periods are developed for both the clusters and for gauged and ungauged sites.
- (g) Making general recommendations regarding adoption of this approach over others and scope for further development.

### **1.3 LAY OUT OF THESIS**

The subject matter of this thesis has been laid out in six chapters. The chapter-I gives introduction of the study as well as a broad objective. Literature surveyed on regionalization approach and clustering techniques are incorporated in chapter-II. Chapter-III gives a brief note on study area, data availability and preliminary processing of data. The methodologies applied in this study have been described in chapter-IV. The results are presented in chapter-V and chapter-VI presents the conclusions and limitations of the study and also the scope for further work.

## **CHAPTER-II**

### **LITERATURES REVIEW**

#### **2.1 GENERAL**

The estimation of flood magnitude associated with a given return period (recurrence interval) is a crucial task for designing of variety of engineering works and hydraulic structures. For flood frequency estimation, the work on national as well as international level, has progressed processed along two main approaches viz. (i) direct analysis of peak flows and (ii) indirect estimation of flood frequencies through stochastic models of rainfall and deterministic model of rainfall and runoff. Statistical analysis of peak flows remains the first choice, when there is a long record of gauged floods close to the site of interest. Indirect estimation of flood frequency has been attempted in the past through rainfall runoff models (NERC, 1975 and Robson and Reed 1999), synthetic unit hydrograph (CWC , 1973, 1993) and physically based flood frequency models or derived distributions (Goel et al., 2000 and Kurothe et al., 2001).Some of the works based on direct analysis of peak flows are presented in section 2.2. The regionalization has been based on clustering . A brief review of clustering is presented in section 2.3.

#### **2.2 DIRECT ANALYSIS OF FLOOD PEAKS**

The research on flood frequency analysis has taken place with varying intensity over couple of decades. During seventies and eighties much effort was directed on developing efficient at-site flood frequency procedures. Many new distributions and estimation methods were developed and reported in literature. Research in nineties was mainly dominated by L-moments. During nineties, number of technical papers based on artificial neural network (ANN) also appeared in hydrologic literature.



Regional flood frequency analysis typically begins with a “region”. A region can be considered to comprise a group of sites from which the extreme flow information can be combined for improving the estimation of extreme flows at any site in the region. The three requirements that a region should possess to ensure effective information transfer (Burn and Goel, 2000) are:

- (i) The region should be hydrologically homogeneous. This requirement arises from the need to ensure that extreme flow that is transferred to target site is similar to the extreme flow information of that site.
- (ii) The region should be identifiable which implies that a regional home can be readily determined for a new catchment, which may be ungauged.
- (iii) The region should be sufficiently large. Larger regions imply that more extreme flow information is incorporated into estimation of extreme flow quantiles.

As the size of the region is to be increased, there is a tendency for homogeneity of a region to decrease. So there is a trade-off between the first and third requirement.

Two most commonly considered measures on which regional homogeneity is judged are dimensionless scale and shape parameters, usually expressed as  $C_v$  and  $C_s$ . Alternatively a particular, standardized by division by a particular index flood, may be the measure on which homogeneity is judged.

A particular quantile, standardized by division of a particular index flood, may be measure on which homogeneity is judged. For instance  $X_{10}=Q_{10}/Q$  was used by Dalrymple (1960) to test if all stations in a given geographic region could be considered homogeneous. That test is based on assumption of an underlying EV1 distribution.

The regional flood frequency analysis for the British Isles reported in the flood studies report (NERC, 1975), employed a similar approach to that described by Dalrymple. However, the procedure for estimation of region curve and the relationship between mean annual flood and catchment characteristics is different from that of USGS method.

When the sample size is small, the sample moments can be very different from those of the population from which the sample was drawn. Moreover, the estimated parameters of the distribution fitted by this method are often found to be less accurate as compared to the method of maximum likelihood. On the other hand, the maximum likelihood estimators even though supposed to be efficient, suffer from the limitations concerning convergence due to complex nonlinear formulations. Keeping these limitations in view, Greenwood et al., (1979) suggested a more elegant, probability weighted moment (PWM) approach for estimating the parameters of the distributions, which can be expressed in inverse form. The comparison of PWMs with traditional techniques for Gumbel distribution was presented by Landwehr et al., (1979a).

Estimation of parameters and quantiles for Wakeby distribution was presented by Landwehr et al.,(1979b). This methodology appears suitable for situations where records are extremely short; and stream flow observations are highly skewed and highly kurtotic. A number of papers on PWM concept in Indian context are Singh and Seth (1985), Goel (1998), NIH reports (1985-86, 1990-91 and 1997-98) are worth mentioning.

Hosking and Wallis (1997) state that L-moments are an alternative system of describing the shapes of probability distribution. They are robust to outliers and virtually unbiased for small samples, making them suitable for flood frequency analysis including identification of distributions and parameter estimation. Historically they arose as

modifications of the PWMs of Greenwood et al. (1979). Zafirakou - Kouloris et al. (1998) mention that like ordinary product moments, L-moments summarize the characteristics or shapes of theoretical probability distributions and observed samples. They also suggested significant advantage over ordinary product moments, especially for environmental data sets. Hosking and Wallis (1997) suggested that the goodness-of fit criterion for each of various distributions is defined in terms of L-moments and is termed as Z-statistics. Kumar and Chatterjee (2005) decided the robust distribution based on the goodness-of-fit criteria. The aim of this measure is to identify a distribution that fits the observed data acceptably close.

Hosking (1990) introduced L-moment ratio diagram for the purpose of selecting a suitable distribution. Vogel and Fennessy (1993) recommended the use of L-moment Ratio diagram as compared to moment diagrams. L-moments are more convenient as these are more directly interpretable as measures of the scale and shape of probability distribution. L-moment ratio diagram has been widely used in the number of studies like Choudhury et al. (1999), Pilon and Adamowski (1992), Vogel et al. (1993 a, b), Peel et al. (2001).

### **2.3 CLUSTER ANALYSIS**

Present study will focus on gradual development of regional flood frequency analysis with clustering techniques. As far as homogeneity is considered, it cannot always be ensured. So hydrologists also tried with other approaches, which dispensed with fixed regions. These approaches included cluster analysis, region of influence (ROI) based approaches, and flood seasonality based approaches. These approaches are based upon flexible boundary approaches where a set of catchments can be separated to any one group differ from one

another as little as possible, and catchments of different groups are dissimilar. A brief review of selected works in the area of flexible boundary approaches is presented as follows:

DeCoursey (1973) applied cluster analysis to site characteristics of streams, in Okalhama to form groups of sites having similar flood response. The essence of cluster analysis is to identify clusters (groups) of gauging stations such that the stations within a cluster are similar while there is dissimilarity between the clusters.

DeCoursey and Deal (1974) used cluster analysis to define the region. A discriminant analysis was performed and stations which were misclassified are switched and discriminant analysis was run again.

Mosley (1981) applied cluster analysis to selected catchments of New Zealand to identify groups of catchments characterized by specific mean annual flood and coefficient of variation. He opined that cluster analysis does not entirely eliminate subjective decisions, but greatly facilitates interpretation of data set.

Tasker (1982) identified homogeneous regions based on cluster analysis of catchments characteristics and used discriminant analysis to determine the probability of an ungauged site belonging to a particular cluster of stations. According to him cluster analysis is a more objective method of creating regions.

Acreman and Sinclair (1986) used Hierarchical clustering algorithm for clustering the basins for flood frequency in Scotland according to their physical characteristics.

Burn (1989), Nathan and McMohan (1990) used cluster analysis to group sites into homogeneous regions. Both these papers emphasize in careful selection of variables.

Fovell and Fovell (1993) used hierarchical clustering (HC) in combination with Principal component analysis (PCA) for identifying climatic regions of the United States based on

monthly rainfall and temperature data. They used PCA to deduce the prominent features (dimensions) and HC to group the basins based on the identified dimensions.

Burn and Arnell (1993) used HC to identify the group of stations with similar flood responses, derived from the time series of annual maximum discharge values.

Bogardi et al. (1994) used the K-means clustering algorithm to derive clusters utilizing the atmospheric circulation pattern, which in turn is used for developing a hydrometeorological model for areal drought.

Burn and Goel (2000) used K-means algorithm for identification of similar group of catchments for Zone-3 of Central India. They used overlapping groups to estimate flow quantiles for gauged or ungauged catchments.

Parida B.P (2000) used K-mean clustering method for a Bulgarian catchments and suggested that catchments characteristics (attributes) are represented through a common measure such as Euclidean distance have been used to partition basins into regions which yield minimum partitioning errors. He has also suggested for Post-clusterisation operations like verification of geographic continuity, computation of  $C_v$  of  $C_v$ .

Thandaveswara and Sajjikumar (2000) have used ART-II technique in Artificial Neural Network in classification of river basin for finding homogeneous regions. They suggested that if  $C_v$  of  $C_v$  of a cluster is greater than 0.4 the region is highly heterogeneous. They suggested for overall objective of clustering as

- i) to have statistically acceptable homogeneity and
- ii) to have sufficient data in each cluster for further hydrologic studies.

Bhatt (2003) in his study has made critical evaluation of conventional techniques such as Ward's method and K-mean method and modern techniques like Fuzzy C-mean and Kohonen (ANN based) method.

Lim and Lye (2003) used Hierarchical clustering (average linkage method) in order to delineate homogeneous sub-regions in Sarawak, Malaysia. They have also done appropriate scaling of catchment characteristics to ensure that these factors fell between zero and unity.

Shi (2002) has applied clustering technique by using Fuzzy clustering technique and neural network.

Jingyi and Hall (2004) used K-mean, Fuzzy C-mean, Hierarchical clustering (Ward's method) and Kohonen Self Organising Feature Map for clustering the Gan -Ming river basin of China. He described that the Ward's method manipulates the data points to form the initial clusters and then uses the K-means method to adjust inaccurately assigned sites. The Fuzzy method assigns each data point to a particular class by hardening the Fuzzy partition matrix (U). For these methods the expected number of classes must be specified. But Kohonen method can both select the number of clusters and allocate each site to a cluster. So the Kohonen method describes an objective estimate of number of clusters.

Stambuk et al. (2007) have investigated possible application of the Kohonen Self Organizing Maps (SOM) to the social sciences data clustering, and compare the results of the procedure to the Principal Component Analysis (PCA) and Hierarchical clustering methods.

Modarres (2007) in his study indicated the advantage of multivariate methods with L-moment methods for hydrologic regionalization.

## **CHAPTER-III**

### **STUDY AREA, DATA AVAILABILITY AND PRELIMINARY PROCESSING OF DATA**

#### **3.1 GENERAL**

The chapter gives details of study area and data used. The results of preliminary processing of data are also presented in this chapter.

#### **3.2 STUDY AREA**

The study area Mahanadi basin is a major east flowing river in peninsular river system. It originates near Pharasiya village of Raipur district of Chhatisgarh state. It is an interstate river covering Chhatisgarh, Orissa, Jharkhand, MadhyaPradesh and Maharashtra. The total drainage area of river basin is 1,41,589 sq.km. The basin is en-compassed within the geographical co-ordinate of  $80^{\circ}-30'$  to  $86^{\circ}-50'$  of East Longitude and  $19^{\circ}-20'$  to  $23^{\circ}-35'$  of North Latitude (Figure 3.1). The basin is physically bounded by Central Hills in the North, Eastern Ghat in the South, Maikala hill range in the West and by Bay of Bengal in the East. The basin is largely divided into four parts such as Central table land, Northern plateau, Eastern ghats and Coastal plain. As the basin is comprised of very vast area there is a large variation in geographic and climatic conditions throughout the basin.

The state wise coverage of drainage area of the river Mahanadi is as follows in which states like Orissa and Chhatisgarh holds the largest share upto more than 99%.

**Table 3.1. Statewise catchment details Mahanadi basin.**

Sl.No.	State	Catchments Area (sq.km.)	% to total basin
1	Madhya Pradesh and Chhatisgarh	75,336	53.21
2	Orissa	65889	46.53
3	Maharastra	238	0.17
4	Jharkhand	126	0.09
Total		1,41,589	100

There are 14 major tributaries of river Mahanadi are Seonath, Hasdeo, Mand, Kelo, Birai, Pairi, Jonk, Sukha, Kanki, Lialr, Lath, Ong, Tel and IB (NWDA, 2004) .The river Mahanadi shares major part of subzone-3(d).Total area covered under this zone is 1,95,256. sq.km. Besides Mahanadi two other basins Brahmani and Baitarani are the parts of this zone Table 3.2 shows the catchment area of basin under subzone-3(d) and Figure 3.2 shows location of this subzone over all hydrometeorological subzones.

**Table 3.2 Basin details of subzone-3(d)**

Basin name	Catchment area (sq.km.)	Stream length (km)
Mahanadi	1,40,628	850
Brahmani	35,337	705
Baitarani	19,291	333

(Source: N.I.H. Report, 1994-95)





**Figure 3.1 Location map of Mahanadi Basin**

(Source: [http://Encarta.msn.com/map\\_701605802/Mahanadi\\_Basin.html](http://Encarta.msn.com/map_701605802/Mahanadi_Basin.html).)

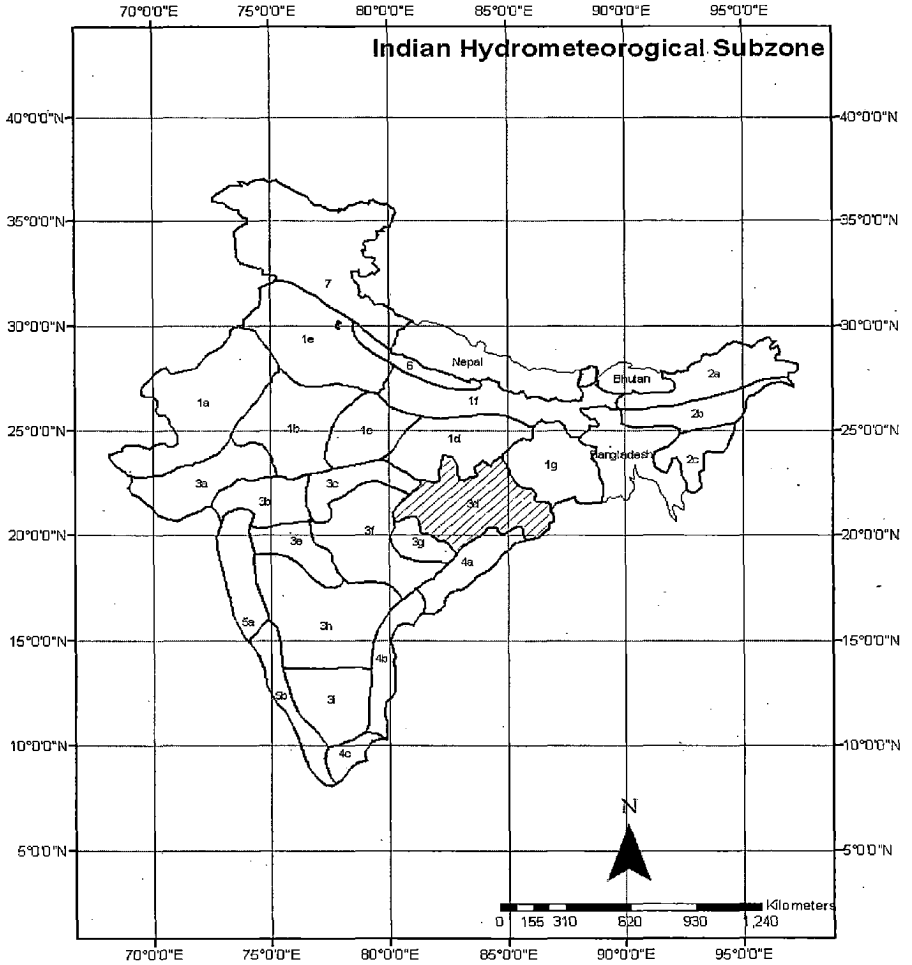


Figure 3.2 Location of subzone 3(d)

### **3.2.1 Basin shape**

The basin is roughly circular in shape with a diameter of about 400 km. and a 60 km. wide and 160 km. long exit passage. The basin is fan shaped with Horton form factor as 0.66.

### **3.2.2 Topography**

The upper reaches of the basin lies in a very undulating plateau with hillocks eroded moulds. The southern part of the plateau is open but to the east and west there are a number of hill ranges which have steep slopes resulting in water draining directly into the Mahanadi river. The basin continuously slopes towards the main valley with no congestion

### **3.2.3 Physiography**

The basin upto Hirakud can be divided into five groups physiographically viz. (i) Hill top and slope (ii) Upland (iii) Medium land (iv) Low land and (v) River banks.

### **3.2.4 Climate**

The climate in the basin area is tropical monsoon type with distinct seasons viz, Summer from March to May, the monsoon season from June to September and winter from October to February. The hottest and coldest months of the year are May and December respectively.

### **3.2.5 Rainfall**

The basin receives about 90% of its rainfall during the monsoon season. Generally, the southwest monsoon sets by the middle of the June over the entire basin and remains active till the end of September. The spatial variation in rainfall is moderate in the basin. The

formation of depressions in the Bay of Bengal cause cyclones which bring about wide spread heavy rains resulting in floods and destructions. The basin falls in the south-west monsoon track thus receives heavy rainfall during monsoon periods.

### 3.2.6 Temperature

The coldest and hottest months in the sub-basin are December and May respectively. The highest monthly mean maximum temperature is recorded as 42.1<sup>0</sup>C while lowest mean monthly mean temperature is 8.2<sup>0</sup>C. The highest single point temperature is 47.7<sup>0</sup>C and lowest is 0<sup>0</sup>C. The table 3.3 shows the variation of temperature over the entire basin. The maximum temperature occurs at central table land whereas lowest temperature in Eastern ghat is 8.7<sup>0</sup>C. Table 3.3 shows the variation of rainfall and temperature in Mahanadi basin.

**Table 3.3 Variation of rainfall and temperature in Mahanadi basin**

Region	Rainfall(mm)	Temperature (in <sup>0</sup> C)	
		Max	Min
Central table land	1394	42.1	12.2
Northern plateau	1327	41.5	13.2
Eastern ghat	1394	39.0	8.7
Coastal plain	1720	38.8	15.0

### 3.2.7 Soil

The soils of the basin can be grouped into five types viz. red soils, laterite soils, black soils, alluvial soils, red and yellow soils.

### 3.3 DATA AVAILABILITY

In the present study, annual maximum flood series data of 22 gauging stations spread over entire basin are collected. The locational details of all 22 G & D sites with respective site ids, names of the streams, catchment area are given in Table 3.4 and the map of Mahanadi basin with river network and G & D sites in Figure 3.3.

**Table 3.4 Locational details of Mahanadi basin G&D sites**

Station Id.	Station Name	River	Catchment Area(sq.km.)	Latitude	Longitude
1	Sundergarh	Ib	5870	22-06-55	84-00-40
2	Kurubhata	Mand	4625	21-59-15	83-12-15
3	Ghatora	Kuruvu	8035	22-02-04	82-13-34
4	Jondhra	Litaguru	29645	21-43-00	82-20-34
5	Basantapur	Son	57780	21-43-18	82-47-27
6	Andhiyarkore	Hamp	2210	21-50-02	81-36-21
7	Bamanidhi	Son	9730	21-53-55	82-42-29
8	Rampur	Jonk	5719	21-39-57	82-31-30
9	Salebhata	Ong	4650	20-59-00	83-32-22
10	Baronda	Pairi	3225	20-55-06	81-52-56
11	Rajim	Sukha	8760	20-58-00	81-52-30
12	Kotni	Seonath	6990	21-13-02	81-14-19
13	Simga	Seonath	16060	21-37-33	81-41-36
14	Kantamal	Tel	19600	20-39-00	83-43-55
15	Kesinga	Tel	11960	20-11-51	83-13-30
16	Tikarpara	Mahanadi	124450	20-38-00	84-37-08
17	Kelo	Kelo	1150	21-53-19	83-24-10
18	Mahendragarh	Hasa	1100	23-12-10	82-12-54
19	Pandigaon		6060	20-05-35	83-05-00
20	Pathardihi	Kharuan	2511	21-20-28	81-35-48
21	Seorinarayan	Mahanadi	6563	21-43-00	82-20-34
22	Sukuma		1365	20-48-52	83-30-20

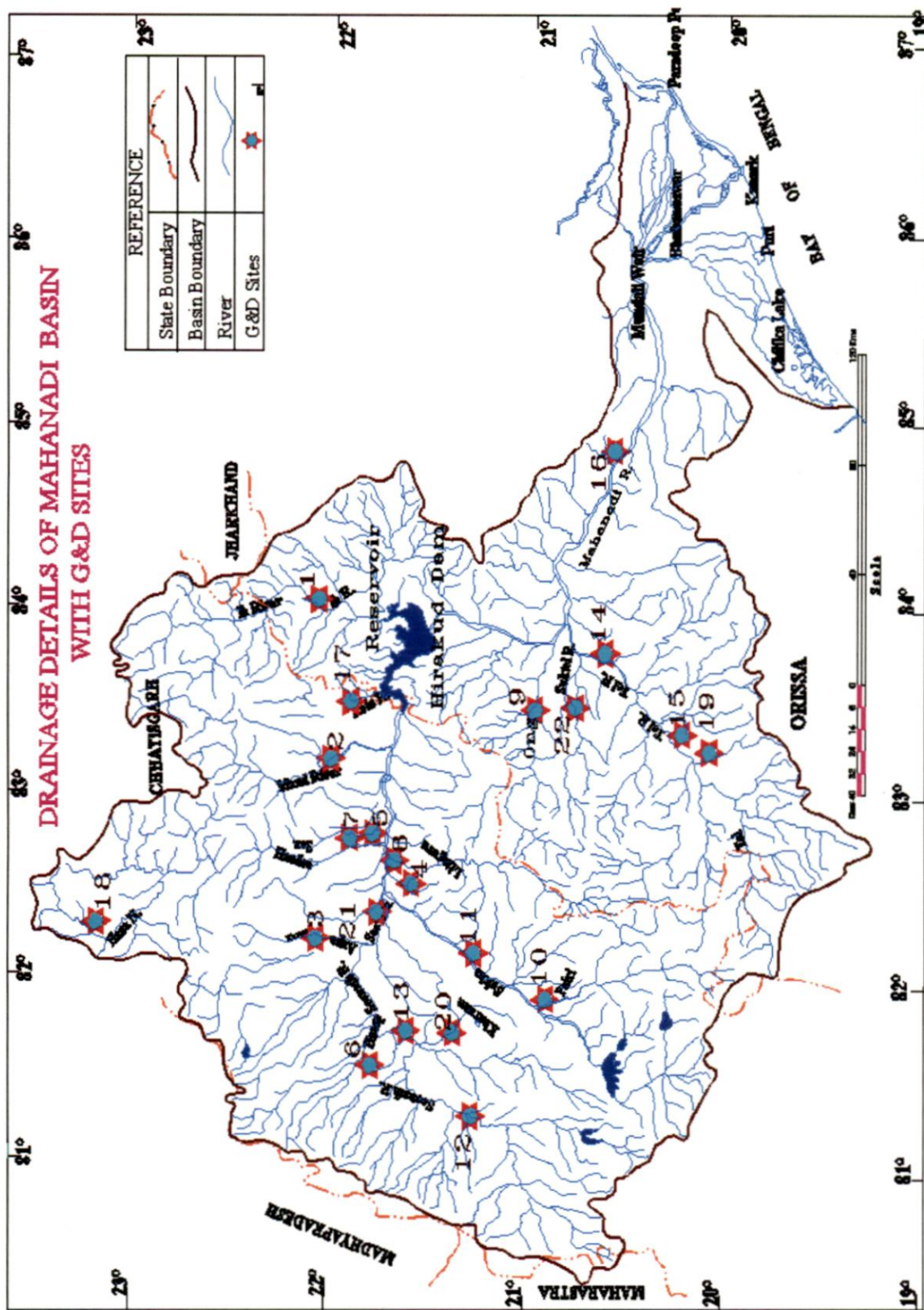


Figure 3.3 Drainage details of Mahanadi basin with G&D sites

Besides AMS data, catchments area, mean rainfall, maximum 1-day precipitation, elevation of individual stations are also collected. The longest stream length has been measured by using GIS software. All these catchment characteristics (variables) describes the property of that particular catchment and relates / distinguishes from that of other catchments. The catchment variables are recorded in Table No.3.5.

**Table 3.5 Site characteristics (variables) of Mahanadi basin**

<b>Site Id.</b>	<b>Q<sub>max</sub> (cumecs)</b>	<b>C.A (sq.km.)</b>	<b>Station Elev(m)</b>	<b>Nor.rain (mm)</b>	<b>Max 1-day ppt(mm)</b>	<b>Stream length (km.)</b>
1	10400.00	9809	214.00	1193.72	292.00	150.40
2	2160.00	3019	215.00	1186.22	272.80	168.55
3	2281.00	3112	246.00	938.06	235.00	49.33
4	11033.30	9506	219.00	1105.50	285.00	90.96
5	33087.95	57780	206.00	1060.52	176.80	58.35
6	851.98	2210	283.83	915.62	149.20	102.90
7	9583.10	9730	223.00	891.76	174.00	45.52
8	7095.80	5719	231.88	876.20	135.40	152.38
9	9916.00	4538	140.00	1039.96	267.80	48.84
10	3456.00	3225	289.87	943.20	199.40	108.24
11	8620.20	8413	283.83	875.40	155.00	56.42
12	5269.00	6990	283.03	1235.00	236.00	139.51
13	11703.00	16456	254.46	911.10	310.40	141.43
14	16263.00	19600	118.00	1052.10	204.80	226.47
15	12822.00	11960	166.00	1249.00	271.00	154.71
16	33800.00	124450	50.00	1192.96	305.20	601.55
17	1403.00	1266	230.00	900.00	160.00	53.34
18	2088.00	1100	411.00	1068.42	240.60	55.01
19	4217.00	6060	180.00	1240.20	265.00	47.92
20	1695.00	2511	279.91	1254.40	247.80	114.05
21	10958.40	6563	209.50	1135.54	289.40	164.75
22	2315.00	1365	156.94	1205.00	265.00	36.17
<b>Max</b>	<b>33800.00</b>	<b>124450.00</b>	<b>411.00</b>	<b>1254.40</b>	<b>310.40</b>	<b>601.55</b>
<b>Min</b>	<b>851.98</b>	<b>1100.00</b>	<b>50.00</b>	<b>875.40</b>	<b>135.40</b>	<b>36.17</b>
<b>Max-Min</b>	<b>32948.02</b>	<b>123350.00</b>	<b>361.00</b>	<b>379.00</b>	<b>175.00</b>	<b>565.38</b>

### 3.4 PRELIMINARY PROCESSING OF DATA

Fitting of frequency distribution involves the test of independence of the data, as it is a basic assumption for frequency distribution. The statistical parameters such as coefficient of variation ( $C_v$ ) and co-efficient of skewness ( $C_s$ ) were determined from original AFS as well as Log-transformed series to check the preliminary fitting of appropriate distribution function (Rai,et al.,2003). The statistical parameters like mean (M),standard deviation (S),coefficient of skewness ( $C_s$ ),coefficient of kurtosis ( $C_k$ ) and lag serial correlation coefficients ( $r_1, r_2, r_3$ ) of original series Table 3.7 as well as for Log transformed series were computed and shown in Table 3.6.

**Table 3.6 Statistical parameters of AFS (Original series)**

Site	STATISTICAL PARAMETERS (ORIGINAL SERIES)						CORR.COEFF		
	MAX	MIN	MEAN	STD.	$C_s$	$C_k$	$r_1$	$r_2$	$r_3$
1	10400	685.52	2442.1	2334.7	2.64	10.62	-0.187	0.321	-0.085
2	2160	300.03	1368.3	529.77	-0.28	2.48	-0.334	-0.336	0.495
3	2281	137.2	740.38	514.39	1.68	6.94	-0.077	0.076	0.125
4	11033.3	1600	4939.1	2367.6	0.98	4.48	0.015	-0.148	-0.093
5	33087.95	2741	13496	7121.7	0.79	4.2	-0.098	0.179	-0.003
6	851.98	46.74	373.66	216.1	0.99	4.09	-0.243	-0.373	-0.167
7	9583.1	689.8	3478	2399	0.76	3.23	0.023	0.173	0.252
8	10958.4	88.5	1839.8	2111.6	3.39	17.17	-0.182	0.045	-0.32
9	9916	140	2101.4	1919	2.81	13.73	-0.18	0.127	-0.383
10	3456	196.32	781.06	896.72	2.42	8.68	0.418	-0.151	-0.256
11	8620.2	254.5	3787.7	3083.1	0.37	1.78	-0.208	-0.003	-0.036
12	5269	463.3	1600	1142	1.85	8.2	-0.029	-0.116	0.342
13	11703	843.2	4040.3	2390.3	1.34	6.4	-0.132	0.098	0.037
14	16263	891.47	7884.2	5095.2	0.04	1.79	-0.146	-0.079	0.106
15	12822	600	5646.1	4164.5	0.44	2.21	-0.219	-0.16	0.024
16	33800	4774	20527	7830.7	-0.32	2.56	-0.169	-0.107	0.214
17	7920	125.8	1308.2	2221.9	3.16	13.21	-0.164	0.454	0.107
18	2088	97.51	447.98	480.95	2.81	11.86	-0.152	-0.335	0.394
19	4217	89.9	2362.1	1246.7	-0.54	4.19	-0.341	-0.082	-0.167
20	1695	235.1	960.15	466.51	-0.13	2.48	-0.108	0.125	-0.079
21	10958.4	113	2263.6	2524	2.75	12.28	-0.23	0.185	-0.378
22	2315	67.7	815.23	639.92	1.31	4.96	0.115	-0.207	-0.019



**Table 3.7 Statistical parameters of AFS (Log transformed Series)**

Site ID	MEAN	STDEV.	SKEW	KURTOSIS
1	7.53	0.69	0.93	4.24
2	7.13	0.49	-1.41	5.99
3	6.4	0.69	-0.23	3.65
4	8.39	0.5	-0.36	3.64
5	9.36	0.6	-0.7	3.52
6	5.77	0.62	-0.75	5.41
7	7.89	0.78	-0.22	2.15
8	7	1.09	-0.8	3.37
9	7.3	0.93	-0.74	4.51
10	6.18	0.92	0.67	3.63
11	7.75	1.15	-0.48	2.25
12	7.17	0.67	0.15	2.65
13	8.13	0.62	-0.4	3.17
14	8.68	0.88	-0.62	2.45
15	8.29	0.94	-0.43	2.47
16	9.83	0.49	-1.27	4.66
17	6.3	0.66	-0.93	6.18
18	5.77	0.78	0.8	4.14
19	7.72	0.48	-0.55	3
20	6.71	0.63	-0.89	3.1
21	7.21	1.15	-0.56	3.73
22	6.37	0.93	-0.81	4.46

# **CHAPTER-IV**

## **METHODOLOGY**

### **4.1 GENERAL**

The methodologies for the following are presented in this chapter:

- (i) Screening of datasets.
- (ii) L-moments and its advantages.
- (iii) Statistical measures based on L-moments.
- (iv) Note on size and modification of pooling groups, and
- (v) Clustering methods.

### **4.2 SCREENING OF DATASETS**

In order to have meaningful estimates from flood frequency analysis, the peak flood data used for analysis should satisfy the following assumptions:

- i) the data should be random
- ii) the data should be homogeneous
- iii) the sample size should be such that the population parameters can be estimated from it
- iv) the data should be of good quality

If the data available for analysis do not satisfy any of the above listed assumptions, then much reliability cannot be attached to the estimates. Data related problems may arise due to unreliable flow estimates, broken record, zero flood years, presence of outliers etc.(Goel and Seth,1985).

In the present study tests like Anderson's correlogram test for randomness, Kendall's rank test for trend and Grubb and Beck test for outliers are applied.

#### **4.2.1 Anderson's correlogram test**

This test is used to test the randomness of the data as for the subsequent analysis the series should be independent. In our study the test is applied at 5% significance level with a Z value of 1.96.

#### **4.2.2 Kendall's rank test**

The test is used to test the presence of trend in data series. At 5% significance level the Z value is 1.96. In our study the data series is tested for 5% significance level.

#### **4.2.3 Grubbs and Beck test**

An outlier is an observation that deviates significantly from the bulk of the data, which may be due to errors in data collection, or recording or due to natural causes. The presence of outliers in the data causes difficulties when fitting a distribution to the data. Low and high outliers are both possible and have different effects on the analysis. The Grubb and Beck test used at 10 % significance level is applied in this study for testing of outliers (Hamed and Rao, 2000).

### 4.3 L-MOMENT AND ITS ADVANTAGE

L-moments are a recent development within statistics. It is the “linear combination of probability weighted moments”. They form the basis of an elegant mathematical theory in their own right and can be used to facilitate the estimation process in regional frequency analysis. L-moment methods are demonstrably superior to those that have been used previously and are now being adopted by many organizations worldwide.

Here the fundamental component of flood frequency analysis is to fit a flood frequency distribution in region basis. Common approaches for this purpose are like Method of Moments, Maximum likelihood estimates, L-moment approach etc. The present study utilises the suitability of L-moment approach for this purpose. The L-moment as presented by Hosking Wallis (1997) is a development of PWM and is computationally convenient and robust to outliers. The theory of PWM (Greenwood et. al. 1979) are summarized as follows

$$\beta_r = E\{X[F(X)]^r\} \quad (4.1)$$

where,  $\beta_r$  is the  $r^{\text{th}}$  order PWM and  $F(x)$  is the cumulative distribution function of the random variable  $X$ . The unbiased estimator ( $b_r$ ) of the PWM are given by Hosking and Wallis (1997) as

$$b_r = n^{-1} \sum_{j=r+1}^n \frac{(j-1)(j-2)\dots\dots(j-r)}{(n-1)(n-2)\dots\dots(n-r)} x_{j:n} \quad (4.2)$$

where  $n$  is the sample size and  $x_{j:n}$  represents an ordered sample  $x_{1:n} \leq x_{2:n} \leq \dots \leq x_{n:n}$  from distribution of  $X$ .

Zafirakou-Koulouris et al.(1998) mention that like ordinary product moments, L-moments summarize the characteristics or shapes of theoretical probability distributions and observed samples. Both moment types offer measures of distributional location (mean), scale (variance), skewness (shape) and kurtosis (peakedness).The L-moments offer significant advantages over ordinary product moments, especially for environmental data sets, because of the following:

1. L-moment ratio estimators of location scale, shape are nearly unbiased, regardless of the probability distribution from which the observations arise. (Hosking 1990);
2. L-moment ratio estimators such as L-CV, L-skewness and L-kurtosis can exhibit lower bias than conventional product moment ratios, especially for highly skewed samples;
3. The L-moment ratio estimators of L-CV and L-skewness do not have bounds, which depend on sample size as do the ordinary product moment ratio estimators of CV and skewness;
4. L-moment estimators are linear combination of the observations and thus are less sensitive to the largest observations in a sample than product moment estimators, which square or cube the observations;
5. L-moment ratio diagrams are particularly good at identifying the distributional properties of highly skewed data, where as ordinary product moment diagrams are almost useless for this task (Vogel and Fennessey 1993).

#### 4.4 STATISTICAL MEASURES OF L-MOMENT

Three statistical measures discordancy measure, heterogeneity measure and goodness of fit measure are used in regional studies. These measures, as explained by Hosking and Wallis (1997) are presented in following sections.

##### 4.4.1 Discordancy Measure

The Discordancy measure is a test to decide the unsuitability of a site to a region .It is defined as

$$D_i = \frac{1}{3} (\mathbf{u}_i - \bar{\mathbf{u}})^T \mathbf{S}^{-1} (\mathbf{u}_i - \bar{\mathbf{u}}) \quad (4.3)$$

Where  $\mathbf{u}_i$  is the vector of L-moments,  $L_{cv}$ ,  $L_{cs}$  and  $L_{ck}$  for a site.

$$\bar{\mathbf{u}} = N^{-1} \sum_{i=1}^N \mathbf{u}_i \quad (4.4)$$

$$\mathbf{A} = \begin{pmatrix} \mathbf{u}_i - \bar{\mathbf{u}} \end{pmatrix}^T \begin{pmatrix} \mathbf{u}_i - \bar{\mathbf{u}} \end{pmatrix} \quad (4.5)$$

A critical D- statistics of a site is 3 when the number of sites in a region is 15 or more and it shows the site is not fitting properly to the group. So that particular site may be removed / shifted to other group or checked for errors. The Flood Estimation Handbook (Robson and Reed) says that the Discordancy measure is only useful when there are minimum seven sites in a group. As  $D_i$  test is a highly essential test, care is taken to keep atleast seven or more sites in a group. The critical values for different number of sites as

described in Flood Estimation Handbook based on a 10 % significance level as given in table -1.

**Table 4.1, Critical values for Hosking and Wallis discordancy test**

Sites in a pooling group	7	8	9	10	11	12	13	14	≥15
Critical values of $D_i$	1.917	2.14	2.329	2.491	2.632	2.757	2.869	2.971	3

However, Hosking and Wallis (1996) has given critical  $D_i$  values for groups having 5 and 6 sites as 1.333 and 1.648 besides other values are as mentioned in above table.

#### 4.4.2 Heterogeneity measures

It is used to estimate the degree of heterogeneity and to assess whether they might reasonably be treated as homogeneous. Specifically, the heterogeneity measure compares the between site variations in sample L-moments for the group of sites with that expected for a group of region. Hosking's Heterogeneity test fits 4 parameter Kappa distribution. A series of 500 simulations done and L-statistics of actual region is compared with a simulated series. The H-statistics defined as

$$H_i = (V_i - \mu_v) / \sigma_v \quad (4.6)$$

Heterogeneous is evaluated using the L-moment ratios and can be based on

L-CV alone ( $H_1$  statistics)

L-CV and L- skewness ( $H_2$  statistis)

### L-skewness and L-kurtosis ( $H_3$ statistic)

For each simulated region, the measures of variability  $V_i$  (where  $V_i$  is any of three measures  $V_1$ ,  $V_2$  and  $V_3$ ) is calculated. From the simulated data the mean  $\mu_v$  and standard deviation  $\sigma_v$  of the  $N_{sim}$  values of  $V_i$  are determined.

The critical H statistics for a region to be homogeneous is as mentioned below

$H < 1$	homogeneous
$1 \leq H \leq 2$	possibly heterogeneous
$H > 2$	definitely heterogeneous

Hosking and Wallis (1991) observed that statistics  $H_2$  and  $H_3$  based on measure of  $V_2$  and  $V_3$  lack the power to discriminate between homogeneous and heterogeneous regions but  $H_1$  based on  $V_1$  has much better discriminating power. So  $H_1$  is treated as a much better indicator of heterogeneity measure. Also,  $H_1$  was found to be a better indicator of heterogeneity in large regions, but has a tendency to give false indication of homogeneity for small regions (Rao and Hamed,2000).

The measure  $H_2$  indicates whether at-site and regional estimates will be close to each other. A large value of  $H_2$  indicates whether or not the at-site and regional estimates will be in agreement, whereas a large value of  $H_3$  indicates a large deviation between at-site estimates and observed data.



### 4.4.3 Goodness of fit measure

#### 4.4.3.1 L-moment ratio diagram

Hosking (1990) introduced L-moment diagram for the purpose of selecting a suitable distribution. In L-moment Ratio Diagram, the goodness of fit is judged by how well the L-skewness and L-kurtosis of the fitted distribution match the L-skewness and L-Kurtosis of the observed data. Observed L-moment ratios ( $\tau_3$  and  $\tau_4$ ) are plotted on L-moment ratio diagram for all sites in the region. Looking at the position of observed points ( $\tau_3$  and  $\tau_4$ ) of the various sites on theoretical L-moment ratio diagram, the most appropriate distribution for the given region can be selected. A significant advantage of L-moment ratio diagram is that one can compare the fit of several distributions using a simple graphical instrument (Goel and Arya, 2006).

#### 4.4.3.2 Z-statistics

It indicates suitability of a candidate distribution to a data series and is appropriate for evaluating and comparing 3-parameter distribution. The Z-statistics for the goodness of fit measure as defined by Hosking is

$$Z^{\text{DIST}} = (Z_4^{\text{DIST}} - Z_4 + B_4) / \sigma_4 \quad (4.7)$$

DIST = a particular distribution

$Z_4^{\text{DIST}}$  = L-kurtosis for fitted distribution

$Z_4$  = pooled L-kurtosis

$B_4$  = bias correction

$\sigma_4$  = estimate of sample variability of L-kurtosis.

The  $Z^{DIST}$  value should be as closely to zero. However a value between -1.64 and 1.64 is considered to be suitable for a fitting distribution at 10 % significance level. While a number of distribution may qualify the goodness-of-fit criteria, the most potential will be one that has minimum  $|Z^{DIST}|$  value. The growth curve parameters are obtained using L-CV and L-skewness whereas, goodness of fit exercises L-kurtosis as a check on how well the distribution fits. Besides the Z-statistics the L- Moment Ratio Diagram also decides the suitability of a distribution.

#### **4.5 NOTE ON SIZE AND MODIFICATION OF POOLING GROUP**

##### **4.5.1 SIZE**

Choosing an appropriate size of pooling –group requires compromise. If the pooling group is too small, then the pooled L-moments could be highly variable and predictions of rare flood events uncertain. If it is too large, it could include sites that are rather different from the subject site. Pooling group size is defined in terms of the number of station years of data rather than number of stations because of the large variations in record length.

The FEH recommendation is that the number of station years in the pooling group should be set at approximately five times the return period, the 5T rule. This is a rule of

thumb selected as a compromise between large indiscriminately pooled regions and excessive reliance on a small number of station - years of data.

#### **4.5.2 MODIFICATIONS**

Pooling groups that are heterogeneous should be investigated with a view to possible modification. The greater the heterogeneity, the greater the need for the pooling group to be reviewed. The FEH says that pooling group showing  $H_2$  value higher than 4 should be investigated for around 10% of sites. However, it is very important that sites should not be removed from the pooling group just because they reduce the heterogeneity.

A heterogeneous pooling group is acceptable for flood frequency estimation as long as it has been thoroughly investigated and any unsuitable sites are removed. A representative heterogeneous pooling group will give a better estimate than a non representative homogeneous pooling group.

#### **4.6 CLUSTERING METHODS**

The overall objective of clustering (Thandeeswara and Sajikumar, 2000) are

- (1) to have statistically acceptable homogeneity, and
- (2) to have sufficient data in each cluster for further hydrologic studies.

An important feature of clustering methods is the demarcation of hydrologically homogeneous regions. In the present study four clustering techniques namely agglomerative Hierarchical (Ward's method), K-means, Fuzzy C-means and Kohonen's Self Organising map (SOM) have been used. The following sections give the general

framework of clustering technique as well as four clustering technique used in this study. The performance has been evaluated on the basis of L-moments based statistical measures (Hosking and Wallis, 1993, 1997). In all these methods selection of attributes and standardization is important. These are explained in the subsequent sections and are followed by the description of different methods used in the study.

#### 4.6.1 Selection of attributes

For application of clustering methods each site should be presented with its individual characteristics (Attributes). Some common attributes under physical ,hydrological and meteorological categories (Parida ,2000) are listed in Table 4.2.

**Table 4.2 Types of Attributes**

Sl.no	Attribute category	Name of the Attributes
1	Physical	Basin area, Average slope of basin, Elevation of gauging site, Length of main channel, General soil characteristics of the basin.
2	Hydrological	Annual average flows, Coefficient of variation and Coefficient of skewness of annual peak floods
3	Meteorological	Annual average rainfall, Other critical rainfall values (say 50 year-3 hour) rainfall or some such identified characteristics.

#### 4.6.2 Standardization of the data

The variables for consideration in clustering are derived by transformation of site characteristics that are measured at different scales. Appropriate transformation by

scaling is necessary to ensure that these factors fall between zero and unity (Lim and Lye,2003).Before applying the data in any of the clustering methods the catchments characteristics (variables/attributes of each site) are to be rescaled by the formula

$$X_{kn} = \frac{Y_{kn} - Y_{n(\min)}}{Y_{n(\max)} - Y_{n(\min)}}$$

in order to make all values dimensionless (Zhang, Hall). Where  $Y_{kn}$  is the  $n^{\text{th}}$  feature at site  $k$ .  $Y_{n(\max)}$  and  $Y_{n(\min)}$  are the maximum and minimum of the  $n^{\text{th}}$  feature within the data set. The process is also known as normalization of data.

#### 4.6.3 Ward's method (HC)

In Hierarchical clustering, the data are first separated into a few broad classes. Each Class is further divided into smaller classes, and each of these is further partitioned, until terminal classes are generated to form a tree (dendrogram). It is an iterative procedure in which  $N$  data points are partitioned into groups which may vary from a single cluster containing all  $N$  points to  $N$  clusters. This technique can be divided into two methods, Agglomerative and Divisive. In agglomerative, clusters initially containing one element each are successively fused to generate large clusters. In divisive, a large cluster is divided into successively smaller clusters. Here the clustering method is well established by formation of a dendrogram and one has to

- i) identify the closest elements in the distance matrix.
- ii) fuse them into clusters
- iii) compute the new distance matrix

and the steps continue till one gets the desired number of clusters.

Four of the better known algorithms for hierarchical clustering are average linkage, complete linkage, single linkage and Ward's linkage.

**Average linkage** clustering uses the average similarity of observations between two groups as the measure between the two groups. **Complete linkage** clustering uses the farthest pair of observations between two groups to determine the similarity of the two groups. **Single linkage** clustering, on the other hand, computes the similarity between two groups as the similarity of the closest pair of observations between the two groups. Ward (1963) proposed a clustering procedure seeking to form the partitions  $P_n, P_{n-1}, \dots, P_1$  in a manner that minimizes the loss associated with each grouping, and to quantify that loss in a form that is readily interpretable. At each step in the analysis, the union of every possible cluster pair is considered and the two clusters whose fusion results in minimum increase in 'information loss' are combined. Information loss is defined by Ward in terms of an error sum-of-squares criterion, ESS ([http:// www. resample.com / xlminer/help/HClst/HClst\\_intro.htm](http://www.resample.com/xlminer/help/HClst/HClst_intro.htm)).

**Ward's linkage** is distinct from all the other methods because it uses an analysis of variance approach to evaluate the distances between clusters. In short, this method attempts to minimize the Sum of Squares (SS) of any two (hypothetical) clusters that can be formed at each step. In general, this method is regarded as very efficient, however, it tends to create clusters of small size. <http://www.gseis.ucla.edu/courses/ed231a1/notes2/cluster.html>

#### 4.6.4 K-mean clustering (KM)

This method was developed by MacQueen (1967). It is best described as a partitioning method. It partitions the data into K mutually exclusive clusters and returns a vector of indices indicating to which of the K-clusters it has assigned each observation. The algorithm clusters N objects based on attributes into K partitions where  $K < N$ . The optimization function

$$V = \sum_{i=1}^k \sum_{x_j \in S_i} (x_j - \mu_i)^2 \quad (4.8)$$

It tries to achieve minimum intra cluster variance or the squared error function. Where there are K clusters  $S_i=1,2,\dots,K$ , and  $V_i$  is the centroid or mean point of all the points  $X_j \in S_i$ . (<http://en.wikipedia.org/wiki/K-means>).

The K-mean algorithm shall follow three steps until it converges. The iteration will continue till it stabilizes i.e. no object changes the group.

It determines the

- co-ordinate of centroid.
- distance of each object to the centroid.
- group the object based on minimum distance.
- repeat the steps until no more assignment takes place.

To get an idea how well-separated the resulting clusters are, a silhouette plot has to be made using the cluster indices output from k-means. The silhouette plot displays a measure of how close each point in one cluster is to points in the neighboring clusters. This measure ranges from +1, indicating points that are very distinct from neighboring clusters, through 0, indicating points that are not distinctly in one cluster or another, to -1, indicating points that are probably assigned to the wrong cluster. Silhouette returns these values in its first output. Clusters having negative silhouette values indicates that these are not well separated (MATLAB Toolbox).

#### 4.6.5 Fuzzy C-mean (FC)

In this method the affinity of a site to undergo either two or more clusters are visualized. Earlier developed by Dunn in 1973 and improved by Bezdeck (1981) is basically used for pattern recognition. Here the data are bound to each clusters by means of a membership function which represents the Fuzzy behavior of this algorithm. It shows how to group data points that populate some multidimensional space into a specific number of different clusters. The objective function

$$J_m = \sum_{i=1}^N \sum_{j=1}^c u_{ij}^m \|x_i - c_j\|^2, \quad 1 \leq m < \infty \quad (4.9)$$

$m$ = any real number.

$u_{ij}$ =degree of membership of  $X_i$  in cluster  $J$

$x_i$ =ith of  $d$ -dimensional measured data.

$c_j$ =  $d$ -dimension center of cluster.



#### 4.6.6 Kohonen self organising map (KN)

Self-organizing maps (SOMs) are a data visualization technique invented by Professor Teuvo Kohonen which reduce the dimensions of data through the use of self-organizing neural networks. Kohonen's SOMs are a type of unsupervised learning. The goal is to discover some underlying structure of the data. Kohonen's SOM is called a topology-preserving map because there is a topological structure imposed on the nodes in the network. A topological map is simply a mapping that preserves neighborhood relations. (<http://www.willamette.edu/~gorr/classes/cs449/Unsupervised/SOM.html>).

The Kohonen map based data- clustering technique is applied to show how multi-dimensional datasets can be reduced to 2-D (feature) maps, manifesting clusters of similar data items (Kiang, Kulkarni et al.,1997)

The SOM algorithm is summarized as follows,

- (i) Initialise the weights  $w_{ij}$  from  $V$  input nodes to  $M$  output nodes to small random values. This way, each input node corresponds to a coordinate axis in the document vector space. Each output node is associated with a vector of weight  $w_{ij}$ , so it can be considered as a point in the input vector space.
- (ii) Describe each document as an input vector  $x_i(t)$  of  $V$  coordinates.
- (iii) Compute Euclidean distance  $d_j$  between the input vector at time  $t$ ,  $x_i(t)$ , and each vector of weights  $w_{ij}$  representing an output node as follows:

$$d_j = \sqrt{\sum_{i=1}^n (x_i(t) - w_{ij}(t))^2} \quad (4.10)$$

Select winning node  $j$ , which produces minimum  $d_j$ , update weights to nodes  $j$  and its neighbour to reduce the distance between them and input vector  $x_i(t)$  is as follows:

$$W_{ij}(t+1) = w_{ij}(t) + \eta(t)(x_i(t) - w_{ij}(t)) \quad (4.11)$$

After such updates, the nodes in the neighbourhood of  $j$  become more similar to the input vector  $x_i(t)$ . Here,  $\eta(t)$  is an error adjusting – coefficient ( $0 < \eta(t) < 1$ ) that decreases over time. Being unsupervised neural network, the SOM is known to be more resistant to noisy inputs (Chen, 1994), than statistical clustering technique such as K-means (Jains and Dubes, 1988)

#### 4.7 DISCUSSION ON DIFFERENT CLUSTERING METHODS

Wiltshire, 1985 and Nathan and McMahon, 1990 among others have drawn attention to the large amount of subjectivity involved in drawing the regional boundaries by using KM. Different hydrologists can obtain different groupings and the geographical proximity is no guarantee of homogeneity since neighbouring catchments can be physically different. To avoid geographical proximity, multivariate techniques (HC, FC and SOM) which use partitioning techniques can be applied. One of the primary features of partitioning techniques is that the allocation of a data point to a cluster is revocable, i.e. although an object is assigned initially to a cluster, it can be removed subsequently to other cluster.

Initially all sites are treated as separate clusters. In ward's method any two sites that are closest in terms of Euclidean distance are joined. In the next step, either a third

site joins first two, or two other sites join together into a different cluster. This process continues until all clusters are joined into one.

A dendrogram can effectively summarize the results of the clustering procedure. The similarity level at any step is the percentage of minimum distance at that step relative to the maximum inter-observation distance in the data. The decision on how many groups or regions to use, which essentially determines the cut-off level for similarity, is largely heuristic. However, the pattern of how similarity or distance values change from step to step can assist in choosing the final grouping. The step where the values change abruptly may indicate a good point for cutting the dendrogram.

The HC method manipulates the data points to form the initial clusters and then uses the *K*-means method to adjust inaccurately assigned sites. The FC method assigns each data point to a particular class by ‘hardening’ the fuzzy partition matrix (*U*). For these two methods, the expected number of classes must be specified. However, the Kohonen network can both select the number of clusters and allocate each site to a cluster. Therefore, *only the KN-SOM method as applied in this study produces an objective estimate of the number of clusters* (Zhang and Hall, 2004).

After clusterisation to further confirm this some other aspects related with homogeneity were also undertaken like (Parida, 2000),

- Verification of geographical continuity.
- Computation of  $C_c$  (Co-efficient of variation of the  $C_v$ s)
- Computation of error in partitioning other than into two groups.

- Test of homogeneity using L-moments (all 3 statistical measures )

In this study all four clustering methods described are applied. The post-clustering test like verification of geographical continuity on the basis of fixed / flexible boundary approach are considered. In fixed boundary approach the boundary line between clusters is fixed but the sites can switch over from cluster to cluster subject to suitability. But in flexible boundary approach the boundary line is demarcated according to the position of sites in cluster, no switch over is possible. We have used fixed boundary approach in our study.

## CHAPTER-V

### RESULTS AND DISCUSSION

#### 5.1 GENERAL

The methodology of clusterisation was applied to the entire Mahanadi basin. Regional flood formulae for different clusters have been developed. The results of different clustering methods and flood frequency analysis are presented in this chapter.

#### 5.2 REGIONAL HOMOGENEITY

##### 5.2.1 Screening of Dataset

The annual daily maximum discharge data of 22 gauge and discharge sites derived from daily discharge data are subjected to following tests Anderson's correlogram test for randomness, Kendall's rank test for trend and Grub and Beck test for outliers for initial screening. The results of each test are described separately and results mentioned Table 5.1 to Table 5.3.

##### Anderson's correlogram test

For checking the randomness of the AFS original series of individual stations are first undergone through Anderson's correlogram test for  $r_1$  only at 5% significance level for a Z value of 1.96. The range of the  $r_k$ (upper) and  $r_k$ (lower) values are tabulated for  $k=1$  at 95% confidence limits. In all cases the  $r_1$  values obtained remain within the upper and lower limits. So all the 22 datasets are tested as random. The results of this test are shown with the range of  $r_k$  upper and lower values obtained for individual sites respectively in Table No.5.1.

**Table 5.1 Results of Anderson's correlogram Test**

Site Id	Station Year	$r_1$	95% confidence limits		Remark
			( $r_1$ ) Upper	( $r_1$ ) Lower	
(1)	(2)	(3)	(4)	(5)	(6)
1	19	-0.187	0.420	-0.531	Random
2	18	-0.334	0.431	-0.549	Random
3	18	-0.077	0.431	-0.549	Random
4	19	0.015	0.420	-0.531	Random
5	26	-0.098	0.360	-0.440	Random
6	15	-0.243	0.472	-0.615	Random
7	24	0.023	0.374	-0.461	Random
8	26	-0.182	0.360	-0.440	Random
9	26	-0.18	0.360	-0.440	Random
10	16	0.418	0.457	-0.590	Random
11	26	-0.208	0.360	-0.440	Random
12	19	-0.029	0.420	-0.531	Random
13	24	-0.132	0.374	-0.461	Random
14	22	-0.146	0.391	-0.486	Random
15	19	-0.219	0.420	-0.531	Random
16	26	-0.169	0.360	-0.440	Random
17	10	-0.164	0.582	-0.804	Random
18	17	-0.152	0.444	-0.569	Random
19	9	-0.341	0.616	-0.866	Random
20	17	-0.108	0.444	-0.569	Random
21	17	-0.23	0.444	-0.569	Random
22	14	0.115	0.489	-0.643	Random

**Kendall rank correlation test**

Applying Kendall rank correlation test on above series the maximum and minimum z values varies between 1.93 and 0.11 which is within the 5% significance level. So all the serieses are used here show absence of any trend. The results of Kendall rank test shown in Table 5.2.

**Table 5. 2 Results of Kendall rank correlation test**

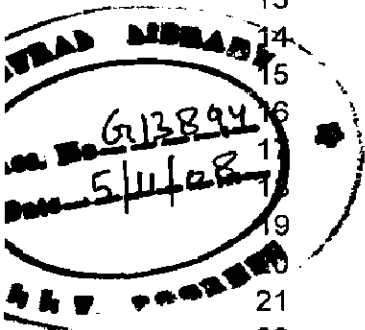
Site ID	Station Year	Value of p	Z computed	Remarks
1	19	74	-0.804	No Trend
2	18	87	0.795	No Trend
3	18	51	-1.931	No Trend
4	19	68	-1.224	No Trend
5	26	126	-1.6	No Trend
6	15	38	-1.435	No Trend
7	24	106	-1.587	No Trend
8	26	153	-0.418	No Trend
9	26	159	-0.154	No Trend
10	16	57	0.72	No Trend
11	26	155	0.11	No Trend
12	19	69	-1.154	No Trend
13	24	115	-1.141	No Trend
14	22	107	-0.479	No Trend
15	19	78	-0.524	No Trend
16	26	148	-0.639	No Trend
17	10	22	-0.268	No Trend
18	17	52	-1.318	No Trend
19	9	23	1.042	No Trend
20	17	54	-1.153	No Trend
21	17	48	-1.647	No Trend
22	14	44	-0.164	No Trend

**Grubbs and Beck test**

The approximation proposed by Pilon et al.(1985) at 10 % significance level is applied in order to find the high and low values with respect to the high and low outliers as per G & B test. The results for individual sites are mentioned in Table 5.3 and it shows that none of the values of entire 22 series are having any high or low outliers. After the above preliminary tests the datasets are subjected to following tests for checking the homogeneity of the region.

**Table 5. 3 Results of Grubbs and Beck Test**

Site ID	Station Year	K <sub>n</sub>	Grubbs and Beck Test		Observed values		Remarks
			X <sub>n</sub>	X <sub>i</sub>	Max	Min	
1	19	2.36	10545.2	377.9	10400	685.5	No outliers
2	18	2.34	3940.9	392.9	2160	300	No outliers
3	18	2.34	2975.3	120.7	2281	137.2	No outliers
4	19	2.36	14288.1	1368.6	11033	1600	No outliers
5	26	2.50	52341.4	2561.6	33088	2741	No outliers
6	15	2.25	1113.6	97.6	851.98	46.74	No outliers
7	24	2.47	18251.4	392.2	9583.1	689.8	No outliers
8	26	2.50	18589.7	66.2	10958	88.5	No outliers
9	26	2.50	12830.2	181.1	9916	140	No outliers
10	16	2.28	3580.6	77.6	3456	196.3	No outliers
11	26	2.50	30589.3	295.8	8620.2	254.5	No outliers
12	19	2.36	6283.2	266.3	5269	463.3	No outliers
13	24	2.47	15886.8	728.6	11703	843.2	No outliers
14	22	2.43	49506.9	693.9	16263	891.5	No outliers
15	19	2.36	36345.1	438.7	12822	600	No outliers
16	26	2.50	58518.4	6122.2	33800	4774	No outliers
17	10	2.09	1778.9	205.4	7920	125.8	No outliers
18	17	2.31	1825.7	55.4	2088	97.51	No outliers
19	9	1.98	6426.6	734.8	4217	89.9	No outliers
20	17	2.31	3493.8	194.2	1695	235.1	No outliers
21	17	2.31	19032.3	96.2	10958	113	No outliers
22	14	2.21	3626.7	107.9	2315	67.7	No outliers



**5. 2.2 Discordancy measure**

Applying the program of Hosking and Wallis (1997) on the data set the Discordancy Test shows that no value is Discordant as the maximum D<sub>i</sub> value is 1.96. Again the weighted means of L<sub>cv</sub>, L<sub>skew</sub>, L<sub>kurt</sub> are 0.3663, 0.2088, 0.1619 respectively, which are within their critical values as shown below and detail shown in Table 5.4.

$$L_{cv} \quad (t_2) \quad 0 < t_2 < 1, L_{skew} \quad (t_3) \quad -1 < t_3 < 1 \text{ and } L_{kurt} \quad (t_4) \quad -1 < t_4 < 1$$



**Table 5. 4 Result of L-statistics and  $D_i$  values**

<b>SITE ID</b>	<b>N (Station Year)</b>	<b>L-CV</b>	<b>L-SKEW</b>	<b>L-KURT</b>	<b>D(i)</b>
1	19	0.4303	0.5285	0.3917	1.21
2	18	0.2243	-0.0714	0.0077	0.97
3	18	0.3672	0.3029	0.2256	0.23
4	19	0.2665	0.1994	0.2171	1.1
5	26	0.2962	0.1203	0.159	0.26
6	15	0.3145	0.2769	0.2078	0.97
7	24	0.39	0.189	0.0138	0.96
8	26	0.4966	0.3748	0.3109	1.5
9	26	0.4244	0.3221	0.2712	0.29
10	16	0.5095	0.5682	0.4175	1.24
11	26	0.3951	0.0811	0.0633	1.32
12	19	0.366	0.2677	0.1389	0.51
13	24	0.3193	0.1688	0.1418	0.16
14	22	0.3724	0.0152	0.0902	1.36
15	19	0.4264	0.1554	0.0528	1.73
16	26	0.2212	-0.0869	0.0643	1.23
17	10	0.3263	0.2021	0.195	0.14
18	17	0.4469	0.5496	0.4248	1.19
19	9	0.2663	-0.053	0.1252	1.66
20	17	0.2848	-0.046	0.0388	0.78
21	17	0.5149	0.3954	0.3312	1.98
22	14	0.3466	0.218	0.3281	1.2
<b>Weighted means</b>		0.3663	0.2088	0.1619	

### 5. 2.3 Heterogeneity Measure

The Heterogeneity measures i.e  $H_1$ ,  $H_2$  and  $H_3$  for the entire Mahanadi basin are 2.85, 4.19 and 4.72 respectively. This clearly indicates that the region is heterogeneous as both  $H_1$  and  $H_2$  statistics are more than 2. Hence different clustering techniques have been applied to form homogeneous regions.

## 5.3 CLUSTER FORMATION

### 5.3.1 Normalisation of variables

Prior to application of in any of the clustering methods the catchment characteristics (variables) is to be normalized. In our study variables like annual daily maximum discharge ( $Q_{max}$ ), catchment area (CA), longest stream (RL), station elevation (SH), normal rainfall (NR), maximum 1-day precipitation (MP) are used. Out of all variables the length of longest stream (RL) is derived from the river network map of the basin by using GIS where as all other variables are collected from different relevant sources. The normalized values are shown in Table- 5.5

**Table 5. 5 Normalised values of site Characteristics**

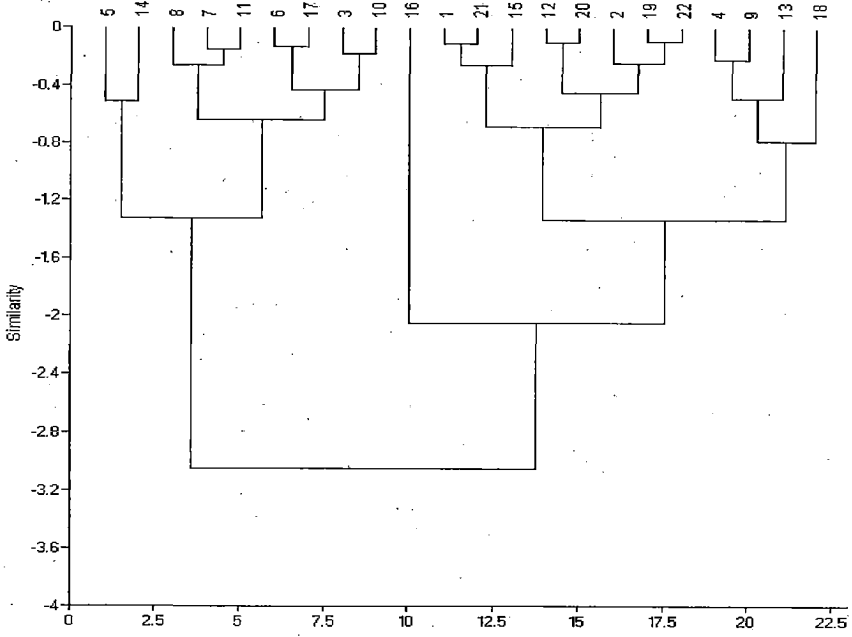
SITE	Qmax	C.A	SH	NR	MP	RL
1	0.290	0.071	0.454	0.840	0.895	0.202
2	0.040	0.016	0.457	0.820	0.785	0.234
3	0.043	0.016	0.543	0.165	0.569	0.023
4	0.309	0.068	0.468	0.607	0.855	0.097
5	0.978	0.460	0.432	0.488	0.237	0.039
6	0.000	0.009	0.648	0.106	0.079	0.118
7	0.265	0.070	0.479	0.043	0.221	0.017
8	0.190	0.037	0.504	0.002	0.000	0.206
9	0.275	0.028	0.249	0.434	0.757	0.022
10	0.079	0.017	0.664	0.179	0.366	0.127
11	0.236	0.059	0.648	0.000	0.112	0.036
12	0.134	0.048	0.646	0.949	0.575	0.183
13	0.329	0.124	0.566	0.094	1.000	0.186
14	0.468	0.150	0.188	0.466	0.397	0.337
15	0.363	0.088	0.321	0.986	0.775	0.210
16	1.000	1.000	0.000	0.838	0.970	1.000
17	0.017	0.001	0.499	0.065	0.141	0.030
18	0.038	0.000	1.000	0.509	0.601	0.033
19	0.102	0.040	0.360	0.963	0.741	0.021
20	0.026	0.011	0.637	1.000	0.642	0.138
21	0.307	0.044	0.442	0.686	0.880	0.227
22	0.044	0.002	0.296	0.870	0.741	0.000

Here 4 clustering techniques are applied one by one, using PAST software for making Hierarchical clustering through Ward's method, MATLAB software for doing K-mean and fuzzy C-mean clustering and ANN (NNClust) software for doing clustering through self organization map.

## **5.4 APPLICATION OF CLUSTERING METHODS**

### **5.4.1 Result of Ward's method (HC)**

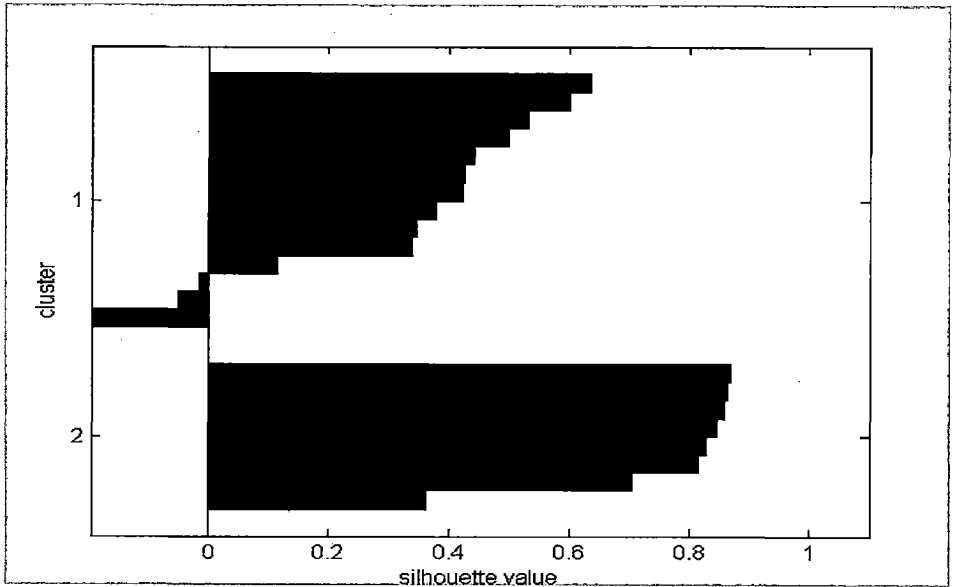
In Ward's method (Hierarchical clustering) a group of dendrogram is formed as per the respective site characteristics. By having some cut-off in similarity measure the dendrograms can be divided into two or more groups and their respective site ids with certain identity like cluster 1 and cluster 2 and so. The dendrogram obtained as an identification of individual sites are tried for different number of clusterings. A straight line has been drawn against the similarity measures to show at what similarity measure maximum how many cluster can be formed with how many number of sites.. Trial has been taken for 3 and 4 clusters and as the requirement confined to two clusters the straight line drawn has been positioned at such a place, where the entire sites can be divided into two groups. The formation of two clusters with respective sites (the site ids at reverse X-axis) are shown in Figure No 5.1. It shows the sites with id 5, 14, 8, 7, 11, 6, 17, 3, 10 are in one cluster and rest in another cluster.



**Figure 5.1 Dendrogram of Hierarchical Clustering (Ward's method)**

### 5.4.2 Result of K-mean method

The K-mean method when applied in MATLAB by using the developed program with respect to the site characteristics it produces the silhouette value map showing the affinity of sites to remain in a group is shown in Figure 5.2.



**Figure 5.2 Representation of Silhouette value of each site (K-mean)**

Out of 22 sites 3 sites show negative silhouette values which are again transferred to either of the suitable group. The required number of clusters is fed into the software initially to get the justified result. In our study the site is divided for two representative clusters. Allocation of different sites into either of the clusters are generated from the software is mentioned in Table 5.6. It shows that there are 8 sites allotted to cluster-1 and 14 sites two cluster-2.

**Table 5. 6 Site allocation in K-mean Clustering Methods**

<b>Result of K-mean</b>	
<b>Site ID</b>	<b>Cluster ID</b>
1	2
2	2
3	1
4	2
5	2
6	1
7	1
8	1
9	2
10	1
11	1
12	2
13	2
14	2
15	2
16	2
17	1
18	1
19	2
20	2
21	2
22	2

#### **5.4.3 Result of Fuzzy C-mean**

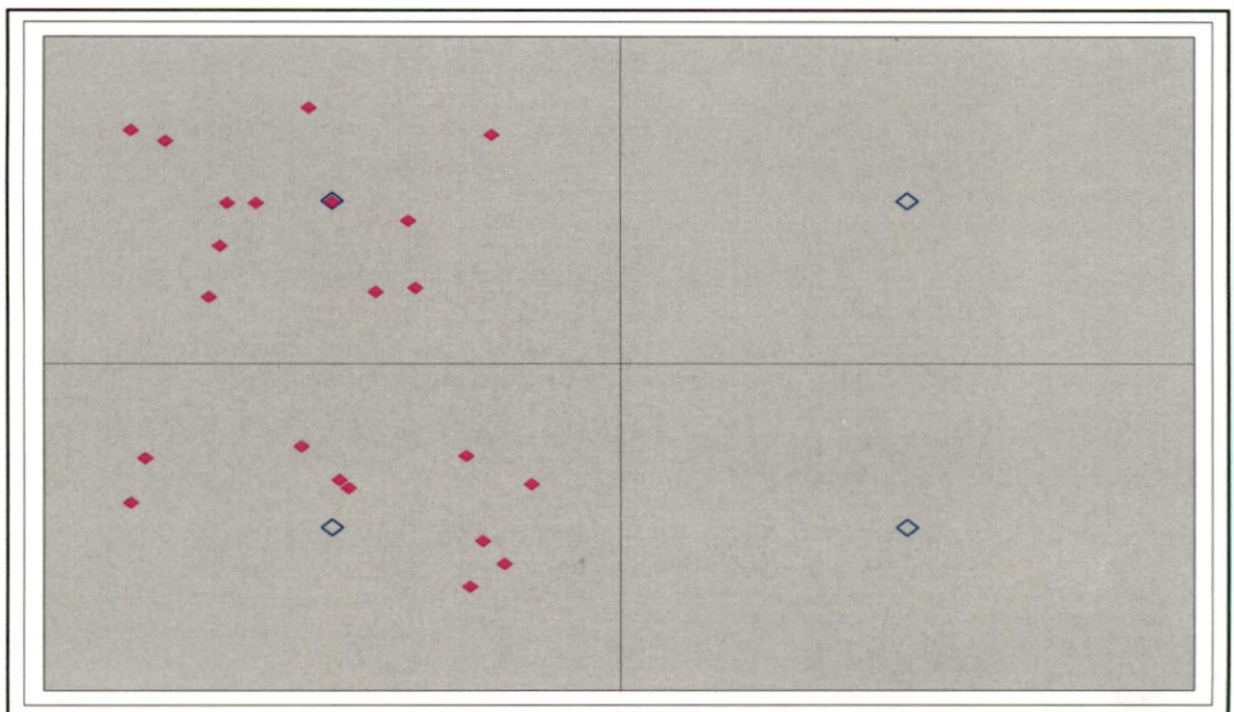
This method when applied with MATLAB software and the result generated from the program depicts the which site is allotted to which cluster as per their fuzzy characteristics of site variables. It shows that there are 12 sites in cluster-1 and 10 sites in cluster-2. Above all when applied Self Organisation Method it clearly shows that there are two clusters exists and their respective ids shown in the SOM . The results of K-mean, Fuzzy c-mean and Self organization maps are shown below The comparative results of different methods are shown in the table 5.7.

**Table 5. 7 Site allocation in Fuzzy C-mean Clustering Method**

Result of Fuzzy C-mean	
Cluster-1	Cluster-2
1	3
2	5
4	6
9	7
12	8
14	10
15	11
16	13
19	17
20	18
21	
22	

**5.4.4 Result of Self Organisation Map method**

When the site characteristics are applied to NN clust. software different combination of sites are generated . By adjusting the parameters it clearly indicates that formation of two cluters only can be possible by using the site characteristics for all 22 sites. The output shows the allocation of sites to each cluster pictorially in Figure 5.3.



**Figure 5.3 Representation of Cluster Allocation by SOM**

The self organizing map method also provides the allocation of site id to individual clusters generated by the software as per its id. The output is mentioned in Table 5.8 which is the direct representation of pictorial maps in numerical form. It shows that there are 10 sites in cluster-1 and 12 sites in cluster-2.

**Table 5. 8 Site allocation in SOM Clustering Methods**

Result of SOM	
Site ID	Cluster ID
1	1
2	1
3	2
4	2
5	2
6	1
7	2
8	2
9	2
10	2
11	2
12	1
13	2
14	2
15	1
16	2
17	1
18	1
19	1
20	1
21	2
22	1

### 5.5 ANALYSIS OF CLUSTERING METHODS

In this study four clustering methods namely HC, KM, FC, KN (SOM) were applied to identify grouping of sites and the results of different clustering methods are discussed separately. The sites allotted to different clusters by different methods are tested for heterogeneity by applying the same Hosking program shown in Table 5.9.



**Table 5.9 Results of different clustering methods with heterogeneity measure**

	HC		KM		FC		SOM	
<b>Cluster-1</b>	1,2,4,9,12,13,15,16, 18,19,20,21, 22		1,2,4,5,9,12,13,14, 15,16,19,20,21,22		1,2,4,9,12,14,15, 16,19,20,21,22		1,2,6,12,15,17,18, 19,20,22	
<b>Cluster-2</b>	3,5,6,7,8,10,11,14, 17		3,6,7,8,10,11,17,18		3,5,6,7,8,10,11,13,17,18		3,4,5,7,8,9,10,11,13, 14,16,21	
<b>Cluster</b>	1	2	1	2	1	2	1	2
<b>H1</b>	2.46	1.41	2.84	0.61	3.13	1.61	0.77	3.28
<b>H2</b>	3.32	2.30	3.98	1.55	5.06	1.72	2.29	3.41
<b>H3</b>	3.23	3.35	3.87	2.18	5.26	1.97	2.72	3.76

Here it shows the H-statistics has improved a lot for cluster-1 and for cluster 2 than when it was tested for entire basin as one region. The Table 5.9 indicates that the clusters generated by FC method shows a good result for where as cluster-1 does not agree to homogeneity. In case of K-mean cluster-2 shows good result but cluster-1 does not. However HC and SOM method results are heterogeneous for both clusters. So it is difficult now to follow either of the four methods directly.

After getting the sites allotted to different clusters it is not clear that which method we are going to apply for clustering .in different methods our job is to find the inter-relationships between the results of different clustering methods shown in Table 5.10.

**Table- 5.10 Inter relationship between different clustering methods**

	HC		KM		FC		SOM	
<b>Cluster No. of sites</b>	1	2	1	2	1	2	1	2
<b>HC</b>	X	X	12	7	11	8	8	7
<b>KM</b>	12	7	X	X	12	8	7	5
<b>FC</b>	11	8	12	8	X	X	7	7
<b>SOM</b>	8	7	7	5	7	7	X	X

In HC there is a good resemblance between KM and FC but not with SOM. There is a good combination between the results of KM and FC. Where as SOM show similarity with

HC than other two methods. The maximum resemblance is between KM and FC. So the results of KM and FC are taken as the base for further application.

The Flood Estimation Handbook (Robson and Reed, 1999) says that the Discordancy measure is only useful when there are minimum seven sites in a group. As  $D_i$  test is a highly essential test, care is taken to keep atleast seven or more sites in a group. Keeping in view the minimum sites in a group, result of SOM method, and 5T rule shifting of sites are tried by fixed boundary approach where we can interchange sites between clusters in order to make a statistically homogeneous region (Burn and Goel, 2000). Different combinations are tried and the results are tested for discordancy and heterogeneity. Finally removing the **site no-16** from either cluster and shifting sites from cluster to cluster (4 sites shifted from cluster-1 and two sites received from cluster-2) the entire basin is divided into two parts shown in (Fig-5.3). One of the primary features of clustering technique is that allocation of a data point to a cluster is revocable i.e. although an object is assigned initially to a cluster, it can be removed to other cluster. Again by applying Hosking program for two rearranged clusters the D-statistics, H-statistics and Goodness of fit is within the range of critical value shown in Table-5.11. Here the cluster -1 represents mostly the upper Mahanadi parts and cluster-2 the lower Mahanadi part almost closer two the border of Orissa as a divider.

**Table-5.11 Final allocation of sites to both clusters**

	Cluster-1	Cluster-2
Site Id.	3,4,5,6,7,8,11,12,13,18,20,21	1,2,9,10,14,15,17,19,22
Maximum $D_i$ value	1.85	1.48
$H_1$	1.84	1.25
$H_2$	1.61	3.09
$H_3$	1.61	4.12

The H-statistics for both clusters (Table 5.11) also allows both clusters are homogeneous.

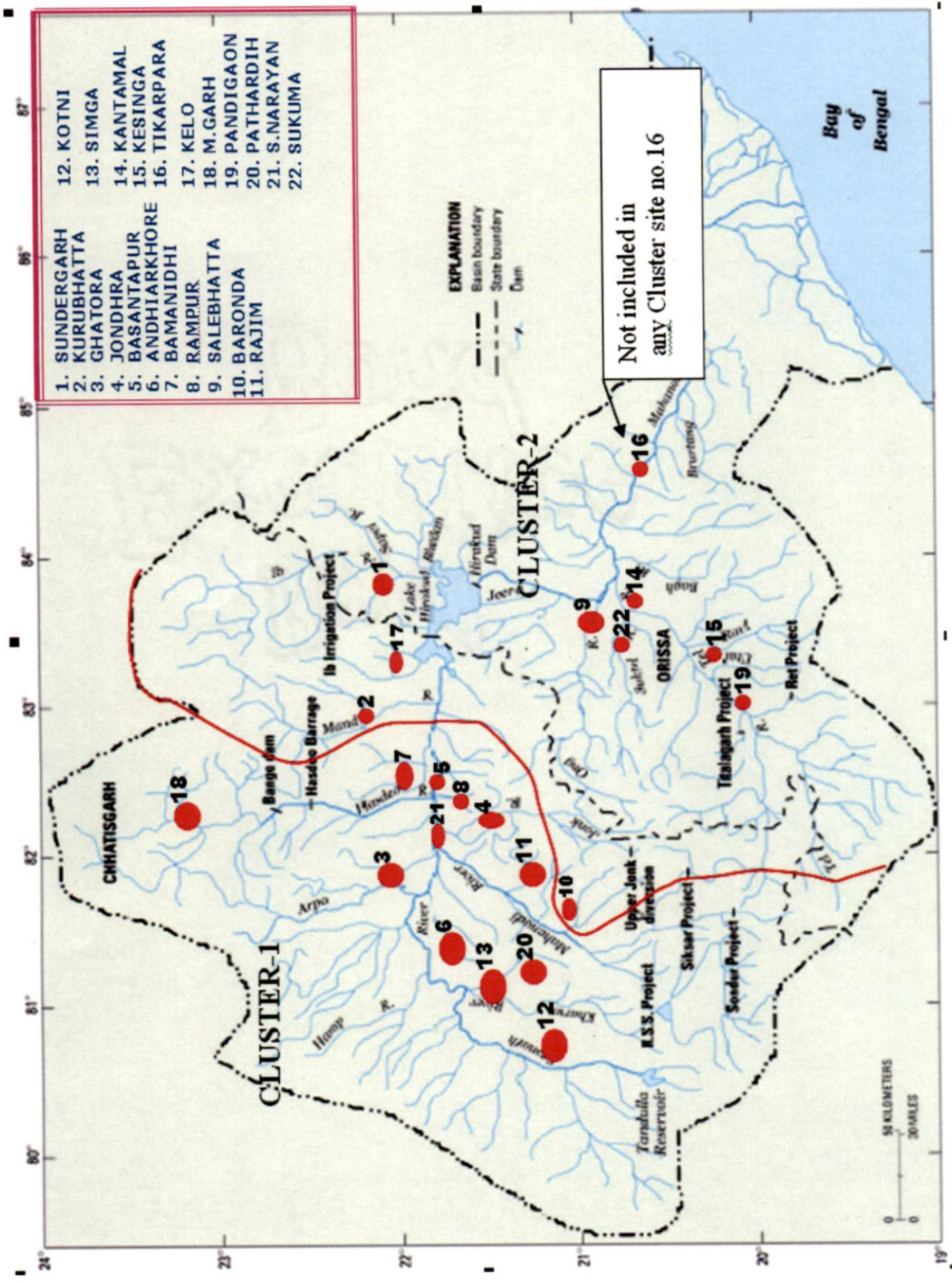


Figure 5.4 Map of Mahanadi basin in two clusters

## 5. 6 FLOOD ESTIMATION

The goodness of fit measure is considered on both Z-statistics and through moment ratio diagram. The Z-statistics as obtained from Hosking's program are shown in Table 5.12 which depicts that distributions like GL,GEV, GN and PT-III are suitable for both the clusters as Z-values of these four distributions are within -1.64 to +1.64 . Where as the GP distribution is not fitting I both clusters.

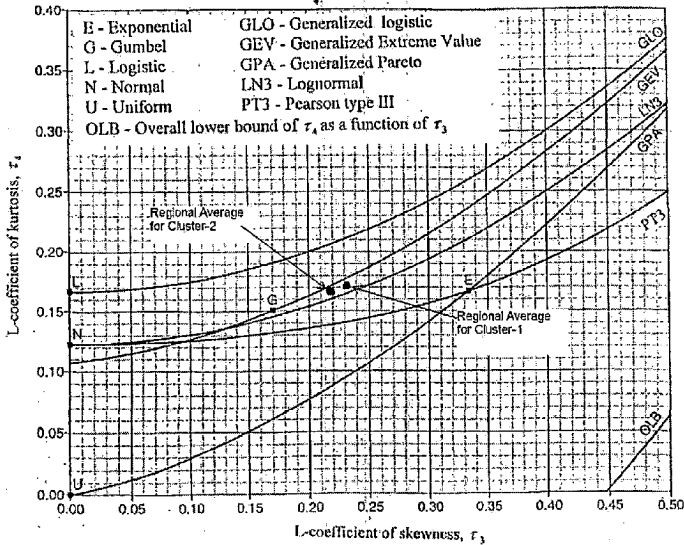
**Table-5.12 Final features of two clusters**

	Cluster-1	Cluster-2
Goodness of Fit (as per Hosking program)	GL (1. 26)	GL (0. 85)
	<b>GEV (0.09)</b>	<b>GEV (-0. 07)</b>
	GN (- 0. 34)	GN (-0. 37)
	PT-III (-1.15)	PT-III (- 0.95)
	GP (-2.72)	GP (-2. 24)
Goodness of Fit (as per Moment Ratio Diagram)	<b>GEV</b>	<b>GEV</b>

Out of all four fitted distribution the Z-value of GEV is the least one in both the cases. Similarly, taking the L-skew and L-kurt value (Table 5.13) obtained from the Hosking's program when plotted in L-moment ratio diagram (Figure5.4) also its closeness towards the GEV line. So it is decided on both the considerations that GEV is the robust distribution for both the clusters. The Moment Ratio Diagram plotted in Figure 5. 4 with values regional parameters in Table 5.13.

**Table 5. 13 Regional Weighted Parameters**

Cluster	L-skew	L-kurt
1	0.2312	0.1684
2	0.2229	0.1681



**Figure 5.5 L-moment ratio diagram showing robust distribution for Mahanadi Basin**

After the cluster delineation, homogeneity checking and finding suitable distributions extreme flows are estimated from the  $Q_t / Q_m$  relationship. As the numbers of station years are around 248 for cluster 1 and 153 for cluster 2, the  $Q_t$  value calculated for 50 and 30 year return period respectively will be most suitable for application on 5T station year basis.

For the gauged sites  $Q_m$  is the average of the available flow data and for ungauged catchments it is calculated from the catchments variables by regression.  $Q_t$  for different

return periods are calculated according to GEV equation established as it is the robust distribution for both the clusters.

The relationship for gauged catchments is shown below as per GEV distribution.

$$Q_t / Q_m = \xi + \alpha / \kappa [1 - \{-\ln(1-1/T)\}^\kappa] \quad (5.1)$$

Where,  $Q_t$  = flood quantile for different return period.

$Q_m$  = at site mean maximum discharge

$T$  = return period

**Table 5. 14 Regional Parameters of GEV Distribution (Mahanadi Basin)**

Regional Parameters	Cluster-1	Cluster-2
$\xi$	0.668	0.664
$\alpha$	0.490	0.506
$\kappa$	-0.093	-0.081

Applying above regional parameters to equation-1, the  $Q_t/Q_m$  for cluster-1 and 2 becomes as shown below

$$(Q_t / Q_m)_1 = - 4.60 + 5.268 (-\ln(1-1/T))^{-0.093} \quad (5.2)$$

$$(Q_t / Q_m)_2 = - 5.583 + 6.247(-\ln(1-1/T))^{-0.081} \quad (5.3)$$

For finding the  $Q_t$  for ungauged catchments value of  $Q_m$  is regressed for all five variables by using Origin - 50 software. In this study mean discharge is used as the index flood instead of using the median flood. Again for simplicity  $Q_m$  is also regressed with the catchments area only for both the clusters.

The mean discharge in cumecs for both clusters in power form is as follows.

$$Q_{m1} = 0.10011 * (CA^{0.95991}) * (SH^{-0.08882}) * (NR^{0.24187}) * (MP^{0.22797}) * (RL^{-0.20386}) \quad (5.4)$$

(with coefficient of correlation 0.95)

$$Q_{m2} = 0.07583*(CA^{0.6801})*(SH^{0.9787})*(NR^{2.17307})*(MP^{-1.06})*(RL^{0.02115}) \quad (5.5)$$

(with coefficient of correlation 0.98)

But, it is always not possible to go for all these 5 variables and put in the equation in order to find the mean discharge for a ungauged catchment. So for a ready reckoner or a hand rule  $Q_m$  is also regressed for catchments area only (when it is difficult to find other catchment characteristics) for both the clusters and  $Q_t / Q_m$  ratio is established accordingly,

$$Q_{m1}=0.4177*(CA)^{0.9669} \quad (5.6)$$

(with coefficient of correlation 0.90)

$$Q_{m2}=1.5242*(CA)^{0.8456} \quad (5.7)$$

(with coefficient of correlation 0.87)

Where,  $Q_{m1}$ = mean flood for cluster-1,  $Q_{m2}$ =mean flood for cluster-2, CA= catchments area, SH= station height, NR = normal rainfall, MP = maximum 1-day precipitation, RL= river length and all the values are their in their respective units.

The regional growth factors also developed for both the clusters which will be helpful for finding the  $Q_t / Q_m$  values are tabulated in Table No 5.15. The  $Q_m$  value multiplied with growth factors will give the extreme flood values for different return periods.

**Table No 5.15 Growth factors for both clusters.**

Return period	Growth factors	
	Cluster-1	Cluster-2
2	0.833	0.839
10	1.889	1.911
20	2.349	2.368
50	2.989	3.000
100	3.502	3.503
200	4.043	4.029
500	4.800	4.760
1000	5.405	5.342

## **5. 7 PREDICTION OF DISCHARGE FOR DIFFERENT RETURN PERIODS**

The 5T station year method says that the prediction for different return periods is based on the total station years present in that pooling group divided by 5. So our cluster-1 has station year 248 and cluster-2 has 153, which can suitably predict for return periods 50 and 30 respectively. However, for making it more reasonable and workable,  $Q_t$  value for different range of gauged (21 sites) and ungauged catchments area and return periods of 10,20,30,40,50 has been calculated by using equation (5.2) and (5.3) and using equation (5.6) and (5.7) which are presented in Table 5.16 and 5.17.



Table 5.16 Extreme flood values (cumecs) for different sites of Mahanadi basin for gauged catchments

Site Id	Belongs to Cluster	Mean Q	C.A(sq.Km.)	Return Period (yrs)					
				2	10	20	30	40	50
Mahanadi-1	2	2442.08	9809	2081.2	4671.9	5771.0	6432.9	6913.1	7292.2
Mahanadi-2	2	1368.32	3019	1166.1	2617.7	3233.5	3604.4	3873.5	4085.9
Mahanadi-3	1	740.38	3112	629.8	1402.5	1735.5	1937.3	2084.4	2200.8
Mahanadi-4	1	4939.12	9506	4201.5	9356.5	11577.3	12923.9	13905.0	14681.8
Mahanadi-5	1	13496.26	57760	11480.7	25566.8	31635.3	35314.9	37995.7	40118.4
Mahanadi-6	1	374.99	2210	319.0	710.4	879.0	981.2	1055.7	1114.7
Mahanadi-7	1	3478.04	9730	2958.6	6588.7	8152.5	9100.8	9791.6	10338.7
Mahanadi-8	1	1839.84	5719	1565.1	3485.3	4312.6	4814.2	5179.7	5469.0
Mahanadi-9	2	2105.25	4538	1794.2	4027.5	4975.0	5545.7	5959.6	6286.4
Mahanadi-10	2	781.08	3225	665.6	1494.2	1845.7	2057.5	2211.0	2332.3
Mahanadi-11	1	4172.36	8413	3549.3	7904.0	9780.0	10917.6	11746.3	12402.6
Mahanadi-12	1	1600.03	6990	1361.1	3031.0	3750.5	4186.7	4504.5	4756.2
Mahanadi-13	1	4040.33	16456	3436.9	7653.9	9470.5	10572.1	11374.6	12010.1
Mahanadi-14	2	7884.15	19600	6719.2	15083.1	18631.4	20768.4	22318.7	23542.6
Mahanadi-15	2	5646.14	11960	4811.9	10801.5	13342.7	14873.0	15983.3	16859.7
Mahanadi-17	2	684.00	1266	582.9	1308.6	1616.4	1801.8	1936.3	2042.5
Mahanadi-18	1	465.63	1100	396.1	882.1	1091.4	1218.4	1310.9	1394.1
Mahanadi-19	2	2439.88	6080	2079.4	4667.7	5765.8	6427.1	6906.9	7285.6
Mahanadi-20	1	960.15	2611	816.8	1818.9	2250.6	2512.4	2703.1	2854.1
Mahanadi-21	1	2263.60	6663	1925.5	4288.1	5305.9	5923.0	6372.6	6728.7
Mahanadi-22	2	922.37	1365	786.1	1764.5	2179.7	2429.7	2611.1	2754.3

Table 5.17 Extreme flood values (cumecs) for different return period for ungauged catchments

Catchment Area (sq.km)	Return Period (years)											
	2		10		20		30		40		50	
	Cluster 1	Cluster 2	Cluster 1	Cluster 2	Cluster 1	Cluster 2	Cluster 1	Cluster 2	Cluster 1	Cluster 2	Cluster 1	Cluster 2
100	29.9	65.3	66.3	146.0	82.0	180.4	91.6	201.7	98.6	216.9	104.3	229.2
200	58.6	116.9	130.0	261.2	160.7	322.8	179.7	360.8	193.4	388.1	204.5	410.1
300	86.9	164.3	192.7	367.1	238.3	453.7	266.5	507.1	286.8	545.5	303.2	576.3
400	115.0	209.1	254.9	467.4	315.2	577.6	352.4	645.6	379.3	694.4	401.0	733.7
500	142.8	252.2	316.6	563.6	391.5	696.6	437.8	778.6	471.1	837.5	498.1	884.8
1000	280.0	451.3	620.8	1008.5	767.8	1246.3	858.5	1393.0	923.9	1498.4	976.8	1583.2
2000	549.1	807.4	1217.5	1804.3	1505.6	2229.8	1683.5	2492.3	1811.7	2680.9	1915.5	2832.5
3000	814.3	1134.7	1805.3	2535.8	2232.5	3133.8	2496.3	3502.7	2686.5	3767.7	2840.3	3980.8
4000	1076.8	1444.6	2387.5	3228.3	2952.5	3989.6	3301.3	4459.2	3552.8	4796.7	3756.3	5067.9
5000	1337.6	1742.1	2965.5	3893.2	3667.3	4811.3	4100.6	5377.7	4413.0	5784.6	4665.7	6111.8
6000	1596.8	2030.2	3540.3	4537.0	4378.0	5606.9	4895.3	6266.9	5268.2	6741.1	5570.0	7122.4
7000	1854.8	2310.6	4112.3	5163.6	5085.4	6381.3	5686.2	7132.5	6119.4	7672.2	6469.9	8106.1
8000	2111.7	2584.6	4681.9	5776.0	5789.8	7138.1	6473.9	7978.4	6967.1	8582.1	7366.2	9067.5
9000	2367.7	2853.2	5249.6	6376.2	6491.8	7879.8	7258.8	8807.4	7811.9	9473.8	8259.3	10009.6
10000	2623.0	3117.0	5815.5	6965.7	7191.6	8608.3	8041.3	9621.7	8653.9	10349.8	9149.6	10935.1

## 5.8 COMPARISON OF RESULTS

Earlier in the study of regional flood frequency analysis of Mahanadi basin has been carried out using Wakeby distribution by Singh and Seth in 1985. They have taken annual peak flows for 23 years for 18 stations from the year 1958 to 1980 with drainage area varying between 17 sq. km. to 1150 sq. km. Three stations are used as test stations in that study. The station Kelo is one of the test study which is also station No-17 in this study. The results of flood values for both the studies are compared in Table No. 5.12 and it seems that the values are with very less deviations.

**Table No 5.18 Comparison of flood values for Kelo site (Br. No. 121)**

Return periods	Using regional parameters(L-moment/GEV)	Using regional parameters(Wakeby)	Absolute % difference
2	529.5	559	5.3
10	1188.7	1319	9.9
20	1468.3	1674	12.3
50	1855.3	2173	14.6

(Note- Catchment area of Site-17 of my study is 1266 sq. km and that of Br. No. 121 is 1150. So the flood values are made proportionate accordingly).

Further, the Technical Report of NIH, 1997-98 on sub-zone-3(d) regarding computation of growth factor on L-moment approach is compared with the growth factors obtained in this study.

**Table 5.19 Comparison of results for two studies (between subzone 3(d) and only Mahanadi basin.)**

Return period	NIH-1997-98	This study	
	Subzone-3(d)	Cluster-1	Cluster-2
2	0.828	0.833	0.839
10	1.878	1.889	1.911
20	2.367	2.349	2.368
50	3.086	2.989	3.000
100	3.697	3.502	3.503
200	4.375	4.043	4.029
500	5.389	4.800	4.760
1000	6.257	5.405	5.342

Hence, the results obtained in this study with respect to the previous one are having very less difference .So the application of clustering methods for regionalization of the basin is reasonably acceptable.

## CHAPTER- VI

### CONCLUSIONS AND SCOPE FOR FURTHER STUDY

#### 6.1 CONCLUSIONS

In this study discharge data of 22 G&D sites of the Mahanadi basin have been considered are operated by CWC. Initially the datasets are screened for randomness, trend and outliers. Hosking's program applied to the data set shows the basin is not conforming to the measures of regionalization. Different clustering methods are applied to the basin by taking the catchments variables into consideration. Out of different methods, result showing most common sites are grouped to form clusters. In the process two clusters are formed. Just by interchanging two sites from the clusters two statistically homogeneous clusters are formed. While transferring the sites from one cluster to another the geography of the clusters is so adjusted that almost the basin is divided into upper and lower parts. As both regions are homogeneous and goodness of fit satisfies the distributions like GEV for both clusters, it is very much useful for both gauged and ungauged streams of the entire basin.

For gauged streams  $Q_t$  value for different return periods can be calculated from the growth curve directly or from the quantiles generated from the Hosking's program output. For the ungauged one  $Q_m$  (average discharge) can be calculated from the catchments characteristics by the equation developed and from which  $Q_t$  can be determined. Again to make it more simple  $Q_m$  value for catchment area only is also established. Thus now it becomes more simple to determine  $Q_t$  of any stream if physical location of a site (in order to decide to which cluster it belongs to) along with its catchments area is known.

However, more  $Q_{max}$  values (record length) with sufficient number of reliable gauging stations along with other hydrometeorological characteristics give a better estimate

of  $Q_t$  values and regionalization can be very helpful in pooling flood data such that design flood estimations can be made at ungauged catchments with certain accuracy. With such circumstances and limitations in mind, regional flood growth curves for the state can be developed herein using an approach that is able to minimize the bias due to outliers, shortness of record length (Lim and Lye ). Again pooling of the catchments on the basis of catchments characteristics is more logical as compared to geographic location. On evaluating the performance of different clustering techniques, it is seen that almost 70 to 100 percent catchments are common in clusters formed by different clustering techniques. Out of the four techniques studied none of the techniques is perfect in ensuring the regional homogeneity of clusters. However, the regional homogeneity of different clusters can be achieved by heuristic rearrangement of catchments. The heuristic rearrangement gives additional advantage of more data lengths as some of the catchments may be overlapped in different clusters.

Kohonen Self Organising Map (SOM) is used here to decide the number of clusters. As ANN is fully data dependent a more longer data length would have been given a further better accuracy. This study strives to provide useful results that can be used by those who need to estimate design floods for non-tidally influenced ungauged basins in Orissa.

Although some earlier studies on this aspect is made on subzone -3(d) (Mahanadi basin occupies a major part of subzone-3(d)) by NIH Roorkee, 1994-95 this study is special as far as Mahanadi basin is considered on following grounds.

- (i) This study is uniquely dealing Mahanadi basin alone with as many as 22 G&D sites .
- (ii) L-moment approach is applied.
- (iii) Mahanadi basin is divided into two statistically homogeneous clusters by applying as many as four clustering methods.

- (iv) Six catchments characteristics are derived / collected in order to delineate the clusters and to find the equation. ( $Q_m$ ).
- (v) Individual growth curve and  $Q_t$  equation is developed for clusters for both gauged and ungauged catchments.

## 6.2 SCOPE FOR FURTHER STUDY

Mahanadi basin is such a big basin that it covers almost seventy five thousand sq.km of Chhatisgarh and fifty five thousand sq.km of Orissa as its catchment .The cyclonic storm mostly originates in the Bay of Bengal and moves in a North-Westerly direction and the river flows in the opposite direction. The river falls in the cyclonic track and almost receives a lot of rain in its downstream whereas upstream receives less water as cyclones get weak. If Mahanadi is considered on its flow direction it can be divided into

- upto Hirakud dam ( hilly terrain with steep slope with industrial area)
- Hirakud to Naraj barrage ( flatter slope with heavy industrial and agricultural area)
- Naraj to Bay of Bengal ( more flatter slope with large residential and agricultural area)

As far as vulnerability of flood and residential and industrial growth is concerned the last two parts are more sensitive. It will be better if Mahanadi basin is possible to be divided into atleast three parts so that equations, and growth curves will be more reasonable. For that more number of G&D sites scattered over entire basin with as much record length as possible is to be collected. Data from CWC sites alongwith state government managed sites are to be collected.

The state Orissa is having 11 catchments and all the catchments are subjected to high industrial, agricultural and residential areas. For design of different hydraulic structures as well as a flood forecasting tool determination of flood quantiles for different return periods is

highly essential. So the same regionalization work can be extended to various other basins of Orissa.

The study has a good aspect for application in flood risk mapping of any region. As flood magnitude derived from this study can be related with the corresponding elevation of an area. From this a map can be prepared relating flood magnitude with levels for different return periods. Utility and flood protection (safety measures) of an area can be decided depending upon the flood risk map prepared for that area. So this study can also work like a disaster warning system for flood prone areas.



## REFERENCES

- Acreman, M.C. (1985) Predicting the mean annual Flood from the basin characteristics in Scotland. *Hydrological Science Journal*, Oxford, England ,30 ,37-49.
- Acreman, M.C. and Sinclair, C.D. (1986) Classification of drainage basins according to their physical characteristics: An application for Flood Frequency Analysis in Scotland . *Journal of Hydrology*, Amsterdam, 84,365-380.
- Anderberg, M.R. (1973) *Cluster analysis for applications*, Academic, New York.
- Bezdeck, J.C, Ehrlich, R. and Full, W. (1984) FCM: the fuzzy c-means clustering algorithm. *Comput.Geosci.*102-3.pp.191-203
- Bezdeck, J. C. (1981) *Pattern recognition with fuzzy objective function algorithm*. Plenum Press, New York.
- Bhatt, V.K. (2003) *Estimation of Extreme Flows for Ungauged Catchments*, Ph.D Thesis, I.I.T. Roorkee, India.
- Bobee, B. and Rasmussen, P.F. (1995) *Recent Advances in Flood Frequency Analysis*,*Rev.Geophys.*Vol.33.
- Bogardi, I . and Matyasovszky, I (2002) A hydrometeorological model for drought. *Journal of Hydrology*, Amsterdam 153, 245-264.
- Burn, H. D. and Goel, N. K. (2000) The formation of group for regional flood frequency analysis. *Hydrol. Sci. J* 45(1), 97-112.
- Burn, D. H. (1989) Cluster analysis as applied to regional flood frequency. *J. Water Resour. Plan. Manage.*, ASCE, 115 (5),567-582.
- Chen, H. (1994) *Machine Learning for information Retrieval: Neural Network, symbolic learning and Genetic Algorithms*. *Journal of American society for Information Science*, 46(3), 194-216.

- Central Water Commission, (1973) Estimation of design flood peak. A method based on unit hydrograph principle. Report No.1/73. Hydrology for small catchments directorate . Govt. of India, New Delhi.
- Central Water Commission, (1993) Workshop on rationalization of design storm parameters for design flood estimation. Recommendation. Hydrology studies organization. New Delhi.
- Cunane, C. (1988) Methods and Merits of Regional Flood Frequency Analysis. *J.hydrol.*100.pp.269-290.
- Cunnane, C. (1989) Statistical distributions for flood frequency analysis, WMO, Operational Hydrology Report No.33, WMO No.-718, World Meteorological Organization , Geneva, Switzerland.
- Chowdhury, J.U., Stedinger J.R, and Lu, L.-H. (1989), Goodness-of-fit tests for regional generalized extreme value flood distribution, *Water Resour. Res.*, 27(7), 1765-1776,
- Dalrymple, T. (1960) Flood frequency analysis. Water Supply paper,1543- A US Geol. Survey, Reston Virginia.
- DeCoursey, D. G. (1973) Objective regionalization of peak flow rates. In flood and Droughts, E. F. Koelzer and K. Mahmood (Editors). Proceedings of the second International Symposium in Hydrology, September 11-13,1972, fort Collins, Colorado. Water Resources Publications, pp.395-405.
- Decoursey, D.G. and Deal, R.B.(1974) General aspects of multivariate Analysis with Application to some problems in Hydrology. Proceedings of symposium on Statistical Hydrology. Misc.Pub.No.1275, USDA-ARS,pp.47-68.
- Fovell, R. G. and Fovell .M.Y.C.(1993) Climate zones of the conterminous United States using cluster analysis defined using cluster analysis, *J. climate* , 6(11),2103-2135.
- Goel,N.K., (1998) Flood estimation in sub-Himalayan region using L-moments. International Symposium on 'Hydrology of ungauged streams in hilly regions for small hydro power development'. New Delhi, March 9-10, 1998.

- Goel, N.K. and Seth, S.M. (1985) Data related problems in frequency analysis. Proceeding of Seminar On Flood frequency analysis, N.I.H. Roorkee.
- Goel, N. K and Chander, S. (2002) Regionalisation of Hydrological Parameters. INCOH, Roorkee.
- Goel, N.K., and Arya, D. S. (2006) Completion Report for Development of Dynamic Flood Frequency Model, Indian National Committee on Hydrology (MOWR, New Delhi).
- Government of Orissa, Department of Water Resources, 3<sup>rd</sup> Spiral Study Report of Mahanadi Basin Plan (Volume-1).
- Greenwood, J. A. , Landwehr J. M., Matalas, N.C. and Wallis, J.R. (1979) Probability Weighted Moments : Definition and Relation to parameters of several distributions expressible in inverse form., *Water Resour Res*, 15(5), 1049-1054.
- Hall, M. J. and Minns, A.W. (1999) The Classification of Hydrological Homogeneous Regions. *Journal of Hydrology*.
- Hierarchical Clustering . (Dt.10.06.2008) , [http://www.resample.com/xlminer/help/HClst/HClst\\_intro.htm](http://www.resample.com/xlminer/help/HClst/HClst_intro.htm).
- Hosking, J. R. M. (1990), L-moments : analysis and Estimation of distributions using linear combinations of ordered statistics. *J.R. Stat. Soc. Ser B. Methodol.*, 52(2), 105-124.
- Hosking, J. R. M. (1991a), Approximations for use in constructing L-moment ratio diagrams. Res. Rep. RC-16635,3, IBM Research Division T.J Watson Research center, York Town Heights, New York.
- Hosking, J. R. M. and J. R. Wallis (1991), "Some statistics useful in Regional frequency analysis", Res. Rep. RC 17096, IBM Research Division, York Town Heights, NY 10598.
- Hosking, J. R. M. and Wallis, J. R. (1993), Some statistics useful in regional frequency analysis., *Water Resources Res.*, 29(2), 271-281.

- Hosking, J. R. M. and Wallis J. R. (1997), *Regional Frequency Analysis: An approach based on L-moments*, Cambridge University Press, Cambridge(1997).
- <http://en.wikipedia.org/wiki/K-means>, K-means Clustering.
- Jains, A. K. and Dubes, R. C.(1988) *algorithm for clustering data*. Prentice Hall.
- Kaufman L. and Rousseeuw, P. J(1990) *Finding Groups in Data: An Introduction to Cluster Analysis*, Wiley,(MATLAB).
- Kiang, M. Y , Kulkarni U. R,Goul, M. R, Chi,R. T, Terban, E and Philppakkis, A. (1997) *Improving the Effectiveness of Self Organising Map Networks using a Circular Kohonen Layer*.
- Kumar, R. and Chatterji C., (2005), *Regional Flood Frequency Analysis using L-moments for North Brahmaputra Region of India*. *Journal of Hydrol. Engg*
- Kumar, R., Chatterjee, C., Kumar, S., Lohani, A. K., & Singh, R. D. (2003) *Development of regional flood frequency relationships using L-moments for middle Ganga Plains subzone 1(f) of India*. *Water Resources Management*. 17, 243-257.
- Kohonen, T . (1997), *Self Organisation Map (2<sup>nd</sup> Edition)* Springer.Berlin ISBN3-540-62017-6.
- Kurothe, R. S., Goel, N. K. and Mathur, B.S. (2001) *Derivation of curve number and kinematic wave based flood frequency distribution*. *Hydrol. Sci. J.*, 46(4),571-584.
- Landwehr, J.M., Matalas, N.C. and Wallis, J.R. (1979a) *Probability Weighted Moments compared with some traditional techniques in estimating Gumbel Parameters and quantiles*. *Water Resour.Res.*,15(5),1055-1064.
- Landwehr, J.M., Matalas, N.C. and Wallis, J. R. (1979 b) *Estimation of parameters and quantiles of Wakeby distribution*. *Water Resour.Res.*,15(6),1361-1379.
- Lim, Y. H. , Lye L. M., (2003) *Regional Flood Estimation for Ungauged Basins in Sarawak, Malaysia*.
- MacQueen, J. B. (1967) *Some methods for classification of multivariate observation*.Proc.5<sup>th</sup> Berkeley Symp. on Probability and Statistics. University of California Press, Berkely, 281-297.

- Modares, R (2007) Pooled dry spells of frequency analysis: L-moment vs Multivariate analysis, Draft manuscript.
- Mosley, M. P. (1981) Delineation of New Zealand hydrologic regions. *J. Hydrol.* 49, 173-192.
- Nathan, R. J. and McMahon, T. A., (1990) Identification of Homogeneous Regions for the Purpose of Regionalisation. *J. Hydrol.* 121, pp. 217–238.
- National Water Development Agency, (2004) Technical Study No.174, Water Balance Study of Mahanadi basin upto Hirakud dam (Revised).
- NERC, (1975) Flood Studies Report. Natur. Environ. Res. Council, London, Vols.1-5, 1100 pp.
- N.I.H., Report (1985-86) Regional Flood Frequency Analysis, CS-9, NIH, Roorkee.
- N.I.H., (1987-88) Workshop on Flood Frequency Analysis, NIH, Roorkee.
- N.I.H., (1994-95), Development of Regional Flood Formula for Mahanadi Subzone-3(d). NIH, Roorkee.
- N.I.H., Report (1997-98) Regional flood Frequency Analysis using L-moment, Technical Report, TR (AR), NIH, Roorkee.
- Parida, B. P., (2000) A Partitioning Methodology for Identification of Homogeneous Regions in Regional Flood Frequency Analysis.
- Parida, B.P., Kachroo R.K. and Shrestha D.B. (1998) Regional Flood Frequency Analysis of Mahi Sabarmati Basin (subzone 3-a) fusing index flood procedure with L-moments., *Water Resource Management*, 12, 1-12.
- Peel, M. C. et al, (2001) The utility of L-moment ratio diagram for selecting a regional probability distribution. *Hydrol. Sci. J.*, 46(1), 147-155.
- Pilon, P.J. and Adamowski, K. (1992) the value of regional information to flood frequency analysis using the method of L-moments. *Can. J. civil Eng.*, 19, 137-147.
- Potter, K.W. and Lettenmaier, D.P. (1990) A comparison of regional flood frequency estimation methods using a resampling method, *Water Resour. Res.*, 26(3), 415-424.

- Pilgrim, D. H. and Cordery, I. (1992) Flood runoff – Handbooks of Hydrology, D.R.Maidment,ed.,McGraw-Hill, New York.
- Rai, R. K, Srivastava, S. K. and Jain, M. K (2003) Fitting of frequency distribution function of rainfall for Midnapore district of West Bengal (India ) –A case study, International conference on Water and Environment, Bhopal.
- Rao, A. R. and Srinivas, V.V., (2006) Regionalisation of Watersheds by Fuzzy Cluster Analysis, Journal of Hydrology,Netherland.
- Rao A. R. and Hamed, H.K. (2000) Flood Frequency Analysis. CRC Press, Boca Raton Florida, U.S.
- Research Design and Standards Organization (RDSO), (1991) Estimation of design discharge based on regional flood frequency approach for sub zones 3 (a), 3(b), 3(c) and 3(e) ., Bridges and Floods wing rep. No.20, Lucknow.
- Robson, A. and Reed D. (1999) Statistical procedure for flood frequency estimation, Flood Estimation Handbook.
- Self organization map (10.06.2008) [http:// www. willamette. edu/ ~gorr/ classes/ cs449/ Unsupervised /SOM.html](http://www.willamette.edu/~gorr/classes/cs449/Unsupervised/SOM.html), Self Organisation Map.
- Sengupta, S. K., Bales, J. D., Juback, R., Scott, A. C. and Kane M.D.(2006) Flood forecasting and inundation mapping in the Mahanadi river basin – A collaborative effort between India and U.S.
- Shi, J. J., (2002) Clustering technique for evaluating and validating neural network performance, Journal of Computing in Civil Engineering, Vol.16,No.2,2002.
- Singh,R.D. and Seth,S.M. (1985) Regional flood frequency analysis for Mahanadi basin using Wakeby distribution Proceedings, Seminar on flood frequency analysis, New Delhi
- Stambuk, A., Stambuk N., Konjevoda P., (2007) Application of Kohonen Self-Organising Maps (SOM) based Clustering for the Assessment of Religious Motivation

- Tasker, G.D. (1982)- Comparing methods of hydrologic regionalization, *Water Resources Bulletin*, 18(6):965-970.
- Thandaveswara, B. S. and Sajjikumar, N. (2000) Discussion of "Classification of River Basins using Artificial Neural Network", Vol. 5 No. 3 pp. 290 -298.
- Vogel, R. M. and Fennesy, N. M. (1993) L-moment diagrams should replace product moment diagrams, *Water Resour. Res.*, 29(6),1745-1752.
- Vogel, R. M., McMohan, T. A. and Chiu, F. H. S.(1993a) Flood flow frequency model selection in Australia. *J. Hydrol.*, 146, 421-449.
- Wiltshire, S.E. (1985) Grouping of Catchments for Regional Flood Frequency Analysis. *Hydrol. Sci. J.* 30, pp. 151–159.
- World Meteorological Organisation,(1989) Statistical distributions for flood frequency analysis. (operational hydrology report no.33),Geneva,Switzerland.
- Zafirakou - Kouloris, A., Vogel R.M., Craig, S.M. and Habermeier, J. (1998) L-moment diagrams for censored observations, *Water Resources*, 34(5),1241-1249.
- Zhang, J. and Hall, M. J., (2004) Regional flood frequency analysis for the Gan-Ming river basin in China, *Journal of Hydrology*.