FINAL REPORT

FHWA/IN/JTRP-2002/2

REGIONALIZATION OF INDIANA WATERSHEDS FOR FLOOD FLOW PREDICTIONS (PHASE I)

Studies in Regionalization of Watersheds

by

A. Ramachandra Rao
Professor
Principal Investigator

School of Civil Engineering
Purdue University

Joint Transportation Research Program
Project No. C-36-62K
File No. 9-8-11
SPR-2476

Prepared in Cooperation with the
Indiana Department of Transportation
Federal Highway of Administration
U.S. Department of Transportation

The contents of this report reflect the views of the authors who are responsible for the facts and the accuracy of the data presented herein. The contents do not necessarily reflect the official views or policies of the Federal Highway Administration and the Indiana Department of Transportation. The report does not constitute a standard specification, or regulation.

School of Civil Engineering
Purdue University
February 2004

TECHNICAL REPORT STANDARD TITLE PAGE

1. Report No. FHWA/IN/JTRP-2002/2	**2. Government Accession No.**	**3. Recipient's Catalog No.**

4. Title and Subtitle
Regionalization of Indiana Watersheds for Flood Flow Predictions Phase I
Studies in Regionalization of Indiana Watersheds

5. Report Date
February 2004

6. Performing Organization Code

7. Author(s)
A. R. Rao

8. Performing Organization Report No.
FHWA/IN/JTRP-2002/2

9. Performing Organization Name and Address
Joint Transportation Research Program
1284 Civil Engineering Building
Purdue University
West Lafayette, IN 47907-1284

10. Work Unit No.

11. Contract or Grant No.
SPR-2476

12. Sponsoring Agency Name and Address
Indiana Department of Transportation
State Office Building
100 North Senate Avenue
Indianapolis, IN 46204

13. Type of Report and Period Covered
Final Report

14. Sponsoring Agency Code

15. Supplementary Notes

Prepared in cooperation with the Indiana Department of Transportation and Federal Highway Administration.

16. Abstract

The following five methods of regionalization of watersheds were tested with Indiana watershed and annual maximum flood data: (1) the L-moment based method, (2) the method based on hybrid cluster analysis, (3) the hybrid cluster method using rainfall data, (4) the fuzzy cluster method, and (5) the method based on artificial neural networks.

The results of the L-moment based method and the hybrid cluster method with rainfall data were unacceptable because of the subjectivity involved with the former and the heterogeneity of the of the results obtained by the latter. The remaining three methods gave very similar results. The fuzzy cluster and artificial neural network based methods are much easier to use and hence are recommended.

The results from any of these methods will not give homogeneous regions. The results from the clustering methods must be tested and revised to get homogeneous watersheds.

The data from each of the regions were investigated by using tests based on simple scaling. The results from these tests confirm all the regions, except one, to be homogeneous.
.

17. Key Words

regionalization, watersheds, flood frequencies, scaling, cluster analysis.

18. Distribution Statement

No restrictions. This document is available to the public through the National Technical Information Service, Springfield, VA 22161

19. Security Classif. (of this report)	**20. Security Classif. (of this page)**	**21. No. of Pages**	**22. Price**
Unclassified	Unclassified	93	

TECHNICAL *Summary*

Technology Transfer and Project Implementation Information

INDOT Research

TRB Subject Code: 22 Hydrology and Hydraulics *February 2004*
Publication No.: FHWA/IN/JTRP-2002/2, SPR-2476 *Final Report*

REGIONALIZATION OF INDIANA WATERSHEDS FOR FLOOD FLOW PREDICTIONS PHASE I

Studies in Regionalization of Watersheds

Introduction

Several studies have claimed that regionalization of watersheds is essential to develop regional flood flow equations. These flood flow equations would be used to estimate flood magnitudes at locations where actual flood data are not available.

Although several regionalization methods have been proposed, there is no agreement about the method or methods which are to be used. In this study of regionalization of Indiana watersheds, a two-step procedure was adopted. In the first step, regionalization methods in use were reviewed and the most promising of these were selected for testing. In the second step, the selected methods were tested by using the watershed and flow data. The following regionalization methods were tested:

- The L-moment based method
- The method based on hybrid cluster analysis
- The hybrid cluster method using rainfall data
- The method based on fuzzy cluster analysis
- The method based on artificial neural networks.

Findings

The L-moment based method requires subjective judgment in regionalizing watersheds. Consequently, the results would not be unique and hence unacceptable. The hybrid cluster method is superior to the L-moment method, but is computationally quite involved. The hybrid cluster method in which rainfall data were used gave unacceptable results. The fuzzy cluster and artificial neural network based methods were the easiest methods. The regionalization results from the hybrid cluster, fuzzy cluster, and artificial neural network methods were identical.

Another important finding of the study is that the results from any of these methods will not give statistically homogeneous regions. The results from cluster analysis will have to be tested and the regions revised before arriving at statistically homogeneous regions.

The characteristics of flood data from these regions were tested by noting tests based on simple scaling. The data from the homogeneous regions were found to behave as expected.

Contacts

For more information:

Prof. A. R. Rao
Principal Investigator
School of Civil Engineering
Purdue University
West Lafayette IN 47907
Phone: (765) 494-2176
Fax: (765) 496-1988

Indiana Department of Transportation
Division of Research
1205 Montgomery Street
P.O. Box 2279
West Lafayette, IN 47906
Phone: (765) 463-1521
Fax: (765) 497-1665

Purdue University
Joint Transportation Research Program
School of Civil Engineering
West Lafayette, IN 47907-1284
Phone: (765) 494-9310
Fax: (765) 496-7996

TABLE OF CONTENTS

Page

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

This is the final report of the project “Regionalization of Indiana Watersheds for Flood Flow Predictions (Phase I)” supported by the Joint Transportation Research Program (Indiana SPR-2476, 656-1284-0137, file: 9-8-11).

Several investigators worked very enthusiastically on this project. They are: graduate students Steve Ernst, Michelle Iblings and Heather Finfrock; post-doctoral fellow, Dr. V. V. Srinivas and my colleagues, Prof. G.D. Jeong and Prof. Rao S. Govindaraju. It was a pleasure to work with them, and I would like to acknowledge their contribution.

We would like to acknowledge the support by JTRP, which made the study possible. Messers Merril Dougherty, David Finley, and Dave Ward of INDOT, Keith Hoernschemeyer of FHWA, David Knipe and Raj Gosine of IDNR served on the Study Advisory Committee. They were also of considerable help in conducting the study. Ms. Karen Hatke, Administrator of JTRP was of considerable and cheerful help in administrative matters. Mrs. Dinah Hackerd, Ms. Jan Jennings and Mr. Jack Chan assisted in numerous ways, especially in preparing this report. It is a pleasure to thank them for all their help.

I INTRODUCTION

Better estimation of flood magnitudes from ungaged catchements corresponding to specific recurrence intervals is an important common problem in hydrologic design. In order to develop these regression relationships relating the magnitude of floods to the physiographic and meteorologic characteristics, accurate definition of hydrologically homogeneous watersheds is needed. Such a definition does not presently exist for Indiana watersheds. Also, there is no general consensus about the performance of different regionalization methods. Consequently there are two research needs. (1) The first one relates to selecting a regionalization method or methods, based on objective criteria, which may be used. (2) The second one is the application of the selected regionalization procedures to Indiana watersheds to identify hydrologically homogeneous watersheds. The objectives of research disclosed herein are:

- Selection of a regionalization method or methods, which would be used for regionalization of watersheds. The results of regionalization would be used to develop regional flood frequency models.

- Apply the methods selected under the first objective to Indiana watersheds so that watersheds which are homogeneous in their properties – with respect to flood response – are identified.

1.1 Models Used for Regionalization

By a thorough literature search the regionalization models which are being used for regionalization were examined. The theoretical soundness and the practical use of these models were considered. The models and methods which were examined included:

(a) The method of residuals – U.S.G.S. approach.

(b) The L-moment based regional analysis

(c) Other methods based on cluster algorithms and neural network approaches.

The index flood method was the first method developed for regionalization of watersheds (Dalrymple, 1960). Condie (1980) used it, along with the three parameter log normal distribution for regionalization of Canadian watersheds. Stedinger and Tasker (1985) and Tasker (1989) developed the method of residuals which is used by the USGS for regionalization. Waylon and Woo (1981) developed a regionalization procedure which was tested by using data from Canadian watersheds.

Bhaskar et al. (1989) and Bhaskar and O'Connor (1989) compared the results by the method of residuals and by the clustering algorithms for data from Kentucky. Cluster analysis yielded regions that were not similar to those defined by the method of residuals nor coincidental with geographical boundaries. However they were more distinguishable and better defined in terms of the hydrologic response than the USGS regions.

Burn (1988, 1990a, 1990b, 1997), Burn et al. (1997), Zrinji and Burn (1994, 1996, 1997) have been developing procedures for regionalization. The region of influence approach which they are developing defines regions such that each site has a potentially unique combination of regions. The similarity of the selected regions is assured by using homogeneity tests such as those used by Rao and Hamed (1997). A hierarchical feature, which uses the spatial similarity scales which are observed for different moment orders of flood frequency distributions, has also been used in the procedure. Monte Carlo tests have demonstrated that flood quantile estimation is substantially improved with the region of influence approach.

The region of influence approach by Burn and his associates is an implementation of technique of regionalization without a fixed region which was developed by Acreman (1987) and

Acreman and Wiltshire (1987). Whiltshire (1986a,b,c) also developed a procedure for regionalization based on basin characteristics. Provoznik and Hotchkiss (1998) have used the region of influence approach for Nebraska watersheds.

There have been other techniques of regionalization also. Nathan and McMahon (1990) have compared some of the regionalization techniques using Australian low flow data. The approaches they have tested are based on combination of cluster analysis, multiple regression and principal component analysis. Cavadias (1990) has used the canonical correlation approach to regional flood frequency analysis. Nguyen et al. (1997) have used the scaling approach to regionalization. Ouarda et al. (1997) have used the canonical correlation for regionalization. Smith (1989) has used the extreme order statistics for regional flood frequency analysis.

Another regionalization procedure based only on the L-moments has been proposed by Hosking and Wallis (1997). The procedures in Hosking and Wallis (1997) deserve to be used in any regionalization study.

Although these methods are available, most of the regionalization studies in the United States have been based only on the method of residuals (Curtis (1987), Choquette (1988), Eash (1993), Flippo (1990), Guimaraes and Bohmann (1992), Koltun and Roberts (1990), Lara (1987) and Reich (1988)). The main reason for this situation is that all of these studies are conducted by U.S.G.S and the procedure used reflects the institutional preference. However, several of the studies cited above indicate that better regionalization methods yield more accurate flood frequency estimates and hence they deserve to be investigated and used.

The basic question that arises in using these newly developed techniques is the selection of one or two of these. A striking fact of research in regionalization is that very few (Bhaskar and O’Connor (1989), Nathan and McMahon (1990)) studies have been undertaken to compare regionalization methods to arrive at superior methods based on objective measures. In fact, procedures which have proved to be very good in classification of data, based on Neural Networks (Govindaraju and Rao, 2000) have not even been investigated. These considerations have led to the research objectives proposed herein.

After a review of the literature it was decided to test several methods of regionalization with Indiana data. These included (a) a method based on L-moments only, (b) a method based on Hybrid cluster analysis with flood and geomorphologic data, (c) a hybrid cluster method in which flood and rainfall data are used, (d) a method based on fuzzy cluster analysis and (e) a method based on neural networks. The results from methods based on hybrid cluster analysis, fuzzy cluster analysis and neural network analysis were refined by using homogeneity tests. These methods also gave consistent results. Hence the results from these methods have been accepted as being valid for Indiana. The details of these studies are available in six interim reports. These are:

1. A.R. Rao, S. Ernst and G.D. Jeong (2002). “Results from L-moment based method”, Interim report FHNA/JTRP-2002-2, Joint Transportation Research Program, School of Civil Engineering, Purdue University, W. Lafayette, IN 47907, pp. 98.

2. V.V. Srinivas and A.R. Rao (2002). “Regionalization of Indiana Watershed by Hybrid Cluster Analysis”, Interim report FHNA/IN/JTRP-2002-2, Joint Transportation Research Program, School of Civil Engineering, Purdue University, W. Lafayette, IN 47907, pp. 112.

3. M.L. Iblings and A.R. Rao (2003). “Use of Precipitation and Flow Data for Regionalization of Watersheds”, Interim report FHNA/JTRP–2002–2 Joint Transportation Research Program, School of Civil Engineering, Purdue University, W. Lafayette, IN 47907, pp. 191.

4. V.V. Srinivas and A.R. Rao (2003). “Regionalization of Indiana Watershed by Fuzzy Cluster Analysis”, Interim report FHNA/JTRP–2002–2 Joint Transportation Research Program, School of Civil Engineering, Purdue University, W. Lafayette, IN 47907, pp. 123.

5. V.V. Srinivas, A.R. Rao and R.S. Govindaraju (2003). “Regionalization of Indiana Watersheds Using Artifical Neural Networks”, Interim report FHNA/JTRP–2002–2 Joint Transportation Research Program, School of Civil Engineering, Purdue University, W. Lafayette, IN 47907, pp. 61.

6. H.L Finfrock and A.R. Rao (2003). “Studies in Flood Frequency Analysis”, Interim report FHNA/JTRP–2002–2 Joint Transportation Research Program, School of Civil Engineering, Purdue University, W. Lafayette, IN 47907, pp. 87.

II REGIONALIZATION BY L-MOMENT METHOD

2.1 Introduction

The first objective of the regionalization study by the L-moment method is the determination of the accuracy of Glatfelter's (1984) equations for flood magnitudes in Indiana. The accuracy of Glatfelter's (1984) equations is investigated by inspecting and quantifying the residuals between flood estimates computed by USWRC (1981) methods and by Glatfelter's regression equations.

The investigation of the approach by Burn et al. (1997) for regionalization of Indiana watersheds is the second objective of this present study. Burn et al.'s approach considers the dates of occurrence of annual maximum flood events. According to Burn et al., a group of stations in which annual maximum floods in a region occur simultaneously should be considered as a potential homogeneous group for regional flood frequency analysis.

The third objective of the study discussed in this chapter is to use Hosking and Wallis' (1997) approach for the regionalization of Indiana watersheds. Hosking and Wallis' approach considers L-moments calculated from annual maximum flows. Groups of stations having acceptably homogeneous L-moment characteristics are identified as regions for flood frequency analysis.

2.2 Data Used in the Study

The data used in this study are obtained from several sources. Equations for peak flows are given in a report by Glatfelter (1984) to compute flood magnitudes for recurrence intervals of 2, 10, 25, 50, and 100 years. Results from flood frequency analyses were obtained from the Indiana Department of Natural Resources (IDNR) Division of Water (2001). The IDNR analysis was performed on all past and present gaging stations in the State of Indiana and for neighboring stations in Illinois, Michigan, and Ohio

The IDNR file contains output from flood frequency analyses performed by IDNR. Annual peak flow information is contained in the file. Information in this file includes day and month of occurrence, as well as magnitudes of peak flow events used by IDNR to calculate flood magnitudes for 2, 10, 25, 50, and 100-year recurrence intervals. The results from the flood frequency analysis performed by IDNR are adjusted for outliers and historic events.

The United States Geological Survey (USGS) website (http://water. usgs.gov/nwis/) for the National Water Information System (NWIS) is used to obtain site-specific geographic information and characteristics. Site information obtained from the USGS website includes latitude and longitude values, station elevation, and drainage area information.

Two sets of data are used in this study. The first set includes stations that are used for the purpose of determining the accuracy of Glatfelter's equations discussed in Section 2.3. Locations of the 160 stations within the State of Indiana used for the determination of the accuracy of the Glatfelter equations are shown in Figure 1.

The second set of data used in the present study consists of annual maximum flow recorded at stations in Indiana, Illinois, Michigan, and Ohio. These data are used to perform a regionalization study for Indiana watersheds discussed in Sections 2.4 and 2.5. The locations of the 264 stations in Indiana, Illinois, Michigan, and Ohio, the data from which are used in the regionalization study, are shown in Figure 2.

2.3 Evaluation of Glatfelter Equations

The accuracy of Glatfelter's equations is determined by comparing residuals between flood magnitudes calculated by IDNR (2001) and by Glatfelter (1984). Flood magnitudes calculated by IDNR follow United States Water Resources Council (USWRC, 1981) guidelines. Flood magnitudes calculated by Glatfelter are based on equations derived from a multiple regression analysis of basin characteristics. A comparison of residuals is performed for 2, 10, 25, 50, and 100-year recurrence intervals. The distributions of residual values are analyzed by using spatial plots.

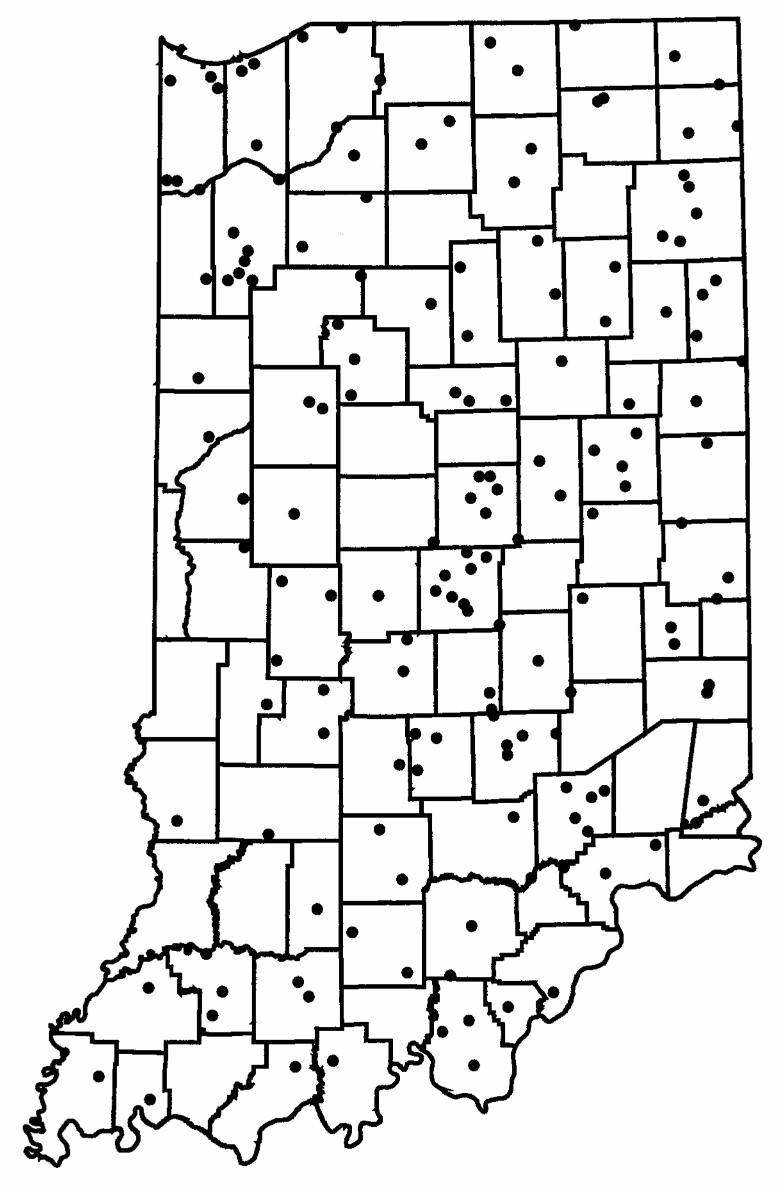

Figure 1. Locations of stations used to estimate of the accuracy of the Glatfelter equations

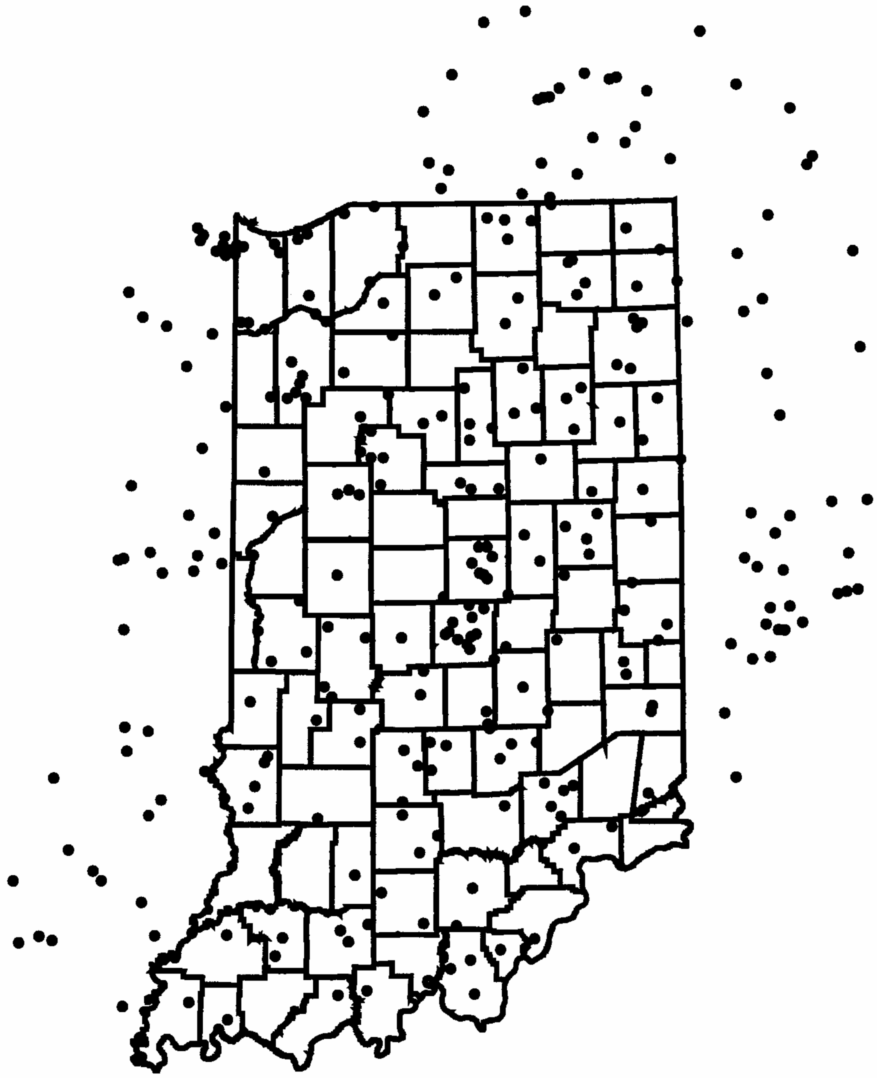

Figure 2. Locations of stations used for the regionalization of Indiana watersheds using L-moments

The IDNR flood frequency analyses were performed using the HEC-FFA (1982) software package produced by the United States Army Corps of Engineers. HEC-FFA follows United States Water Resources Council (USWRC, 1981) guidelines for flood frequency analysis. USWRC guidelines recommend a Log-Pearson Type III distribution be used as the distribution to characterize flood flow frequency and magnitudes of instantaneous annual peak flows (USGS, 1998). IDNR assumed generalized station skew to be –0.2 for all stations in their analyses (INDR, 2001). The reader is referred to USWRC (1981) for further details on the methods used in the INDR flood frequency analysis.

Equations for estimating the magnitude and frequency of floods at ungaged sites on unregulated non-urban streams in Indiana are presented in Glatfelter (1984). The equations were developed using a multiple regression analysis of peak flow data and basin characteristics. Information from 242 stations was used by Glatfelter to develop the multiple regression equations.

Glatfelter equations for the 100-year recurrence interval are listed below as examples of these multiple regression equations. Equations for other recurrence intervals are similar to the 100-year equations but with different coefficients and exponents.

$$Q_{100} = 13.8 DA^{0.695} (STOR + 1)^{-0.243} (PREC - 30)^{1.132} \qquad \text{Region 1}$$

$$Q_{100} = 127 DA^{0.608} (STOR + 1)^{-0.418} RC^{0.902} (PREC - 30)^{0.708} \qquad \text{Region 2}$$

$$Q_{100} = 181 DA^{0.779} SL^{0.466} (I_{24,2} - 2.5)^{0.831} \qquad \text{Region 3}$$

$$Q_{100} = 32.0 DA^{0.565} SL^{0.705} L^{0.730} (I_{24,2} - 2.5)^{0.464} \qquad \text{Region 4}$$

$$Q_{100} = 45.5 DA^{0.760} SL^{0.529} \qquad \text{Region 5}$$

$$Q_{100} = 4734 DA^{0.570} RC^{0.834} (I_{24,2} - 2.5)^{2.068} \qquad \text{Region 6}$$

$$Q_{100} = 70.1 DA^{0.285} SL^{0.488} L^{0.785} RC^{0.854} \qquad \text{Region 7}$$

In the Glatfelter equations, DA is the basin drainage area, STOR is the basin storage, PREC is the average annual precipitation, RC is the basin runoff coefficient, SL is the basin slope, L is the basin length, and $I_{24,2}$ is the two-year, 24-hour precipitation intensity. Glatfelter presents the equations for seven areas within the State of Indiana. The region boundaries used by Glatfelter are essentially the watershed boundaries. The seven Glatfelter regions and watershed boundaries for Indiana are found in Glatfelter's (1984). Standard errors of estimate of flood magnitudes given by these equations range from 24 to 45 percent (Glatfelter, 1984)

The accuracy of these multiple regression equations were determined by computing their standard errors. The standard error is determined by comparing multiple regression based flood magnitudes with flood magnitudes estimated by using USWRC guidelines. The analyses by Glatfelter (1984) were performed 17 years ago. Using the additional years of peak flow data since the Glatfelter (1984) study, the accuracy of the Glatfelter equations is reevaluated in this study. Flood magnitudes for Indiana watersheds are computed by IDNR (2001) by using updated peak flow records.

Regional flood frequency analysis uses the assumption that gaging stations within a region represent catchments with homogeneous hydrologic responses. In order for the Glatfelter regional flood frequency equations to be applied to a ungaged catchment within a region, the watersheds used to develop the regression equations must be sufficiently homogeneous in hydrologic response. However, the Glatfelter equations make estimates on a regional basis without checking the homogeneity of their watersheds.

The accuracy of the Glatfelter equations is assessed by comparing flood magnitudes from Glatfelter multiple regression equations with flood magnitudes from the IDNR flood frequency analysis. The Glatfelter equations are investigated for the seven regions defined by him. The recurrence intervals of 2, 10, 25, 50, and 100-years are investigated for each region. The regions considered by Glatfelter would be considered homogeneous if significant bias is not observed between INDR and Glatfelter flood estimates.

Residuals are used to assess the accuracy of the Glatfelter equations. A residual is the difference between flood magnitudes calculated by IDNR, (Q_I), using USWRC recommended methods and flood magnitudes calculated using Glatfelter (Q_G) regression equations. For the purpose of this study, a residual is defined as:

$$e = Q_I - Q_G \qquad (1)$$

An example of residual values calculated for stations used in the assessment of the accuracy of the Glatfelter equations in given in Table 1. Results are presented of the seven Glatfelter regions for the recurrence intervals of 2, 10, 25, 50, and 100 years in Rao et al. (2002). Large ranges of residual values are present for all regions and recurrence intervals. In order to spatially inspect residuals within individual Glatfelter regions, the percent difference between residuals were used. The percent difference is defined in Equation 2.

$$e(\%) = \frac{Q_I - Q_G}{Q_I} \times 100 \qquad (2)$$

The percent difference gives an accurate representation of the bias between IDNR flood estimates and Glatfelter flood estimates. Examples of percent difference residual findings for 2, 10, 25, 50, and 100-year recurrence intervals is provided in Table 2.

It is easy to determine whether an overestimation or an underestimation is occurring at a specific gaging station by examining its individual residual values. General bias for a region is also estimated by counting the number of positive and negative residual values for a given return period. However, it is difficult to understand, by using only the numbers, the situation at neighboring stations. In addition, it is difficult to form conclusions about ungaged locations throughout a region. For this reason, a spatial inspection technique was used.

The residual percent difference data are examined. Residual results are analyzed spatially within the seven regions proposed by Glatfelter. Stations used in the estimation of the

Table 1. Flood magnitudes and residual values for Region 1

Station mber	Station Name	Q_G					Q_I					*e (3.1))*				
		Q_2	Q_{10}	Q_{25}	Q_{50}	Q_{100}	Q_2	Q_{10}	Q_{25}	Q_{50}	Q_{100}	Q_2	Q_{10}	Q_{25}	Q_{50}	Q_{100}
4093000	Deep River at Lake e Outlet at Hobart, Ind.	1050	1810	2180	2450	2720	1610	3080	3850	4430	5020	560	1270	1670	1980	2300
4093500	Burns Ditch at Gary, Ind.	1390	2350	2820	3160	3500	1590	2750	3310	3710	4110	200	400	490	550	610
4094000	Little Calumet River at Porter, Ind.	1130	2090	2560	2940	3280	1160	2370	3080	3650	4260	30	280	520	710	980
4094500	Salt Creek near McCool, Ind.	925	1630	1980	2240	2490	1030	2100	2710	3170	3660	105	470	730	930	1170
4095300	Trail Creek at chigan City, Ind.	1270	2450	3070	3540	4010	1080	2570	3570	4430	5400	-190	120	500	890	1390
4096100	Galena River near LaPorte, Ind.	317	954	819	952	1080	265	573	776	950	1150	-52	-381	-43	-2	70
4099510	Pigeon Creek Nr Angola, Ind.	509	851	1010	1130	1240	376	632	764	864	966	-133	-219	-246	-266	-274
4099750	Pigeon River near Scott, Ind	1060	1770	2100	2370	2600	1210	1890	2210	2450	2680	150	120	110	80	80
4100220	Waldron Lake near osperville, Ind.	513	878	1050	1180	1310	420	654	758	832	902	-93	-224	-292	-348	-408
4100222	Nb Elkhart River at osperville, Ind.	467	812	974	1100	1220	466	734	861	954	1050	-1	-78	-113	-146	-170
4100500	Elkhart River at Goshen, Ind.	2280	3860	4620	5180	5740	2850	4700	5560	6180	6760	570	840	940	1000	1020
4101000	St. Joseph River at Elkhart, Ind	7340	12300	14700	16400	18200	9720	15200	17900	19900	21900	2380	2900	3200	3500	3700
											Mean	293.8	458.2	622.2	739.8	872.3
											SD	701.1	901.9	987.8	1099.0	1198.6
											Count S.	7	8	8	8	9

Table 2. Summary of percent difference residual findings for Region 1

Station Number	Station Name	*e*(%) (2)				
		Q_2	Q_{10}	Q_{25}	Q_{50}	Q_{100}
4093000	Deep River at Lake George Outlet at Hobart, Ind.	34.8	41.2	43.4	44.7	45.8
4093500	Burns Ditch at Gary, Ind.	12.6	14.5	14.8	14.8	14.8
4094000	Little Calumet River at Porter, Ind.	2.6	11.8	16.9	19.5	23.0
4094500	Salt Creek near McCool, Ind.	10.2	22.4	26.9	29.3	32.0
4095300	Trail Creek at Michigan City, Ind.	-17.6	4.7	14.0	20.1	25.7
4096100	Galena River near LaPorte, Ind.	-19.6	-66.5	-5.5	-0.2	6.1
4099510	Pigeon Creek Nr Angola, Ind.	-35.4	-34.7	-32.2	-30.8	-28.4
4099750	Pigeon River near Scott, Ind	12.4	6.3	5.0	3.3	3.0
4100220	Waldron Lake near Cosperville, Ind.	-22.1	-34.3	-38.5	-41.8	-45.2
4100222	Nb Elkhart River at Cosperville, Ind.	-0.2	-10.6	-13.1	-15.3	-16.2
4100500	Elkhart River at Goshen, Ind.	20.0	17.9	16.9	16.2	15.1
4101000	St. Joseph River at Elkhart, Ind	24.5	19.1	17.9	17.6	16.9

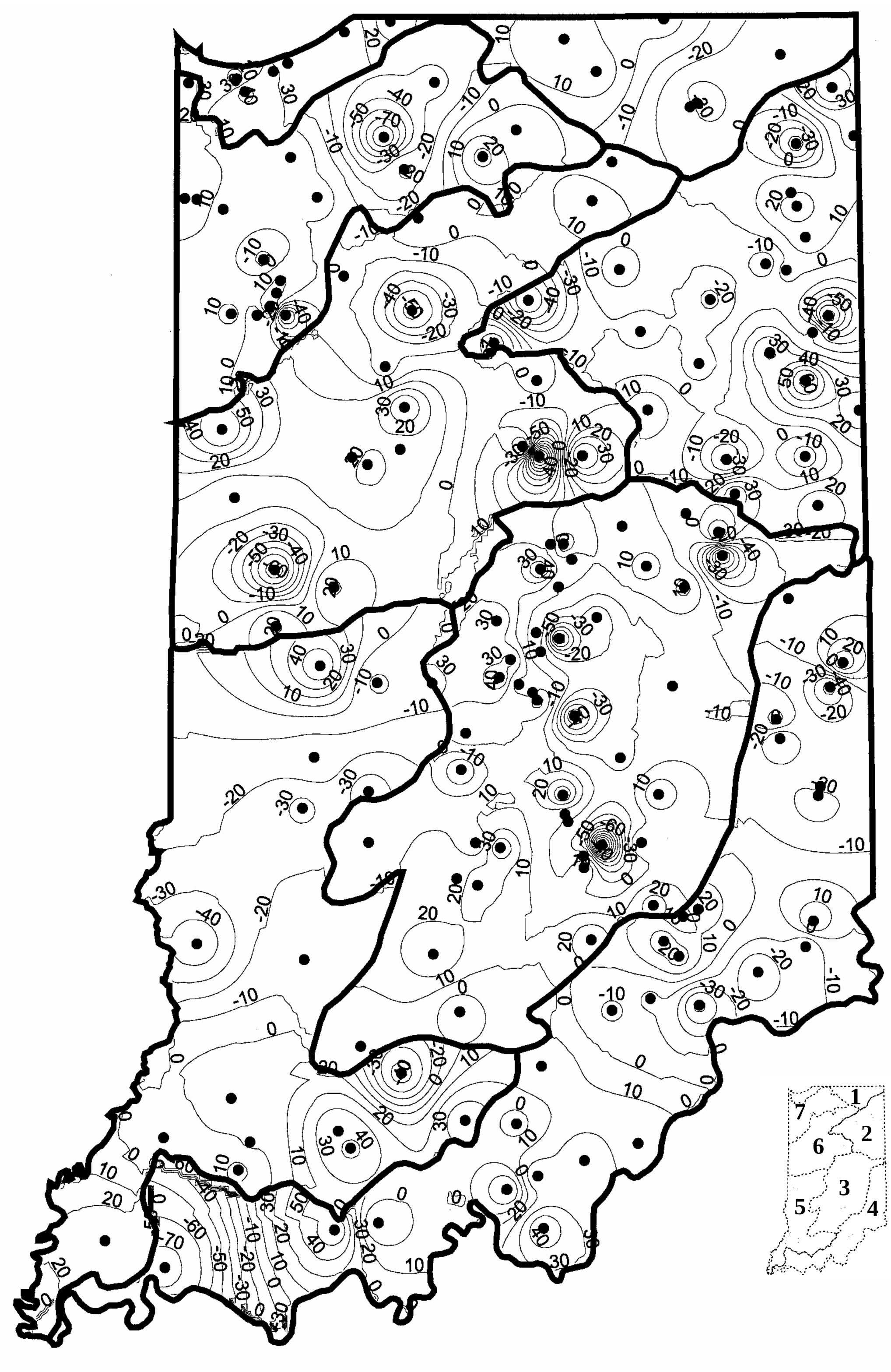

Figure 3. Percent difference contours for 100-year recurrence interval. The in set shows Glatfelter regions.

acuracy of the Glatfelter equations are added to the map using station longitude and latitude values. The spatial analyses are performed by creating contour maps from residual values. An example of these contour maps is given in Figure 3. The results observed for the 100-year recurrence interval are similar to the patterns discussed for other recurrence intervals. Locations observed to have significant bias as indicated by tight contour spacing remain for all recurrence intervals. According to these results, the bias increases with recurrence interval. The pattern of increasing bias is observed in the contour maps as the recurrence interval increases from 2-years to 100-years. In many locations, the contour spacing becomes tighter representing an increase in bias with increasing recurrence interval.

The contour maps strengthen the conclusion that in some areas within the Glatfelter regions there is a significant bias. These maps indicate areas where there is little bias and the Glatfelter multiple regression equations provide comparable results to those calculated by IDNR using USWRC. Low percent error areas, as indicated by contour values, denote an agreement between INDR and Glatfelter flood magnitudes and may be considered potentially homogeneous areas of hydrologic response.

The patterns observed in the contour maps are consistent. The tight contour spacing indicates sharp gradients in percent error observed among neighboring stations. Locations observed with tight contour spacing generally appear near the same map location as recurrence intervals increase from 2-years to 100-years. Therefore, locations in which considerable bias occurs between INDR and Glatfelter flood magnitude remain as the recurrence interval increases.

Areas with wide contour spacing represent a relative agreement of percent error among neighboring stations. In many cases, areas with wide contour spacing change into locations of tight contour spacing as recurrence intervals increase from 2-year to 100 years. This observation supports the finding that bias increases with recurrence interval.

2.4. Directional Seasonality Statistics

Results from the analysis of residuals, discussed in the last section, show that locations within the State of Indiana have considerable heterogeneity and bias. A regionalization approach developed by Burn et al. (1997) is used to determine homogeneous regions for flood frequency analysis of Indiana data. Directional seasonality statistics and the timing of flood events are used in the method by Burn et al. This is a method of classification of catchments with homogeneous hydrologic response.

The directional statistics (Mardia, 1972) used by Burn et al. involve seasonality measures derived from the time of occurrence of peak flow events. The occurrence of peak flow for an event is defined as a directional statistic by converting the occurrence date to the Julian date, where January 1 is Day 1 and December 31 is Day 365 (Bayliss and Jones, 1993; Burn, 1996). The flood event *i* is converted to an angular value using:

$$\theta_i = (Julian\,Date)_i \left(\frac{2\pi}{365}\right) \qquad (3)$$

where θ_i is the angular value in radians for the flood date of event *i*. Every flood event *i* is represented in polar coordinates as a vector with a unit magnitude and direction. The transformation for the determination of the x- and y- coordinates for each flood event *i* is:

$$x_i = \cos(\theta_i) \tag{4}$$
$$y_i = \sin(\theta_i) \tag{5}$$

where the set of points (x_i, y_i) lie on the unit circle. A sample of n flood events (*i 1,2...n)* for a station is summarized to determine the x- and y- coordinates of the mean flood date by using:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n}\cos(\theta_i) \tag{6}$$
$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n}\sin(\theta_i) \tag{7}$$

where x and y represent the x- and y- coordinates of the mean flood date and lie within the unit circle.

Mean flood coordinates $\bar{x}$ and $\bar{y}$ are summarized in two ways. First, the mean direction of flood dates is obtained by using the following:

$$\bar{\theta} = \tan^{-1}\left|\left(\frac{\bar{y}}{\bar{x}}\right)\right| \qquad for\ \bar{x}, \bar{y} > 0 \tag{8}$$

$$\bar{\theta} = \tan^{-1}\left|\left(\frac{\bar{x}}{\bar{y}}\right)\right| \quad +\frac{\pi}{2} \qquad for\ \bar{x} < 0, \bar{y} > 0 \tag{9}$$

$$\bar{\theta} = \tan^{-1}\left|\left(\frac{\bar{y}}{\bar{x}}\right)\right| \quad +\pi \qquad for\ \bar{x}, \bar{y} < 0 \tag{10}$$

$$\bar{\theta} = \tan^{-1}\left|\left(\frac{\bar{x}}{\bar{y}}\right)\right| \quad +\frac{3\pi}{2} \qquad for\ \bar{x} > 0, \bar{y} < 0 \tag{11}$$

The mean direction $\bar{\theta}$ is converted back to the day of the year by using:

$$MD = \bar{\theta}\left(\frac{365}{2\pi}\right) \tag{12}$$

The variable MD represents the average time of occurrence of flood events at a particular gaging station. Burn suggests that for catchments with similar MD values, other hydrologic characteristics may also be similar. MD is expected to be correlated with basin size. Additionally, a relation is expected particular to geographic location, especially for catchments experiencing seasonally influenced flows such as snowmelt (Burn, 1997). The MD variable is useful when considered with the value of the mean resultant. The mean resultant measures the variability of the *n* flood events about the mean date and is defined as the following:

$$\bar{r} = \sqrt{\bar{x}^2 + \bar{y}^2} \qquad (13)$$

The mean resultant essentially provides an indication of the spread of the flood events and the strength of the seasonal dependence of hydrologic response of data from a station. Low values of $\bar{r}$ indicate a hydrologic response that is not greatly dependent upon seasonal characteristics. High values of $\bar{r}$, suggest the hydrologic response has a high seasonal dependence. When the mean resultant is equal to one, the point x , y will lie on the unit circle, indicating a perfect seasonal response in which all peak flow events occur on the same day throughout the annual peak flow series.

Plots of $\bar{x}$ and $\bar{y}$-coordinates of the mean flood date on the unit circle are called to as Julian plots. Julian plots are constructed by using the $\bar{x}$ and $\bar{y}$-coordinates of the mean flood date for each station. A Julian plot of a sample station is shown in Figure 4 which demonstrates the directional statistics discussed about.

Plots of mean date of occurrence, MD, versus mean resultant, $\bar{r}$, are also constructed. A sample MD vs. $\bar{r}$ is shown in Figure 5. These plots allow visualization of the seasonal response of the stations investigated. Watersheds displaying timing and seasonal similarities should be considered as potential members of a hydrologically homogeneous region for flood frequency analysis (Burn, 1997).

The above discussed method is followed to perform the seasonality analysis of Indiana watersheds. The directional seasonality statistics, the mean flood coordinates $\bar{x}$ and $\bar{y}$, mean directions θ and MD, and mean resultant $\bar{r}$ for the 264 stations included in the regionalization procedure are calculated.

Similarities of seasonal data in Julian plots are observed when points representing catchments are clustered in a small area within the unit circle. Visual inspection of the Julian and mean date of occurrence versus mean resultant plots provide similar results for all seven regions analyzed. Although there was a cluster of stations within each region analyzed, distinct clusters of stations are not observed as would be expected for groups of seasonally homogeneous stations. There is not a distinct separation between groups of stations when all the data are plotted together in Figure 5. The main conclusion from the inspection of Julian plots is that the stations analyzed are not distinct in seasonal hydrologic response. Examining the distribution of mean date of occurrence and mean resultant values for each region strengthened this conclusion.

The seasonality approach is proposed as a method to derive homogeneous regions without directly using the magnitude of flood events as similarity variables. The inconclusive results from this approach suggest that the timing and regularity of peak flow events are not distinct for Indiana watersheds.

The climate of Indiana is fairly uniform. Sharp differences in seasonal weather patterns are not present in dates of occurrences of floods in Indiana watersheds used in this study. Burn has success classifying stations based on seasonal response. However,

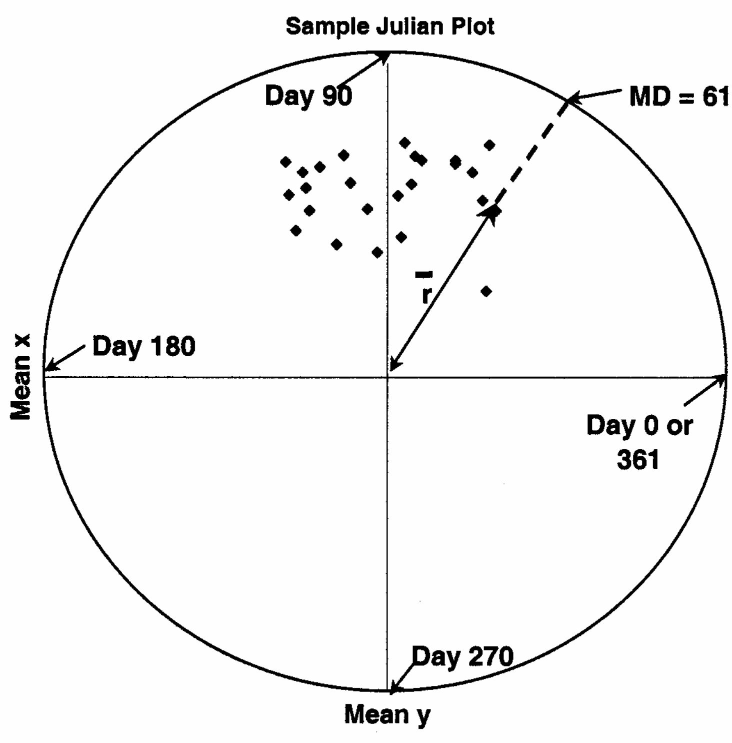

Figure 4a. Sample Julian Plot

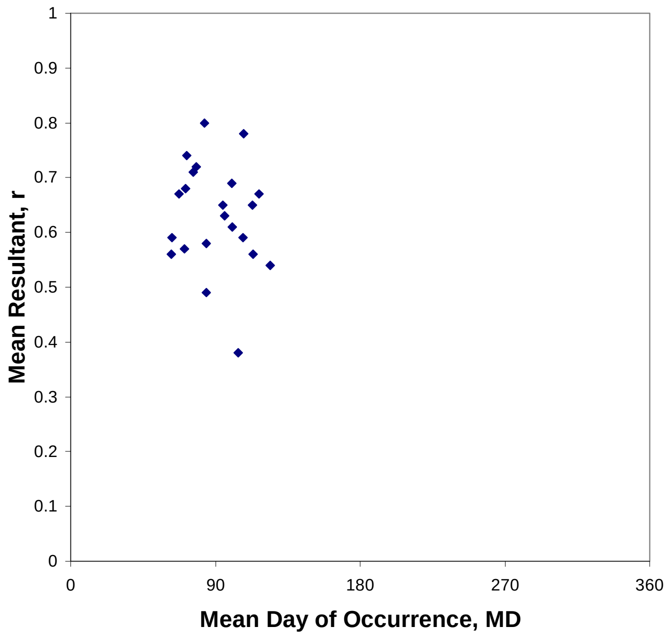

Figure 4b. Sample plot of mean day of occurrence vs. mean resultant

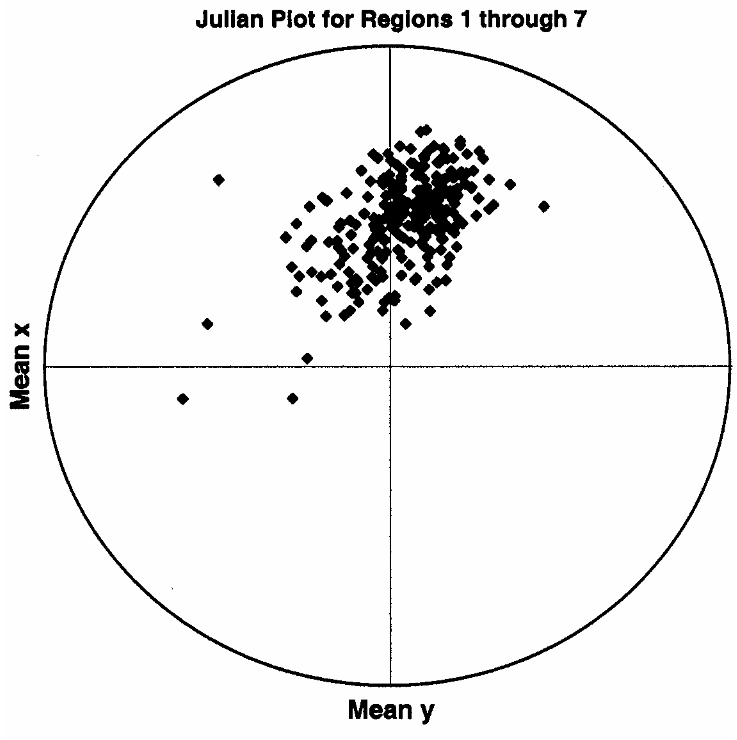

Figure 5a. Julian plot of Indiana flood data

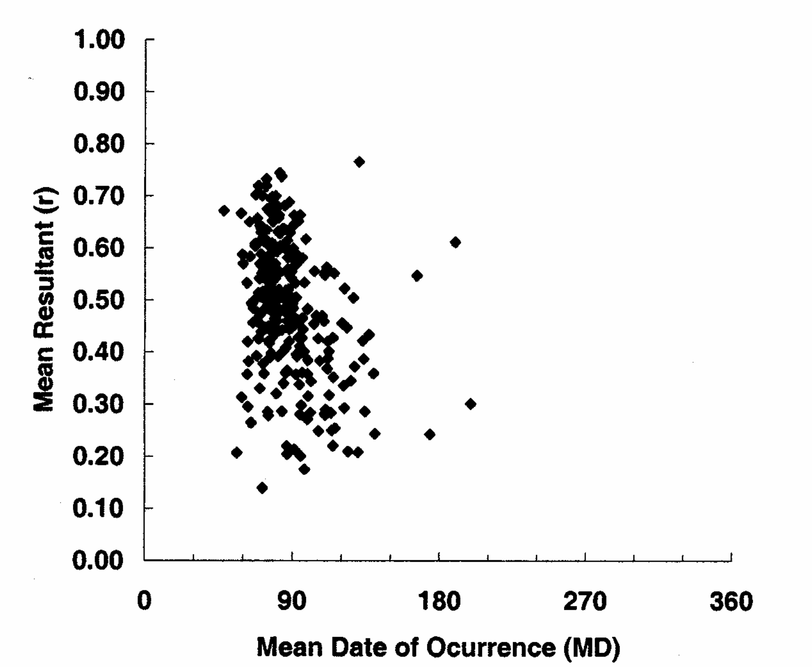

Figure 5b. Mean date of occurrence versus mean resultant for Regions 1 through 7

the area analyzed by Burn is a region in Canada that had distinct climatic zones influenced by snowmelt in the mountainous areas.

2.5 Regionalization with L-Moments

Indiana watersheds are regionalized by using the method developed by Hosking and Wallis (1993). The details of the method are found in Hosking and Wallis (1997). L-moments form the basis for the Hosking and Wallis' approach to regional flood frequency analysis. The homogeneity of potential regions for flood frequency analysis is evaluated by comparing between-site variations of sample L-moments computed from annual maximum flow data. Regions found to be homogeneous are acceptable for regional flood frequency analysis. Regions found to be possibly heterogeneous are marginally acceptable for regional flood frequency analysis. Regions found to be heterogeneous cannot be used for regional flood frequency analysis. The details of L-moments are found in Hosking (1990).

Hosking and Wallis' (1993) approach uses the magnitude of station discordancy measures to indicate potential heterogeneous stations. The removal of significantly discordant stations from a region should result in a decreased heterogeneity measure. As heterogeneity of a region decreases, the homogeneity should increase. A region is considered homogeneous when the heterogeneity test results in H_1 value, the most important statistic, of less than one for the region.

The regionalization procedure used in this study is performed by using Hosking's (1993) L-moment based FORTRAN routine. A group of stations were analyzed with the programs XFIT and XTEST. A discordancy measure is calculated for each station. A heterogeneity measure for each group of stations is also calculated. Groups of stations are considered to be heterogeneous when the statistic H_1 is greater than or equal to one but less than two. Groups of stations were definitely heterogeneous when H_1 is greater than or equal to two. When the H_1 value of a group of stations is H_1 greater than or equal to one, discordant stations are removed from the group of stations. Once the discordant stations are removed, the heterogeneity measure is recalculated. Thus, determination of a homogeneous region is a trial and error procedure.

Throughout the region formation process, heterogeneous groups were analyzed in which no stations are identified as discordant by the critical discordancy value. In this situation three considerations are used. First, geographically contiguous regions are sought during the region forming process. Therefore, if two stations have comparable high discordancy values, the station nearest to the edge of the region is removed.

Contributing drainage area is the second consideration when dealing with a group of stations in which no stations were identified as discordant. For this study, stations with high discordancy measures are eliminated if the contributing drainage area is less than 50 square miles. Hydrologic response of small drainage areas is quite different from that of larger basins. Additionally, small watersheds do not represent a significant percentage of the larger areas that are considered for the regionalization.

The length of the station record is the final consideration in the region forming process. Twenty years of annual peak flow record is used as the minimum for inclusion in the regionalization study. It is possible that L-moments calculated from 20 years of record do not provide a good representation of the hydrologic response of a watershed.

For this reason, stations with approximately twenty years of record with high discordancy values are removed from the L-moment analysis.

The region formation process used for this study began by considering all stations in the state. A trial and error procedure is used to develop geographically contiguous regions. The goal of the regionalization approach is to establish homogeneous regions for regional flood frequency analysis for the State of Indiana. Once a group of stations is found to be homogenous, the region made up of the stations is expanded to incorporate as many stations as possible while maintaining the group to be homogeneous. Six homogeneous regions with an H_1 value less than one are identified (Figure 6). These homogeneous regions are labeled A-1, A-2, A-3, A-4, A-5, and A-6. Two possibly heterogeneous regions with an H_1 value of greater than or equal to one but less than two are also identified. The possibly heterogeneous regions are labeled B-1 and B-2. Three regions are identified as definitely heterogeneous with an H_1 value greater than or equal to two. The definitely heterogeneous regions are identified labeled as C-1, C-2, and C-3. Figure 6 shows the homogeneous regions of Indiana watersheds determined by using Hosking's (1993) L-moment approach.

2.6 Conclusions

The following conclusions are formed on the basis of this study.

1. Regions used by Glatfelter are not homogeneous in their flood response.
2. Burn et al.'s method cannot be used to regionalize Indiana watersheds.
3. Hosking and Wallis' method can be used to clarify Indiana watersheds. The robustness of this classification is questionable.

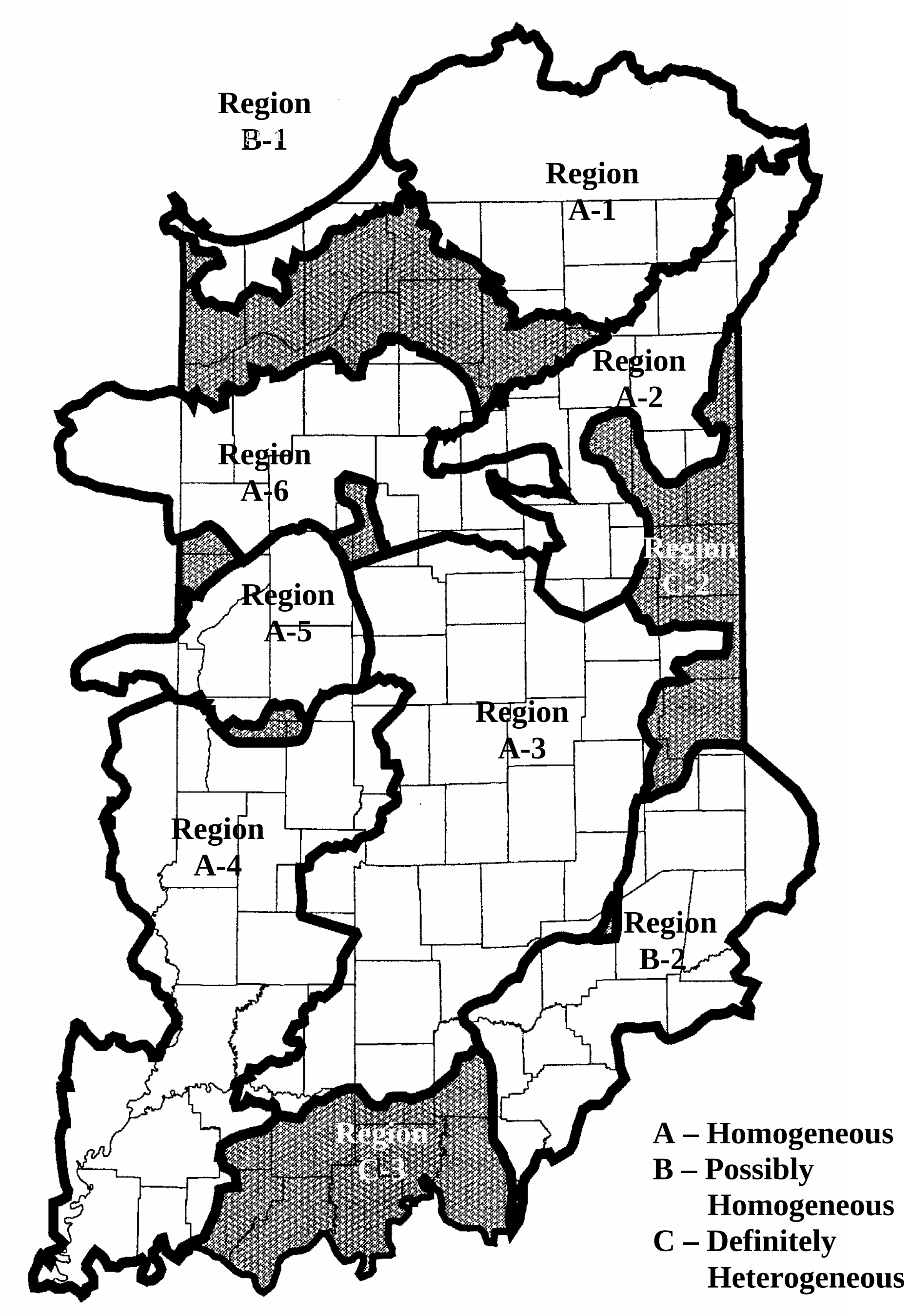

Figure 6. Classification of Indiana watersheds determined by heterogeneity analysis

III. REGIONALIZATOIN BY HYBRID CLUSTER ANALYSIS

3.1 Introduction

The basic objective of the research discussed in this chapter is the regionalization of Indiana watersheds by hybrid-cluster analysis. As no single method has been demonstrated to yield universally acceptable results, several methods of regionalization are in use. No information is available about the relative performance of these methods. Hence, several regionalization methods are investigated in the present study. The first of these is the L-moment based method discussed in chapter two. The second one is the Hybrid Cluster Analysis, which is discussed in this chapter.

Three hybrid-cluster algorithms, which are a blend of agglomerative hierarchical and partitional clustering procedures, are tested in this study to determine their potential in delineating Indiana watersheds into regions that are homogeneous in hydrologic response. The hierarchical clustering algorithms considered for hybridization are Single linkage, Complete linkage and Ward's algorithms, while the partitional clustering algorithm used is the hard K-means algorithm.

3.2 Cluster Analysis

Recently, there has been some interest in the development and use of cluster analysis techniques in different fields. These techniques are recognized with different names in different contexts, such as *unsupervised learning* in pattern recognition, *numerical taxonomy* in biology and ecology, *typology* in social sciences and *partition* in graph theory (Theodoridis and Koutroubas, 1999). Introductory material about cluster analysis and its techniques can be found in Hartigan (1975), Andenderfer and Blashfield (1984), Jain and Dubes (1988), Kaufman and Rousseeuw (1990), Everitt (1993), and Gordon (1999).

A cluster consists of one or more *feature vectors*. In the context of regionalization for flood frequency analysis, a feature vector may comprise of variables representing: (i) physiographic catchment characteristics such as drainage area contributing to flood, average basin slope, main stream slope, stream length, stream density; storage index, soil type index such as infiltration potential, runoff coefficient or effective mean soil moisture deficit, fraction of the basin covered by lakes, reservoirs or swamps; (ii) geographical location attributes such as latitude, longitude and altitude of catchment centroid; (iii) a measure of basin response time such as basin lag or time-to-peak (Potter and Faulkner, 1987); (iv) meteorologic factors such as storm direction, mean annual rainfall, precipitation intensities; (v) at-site flood statistics such as mean, coefficient of variation or skewness coefficient of annual flood series, plotting position estimate of T-year flood event interpolated from the annual flood series (Burn,1990b), flood magnitude corresponding to a T-year recurrence interval (Tasker, 1980). Combination of two or more of the above variables may also constitute an attribute in a feature vector. For example, specific mean annual flood (Mosley, 1981; Wiltshire, 1986), mean annual flow divided by the drainage area (Burn, 1989), the ratio of peak flow for a T-year return period to the drainage area, basin shape defined as the ratio of main stream length to basin area (Acreman, 1985) have been used in the past.

Geologic features of the basin such as fraction of catchment underlain by conglomerates, sandstones, granites, basalts, fine grained sediments and metamorphics, foliated metamorphics, fine-grained igneous rocks, unconsolidated sediments (Nathan and McMahon, 1990) have also been used as attributes.

In the context of climatic applications, a feature vector may comprise of hourly values of (i) air temperature; (ii) dewpoint temperature; (iii) total cloud cover; (iv) wind speed; (v) wind direction; (vi) visibility; (vii) precipitation (Kalkstein et al., 1987). For hydroclimatic applications, monthly values of temperature, precipitation and drought indices may be considered as attributes.

A feature vector is also referred to as "data vector" or "object" in some contexts. A cluster consisting of a single feature vector is referred to as *singleton cluster*. *Clustering*, also known as unsupervised classification, is a process by which a set of feature vectors is divided into clusters or groups such that the feature vectors within a cluster are as similar as possible and the feature vectors of different clusters are as dissimilar as possible.

3.1.1. Agglomerative Hierarchical Clustering

For the given set of N feature vectors, the agglomerative hierarchical clustering procedures begin with N singleton clusters. A distance measure such as those shown in Table 3 is chosen to evaluate the dissimilarity between any two clusters. The clusters that are least dissimilar are found and merged. This results in N –2 singleton clusters and a cluster with two feature vectors. The process of identifying and merging two closest clusters is repeated till a single cluster is left. In general, the number of clusters left at the end of *n* merges is equal to N-*n*. The entire process may be represented as a nested sequence, called dendrogram, which shows how the clusters that are formed at the various steps of the process are related.

Algorithms that are representative of the agglomerative hierarchical method of clustering include: (i) single linkage or nearest neighbor; (ii) complete linkage or furthest neighbor; (iii) average linkage; (iv) Ward's algorithm (Ward, 1963); (v) Lance and Williams flexible method; (vi) density or k-linkage method. These algorithms differ from one another by the strategy used for defining nearest neighbor to a chosen cluster.

Table 3. Dissimilarity measures used for computing distance between feature vectors

Distance measure	Equation
Euclidean	$\sqrt{\sum_{k=1}^{p} (X_{ik} - Y_{jk})^2}$
Squared Euclidean	$\sum_{k=1}^{p} (X_{ik} - Y_{jk})^2$
Manhattan	$\sum_{k=1}^{p} \lvert X_{ik} - Y_{jk} \rvert$
Canberra	$\frac{1}{p} \sum_{k=1}^{p} \frac{\lvert X_{ik} - Y_{jk} \rvert}{X_{ik} + Y_{jk}}$
Chebychev	$\text{Max}_{1 \le k \le p} \lvert X_{ik} - Y_{jk} \rvert$
Cosine	$\frac{\sum_{k=1}^{p} X_{ik} Y_{jk}}{\sqrt{\sum_{k=1}^{p} X_{ik}^2 \sum_{k=1}^{p} Y_{jk}^2}}$

p : number of attributes; X_{ik} : attribute *k* of feature vector *i* in cluster-1;
Y_{jk} : attribute *k* of feature vector *j* in cluster-2;

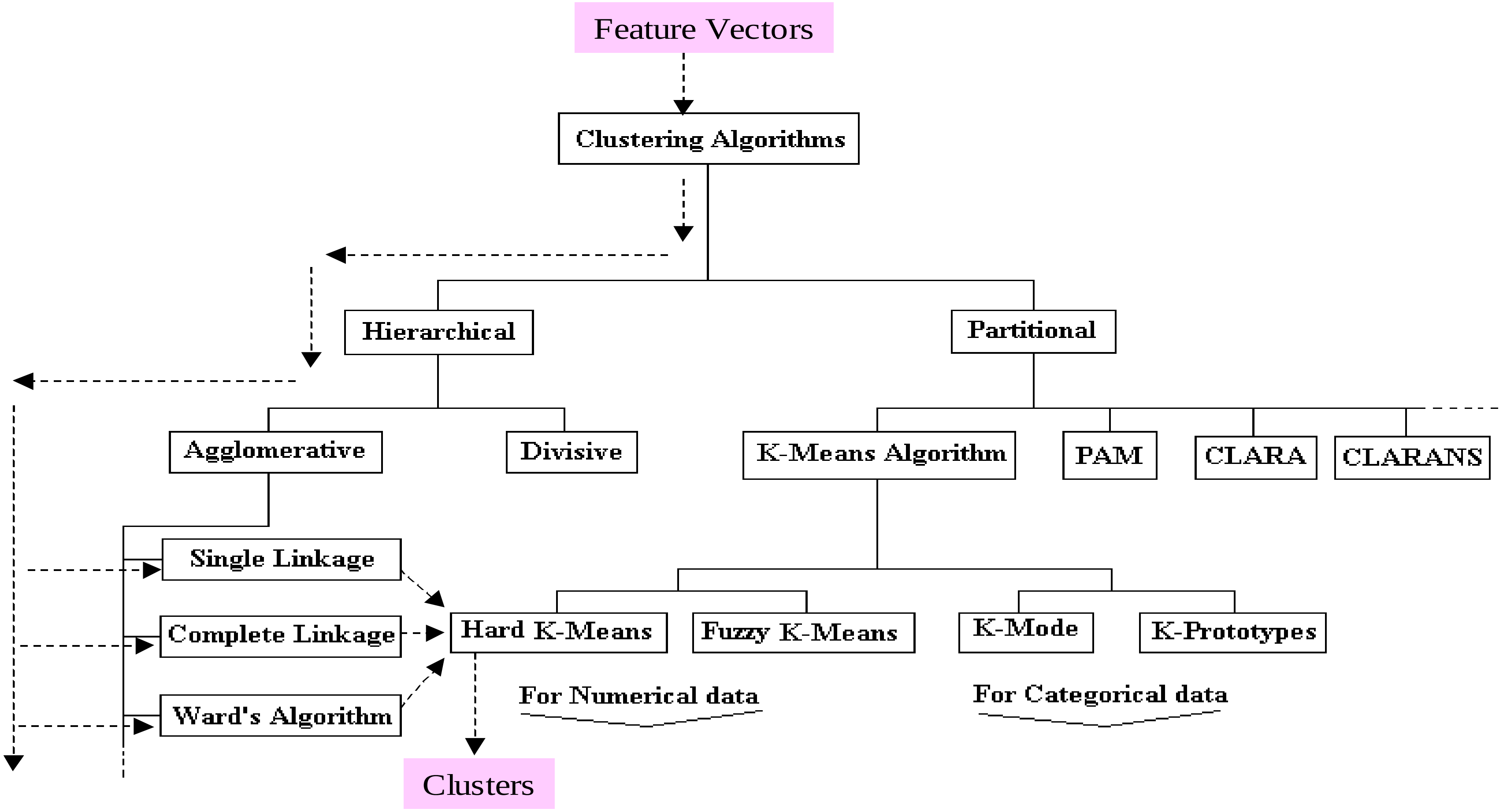

Figure 7. Classification of clustering algorithms. The dotted arrows show the sequence in which agglomerative hierarchical and partitional clustering algorithms are executed to obtain clusters from hybrid clustering procedure.

In the *Single linkage algorithm,* distance between two clusters is the distance between the closest pair of feature vectors, each of which is in one of the two clusters. Clusters with the smallest distance between them are merged. This algorithm tends to form a small number of large clusters, with a few small outlying clusters on the fringes of the space of site characteristics and is not likely to yield good regions for regional flood frequency analysis (Hosking and Wallis, 1997, pp.58-59). The algorithms used in cluster analysis are shown in Figure 7 and discussed below.

In the *Complete linkage* algorithm, distance between two clusters is defined as the greatest distance between a pair of feature vectors, each of which is in one of the two clusters. Clusters with the smallest distance between them are merged. This algorithm tends to form small, tightly bound clusters. It is not suitable for application to large data sets.

In the *Average linkage* algorithm, the distance between two clusters is defined as average distance between them. Clusters with the smallest distance between them are merged. There are several methods available for computing the average distance. These include unweighted pair-group average, weighted pair group average, unweighted pair group centroid and weighted pair group centroid.

Unweighted pair-group average (UPGA): The distance between two clusters is defined as average distance between all pairs of feature vectors, each of which is in one of the two clusters. Clusters with the smallest distance between them are merged.

Weighted pair-group average (WPGA): This method is identical to the *UPGA*, except that in the computations, the size of the respective clusters, i.e, the number of feature vectors contained in them, is used as a weight. This method is preferred when the cluster sizes are suspected to be greatly uneven.

Unweighted pair-group centroid (UPGC): The distance between two clusters is defined as the distance between their centroids. The centroid of a cluster is the mean vector of all the feature vectors contained in the cluster. Clusters with the smallest distance between them are merged. In this method, if two clusters to be merged are very different in their size, the centroid of the cluster resulting from the merger tends to be closer to the centroid of the larger cluster.

Weighted pair-group centroid (WPGC): This method is identical to the *UPGC*, except that feature vectors are weighted in proportion to the size of clusters.

Ward's algorithm (Ward, 1963) is one of the frequently used techniques for regionalization studies in hydrology and climatology (Willmott and Vernon, 1980; Winkler, 1985; Kalkstein and Corrigan, 1986; Acreman and Sinclair, 1986; Nathan and McMahon, 1990; Hosking and Wallis, 1997). It is based on the assumption that if two clusters are merged, the resulting loss of information, or change in the value of objective function, will depend only on the relationship between the two merged clusters and not on the relationships with any other clusters. The governing equation of Ward's algorithm and a detailed explanation of the same are provided in Srinivas and Rao (2002).

In regional flood frequency analysis, Mosley (1981) used agglomerative hierarchical clustering available with BMDP2M cluster analysis program (Dixon, 1975) for regionalization of catchments in Newzealand. Tasker (1982) applied complete linkage algorithm of Sokal and Sneath (1963) for the regionalization of watersheds in Arizona. Nathan and McMahon (1990) compared the performance of single linkage, complete linkage, average linkage, centroid, median and Ward's algorithms of agglomerative hierarchical clustering available with SPSSX (SSPS Inc., 1988) statistical package. Euclidean, squared Euclidean, Manhattan, Chebychev and Cosine distance measures were also considered in their study.

Burn et al. (1997) used agglomerative hierarchical clustering algorithm for the regionalization of watersheds in Canada. Their study used the following dissimilarity measure of Webster and Burrough (1972):

$$D_{ij}^{d} = \frac{D_{ij} + \dfrac{d_{ij}}{d_{\max}} w}{1 + w}$$

where D_{ij} is the Canberra dissimilarity measure of Lance and Williams (1966), whose expression is provided in Table 3; d_{ij} represents the geographic distance between catchments *i* and *j*; $d_{\max}$ denotes the maximum geographic distance between catchment pairs, each of which is in one of the two clusters; *w* is the weighing factor that reflects the relative importance of scaled geographic separation (d_{ij} /$d_{\max}$) and the Canberra dissimilarity value (D_{ij}). Seasonality measures, called mean date of occurrence of flood events and the regularity of the phenomenon at each gauging station have been considered as attributes in the Canberra dissimilarity measure. Seasonality measures may not be useful attributes when the study region consists of catchments that do not show strong seasonal response, as is the case with Indiana watersheds (Rao et al. (2001)).

The divisive hierarchical clustering procedures begin with a single cluster consisting of all the N feature vectors. The feature vector that has the greatest dissimilarity to other vectors of the cluster is then identified and separated to form a splinter group. The dissimilarity values of the remaining feature vectors in the original cluster are then examined to determine if any additional vectors are to be added to the splinter group. This step divides the original cluster into two parts. The largest cluster of the two is subjected to the same procedure in the next step. The algorithm terminates when the clusters resulting from the analysis are all singleton clusters.

3.2. Partitional Clustering Methods

In partitional clustering procedures an attempt is made to recover the natural grouping present in the data through a single partition. Examples of this class of algorithms include K-means clustering (MacQueen, 1967), PAM (Partitioning around medoids, Kaufman and Rousseeuw, 1990), CLARA (Clustering Large Application, Ng and Han, 1994), CLARANS (Clustering Large Applications based on Randomized Search). Of these algorithms only K-means algorithm is suitable for small data sets such as those used for hydrologic modeling. The partitional clustering algorithms that are applicable to categorical data include K-prototypes and K-modes algorithms (Huang et al., 1997).

Burn (1989) used a K-means clustering algorithm to determine appropriate grouping of a network of streamflow gauging stations in southern Manitoba, Canada, for flood frequency analysis. Flood statistics (coefficient of variation of peak flows, mean annual flow divided by the drainage area) and geographic position of catchments (latitude and longitude) were used as attributes in the feature vector. Traditionally, flood statistics such as coefficient of variation are used to test the homogeneity of the derived regions. The use of the same flood related variables to form the regions and subsequently to evaluate the homogeneity of the derived regions leads to formation of regions that are homogeneous but not necessarily effective for regional flood frequency analysis (Burn et al., 1997). If at-site flood statistics are used as attributes in the feature vector, one has to ensure that they do not exhibit a high degree of correlation with the flood quantiles of interest. Moreover, the use of flood statistics in a similarity (or dissimilarity) measure constrains the use of the derived regions for estimating extreme flow quantiles at ungaged sites in the study region.

When cluster analysis is based on site characteristics, the at-site statistics are available for use as the basis of an independent test of the homogeneity of the final regions (Hosking and Wallis, 1997). Burn and Goel (2000) applied the K-means algorithm to site characteristics (catchment area, length and slope of the main stream of river) of a collection of catchments in India to derive regions for flood frequency analysis. Wiltshire (1986) and Bhaskar and O’Connor (1989) used the K-means algorithm. Wiltshire (1986) adopted the iterative relocation algorithm of Gordon (1981), whereas the latter work used the FASTCLUS clustering procedure of SAS package. While Wiltshire (1986) made random partition of data to initiate his clustering algorithm, Bhaskar and O'Connor (1989) specified a limiting value for the minimum distance between initial cluster centers.

3.3 Data Used in the Study

Information related to magnitude of peak flows and the date and time of occurrence of the flood events at the gauging stations in Indiana is extracted from the electronic file of Indiana Department of Natural Resources (IDNR) Division of Water (2001). Peak flow information for the stations located outside Indiana, latitude and longitude values of all the 273 stations and drainage areas for the 28 pooled stations are extracted from United States Geological Survey's national water information system web site *http://water.usgs.gov/nwis/peak*. Details of nine attributes considered by Glatfelter (1984) to assess the degree of similarity between drainage basins in Indiana are available for the 245 stations from Glatfelter (1984). The range of each of these attributes is presented in Table 4. The attributes were screened to extract independent attributes for cluster analysis.

Table 4. Attributes available for the Glatfelter stations.

Attribute	Range
Drainage Area	0.11 – 11125.00 mi^2
Mean Annual Precipitation	34 – 46 in
Main channel Slope	0.90 – 267.00 ft/mile
Main channel Length	0.3 – 315.0 miles
Basin Elevation	412.0 – 1190.0 ft
Storage[1]	0% – 11%
Soil Runoff coefficient	0.30 –1.00
Forest cover in Drainage Area	0.0 – 88.4%
I(24,2)[2]	2.6 – 3.35 in

[1] Storage – percentage of the contributing drainage area covered by lakes, ponds or wetlands

[2]I(24,2) – 24-hour rainfall having a recurrence interval of 2 years, in inches.

The features extracted for cluster analysis are: (i) four physiographic attributes, drainage area, slope of the main channel in the drainage basin, soil runoff coefficient and storage; and (ii) one meteorological attribute, mean annual precipitation. The geographic location attributes Latitude and Longitude are included in the feature vector with a view to identify regions that are geographically contiguous.

Of the seven attributes only drainage area was transformed using logarithmic transformation. Then, each of the seven attributes were standardized. In the first set of trials, a weight of 1 was assigned to all the attributes, implying equal importance to all features.

Three hybrid-clustering procedures, that are a blend of hierarchical and partitional clustering algorithms, are tested to determine their potential in delineating watersheds of Indiana into regions that are homogeneous in hydrologic response. The hierarchical clustering algorithms considered for hybridization were single linkage, complete linkage and Ward's algorithms, while the partitional clustering algorithm considered was the K-means algorithm. The clusters obtained from the hybrid-cluster analysis were subsequently modified, following a heuristic process, to obtain regions that are homogeneous in hydrologic response. The study resulted in delineation of Indiana into five homogeneous regions and one heterogeneous region. In addition, one homogeneous subregion has been identified in the Kankakee river basin that comprises of several sites of region-5. The homogeneous regions identified form the basis for effective transfer of information. In simple terms, for estimation of flood quantiles, the data at the target site (gauged or ungauged) of interest in a homogeneous region can be augmented with data from the gauged sites within the region.

The study resulted in delineation of Indiana into five homogeneous regions, one heterogeneous region and an unallocated residue of 23 stations of which 21 are located in Indiana. The final pictorial representation of the homogeneous regions is given in fig. 8.

Table 5. Characteristics of the regions formed

Region number	N	RS	Heterogeneity measure		
			H_1	H_2	H_3
1	62	1689	0.86	-0.12	-0.94
2	58	1730	0.85	0.43	-0.65
3	30	804	-0.46	0.66	0.28
4	73	3039	0.48	-0.40	-1.78
5	42	1938	0.04	-0.91	-0.85
6	14	519	13.69	6.33	2.94

N: Number of stations

RS: Region size in station years

The results from Table 5 indicate that regions 1 to 5 are all acceptably homogeneous, while region-6 adjoining lake Michigan is highly heterogeneous. All the homogeneous regions identified have more than 5T station-years of data. Region-1 is spread mainly along the course of Wabash river and comprises predominantly of alluvial deposits of the flood plains. Region-2 contains karst formations associated with limestones of the Mississippian age, laid down 320-360 million years ago. Region-3 possesses a karst area consisting of older Devonian and Silurian limestones. The topography of these areas is dominated by sinkholes, sinking streams, large springs and caves. For the ungauged catchments lying at the border between the regions 2 and 3, the possibility of including information from both the regions can be considered. Region-4 is in central Indiana. The soil here is predominantly loamy glacial till. Region 5 is spread over northern part of Indiana. It comprises of a wide range of soil classes (clayey glacial till, sandy and loamy deposits, loamy glacial till) overlying the Mississippian rocks of Michigan basin and Devonian and Mississippian shale.

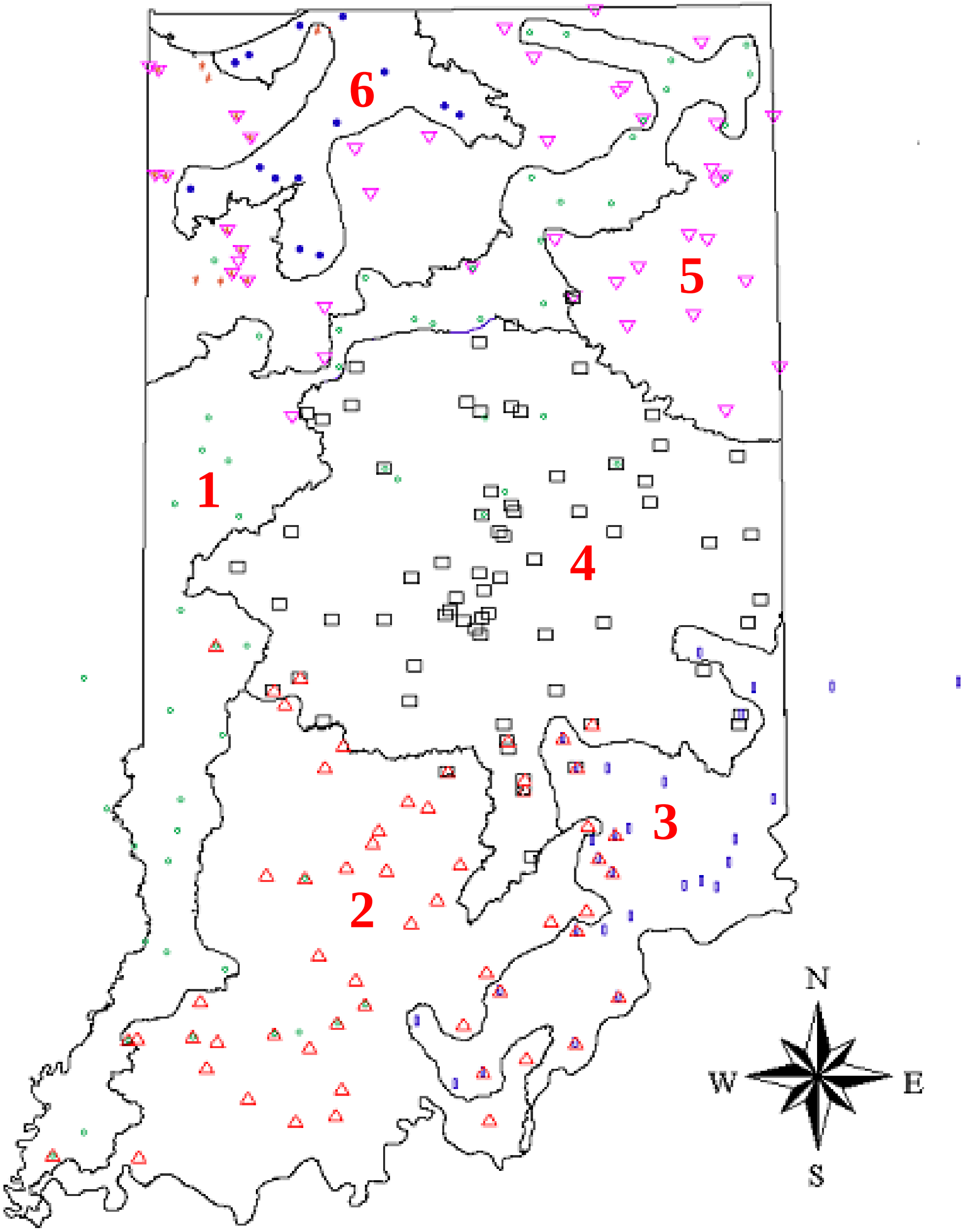

Figure 8. Location of the regions defined using the hybrid cluster analysis.

3.5 Conclusions

This study illustrated the usefulness of the hybrid cluster analysis technique for regional flood frequency analysis. The study also showed that, in general, Ward’s algorithm is better than complete linkage and single linkage algorithm in identifying groups of clusters with optimal value of objective function. However, the clusters obtained by initializing K-means algorithm with the clusters resulting from hierarchical clustering algorithms are quite comparable. Consequently it is not possible to suggest any single hybrid model as the best for regionalization. Further, the process of revising the results from clustering algorithm is important for arriving at the final regions irrespective of the method opted for obtaining the clusters.

IV USE OF PRECIPITATION AND FLOW DATA FOR REGIONALIZATION OF WATERSHEDS

The main objective of this study is to use hard clustering to form regions (or groups of sites) in the state of Indiana, which are homoegeneous in hydrologic response by using precipitation data. Specifically, three types of hierarchical methods, complete linkage, single linkage, and Ward's algorithm, and the K-means method, a non-hierarchical method, were utilized. It is assumed that a unique set of homogeneous regions will be found. However, more than a unique set of homogeneous regions may be obtained depending on the initial partitioning, resulting in significant changes in the region boundaries. Although the loss of confidence due to this inconsistency is difficult to measure, it may be small when compared to the improvements in prediction capability due to regionalization (Moseley 1981). In an effort to reduce this potential inconsistency, different hierarchical methods were used to determine an initial partitioning. A non-hierarchical method is then used to give a final set of groups. Regions resulting from the clustering algorithms are tested for homogeneity, using L-moment ratios. It is attempted to determine an optimal number of regions by maintaining a balance between group size and homogeneity.

4.1 Data Used in the Study

The data used in this study are grouped into four types: (1) Precipitation, (2) Peak Annual Flow, (3) Latitude and Longitude, and (4) Curve Number and Runoff Coefficient. These data are used to perform cluster analysis for sites within the state of Indiana.

Precipitation data for Indiana rain gauges were obtained through the State Climatologist at Purdue University, and is available on CD-Rom for various states within the United States. The data are originally collected by the National Climatic Data Center of the National Weather Service. The data used include daily rainfall values at 136 rainfall gauges within the state of Indiana from 1901 to 1995. The length of record at each gauge varies; the period of record that a maximum number of stations have in common is from 1949 to 1968.

A Fortran program was developed to extract Maximum Annual Rainfall (MAR) values from the observed data. The output of the program gives the maximum rainfall for each year of available record, as well as the year, month, day, and hour that the value was recorded in that year. Similar programs were developed to extract a consecutive 3- and 5-day maximum rainfall for each year of record, for each rainfall gauge. These values are referred to as the 1-day, 3-day, and 5-day maxima.

Sites with maximum annual rainfall (MAR) records less than 20 years were excluded from the cluster analysis. Statistics of the MAR records were calculated and considered as attributes for clustering. These include the mean, standard deviation, skewness, and L-moments such as the Coefficient of Variation (L-CV), L-skewness, and L-kurtosis.

Stations with peak annual flow (PAF) records of length 10 or less years were excluded from cluster analysis, yielding a collection of 255 sites. The mean, standard deviation, and other statistics of the PAF records are calculated and are considered for cluster analysis.

To ensure that the regions obtained are geographically continuous, the geographic location of each site is sometimes considered as an attribute for cluster analysis. The latitude and longitude of each rainfall gauge and streamflow station were obtained and used as attributes in the cluster analysis.

The curve number *CN* and the *C* values in rational formula are listed in tables, usually found in textbooks or common references. To obtain these two parameters, the entire state was assumed to be agriculture land, with good managerial practices and an average antecedent moistures condition. The hydrologic soil group for each site was determined by using the soil maps in for the state. The effect of including these parameters as attributes in cluster analysis is investigated, because they account for characteristics of the land, e.g. soil type, which affect the hydrologic response of a watershed to rainfall events. As before, the assumption is made that the soil characteristics at a particular rainfall gauging site are representative of those of the entire watershed associated with that site.

4.2 Methodology

In the WARD algorithm, cluster membership is assessed by calculating the total sum of squared deviations from the mean of a cluster centroid. The centroid of a cluster is a point whose attribute values are the mean of the attribute values of all the points located in the cluster. The criterion for addition of a single site into a cluster is that the fusion should produce the smallest possible increase in the error sum of squares. Ward's method tends to form clusters of equal size.

The K-means method is a non-hierarchical clustering method and will be used to give a final group of clusters. It seeks to partition *N* data points into *K* different clusters S_j containing N_j data points so as to minimize the sum-of-squares criterion

$$J = \sum_{j=1}^{K} \sum_{n \in S_j} \left\| x_n - \bar{x}_j \right\|^2$$

where x_n is a vector representing the *n*th data point set and $\bar{x}_j$ is the mean of the data points in S_j. The algorithm consists of a simple re-estimation procedure. Once the data points are assigned at random to the *K* sets, the centroid is computed for each set. Then each site is assigned to a cluster depending on the minimum distance between its attribute value and the cluster mean. These two steps are alternated until a stopping criterion is met, i.e., when there is no further change in the assignment of the data points.

Once a partitioning is obtained through cluster analysis, the homogeneity of each cluster is determined based on a regional homogeneity test. The algorithm to calculate homogeneity of a group of sites is taken from Hosking (1990), and is based on L-moment ratios of sample values, for example, the available peak annual flow (PAF) record at each site. If a cluster is homogeneous, each site within the cluster should have identical L-moment ratios; however, their sample L-moment ratios will differ due to varying sample size. Therefore the test examines whether the between-site dispersion of the sample L-moment ratios for a group of sites is larger than the dispersion expected of a homogeneous region. The three measures of dispersion suggested by Hosking and Wallis (1993), H_1, H_2, and H_3, are used in the study.

Different hierarchical clustering methods are investigated to determine the one which is more suitable for Regionalization of Watersheds. Those methods are Single-Linkage (SL), Complete-Linkage (CL), and Ward's method (WARD). The K-means Method is used as the partitioning method. For the purpose of identification, the hierarchical method used in the clustering algorithm is also indicated in the name of a particular combination, following the number. For example, RMSG1SL includes the mean, standard deviation, and skewness of the MAR as attributes, and uses Single-Linkage as the hierarchical method. In contrast, RMSG1CL also includes the mean, standard deviation, and skewness of the MAR, but uses Complete-Linkage as the hierarchical method. If the name of a particular combination does not include an identifier for the hierarchical method used, the default is WARD.

Table 6: Combinations of attributes for precipitation data

	Combinations			
Attributes	RMSG	RMSGL	RMCWL	RCWKL
Mean (M)	x	x	x	
Standard Deviation (S)	x	x		
Skewness (G)	x	x		
L-CV (C)			x	x
L-Skewness (W)			x	x
L-Kurtosis (K)				x
Lat and Long (L)		x	x	x

The MAR data are used to determine the homogeneity of the clusters resulting from the combinations which include statistics of the MAR as attributes. Similarly, the 3-day MAR data are used for the combinations which include statistics of the 3-day MAR (denoted by a 3), and so on (Table 6).

Statistics of Peak Annual Flow (PAF) at 255 streamflow stations are combined in various ways to perform cluster analysis. The four combinations of attributes for this section are listed in Table 7. Each combination for this section begins with the letter P.

Each attribute is represented by a unique letter which is included in the combination name. The PAF records at the streamflow stations are used to determine the homogeneity of the clusters resulting from these combinations.

Table 7: Combinations of attributes for PAF data at streamflow stations

	Combinations			
Attributes	PMS	PMSL	PMSCL	PMSRL
Mean (M)	x	x	x	x
Standard Deviation (S)	x	x	x	x
Curve Number (C)			x	
Runoff Coefficient (R)				x
Lat and Long (L)		x	x	x

Statistics of MAR data at 255 streamflow stations are combined in various ways to perform cluster analysis. The main combinations are listed in Table 8. Each combination for this section begins with the letter X. Again, each attribute is represented by a unique letter and included in the name of the combination. A number is used to differentiate between the MAR, 3-day MAR, and 5-day MAR. There are nine combinations of attributes in this section. The PAF records are used to determine the homogeneity of the clusters resulting from these combinations.

Table 8: Combinations of attributes for precipitation data at streamflow stations.

	Combinations		
Attributes	XMSL	XMSCL	XMSRL
Mean (M)	x	x	x
Standard Deviation (S)	x	x	x
Curve Number (C)		x	
Runoff Coefficient (R)			x
Lat and Long (L)	x	x	x

4.3. Rainfall Analysis Results

A typical cluster map, such as that for RCWKL5 is shown in Figure 9. Some clusters are isolated from other sites, but the majority of the clusters are mixed with sites of different clusters. Less cluster distinction is exhibited in comparison with the 1-day and 3-day cases, and with the previous combination (RMCWL5). This trend is also evident in maps with higher numbers of clusters.

In general, the cluster maps resulting from RCWKL include more defined clusters than those for RMCWL. However, the clusters produced by RCWKL and RMCWL have similar HM's which display a higher degree of correlation between sites than the first two combinations. Since the first two combinations do not include L-moments as attributes, it is concluded that using these attributes results in more heterogeneous regions. The majority of the clusters generated by the analyses presented in this section do not exhibit clear structure. Furthermore, most of the Heterogeneity Measures (HM's) indicate that the clusters are not homogeneous.

The precipitation data were found to be highly correlated. The mean Cross-Correlation Coefficient (CCC) is 0.656 for the 1-day MAR series, 0.742 for the 3-day MAR series, and 0.752 for the 5-day MAR series. These high values indicate a strong dependence between the MAR data. In other words, the probability for a particular MAR value for a give year at a particular site is highly dependent on the MAR value for that year at a correlated site. Cluster analysis is strongly based on the assumption that site attributes are independent and are not correlated. Because the precipitation data are highly correlated, it is not appropriate to use this data to determine the homogeneity of clusters for a particular combination of attributes.

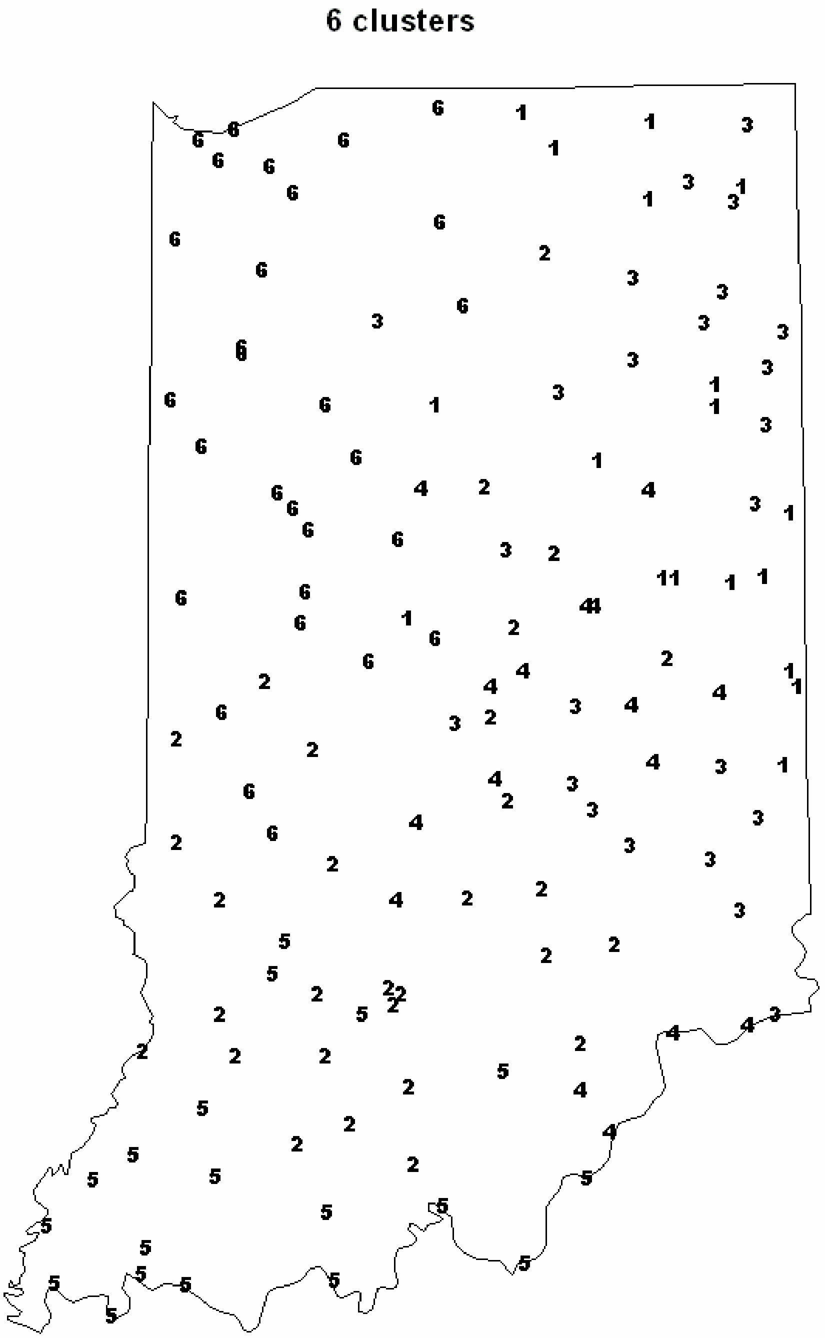

Figure 9: Cluster map for RCWKL5

4.4. Results for Streamflow Data

Due to the addition of the latitude and longitude, the clusters in this case are much more distinct than those resulting from rainfall data. Isolated regions for this combination are clearly visible. The optimum number of clusters was chosen as 7; the cluster map for 7 clusters is shown in Figure 10. Although there is still mixture between clusters, a clear structure is exhibited by the formation of distinguished regions. The addition of the latitude and longitude provides for a better partitioning of sites for the PAF data.

In order to form regions based on Peak Annual Flow (PAF) data, the mean and standard deviation of the PAF data are used as attributes for cluster analysis. The clusters resulting from this analysis (PMS) are not clear when compared to those resulting from the addition of the latitude and longitude (PMSL). Because the addition of the latitude and longitude does not result in significantly higher Heterogeneity Measures (HM's), it is concluded that the latitude and longitude should be included as an attribute for cluster analysis with PAF data. This conclusion is consistent with the results from Maximum Annual Rainfall (MAR).

The addition of the Curve Number as an attribute does not reduce the HM's, nor does it result in more defined clusters than in combination PMSL. Therefore it is not beneficial to include the Curve Number as an attribute. However, the addition of the Runoff Coefficient results in clearly-structured clusters, with similar clarity to those in combination PMSL. Furthermore, since the HM's for combination PMSRL are slightly lower than for PMSL, it is concluded that it is beneficial to include the Runoff Coefficient as an attribute in cluster analysis with PAF data.

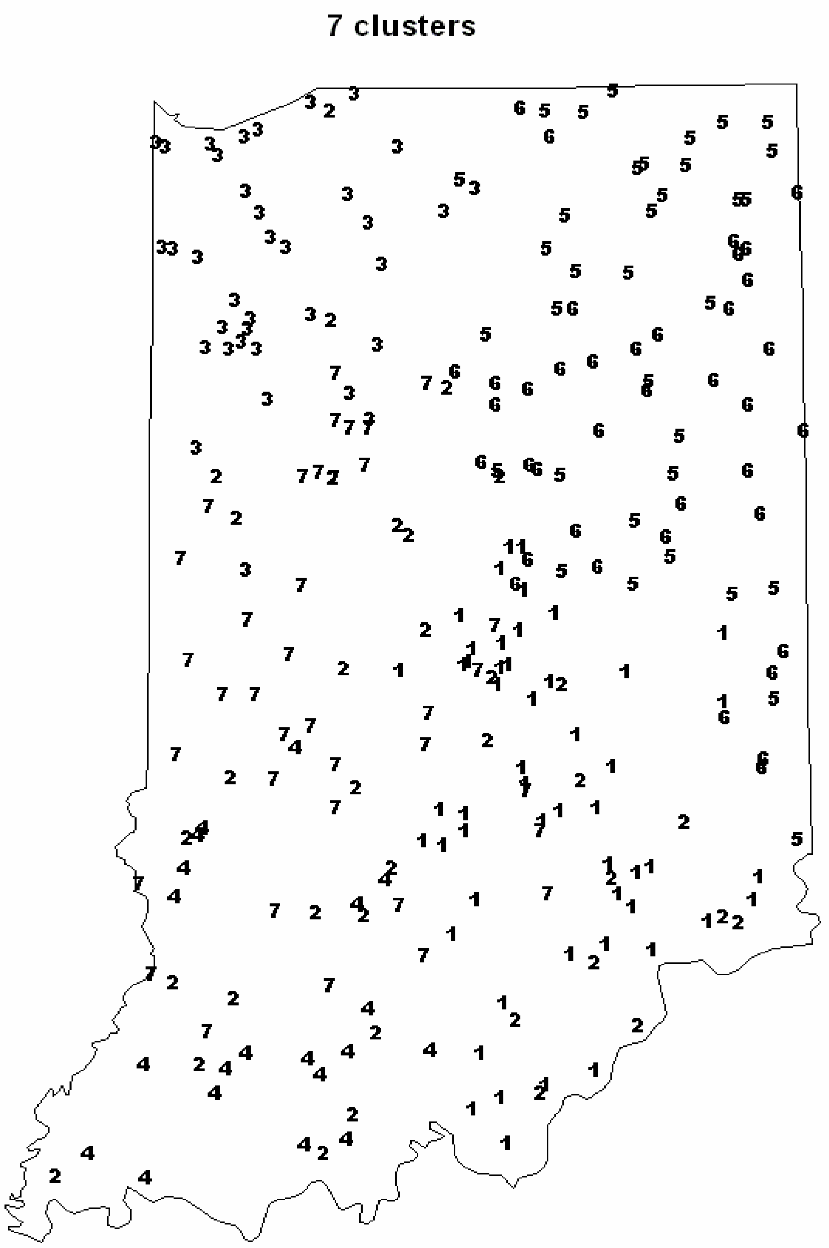

Figure 10: Cluster map for PMSL

The addition of the Curve Number as an attribute did not reduce the HM’s, nor does it result in more defined clusters than in combination PMSL. Therefore it is not beneficial to include the Curve Number as an attribute. However, the addition of the Runoff Coefficient results in clearly-structured clusters, with similar clarity to those in combination PMSL. Furthermore, since the HM's for combination PMSRL are slightly lower than for PMSL, it is concluded that it is beneficial to include the Runoff Coefficient as an attribute in cluster analysis with PAF data.

The clusters in Figure 11 for 5 clusters are also well-defined. Cluster 4 is much more heterogeneous than the others for 5 clusters. In the cluster map, Cluster 4 is also located in the northwest portion of Indiana.

In general, performing cluster analysis by considering precipitation data at streamflow stations produces secluded regions, with Heterogeneity Measures that classify most of the regions as heterogeneous. The single highly heterogeneous cluster is consistently in the same location for this combination, regardless of which MAR data are used.

4.5 Results for Precipitation and Streamflows data

The map for 6 clusters for XMSRL5 is shown in Figure 11. Cluster 3 is again found in the northwest portion of the state. When compared to the map in Figure 11 for XMSL5, there is less mixture between clusters; it is evident that the addition of the Runoff Coefficient also results in clearer divisions of clusters for the 5-day case.

Given that the addition of the Runoff Coefficient produces a clearer partitioning of sites for all MAR data, and that the addition does not increase the heterogeneity of the resulting regions, it is beneficial to include it as an attribute for cluster analysis when considering precipitation data at streamflow stations. The clusters located in the

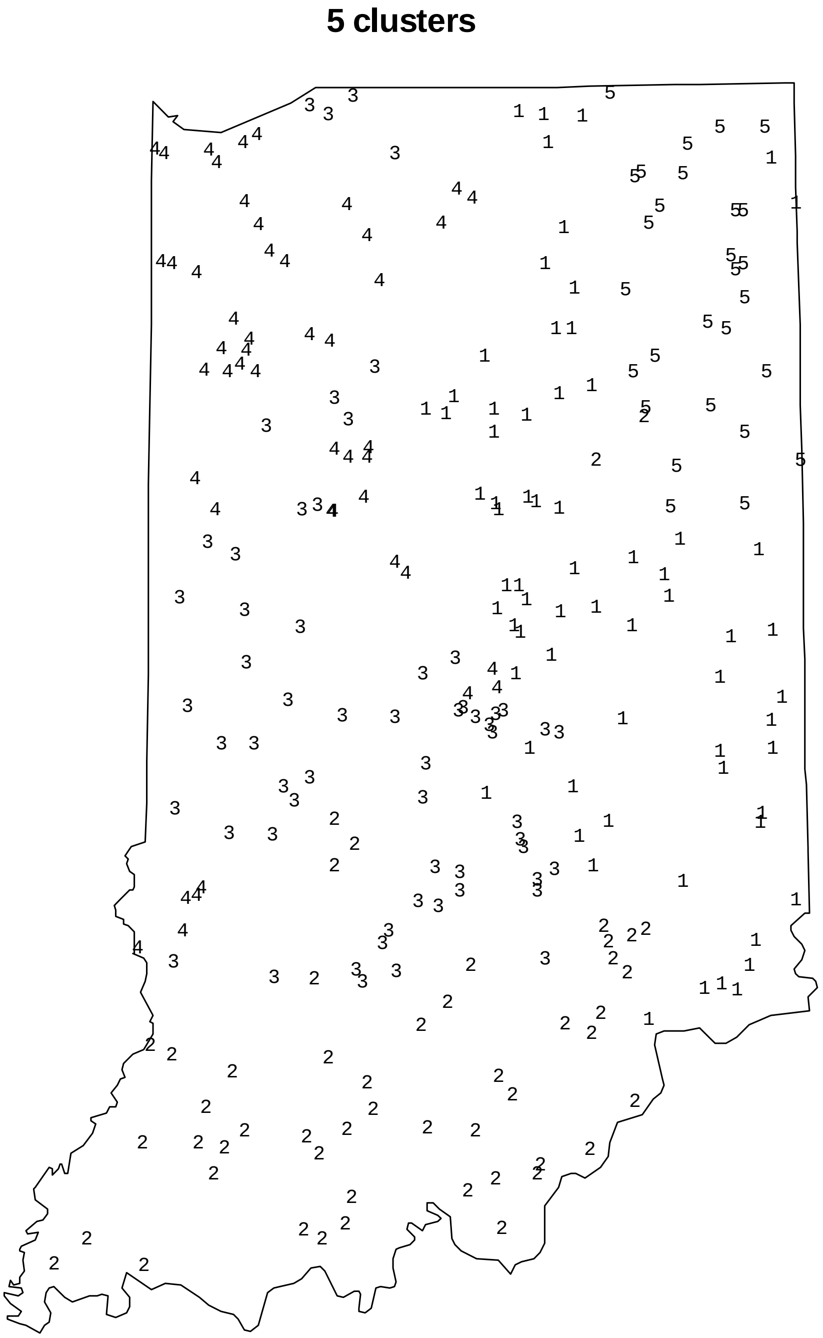

Figure 11: Cluster map for XMSL5

northwest portion of the state display consistently higher Heterogeneity Measures than the other clusters. For this reason, it is determined that this area is highly heterogeneous when considering precipitation data at streamflow stations

Regions were formed based on Maximum Annual Rainfall data. The inclusion of the Runoff Coefficient was again beneficial for this analysis. The most effective combination was determined to be XMSRL. The optimum number of clusters is again 7, and the cluster map for XMSRL5 is shown in Figure 11. Consistently lower HM's were produced by the use of the 3-day MAR data, when compared to the 1-day and 5-day cases. These values are listed in Table 6. Only Region 6 can be considered as possibly homogeneous.

By essentially linking precipitation data to streamflow stations, clear partitioning of sites give a unique set of regions. An undetermined amount of correlation still exists between the sites. Given that two sites with the same cluster membership are highly correlated, one should be removed from the analysis.

Table 9: Heterogeneity measures for 7 clusters for XMSRL3

Region	*H1*	*H2*	*H3*
1	3.29	2.79	1.67
2	2.81	1.56	0.67
3	13.42	4.67	2.36
4	5.92	1.49	0.74
5	2.42	0.17	-0.60
6	1.58	-1.02	-1.52
7	4.71	0.75	-0.37

Srinivas and Rao (2002) used a hybrid approach to cluster analysis to form regions in Indiana for Flood Prediction (Fig. 8). Although there is some similarity between the results in Figs. 12 and Fig. 8, there is one major drawback in all the results presented in this chapter. None of the regions in the results presented in this chapter are homogeneous. All the regions, except one, presented by Srinivas and Rao are homogeneous. This leads us to the conclusion that rainfall does not add much to the regionalization of Indiana watersheds.

4.6. Conclusions

The boundaries between the final regions presented in Figure 12 are approximate, and the Heterogeneity Measures do not fall within the homogeneous range. It is possible that by deleting one or more sites within each cluster, their HM's may decrease. The absence of any structure in the maximum rainfall in Indiana does not significantly contribute to regionalization.

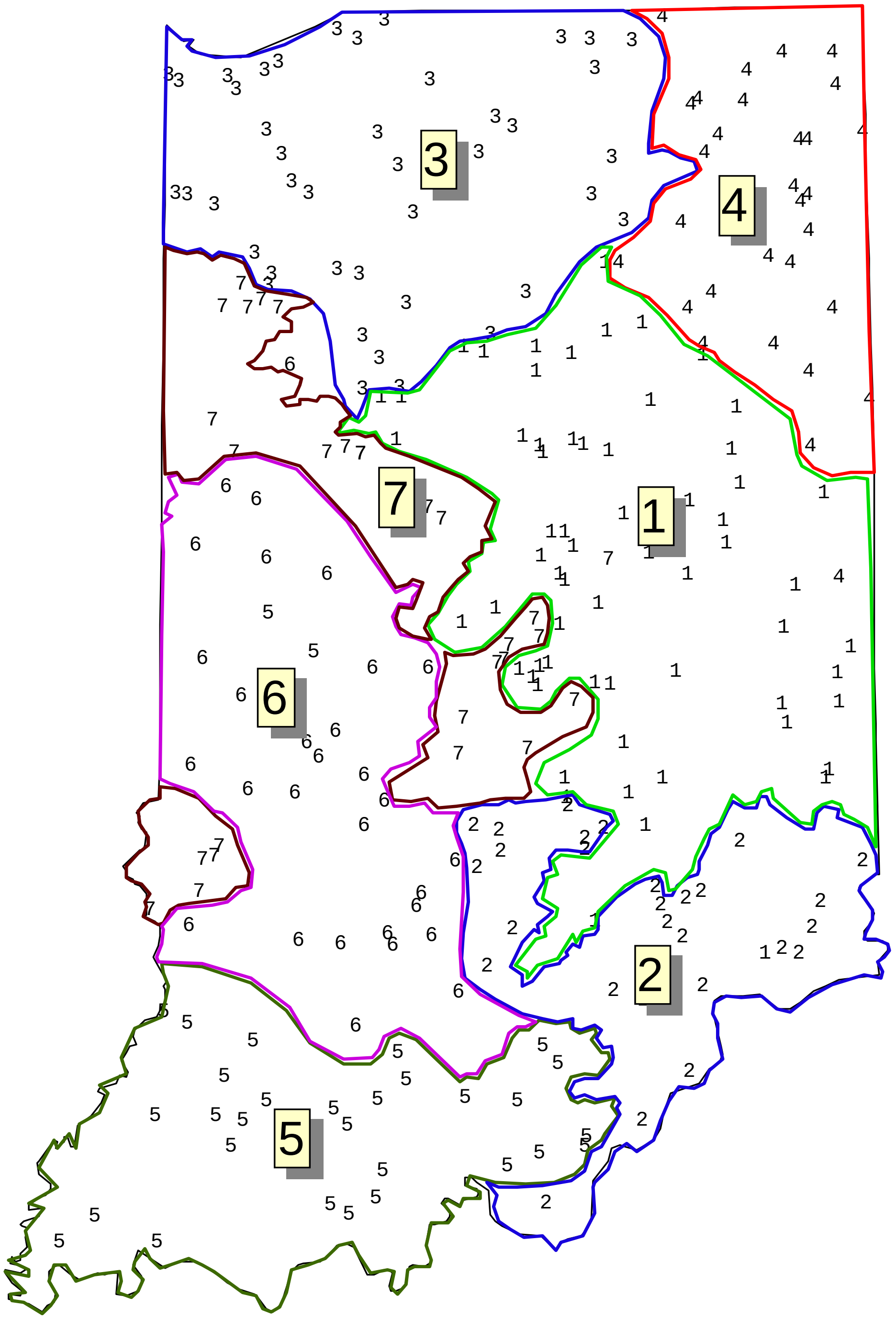

Figure 12: Regions for XMSRL3

V. REGIONALIZATION OF INDIANA WATERSHEDS BY FUZZY CLUSTERING METHODS

The basic objective of the research discussed in this chapter is to investigate the use of fuzzy clustering methods for regionalization of watersheds. In hydrologic literature, several approaches have been proposed for regionalization. However, the fuzzy clustering methods have not been used for regionalization. Consequently its performance in regionalization was tested and the results reported herein.

5.1. Classification of Fuzzy Clustering Agorithms

In the last two decades, several investigators have devoted their efforts to develop a variety of fuzzy cluster analysis techniques. This work received impetus from the developments in high-speed computers, and the fundamental importance of classification as a scientific procedure (Aldenderfer and Blashfield, 1984).

Clustering algorithms may be classified into supervised and unsupervised based on the uncertainty in the number of natural classes (or clusters) and hierarchies present in the data. Supervised clustering algorithms are used when the number of clusters in the input data set is known *a priori*, whereas unsupervised clustering algorithms are used when the number of clusters in the input data set is not known. In the context of regional flood frequency analysis, since the internal structure of the data is not known *a priori*, unsupervised clustering algorithms are the option by default.

The majority of unsupervised clustering algorithms start with two clusters and the number of clusters is increased every time after clustering is performed and cluster validity measures are computed. Cluster validity measures are independently used to evaluate and compare clustering partitions and even to determine optimal number of clusters existing in a data set. Unsupervised clustering algorithms differ from one another in their strategy of computing the new cluster center. The flow charts of supervised and unsupervised clustering algorithms are presented in Figures 13 and 14 respectively.

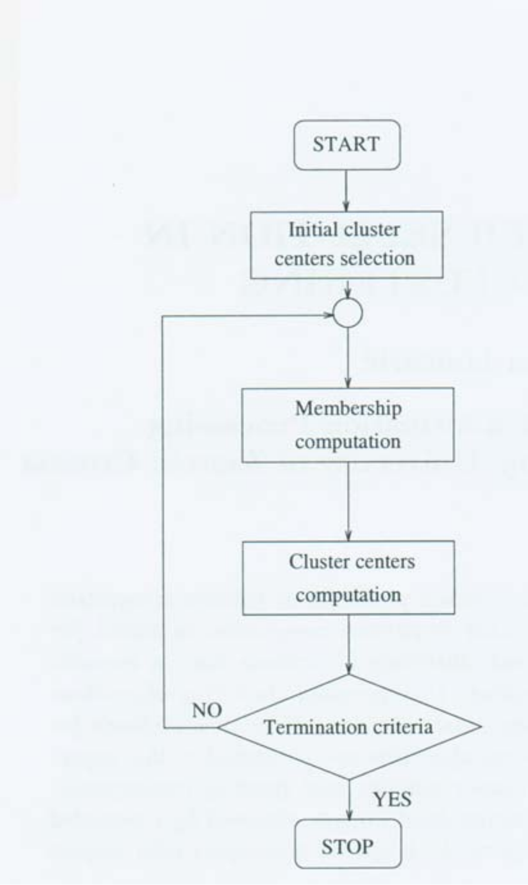
START
Initial cluster centers selection
Membership computation
Cluster centers computation
NO
Termination criteria
YES
STOP

Figure 13. Flow chart of Supervised clustering (Reference: Cosic and Loncaric, 1996).

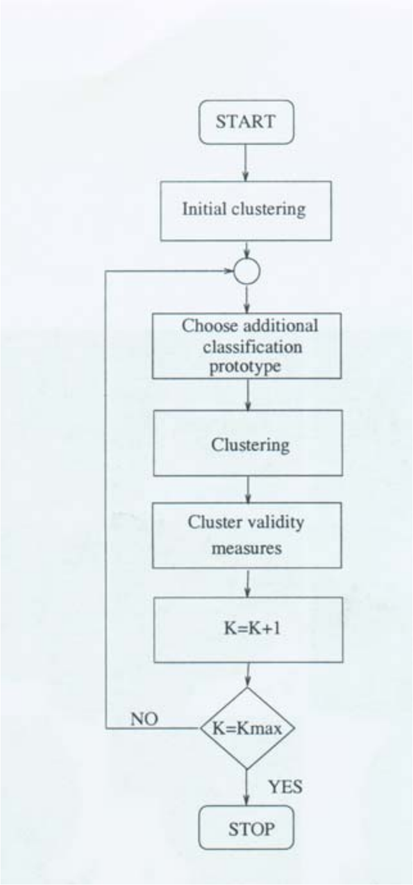
START
Initial clustering
Choose additional classification prototype
Clustering
Cluster validity measures
K=K+1
NO
K=Kmax
YES
STOP

Figure 14. Flow chart of Unsupervised clustering (Reference: Cosic and Loncaric, 1996).

The fuzzy clustering methods can be divided into two types based on the strategy adopted for partitioning the data (Yang, 1993): One that uses a fuzzy relation to perform fuzzy clustering; the other that uses the objective function to determine fuzzy clustering. The grouping achieved from fuzzy relations are separate segments, whereas those resulting from the use of objective functions constitute soft segmentation. The fuzzy clustering based on fuzzy relations were proposed by Tamura et al. (1971). They presented a multistep procedure by using the composition of fuzzy relations beginning with a reflexive and symmetric relation. The description of the original fuzzy clustering algorithm based on objective function dates back to 1973 (Bezdek, 1973; Dunn, 1974). This algorithm was conceived in 1973 by Dunn (1974) and further generalized by Bezdek (1973; 1981). Subsequently, Rouben (1982), Trauwaert (1985; 1988), Gath and Geva (1989), Gu and Dubuisson (1990), and Xie and Beni (1991) among others revised the algorithm by Dunn (1974). A description of these generalizations can be found in Bezdek and Pal (1992).

Among the existing fuzzy clustering methods, the Fuzzy c-means (FCM) algorithm proposed by Bezdek (1981) is the simplest and is the most popular technique of clustering. It is an extension of the hard c-means algorithm to fuzzy framework. A description of hard c-means algorithm is found in Srinivas and Rao (2002). The FCM algorithm has found applications in a variety of areas including agricultural engineering, astronomy, chemistry, geology, image analysis, medical diagnosis, shape analysis and target recognition (Bezdek, 1987). The FCM algorithm was used in this study.

5.2. Cluster Validity

Validity evaluation is a procedure that is oriented to evaluate and compare clusters resulting from a clustering algorithm for different choices of parameters or to compare clusters resulting from different clustering algorithms (Backer and Jain, 1981). Often, a cluster validity function is considered different from the objective function being optimized. However, it has also been incorporated into objective function in a few studies.

The criteria that are considered in cluster evaluation and selection include *compactness* and *separation* of clusters.

- *Compactness*: Optimal partition requires that the members of each cluster should be as close to each other as possible. A common measure of compactness is the variance, which should be minimized. If only compactness is considered as the validation criterion, then the best partition is obtained when each data point is considered as a separate cluster.
- *Separation*: Optimal partition requires that the clusters should be widely spaced. In other words, clusters should be far from each other. If only optimal separation is considered as the validation criterion, then the best partition is obtained when all the data points are included in a single cluster.

Clustering validity has often been used to determine optimal number of clusters in a data set (e.g., Gath and Geva, 1989; Xie and Beni, 1991; Theodoridis and Koutroubas, 1999; Halkidi et al., 2001). The procedure requires fixing all the parameters of a clustering algorithm (except number of clusters, C). Next, the parameter C is varied from 1 to a maximum value C_{max} in increments of 1. Cluster validation index is applied to clusters obtained for each choice of C to compute the validation value. The validation values are then plotted against their respective C values on a graph to determine the optimal number of clusters for a given data set.

5.3. Data Used in the Study

In this study, flow records from 273 gauging stations in Indiana are used. These included all the 245 stations considered by Glatfelter (1984). In pooling additional gauging stations to those considered by Glatfelter, a screening criterion of having a minimum record length of 10 years was used. Twenty-eight stations passed this screening test. The selection of the record length threshold for the screening process is subjective. The threshold value chosen should be large enough to retain the identity of the homogeneous regions even when sites with short record length are excluded. In the following discussion, the 245 stations considered by Glatfelter will be referred to as *Glatfelter stations* and the 28 stations included through the screening criterion will be referred to as *Pooled stations*. Further details about the data used are found in Srinivas and Rao (2002) and in chapter III of this report.

The sensitivity of flood response to variations in the values of the attributes was investigated by Srinivas and Rao (2002). This lead to a selection of the meteorological attribute, mean annual precipitation and four physiographic attributes: drainage area, slope of the main channel in the drainage basin, soil runoff coefficient and storage. The geographic location attributes Latitude and Longitude are included in the feature vector to identify regions that are geographically contiguous.

5.4. Results

The study resulted in delineation of Indiana into five homogeneous regions, one heterogeneous region and an unallocated residue of 23 stations of which 21 are located in

Indiana. The results from the analysis are presented in Tables 6. The final pictorial representation of the regions is presented in Fig. 15.

Table 10. Characteristics of the regions formed

Region number	N	RS	Heterogeneity measure		
			H_1	H_2	H_3
1	62	1689	0.86	-0.12	-0.94
2	58	1730	0.85	0.43	-0.65
3	30	804	-0.46	0.66	0.28
4	73	3039	0.48	-0.40	-1.78
5	42	1938	0.04	-0.91	-0.85
6	14	519	13.69	6.33	2.94

N: Number of stations

RS: Region size in station years

The results from Table 6 indicate that regions 1 to 5 are all acceptably homogeneous, while region-6 adjoining Lake Michigan is highly heterogeneous. All the homogeneous regions identified have more than 5T station-years of data. For the ungauged catchments lying at the border between the regions 2 and 3, the possibility of including information from both the regions can be considered. Region-4 is in central Indiana. The soil here is predominantly loamy glacial till. Region 5 is spread over northern part of Indiana. It comprises of a wide range of soil classes (clayey glacial till, sandy and loamy deposits, loamy glacial till) overlying the Mississippian rocks of Michigan basin and Devonian and Mississippian shale.

Interestingly, the delineated regions bear remarkable resemblance with geological features and soil regions of Indiana (Srinivas and Rao, 2002). Twenty three sites which amount to 506 station years of data or approximately 5.4 % of the 9385 station years of record considered for modeling did not fit in any of the hydrologic regions. Details of the residual stations are found in Srinivas and Rao (2002).

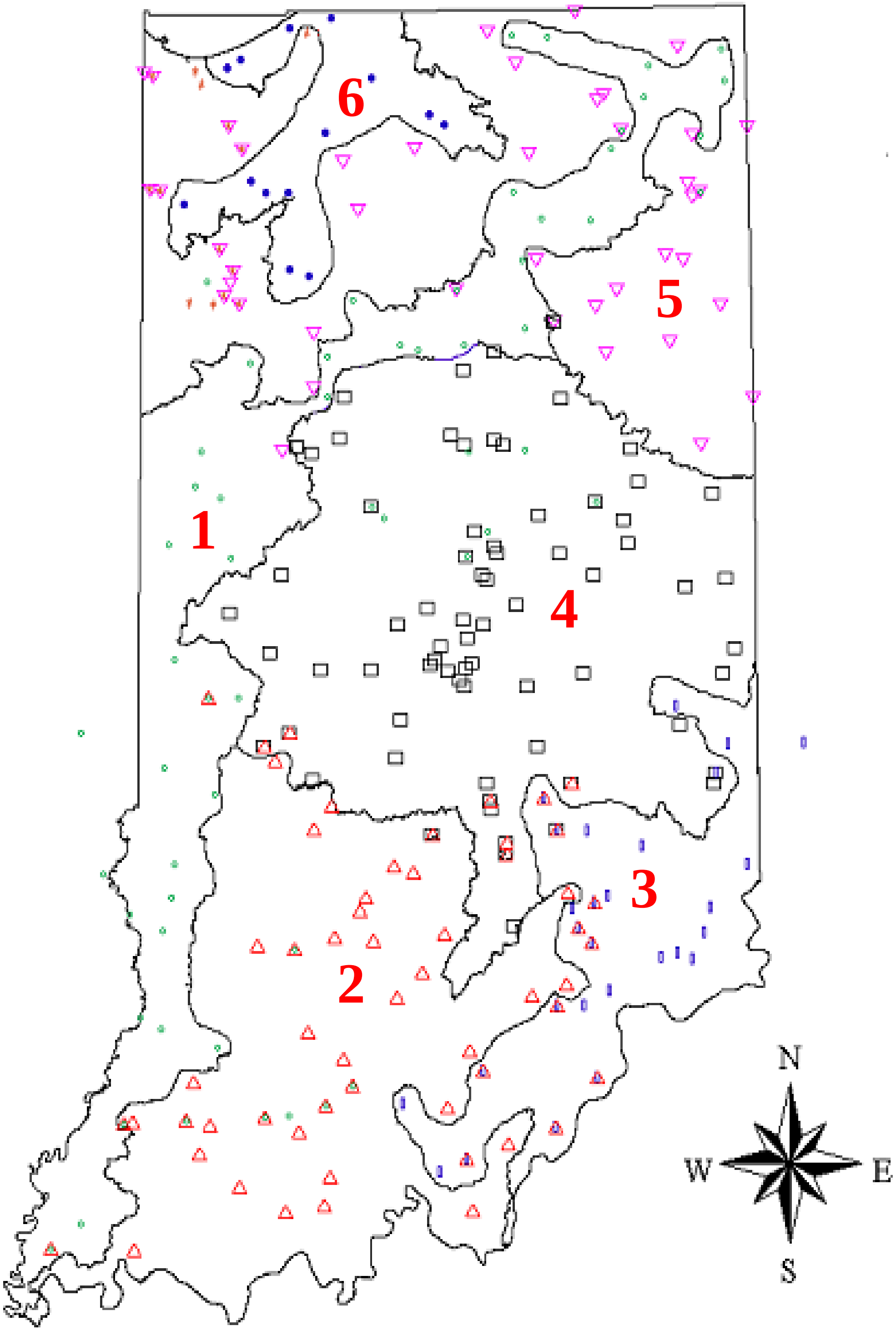

Figure 15. Homogeneous regions formed by fuzzy cluster analysis.

5.5. Conclusions

The study resulted in delineation of Indiana into five homogeneous regions and one heterogeneous region. In addition, one homogeneous subregion has been identified in the Kankakee river basin. The homogeneous regions identified are same as those presented in Srinivas and Rao (2002) where Hybrid clustering algorithms were used to identify homogeneous hydrologic regions in Indiana. Nevertheless, it is worth mentioning that the effort devoted to revise the optimal partitioning achieved by the FCM algorithm to arrive at the final result is comparatively smaller. Also, the result from the FCM algorithm strengthens the conclusions drawn in Srinivas and Rao (2002).

VI REGIONALIZATION USING ARTIFICAL NEURAL NETWORKS

6.1 Introduction

The primary objective of the work presented in this chapter is to partition Indiana watersheds into groups that are homogeneous in hydrologic response by using artificial neural networks based cluster analysis. The nonlinearity and flexibility embedded in artificial neural networks makes them well-suited for complex problems such as those encountered in regionalization.

In the past two decades, artificial neural network based models have been extensively developed and studied by several investigators in an effort to simulate the behavior of the neurons in the human brain. The first artificial neuron was produced in 1943 by the neurophysiologist Warren McCulloch and the logician Walter Pits. However, scientists recognized their real potential only in the early nineteen eighties with the advent of modern computing facilities. Artificial neural networks have found wide applications in hydrological sciences as evidenced by articles in Govindaraju and Rao (2000), Minns (1995), ASCE Task Committee (2000 a,b).

In this study, flow records from 273 gauging stations in Indiana are used. These include all the 245 stations considered by Glatfelter (1984). In pooling additional gauging stations to those considered by Glatfelter, a screening criterion of having a minimum record length of 10 years was used. Twenty-eight stations passed this screening test. The selection of the record length threshold for the screening process is subjective. The threshold value chosen should be large enough to retain the identity of the homogeneous regions even when sites with short record length are excluded. In the following discussion, the 245 stations considered by Glatfelter will be referred to as *Glatfelter stations* and the 28 stations included through the screening criterion will be referred to as *Pooled stations*. The location of these stations in the study region is found in Srinivas and Rao (2002) and in chapter II of this report.

Details of nine attributes used by Glatfelter (1984) to assess the degree of similarity between drainage basins in Indiana are available for the 245 stations from Glatfelter (1984). The range of each of these attributes is presented in Table 7. The values of these attributes for each of the stations considered in the study are presented in Srinivas and Rao, (2002). The attributes are subjected to a screening process with the goal of extracting independent attributes for cluster analysis.

Table 11. Attributes available for the Glatfelter stations.

Attribute	Range
Drainage Area	0.11 – 11125.00 mi^2
Mean Annual Precipitation	34 – 46 in
Main channel Slope	0.90 – 267.00 ft/mile
Main channel Length	0.3 – 315.0 miles
Basin Elevation	412.0 – 1190.0 ft
Storage[1]	0% – 11%
Soil Runoff coefficient	0.30 –1.00
Forest cover in Drainage Area	0.0 – 88.4%
I(24,2)[2]	2.6 – 3.35 in

[1] Storage – percentage of the contributing drainage area covered by lakes, ponds or wetlands

[2] I(24,2) – 24-hour rainfall having a recurrence interval of 2 years, in inches.

6.2. Results from ANN Clustering Algorithm

The selected seven attributes (mean annual precipitation, drainage area, slope of the main channel in the drainage basin, soil runoff coefficient, storage, latitude and longitude) were used to derive clusters for Indiana using ANN clustering algorithm.

To examine the sensitivity of results from ANN clustering algorithm to variation in the number of clusters, the parameter *m* was varied from 1 to 10. Three different scenarios were considered for the cluster analysis: (1) equal weight was assigned to all the seven attributes; (2) more importance was given to drainage area than other attributes; and (3) analysis was conducted without geographic location attributes (latitude and longitude). Comparison of results from scenario-1 with those of scenario-3 demonstrate the role of latitude and longitude in providing clusters with well-defined boundaries for the region.

The prime objective of analyzing patterns/classifications provided by clustering algorithm for a variety of scenarios is to identify plausible solution for which majority of sites are classified into clusters that are as homogeneous as possible. In general, majority of identified clusters tend to be homogeneous with increase in *c*, the number of clusters. However, increase in *m* provides several small clusters that are ineffective for regional flood frequency analysis (RFFA). A region identified for flood frequency analysis should be sufficiently large to provide an effective estimate of the flood quantile of interest. Reed et al. (FEH 1999, p.28, Vol.3) suggested *5T rule* which specifies that the pooled stations should collectively supply five times as many station years of record as the target return period. On the other hand, one may also argue against using very large clusters. The extent to which regional frequency analysis is preferable to at-site analysis depends on the number of sites in a region (Hosking and Wallis, 1997, p.119). From simulation results for four varieties of representative regions and with sites having record length of 30, they concluded that there is little to be gained by using regions larger than about 20 sites, unless extreme quantiles corresponding to nonexceedance probability F ≥0.999 are to be estimated.

For the work presented in this chapter, optimal value of *m* is determined as a tradeoff between decrease in region size and increase in homogeneity of the region with the goal of identifying plausible hydrologic regions that are effective for RFFA.

The heterogeneity measures of Hosking and Wallis (1993), were used to assess the homogeneity of the identified regions and weigh information from each station in proportion to its record length. As a consequence, stations with longer record length would have more influence on the heterogeneity indices than stations with shorter record length. This may have adverse effects, especially when some stations in a region have much longer record lengths than others. To address this effect, the hydrologic regions that have been obtained by region revision process are further examined for their robustness. In this exercise, specifying various threshold values segregates stations with record length significantly different from the rest of the group and the region consisting of rest of the stations was examined for homogeneity. It was also intended to identify and exclude a few stations that have an adverse affect on the homogeneity of the regions. This information helps in judging appropriate threshold values for the analysis. All the homogeneous regions identified are indeed robust. The final regions formed by using ANN is given in fig. 16.

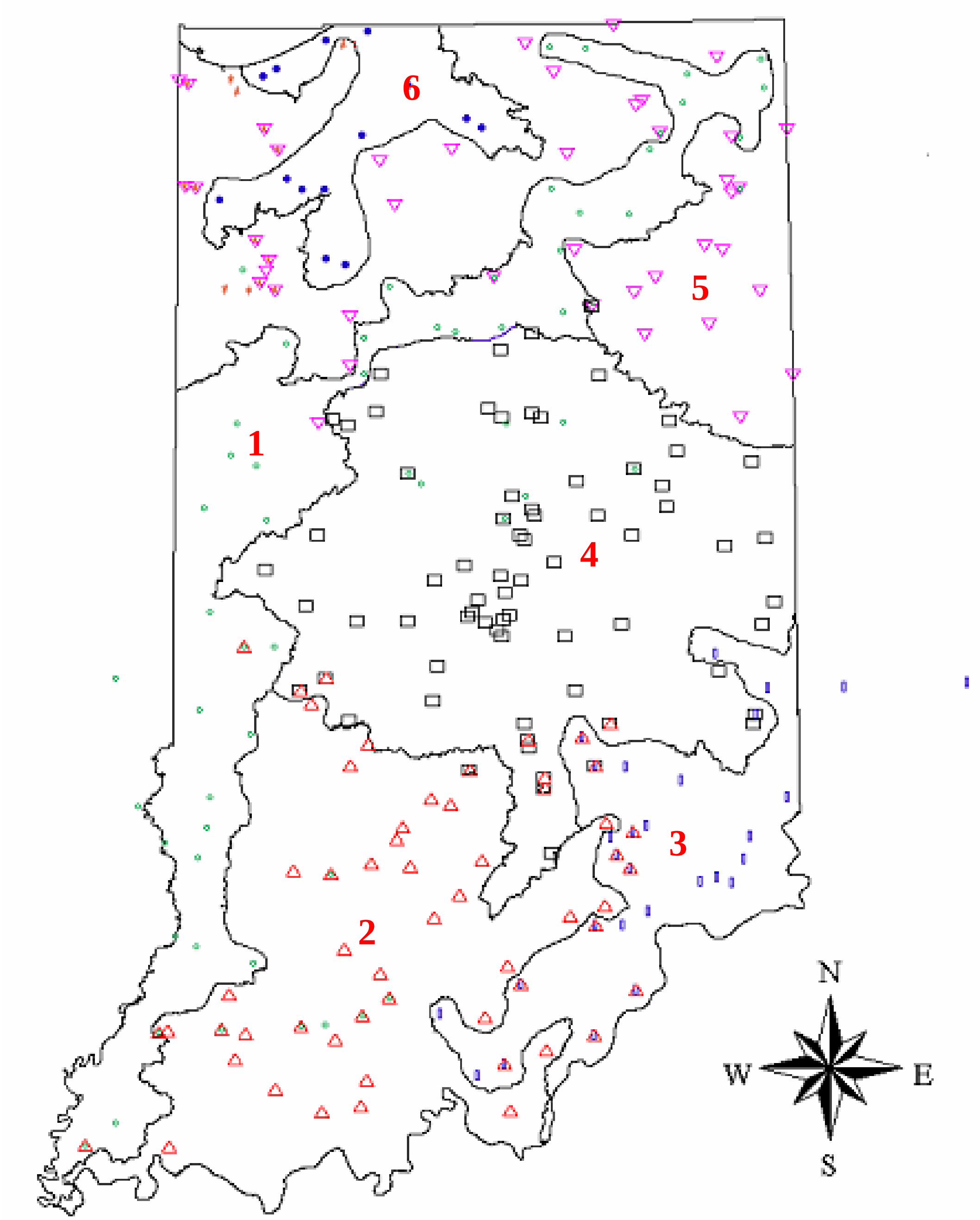

Figure 16. Final results of regionalization by ANN method.

6.3 Conclusions

The results of homogeneity tests indicated that regions 1 to 5 are all acceptably homogeneous, while region-6 adjoining Lake Michigan is highly heterogeneous. All the homogeneous regions identified have sufficient pooled information to be effective for flood frequency analysis. Region-1 is spread mainly along the course of Wabash river and consists predominantly of alluvial deposits of the flood plains. Region-2 contains karst formations associated with limestones of the Mississippian age. Region-3 possesses a karst area consisting of older Devonian and Silurian limestones. The topography of these areas is dominated by sinkholes, sinking streams, large springs and caves. For the ungauged catchments lying at the border between regions 2 and 3, the possibility of including information from both the regions can be considered. Region-4 is in central Indiana. The soil here is predominantly loamy glacial till. Region 5 is spread over northern part of Indiana. It comprises of a wide range of soil classes (clayey glacial till, sandy and loamy deposits, loamy glacial till) overlying the Mississippian rocks of Michigan basin and Devonian and Mississippian shale.

Interestingly, the delineated regions bear remarkable resemblance with geological features and soil regions of Indiana. Twenty three sites which amount to 506 station years of data or approximately 5.4 % of the 9385 station years of record considered for modeling did not fit in any of the hydrologic regions. Readers are referred to Srinivas et al. (2003) for details about the residual stations.

VII TESTING THE RESULTS OF REGIONALIZATION BY SIMPLE SCALING

7.1 Introduction

In this chapter, the simple scaling techniques are used to test the characteristics of data from homogeneous regions. The results are compared to those from a highly heterogeneous region to stress the importance of regionalization. Although the results presented in this chapter are based only on basin area, similar anaylses could be made with other attributes such as average basin slope or average rainfall.

For this study, the data used are the same as those discussed in chapter II. In addition to the peak annual flows, other gage and watershed attributes, such as basin area, are also necessary and can be found at http://water.usgs.gov/nwis/peak.

7.2 Simple Scaling Using Return Period Floods

Some of the theory and techniques used to obtain flood frequency information for watersheds that have incomplete data or no data at all are discussed below. Relationships between floods at several different return periods and basin areas are derived for each region so that the floods of different recurrence intervals can be estimated for watersheds that have no gages from which that information can be drawn.

The theory behind simple scaling discussed in Gupta et al. (194) is used here. Assuming that the peak annual flows are located in a statistically homogeneous region, then the data can be lumped into a set, Q{A}. Then the scaling invariance can be defined as,

$$\frac{Q(\lambda)}{g(\lambda)} \overset{d}{=} Q(1), \qquad (14)$$

where λ is a scale parameter, which can be expressed as,

$$\lambda = \frac{A_i}{A_j}, \qquad (15)$$

where A_i and A_j and are the basin areas of watersheds with different scales, g(λ) can be a random or a nonrandom function, and the d above the equals sign indicates that the probability distributions on both sides are the same. If g(λ) is nonrandom and known, then the distribution of Q(A) for a basin with area A can be determined from the distribution of Q(1) with a basin area of one. Equation 14 can be rewritten in terms of flood quantiles as,

$$q_p(A) = g(A) \cdot q_p(1). \qquad (16)$$

The function g(A) can be determined for processes exhibiting simple scaling invariance, or when CV[Q(A)] = CV[Q(1)], where CV is the coefficient of variance defined as,

$$CV = \frac{\sigma}{\mu}, \qquad (17)$$

where σ and μ are the variance and the mean of the data set, respectively. Gupta and Waymire (1990) show that the function g can be represented by,

$$g(\lambda) = \lambda^{\theta} \qquad (18)$$

where θ is a scaling exponent that can be any real number. Using this relation, it can be said that two distributions $Q(A_i)$ and $Q(A_j)$ are related by,

$$Q(A_i) \overset{d}{=} \left(\frac{A_i}{A_j}\right)^{\theta} \cdot Q(A_j). \qquad (19)$$

Using the expression in Equation 19, a universal relationship between Q and A for any given return period, T, can be written as,

$$Q_T = X \cdot A^{\theta} \qquad (20)$$

where X is some coefficient that denotes the intercept of the power law such that the intercept is equal to log X. X and θ can be determined through a simple regression analysis. To determine visually whether a regression analysis is even necessary, the floods at return periods of 2, 5, 10, 25, 50, 100, 200, and 500 years versus basin area are plotted on a log-log scale. An example is shown in fig. 17. If it appears that the data could be fitted with a straight line, then a power law equation, such as in Equation 20, may be useful and a regression analysis should be done to determine appropriate values for X and θ. If θ is a constant, or very similar for all return periods tested within each region, then simple scaling is valid and useful in predicting peak flows.

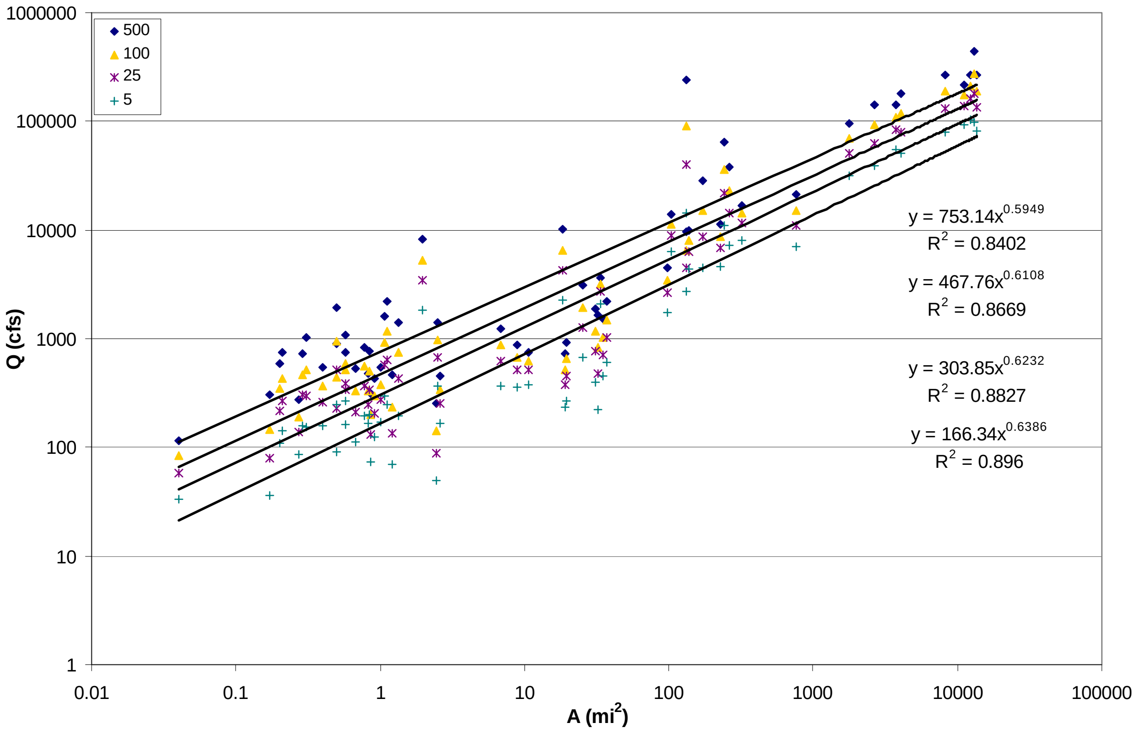

Figure 17: Q vs. A for Return Periods 5, 25, 100, and 500 yr in Region 1

The slopes, or θ values, are very similar for each return period within Regions 1-5. The correlation for these regions is also quite strong, with most R^2 values ranging between 0.85 and 1. Figures 18 shows the poor results for Region 6. This is not surprising considering that it is assumed that the regions should be statistically homogeneous and Srinivas and Rao (2002) found Region 6 to be heterogeneous. The slopes are very different, resulting in power regressions that appear to converge to a point, and the fit to the data is also poor. The R^2 values ranging 0.4 to 0.9 conclusively show that the correlation is poor for most of the power law relationships. These results indicate that simple scaling is valid for the homogenous Regions 1-5 and not valid for Region 6, which is heterogeneous.

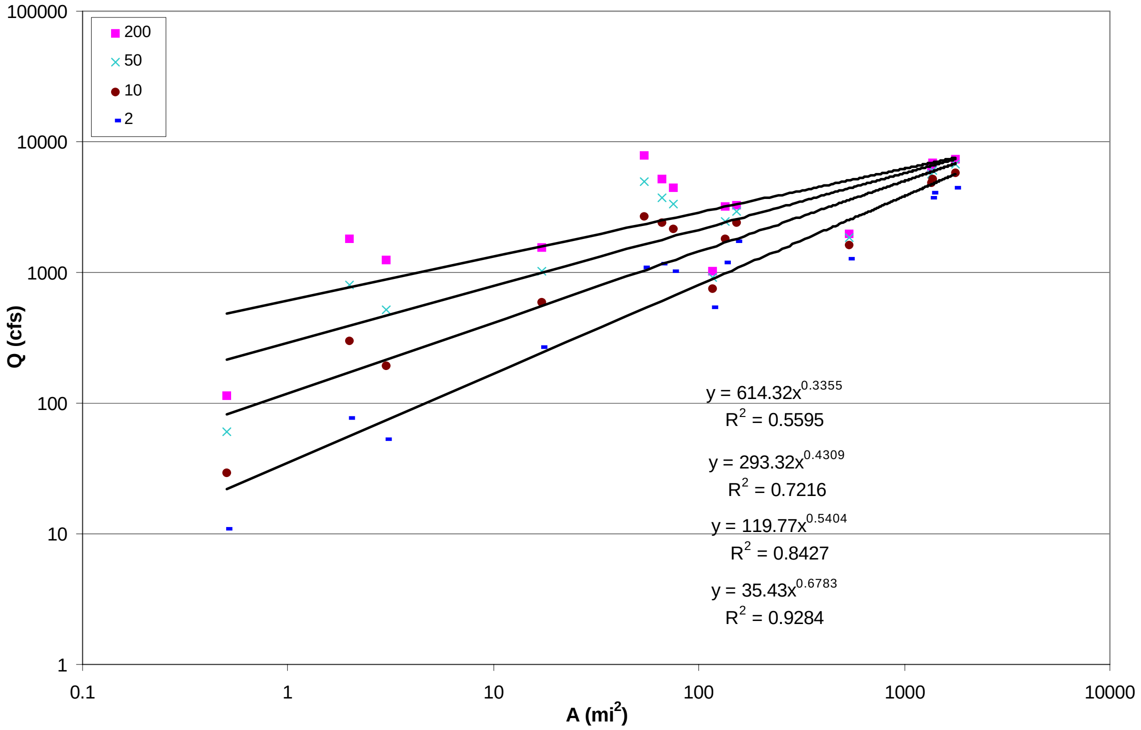

Figure 18: Q vs. A for Return Periods 2, 10, 50, and 200 yr in Region 6

7.3 Scaling By Using Moments

Technically, simple scaling is purely defined by Equation 4.1, which states the distributions of two data sets are equal. To apply the theory of simple scaling to the statistical moments of the distributions, it is assumed that moments of the distributions must be equal. Ribeiro et al. (1998) apply this assumption to obtain the following relationship for simple scaling using statistical moments,

$$\frac{E[Q(A)]^h}{A^{\theta \cdot h}} = E[Q(1)]^h \, , \qquad (21)$$

where h is the order of the moments. This expression can be rewritten using log-transforms as,

$$\log\left\{E[Q(A)]^h\right\} = \theta \cdot h \cdot \log(A) + \log\left\{E[Q(1)]^h\right\}, \qquad (22)$$

to more clearly show that the order of statistical moment, h, is proportional to the slope of the log-log line.

Using the expression in Equation 21, a universal relationship between any statistical moment E[Q(A)]^h of order, h, and A can be written as,

$$E[Q(A)]^h = Y \cdot A^{\theta \cdot h} \qquad (23)$$

where Y is some coefficient that denotes the intercept of the power law such that the intercept is equal to log Y. Y and θ can be determined through a simple regression analysis. For this study, only the first three moments are used. The first moment, E[Q(A)], is the mean, which is defined as,

$$E[Q(A)] = \mu = \frac{1}{N}\sum_{i=1}^{N} Q_i \ . \qquad (24)$$

The second moment, E[Q(A)]^2, is the variance, which is defined as,

$$E[Q(A)]^2 = \sigma = \frac{1}{N-1}\sum_{i=1}^{N}(Q_i - \mu)^2 \ . \qquad (25)$$

The third moment, E[Q(A)]^3, is defined as,

$$E[Q(A)]^3 = \frac{N}{(N-1)(N-2)}\sum_{i=1}^{N}(Q_i - \mu)^3 \ . \qquad (26)$$

The first, second, and third moments are plotted versus basin area and an example is given in fig. 19. If θ is a constant, or the slope of the second moment is nearly twice the slope of the first moment and or the slope of the third moment is nearly triple the slope of the first moment within each region, then simple scaling is valid and useful in predicting the moments of peak flows.

The slopes of the second moment are approximately twice that of the first moment and that the slopes of the third moment are nearly triple that of the first moment within Regions 1-5. The correlation for these regions weakens slightly as the order of the

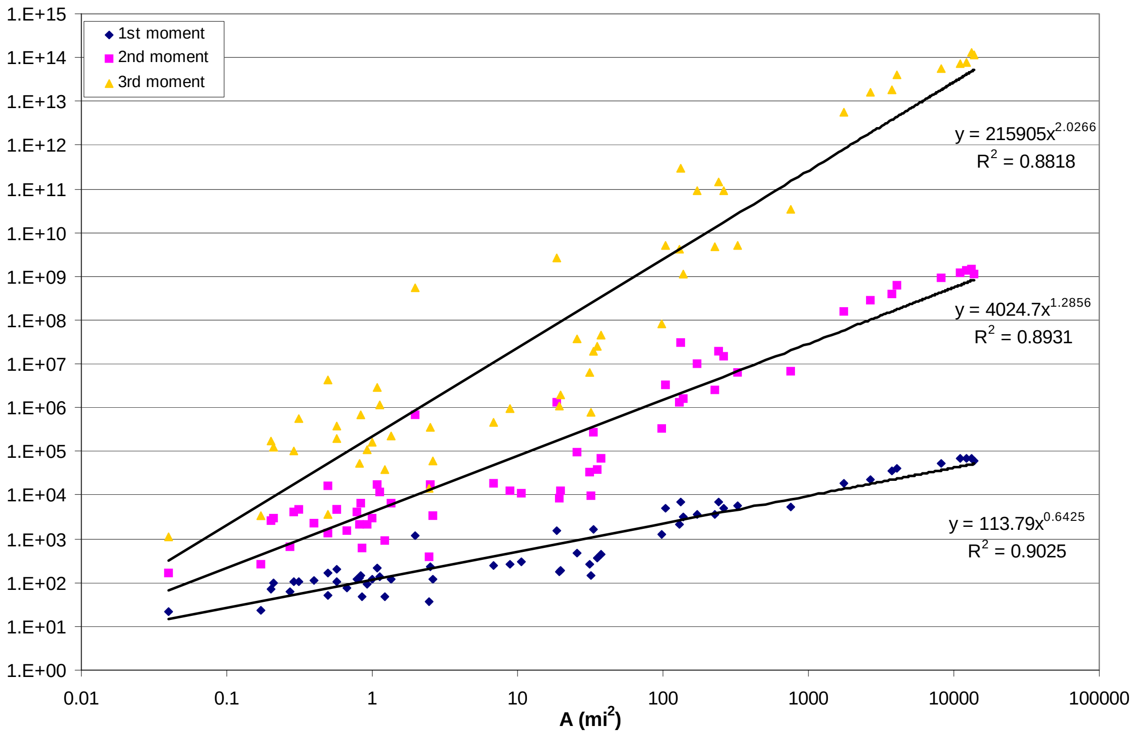

Figure 19: Statistical Moments vs. A for Region 1

moment increases, which explains why the slope of the second moment is closer to twice that of the first moment compared to the slope of the third moment being less close to triple the slope of the first moment. This indicates that for larger orders of moments, simple scaling may not be valid, though more research would necessary to determine a proper threshold. Figure 20 shows the poor results for Region 6. Again, this is not surprising considering that Srinivas and Rao (2002) found Region 6 to be heterogeneous. The slopes of the second and third moments are very different from double and triple the slope of the first moment. The correlation for the second and third moments is also poor, yielding R^2 values of 0.67 and 0.45, respectively. These results indicate that simple scaling using statistical moments is valid for the homogenous Regions 1-5 and not valid for Region 6, which is heterogeneous.

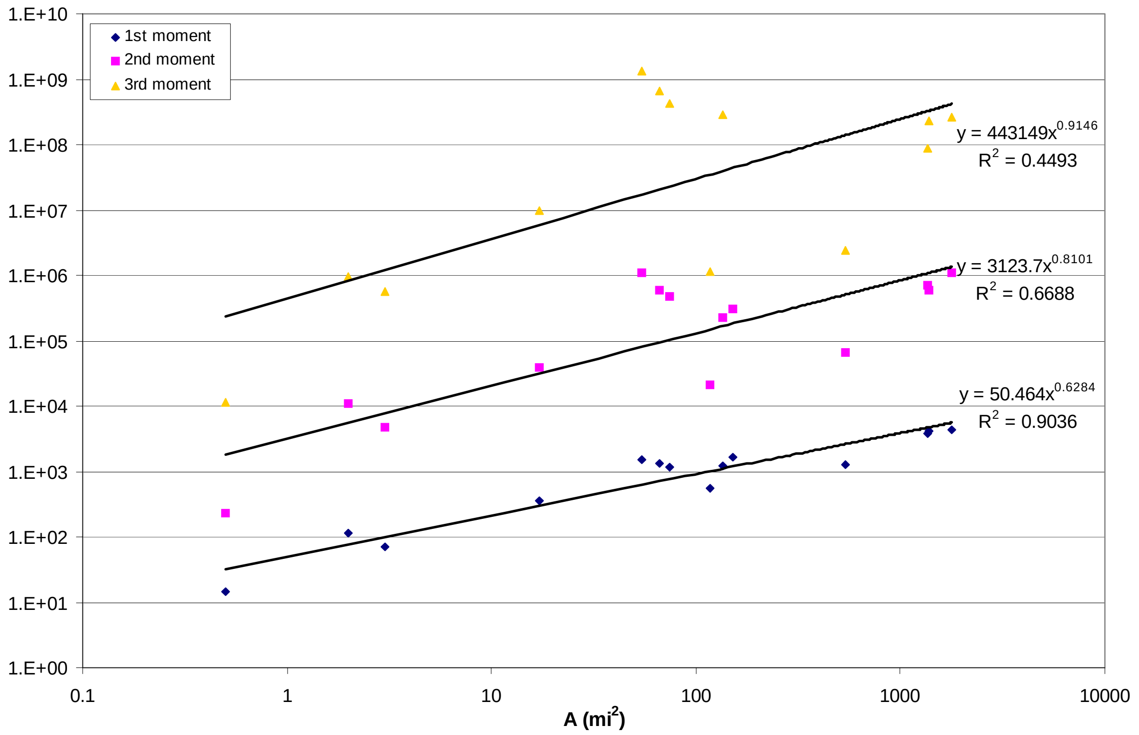

Figure 20: Statistical Moments vs. A for Region 6

7.4 Coefficient of Variation

Checking for a constant coefficient of variation is another check to determine whether simple scaling using statistical moments is valid. This will also serve as an assumption check for the theory discussed above. The coefficients of variation, CV, are estimated using Equation 16 for all gages within each of the six regions. These are grouped by region and then plotted against basin area on log-log scales. A power regression is fitted to the data to check the slope. If the slope is zero, or nearly zero, then it is assumed that the coefficient of variation is constant. If there is a sufficiently large slope, then the coefficient of variation is not constant with that region and simple scaling is not valid for that region. The plots of the coefficient of variation versus basin area were examined and examples are provided in figs. 21 and 22.

For Regions 1–5, the slope of the regression of the coefficient of variation versus basin area is very small. The absolute value of the slopes range between 0.0003 and 0.013, which are small enough to consider the coefficient of variation to be a constant. For Region 6, the slope of the regression of the coefficient of variation versus basin area is visually noticeable. The slope is -0.223, which indicates that the coefficient of variation varies inversely with basin area. This concurs with the findings of Ribeiro et al. (1998), who used similar methods of simple scaling with nonregionalized data from Ontario, Canada. This further supports that Region 6 is highly heterogeneous and simple scaling techniques are invalid for this region.

7.5 Cumulative Distributions

A final check for simple scaling using statistical moments is to determine whether the regionalization data follows a single distribution.

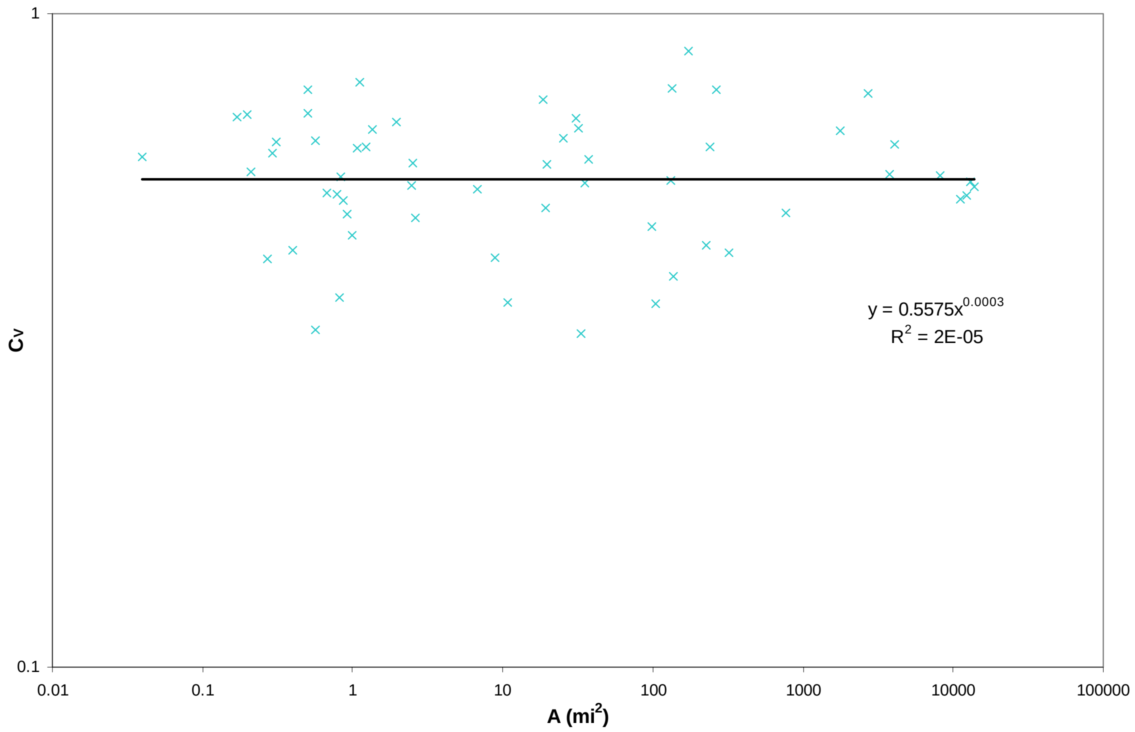

Figure 21: Coefficient of Variation vs. A for Region 1

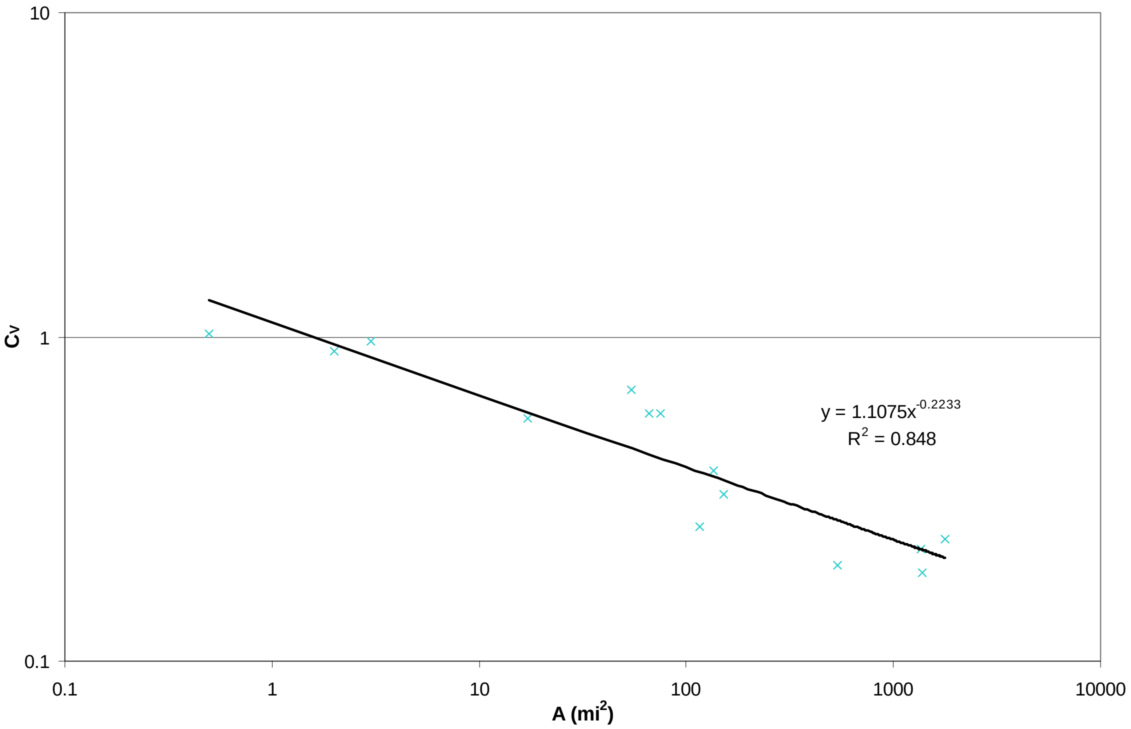

Figure 22: Coefficient of Variation vs. A for Region 6

In order to test this, a standardized random variable, q_{ij}, is used to empirically determine the distribution of the data, and defined in Ribeiro et al. (1998) as,

$$q_{ij} = \frac{Q_{ij}}{A_j^{\theta \cdot h}} \tag{27}$$

where the subscript, j, symbolizes a specific watershed or flood gage and the subscript, i, symbolizes the year.

The flood data are plotted as one set for each region. Examples of these plots for region one and six are shown in Figures 23 and 24. Included on the figures is a comparison to the lognormal distribution, which takes the form,

,

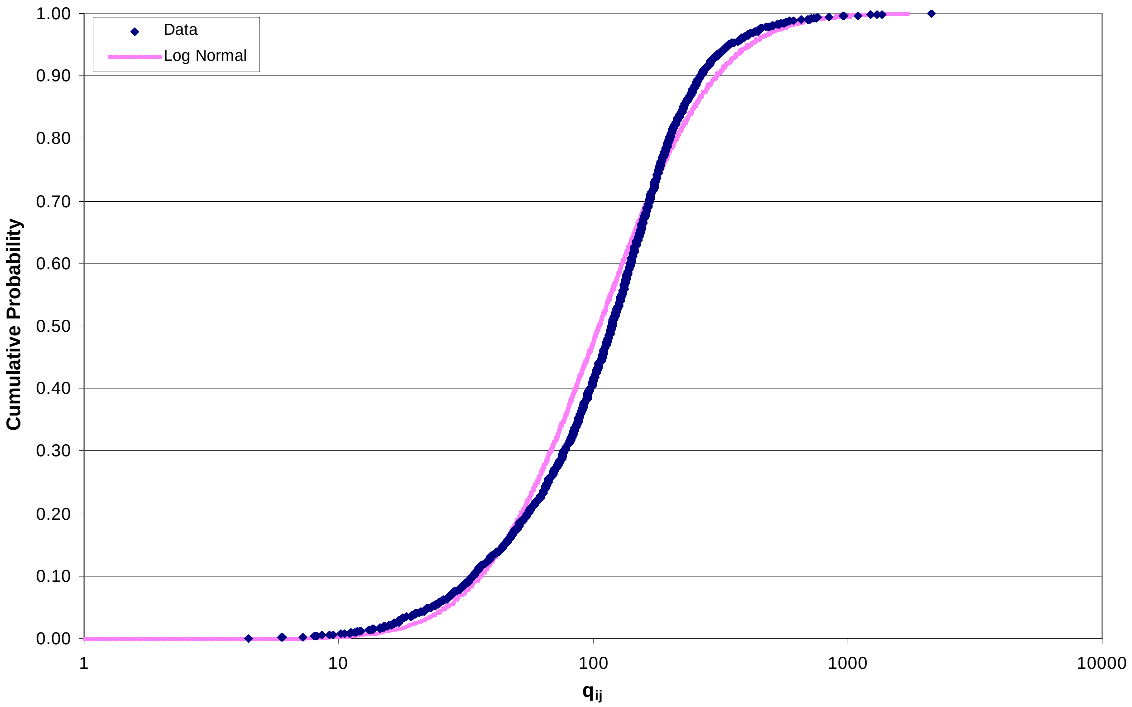

Figure 23: Fitted Log Normal and Empirical Cumulative Distributions for Region 1

$$f(x) = \frac{1}{x\sqrt{2\pi\sigma_Y^2}} \cdot e^{-\frac{1}{2}\left(\frac{\ln x - \mu_Y}{\sigma_Y}\right)^2}, \qquad (28)$$

where Y = ln x, since the shapes of the distributions above appear to take the shape of the lognormal distribution. It is important to note that Equation 28 is the pdf of the lognormal distribution, and the cdf must be found by numerical integration of the pdf.

For Region 1-5, the empirical cumulative distribution is a smooth line that compares well with the fitted lognormal distribution. For Region 6, there is more scatter in the data and the fit of the lognormal distribution is not as good as in the previous regions.

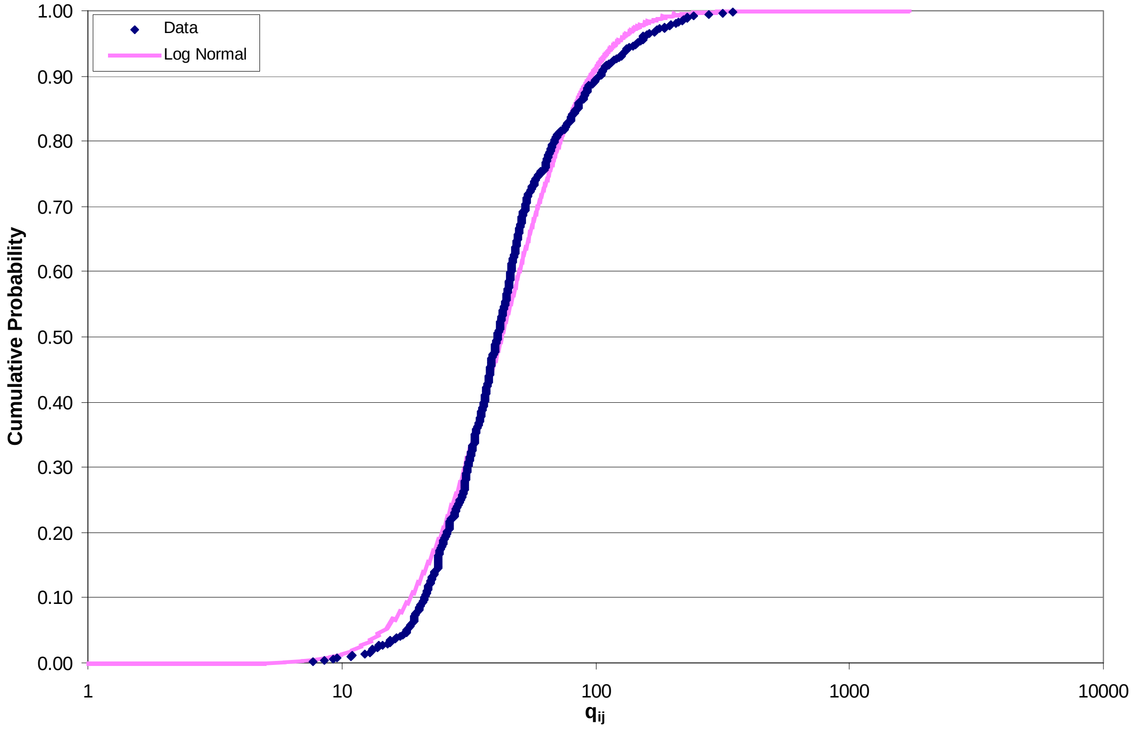

Figure 24: Fitted Log Normal and Empirical Cumulative Distributions for Region 6

When interpreting the plots, the 2, 10, and 50 year floods can be found for each region. These are compared with the estimates obtained from HEC-FFA, which uses the log Pearson type III, LP3, distribution. An example is given in Table 8. For each Region, five gages were selected for the comparison.

Table 12: Comparison of 2, 10, and 50 yr Floods for Region 1

Gage #	A (mi^2)	2 yr			10 yr			50 yr		
		$q_{ij*}A^m$ (cfs)	LP3 (cfs)	% Diff.	$q_{ij*}A^m$ (cfs)	LP3 (cfs)	% Diff.	$q_{ij*}A^m$ (cfs)	LP3 (cfs)	% Diff.
33430	13706	54013	55500	-2.7	120494	103000	17.0	156138	159000	-1.8
33360	8218	38884	51200	-24.1	86743	102000	-15.0	112404	160000	-29.7
33290	3779	23604	37100	-36.4	52658	67000	-21.4	68235	96000	-28.9
41790	763	8444	4750	77.8	18837	8700	116.5	24410	13000	87.8
33425	228	3886	3330	16.7	8669	5600	54.8	11234	7780	44.4

In general, for Regions 1-5, the empirical distribution yields comparable results to the log Pearson type III distribution for larger basin areas. Several smaller basins, less than 1000 square miles in area, shows differences of up to 300%. Generally, regions 1 and 5 underpredict the log Pearson type III distribution, while Regions 2, 3, and 4 overpredict. Region 6 yields comparable results for the 2 year flood; however, the larger return periods yield differences of 100% and above. Again, Region 6 shows the importance of regionalization. For regionalized watersheds, it is clear that small watersheds will yield large errors when using this method, though more research is necessary to determine exact size limit.

VIII CONCLUSIONS FROM THE STUDY

The following conclusions are presented based on the results presented herein.

1. The L-moment based method requires subjective judgment in regionalizing watersheds. Hence the results would not be unique and therefore unacceptable.
2. Of the three methods tested for regionalization, the Fuzzy clustering method and the artificial neural network based methods are easy to use. They give similar results. Either one may be used for regionalization.
3. The regions identified by these methods may not be statistically homogeneous. They may have to be revised to obtain homogeneous regions.

REFERENCES

Acreman, M. C. (1985). “Predicting the mean annual flood from basin characteristics in Scotland.” *Hydrol. Sci. J.*, 30(1), 37-49.

Acreman, M.C. (1987). “Regional Flood Frequency Analysis in the U.K.: Recent Research – New Ideas”. Institute of Hydrology, Wallingford, U.K.

Acreman, M. C., and Sinclair, C. D. (1986). “Classification of drainage basins according to their physical characteristics: An application of flood frequency analysis in Scotland.” *J.Hydrol.*, 84, 365-380.

Acreman, M.C., and S.E. Wiltshire (1987). “Identification of Regions for Regional Flood Frequncy Analysis”, EOS 68(48): 1262 (Abstract).

Aldenderfer, M. S., and Blashfield, R. K., (1984). *Cluster Analysis*. Sage Publications, Beverly Hills, CA.

Andenderfer, M. S., and Blashfield, R. K., (1984). *Cluster Analysis*. Sage Publications, Inc.

ASCE Task Committee on Application of Artificial Neural Networks in Hydrology (2000a). “Artificial Neural Networks in Hydrology I: Preliminary Concepts, ASCE Journal of Hydrologic Engineering, 5(2), 115-123.

ASCE Task Committee on Application of Artificial Neural Networks in Hydrology (2000b). “Artificial Neural Networks in Hydrology II: Hydrologic Applications, ASCE Journal of Hydrologic Engineering, 5(2), 124-137.

Backer, E. and Jain, A. K. (1981). “A clustering performance measure based on fuzzy set decomposition.” IEEE Transactions on Pattern Analysis and Machine Intelligence, 3(1), 66-75.

Bayliss, A.C., Jones, R.C., (1993). Peaks-Over-Threshold Flood Database: Summary Statistics and Seasonality. IH Report No. 121, Institute of Hydrology, Wallingford, UK.

Bezdek, J. C. (1973). “Fuzzy mathematics in pattern classification.” Ph.D. dissertation, Cornell University, Ithaca, NY.

Bezdek, J. C. (1981). “Pattern Recognition with Fuzzy Objective Function Algorithms.” Plenum, New York.

Bezdek, J. C. (1987). “Partition structures: A tutorial.” In: Bezdek, J. C. (Ed.), The analysis of fuzzy information. CRC Press, Boca Raton, FL.

Bezdek, J. C. and Pal, S. K. (1992). “Fuzzy models for pattern recognition.” IEEE Press. New York.

Bhaskar, N.R., C.A. O’Connor, H.A. Myers, and W.P. Puckett (1989). Regionalization of Flood Data using Probability Distribution and Their Parameter”, Res. Rept. No. 173, Kentucky Water Resources Research Inst., 136 p.

Bhaskar, N.R. and C.A. O’Connor (1989). “Comparison of Method of Residuals and Cluster Analysis for Flood Regionalization”, *Jour. Of Water Resources Planning and Management*, Vol. 115, No. 6, pp. 793-808.

Burn D.H. (1988). “Delineation of groups for Regional Flood Frequency Analysis”, *Jour. Of Hydrology*, Vol. 104, pp. 345-361, 1988.

Burn, D.H. (1989). Cluster analysis as applied to regional flood frequency. *Journal of Water Resources Planning and Management*, **115**, 567-82.

Burn, D.H. (1990a). “Evaluation of Regional Flood Frequency Analysis with a Region of Influence Approach”, *Water Res. Research*, Vol. 26, No. 10, pp. 2257-2265.

Burn, D.H. (1990b). "Appraisal of the 'Region of Influence' Approach to Flood Frequency Analysis", *Hydrological Sciences Journal*, Vol. 35, No. 2, pp. 149-165.

Burn, D.H. (1997). "Catchment Similarity for Regional Flood Frequency Analysis using Seasonality Measures", *Journal of Hydrology*, Vol. 202, No. 1-4, pp. 212-230, Dec. 1997.

Burn, D. H., and Goel, N. K. (2000). "The formation of groups for regional flood frequency analysis." *Hydrol. Sci. J.*, 45(1), 97-112, 2000.

Burn, D.H., Z. Zrinji, and M. Kowalchuk (1997). "Regionalization of Catchments for Regional Flood Frequency Analysis", *Jour. Of Hydrologic Engg*., Vol. 2, No. 2, pp. 76-82, April.

Cavadias, G.S. (1990). "The Canonical Correlation Approach to Regional Flood Estimation", *IAHS Publication No. 191*, p. 171-178.

Choquette, A.F. (1988). "Regionalization of Peak Discharges for Streams in Kentucky", Report 87-4209.

Condie, R. (1980). "The Three Parameter Lognormal Distribution Applied to Regional Flood Frequency Analysis by the Index Flood Method", Inland Waters Directorate, Ottawa, Ontario, Canada, Tech. Workshop Series No. 2, pp. 345-253, 1980.

Cosic, D. and Loncaric, S. (1996). New methods for cluster selection in unsupervised fuzzy clustering. Proceedings of the 41th Anniversary Conference KoREMA,vol. 4, pp. 1#3, KoREMA.

Curtis, G.W. (1987). "Technique for Estimating Flood-Peak Discharge and Frequencies on Rural Streams in Illinois USGS Water Resources Investigations", Report 87-4207, 86 p., Urbana, IL.

Dalrymple, T. (1960). "Flood Frequency Analysis", U.S. Geological Survey, Water Supply Paper 1543-A.

Dixon, W. J. (Editor), 1975. "BMDP Biomedical Computer Programs. University of California Press, Berkeley, California.

Dunn, J. C. (1974). "A fuzzy relative of the ISODATA process and its use in detecting compact, well-separated clusters." *Journal of Cybernetics*, 3(3), 32-57.

Eash, D.A. (1993). "Estimating Design-Flood Discharges for Streams in Iowa using Drainage-Basin and Channel-Geometry Characteristics", USGS Water Resources Investigation Report 93-4062, pp. 102 p.

Ernst, S., Rao, A.R., and Jeong, G.D. (2002). Results from L-Moment Based Method, Interim Report No. 2, Joint Transportation Research Program, Project No. C-36-62K, School of Civil Engineering, Purdue University, West Lafayette, IN 47906, pp. 191.

Everitt, Brian. S. *Cluster Analysis*. Third Edition, Halsted Press, New York, 1993.

Flippo, H.N. (1990). "Technique for Estimating Depths of 100-Year Floods in Pennsylvania", USGS Water Resources Report 86-4195, 20 p.

Gath, I. and Geva, A. B. (1989). "Unsupervised optimal fuzzy clustering." IEEE Transactions on pattern analysis and machine intelligence." 11(7), 773-781.

Glatfelter, Dale, (1984). "Techniques for Estimating Magnitudes and Frequency of Floods of Streams in Indiana", WRI 84-4134, USGS, Indianapolis, IN.

Gordon, A. (1999). *Classification*. London: Chapman and Hall/CRC Press.

Govindaraju, R.S. and A.R. Rao (eds.)(2000). "Artificial Neural Network Applications in Hydrology", Kluwer Pub., The Netherlands.

Gu, T., and Dubuisson, B. (1990). “Similarity of classes and fuzzy clustering.” Fuzzy sets and systems, 34, 213-221.

Guimaraes, W.B. and L.R. Bohmann (1992). “Techniques for Estimating Magnitude and Frequency of Floods in South Carolina”, USGS Water Resources Investigation Report 91-4157, 174 p. 1992.

Halkidi, M., Batistakis, Y. and Vazirgiannis, M. (2001). “On clustering validation techniques.” *Journal of Intelligent Information Systems*, Vol. 17, 2001, pp.107-145.

Hartigan, J. (1975). *Clustering Algorithms*. New York: Wiley.

Hosking, J. R. M. (1990). L-moments: analysis and estimation of distributions using linear combinations of order statistics. *Journal of the Royal Statistical Society, Series B,* **52**, 105-124.

Hosking, J.R.M., (1991). “FORTRAN routines for use with the method of L-moments, version 2, *Res. Rep. RC17097*, IBM Research, Yorktown Heights, N.Y.

Hosking, J.R.M., Wallis, J.R., (1993). Some statistics useful in regional frequency analysis, *Water Resources Research*, 29, 271-281.

Hosking, J.R.M. and J.R. Wallis (1997). “Regional Frequency Analysis”, Cambridge University Press, Cambridge CB2 2RU, U.K.

Huang, Z., (1997). “ A fast clustering algorithm to cluster very large categorical data sets in data mining.” DMKD.

Iblings, M.L. and Rao, A.R. (2003). “Use of Precipitation and Flow Data for Regionalization of Watersheds”, Interim Report No. 3, Joint Transportation Research Program, Project No. C-36-62K, School of Civil Engineering, Purdue University, West Lafayette, IN 47906, pp. 191.

Indiana Department of Natural Resources (INDR) (2001) Flood Frequency Analysis Input File.

Indiana Department of Natural Resources (INDR) (2001) Flood Frequency Analysis Output File.

Jain, A. K., and Dubes, R. C. (1988). *Algorithms for clustering data*. Englewood Cliff, NJ: Prentice Hall, Inc.

Kalkstein, L. S., and Corrigan, P. (1986). A synoptic climatological approach for geographical analysis: Assessment of sulfur dioxide concentrations. Ann. Assoc. Amer. Geogr., 76, 381-395.

Kalkstein, L. S., Tan, G., Skindlov, J. A. (1987). An evaluation of three clustering procedures for use in synoptic climatological classification." Journal of Climate and Applied Meteorology. 26, 717-730.

Kaufman, L. and Rousseeuw, P. (1990). *Finding groups in data: An Introduction to Cluster Analaysis*. New York: Wiley.

Koltun, G.F. and J.W. Roberts (1990). "Techniques for Estimating Flood-Peak Discharges of Rural Unregulated Streams in Ohio", Rept. No. FHWA/OH- 90/01, USGS Water Resources Investigation Rept. 89-4126, 74 p.

Lance, G. N., and Williams, W. T. (1966). "Computer programs for hierarchical polythetic classification ('similarity analysis')." *Comp. J.*, 9, 60-64.

Lara, O.G. (1987). "Method for Estimating the Magnitude and Frequency of Floods at Ungaged Sites on Unregulated Rural Streams in Iowa", USGS Water Resources Investigation Report 87-4132, 34 p.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. Proceedings of the 5th Berkeley Symposium 1, 281-297.

Mardia, K.V., (1972). Statistics of Directional Data. Academic Press, New York, NY.

Minns, A.W. (1995). "Applications of artificial neural networks in hydrology" paper presented at the European Geophysical Society General Assembly, Hamburg, 1995 and abstracted in the conference proceedings.

Mosley, M. P. (1981). “Delimitation of New Zealand hydrological regions.” *J. Hydro.*, 49, 173-192.

Nathan, R.J. and T.A. McMahon (1990). “Identification of Homogeneous Regions for the Purposes of Regionalization”, *Jour. Of Hydrology*, Vol. 121, pp. 217-238, Dec.

Ng, R., Han, J. (1994). “Efficient and effective clustering methods for spatial data mining”. Proceeding of the 20th VLDB conference, Santiago, Chile.

Nguyen, V.T.V, G. Pandey and H. Wang (1997). “Scaling Approach to Regional Estimation of Extreme Hydrologic Variables”, Proc. Of the 1997 Congress of the Canadian Society for Civil Engg., V.3, p. 81-90.

Ouarda, T.B.M.J., Boucher, G., Rasmussen, P.F. and B. Bobee (1997). “Regionalization of Floods by Canonical Correlation Analysis”, Proc. 1997 European Water Res. Assn. Conf., Balkema, Netherlands, P. 297-302.

Provoznik, M.K. and R.H. Hotchkiss (1998). “Analysis of Gauging Station Flood Frequency Estimates in Nebraska using L-Moments and Region of Influence Method”, Transportation Research Record, Issue No. 1647, pp. 53-60.

Potter, K. W., and Faulkner, E. B. (1987). “Catchment response time as a predictor of flood quantiles”. *Wat. Resour. Bull.*, 23(5), 857-861.

Rao, A.R. and K.H. Hamed (1997). “Regional Frequency Analysis of Wabash River Flood Data by L-Moments, ASCE Journal of Hydrologic Engineering, Vol. 2, No. 4, pp. 169-180.

Reich, B.M. (1988). "Flood Frequency Methods for Arizona Streams: State of the Art", Report No. FHWA-AZ88-801: ATARC/001, Federal Highway Administration, pp. 56.

Ribeiro-Correa, B., Cavadias, G. S., Clement, B. and Rousselle, J. (1995). "Identification of hydrological neighbourhoods using canonical correlation analysis." Journal of Hydrology. 173, 71-89.

Rouben, M. (1982). "Fuzzy clustering algorithms and their clustering validity." *European Journal of Operational Research.*" 10, 294-301.

Smith, J.A. (1989). "Regional Flood Frequency Analysis Using Extreme Order Statistics of the Annual Peak Record", *Water Resources Research*, Vol. 25, No. 2.

Sokal, R. R. and Sneath, P. H. A. (1963). "Principles of Numerical Taxonomy, W. H. Freeman and Co., San Francisco, California.

Srinivas, V.V. and A.R. Rao. (2002). Regionalization of Indiana Watersheds by Hybrid Cluster Analysis. Research submitted to the Joint Transportation Research Program, School of Civil Engineering, Purdue University. FWHA/IN/JTRP-2002/2, 112.

SSPS, Inc. (1988). SPSS-X User's Guide. McGraw-Hill, New York, NY.

Stedinger, J.R. and G.O. Tasker (1985). "Regional Hydrologic Analysis: 1. Ordinary, Weighted and Generalized Least Squares Compared", *Water Resources Research*, Vol. 21, No. 9, p. 1421-1432.

Tamura, S., Higuchi, S., Tanaka, K. (1971). Pattern classification based on fuzzy relations, *IEEE Trans. Syst. Man Cybern.* 1(1), 61-66.

Tasker, G. D. (1982)."Comparing methods of hydrologic regionalization." *Wat. Resour. Bull.*, 18(6), 965-970.

Tasker, G. D. (1980).“Hydrologic regression with weighted least squares.” *Water Resour. Res.*,16(6), 1107-1113.

Theodoridis, S., Koutroubas, K. (1999). *Pattern recognition*, Academic Press. New York, N. Y.

Trauwaert, E. (1985). “On the meaning of Dunn’s partition coefficient for fuzzy clusters.” International working paper. Vrije Universiteit Brussels.

Trauwaert, E. (1988). “On the meaning of Dunn’s partition coefficient for fuzzy clusters.” Fuzzy sets and systems, 25, 217-242.

U.S. Army Corps of Engineers, (1982). “Flood Flow Frequency Analysis Computer Program”, Users Manual, 723-X6-L7550.

U.S. Geological Survey (USGS), (1998). “Users Manual for Program PEAKFQ, Annual Flood Frequency Analysis using Bulletin 17B Guidelines”, Water Resources Investigation Report # DRAFT – Subject to Revision.

U.S. Water Resources Council, (1981). “Guidelines for Determining Flood Flow Frequency”, Bulletin 17B (revised), Hydrology Committee, *Water Resources Research Council*, Washington.

U.S. Water Resources Council, (1981). “Guidelines for Determining Flood Flow Frequency”, Bulletin 17B (revised), Hydrology Committee, *Water Resources Research Council*, Washington

Ward, J. H. (1963). “Hierarchical grouping to optimize an objective function.” *Journal of the American Statistical Association*, 58, 236-244.

Waylon, P. and M-K. Woo (1981). “Regionalization and Prediction of Annual Floods in the Fraser River Catchment”, British Columbia, *Water Res. Bull.*, Vol. 17, No. 4, pp. 655-661.

Webster, R. and Burrough, P. A. (1972). Computer-based soil mapping of small areas from sample data, II: Classification smoothing. J. Soil Science, 23(2), 222-234.

Whiltshire, S.E. (1986a). "Identification of Homogeneous Regions for Flood Frequency Analysis", *J. Hydro*., 84, 287-302.

Whiltsire, S.E. (1986b). "Regional Flood Frequency Analysis I: Homogeneity Statistics", *Hydro. Sci. J*., 31(3), 321-333.

Winkler, J. A. (1985). "Regionalization of the diurnal distribution of summertime heavy precipitation. *Preprints, Sixth Conference of Hydrometeorology*, American Meteorological Society, 9-16.

Willmott, C. J., and Vernon, M. T. (1980). Solar climates of the conterminous United States: A preliminary investigation. *Sol. Energy*, 24, 295-303.

Wiltshire, S.E. (1986c). "Regional Flood Frequency Analysis II: Multivariate Classification of Drainage Basins in Britain", *Hydrol. Sci. Jour.* 31(3): pp. 335-346.

Xie, L. X., and Beni, G. (1991). "A validity measure for fuzzy clustering." IEEE Transactions on pattern analysis and machine intelligence, 13(8), 841-847.

Yang, M. S. (1993). On a class of fuzzy classification maximum likelihood procedures. Fuzzy Sets and Systems, 57(3): 365-375.

Zrinji, Z. and D.H. Burn (1994). "Flood Frequency Analysis for Ungaged Sites Using a Region of Influence Approach", *Journal of Hydrology*, Vol. 153, No. 1-4, pp. 1-21.

Zrinji, Z. and D.H. Burn (1996). "Regional Flood Frequency with Hierarchical Region of Influence", *Jour. Of Water Res. Planning and Management*, Vol. 122, No. 4, pp. 245-252.

Zrinji, Z. and D.H. Burn (1997) . "Hydrologic Regionalization Using a Homogeneity Test", Proc. Of the Symposium on Engineering Hydrology, San Francisco, CA, pp. 631-646.