Frequency Analysis of Droughts Using Stochastic and Soft Computing Techniques

by
Sara Sadri

A thesis
presented to the University of Waterloo
in fulfilment of the
thesis requirement for the degree of

Doctor of Philosophy
in
Civil Engineering

Waterloo, Ontario, Canada, 2010

© Sara Sadri 2010

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

In the Canadian Prairies recurring droughts are one of the realities which can have significant economical, environmental, and social impacts. For example, droughts in 1997 and 2001 cost over $100 million on different sectors. Drought frequency analysis is a technique for analyzing how frequently a drought event of a given magnitude may be expected to occur. In this study the state of the science related to frequency analysis of droughts is reviewed and studied. The main contributions of this thesis include development of a model in Matlab which uses the qualities of Fuzzy C-Means (FCMs) clustering and corrects the formed regions to meet the criteria of effective hydrological regions. In FCM each site has a degree of membership in each of the clusters. The algorithm developed is flexible to get number of regions and return period as inputs and show the final corrected clusters as output for most case scenarios. While drought is considered a bivariate phenomena with two statistical variables of duration and severity to be analyzed simultaneously, an important step in this study is increasing the complexity of the initial model in Matlab to correct regions based on L-comoments statistics (as apposed to L-moments). Implementing a reasonably straightforward approach for bivariate drought frequency analysis using bivariate L-comoments and copula is another contribution of this study. Quantile es-

timation at ungauged sites for return periods of interest is studied by introducing two new classes of neural network and machine learning: Radial Basis Function (RBF) and Support Vector Machine Regression (SVM-R). These two techniques are selected based on their good reviews in literature in function estimation and nonparametric regression. The functionalities of RBF and SVM-R are compared with traditional nonlinear regression (NLR) method. As well, a nonlinear regression with regionalization method in which catchments are first regionalized using FCMs is applied and its results are compared with the other three models. Drought data from 36 natural catchments in the Canadian Prairies are used in this study. This study provides a methodology for bivariate drought frequency analysis that can be practiced in any part of the world.

Acknowledgements

My undisputed thanks go to my supervisor Professor Don H. Burn for his continuous efforts to provide me feedback and instructions and helping me to increase the depth of my knowledge on stochastic hydrology. It was an honour to be supervised by him and thanks for believing in me and my abilities.

I would like to acknowledge Professor Mahesh Pandey and Professor Fakhreddine Karrey. I gained valuable knowledge taking the courses they have been offering which contributed greatly in enriching the content of my Ph.D. thesis.

I owe thanks to Mr. Hossein Parsaei and Mr. Arash Abghari, graduate students from Departments of Systems Design and Electrical Engineering, respectively, for assisting me with neural networks analysis.

I would like to thank my friend and colleague Dr. Stefano Normani from the Department of Civil and Environmental Engineering for all his technical support on various levels and times, his patience and gentleness in providing me with helpful instructions. Thanks to my friend Dr. David Pritchard from the Department of Math and Combinatorics, for generously spending time to brainstorm with me on the logic of programming and developing algorithms.

My special thanks goes to the people who are sharing their house with me for providing me a peaceful, long-term, and safe accommodation during my stay in Waterloo. I am deeply grateful to Mrs. Donna Robinson and Professor Jim Robinson for everything; it has meant a lot to me.

A badge of honour for my mother, father, and my sister Vida for having faith in me and standing behind me during all times. They are my sunshine and the reason for me to smile every day.

Thanks to good friends: Dr. Katrin Höper, Professor Ebrahim Samei, and Mina Rohanizadegan. Also, thanks to Tim Hortons, Williams, and my coffee machine for being source of unlimited coffee any time.

And finally, thanks to the janitors of the University of Waterloo who have been cleaning my office regularly.

This research was funded by a grant to my supervisor from the Natural Sciences and Engineering Research Council (NSERC) and I am grateful for that as well.

I am very happy to have reached this milestone and for proving to myself, once more, that "no force can sustain itself against the full thrust of a determined human heart."

Table of Contents

List of Tables

List of Illustrations

List of Symbols

Greek

β	$l \times 1$ vector of parameters
β_k	$l \times 1$ regression coefficient in the presence of a weight factor w
ε	$N \times 1$ vector of random disturbances ($\varepsilon : \mathbf{N}(0, \sigma^2\mathbf{I})$)
θ	Parameter hidden in the generating function ϕ
$\Lambda_2^{\star(i)}$	L-covariation coefficient matrix for site i with record length n_i, $i = 1, ..., N$
μ_j	Center of basis function
ν_j	Bandwidth parameter and controls the smoothness of the interpolating function
τ	Degree of correlation between two variables in copulas
$\phi(\cdot)$	Copula generating function
Φ_j	Gaussian basis function
W	Matrix of second-layer weights to be estimated
w_0	Bias parameter
w_j	Weighting vector

Latin

a	Multiplicative error term
$b_k(i)$	Degree of belonging of site i in the kth cluster
C	Copula function
c_k	The centroid or mean point of all points in cluster k
D	Length of consecutive negative deviations followed by positive deviations is defined as negative run-length [T]
$d(\text{centre}_k, i)$	Distance of site i to the centroid of that cluster k
$D(i)$	Discordancy of site i
D_v	Total intra-cluster variance (distance or squared error function)

$F_X(x)$	Non-exceedance probability of event X which can be no greater than x
F_{XY}	Joint cumulative distribution function
H_1	Homogeneity
J	Size of a region (cluster) or sum of $\sum n_i$ in region k
K	Total number of clusters
k	Cluster number
M	Ratio of the negative run-sum and the negative run-length [L^3/T]
m	Number of time intervals where $X_{C_i} \geq X_i$
N	Total number of sites used in modelling
n_i	Size or number of samples in site i
n	Number of catchment characteristics
$\bar{q}$	Mean of at-site estimation
$\hat{q}_i$	Quantile estimation obtained from modelling
q_i	At-site estimation for site i
S	The negative run-sum or the cumulative volume of water deficit [L^3]
S_k	Set of points in the kth cluster
S_T	Model quantile
$\bar{t}$	Group average L-moment ratios, with sites weighted proportionally to their record length
T	Return period and used to express the result from frequency analysis [years]
$t^{(i)}$	Sample L-CV at site i
u	Specific value of U
u_i	A vector containing the L-CV, L-skewness and L-kurtosis values for a site i
$U_k(i)$	Normalized coefficient of site i in the kth cluster ($\in$ [01])
V	Weighted standard deviation of the at-site sample $L-CV$
v	Specific value of V
$w_{0,k}$	The normalized weight value of input target vector into cluster k
$\mathbf{w}$	Normal to the hyperplane
$\mathbf{x}_0$	$1 \times l$ input target vector
x	n-dimensional input vector
X_{C_i}	Truncation level for the ith time interval [L^3]
X_i	Mean flow of the ith time interval [L^3/T]
x_i	Model characteristics
x_j	Standardized value for attribute j from site i
$\mathbf{x}$	$N \times l$ matrix of regressors
Y	Matrix of output values
y	$1 \times N$ vector of natural log of observations
y_j	Output vector

List of Abbreviations

ANFIS	Adaptive Neuro-Fuzzy Inference Systems
ANNs	Artifical Neural Networks
CDF	Cumulative Distribution Function
FCM	Fuzzy C-Means
FES	Fuzzy Expert Systems
FF	Feed Forward
IID	Independent and Identically Distributed
NLR	Non-Linear Regression
NLR-R	Non-Linear Regression with Regionalization
PCA	Principal Component Analysis
RBF	Radial Basis Function
ROI	Region of Influence
SVMs	Support Vector Machines
SVR	Support Vector Machine - Regression

Chapter 1

Introduction

Frequency analysis is a form of hazard or risk assessment based on the fact that in any given period of time, certain events and combinations occur with varying frequencies. Moreover, there is a characteristic distribution of events that is roughly the same for most samples of that event. Frequency analysis has been applied in many different areas of science.

In water resources management and hydrology, frequency analysis involves estimating the expected number of occurrences of a repeating extreme event per unit time. For example, frequency analysis can study the likelihood of recurring severe droughts, floods, rainfalls, and low flows. As a matter of convenience, the frequency of longer duration events such as extreme hydrological events tend to be described by event period (or return period) rather than frequency.

The accuracy of frequency analysis methods in stochastic hydrology has profound significance for economic investment (*Kidson and Richards*, 2005). In this thesis the state of the art of frequency analysis of extreme hydrological events, mostly

droughts, is reviewed and studied. The organization of this thesis is as follows: The following sections of Chapter 1 cover the importance of drought frequency analysis and the objectives of this research followed by some background knowledge useful in the domain of drought frequency analysis. Chapter 2 reviews the various work and research which has been done in the area of drought frequency analysis. Chapter 3 looks into regionalization and bivariate test of homogeneity and discordancy for hydrological regions for the purpose of frequency analysis. Frequency analysis of bivariate droughts using a copula is studied in Chapter 4. Issues such as frequency analysis of ungauged sites and nonparametric analysis of drought data using different statistical approaches and neural networks are addressed in Chapter 5. Chapter 6 summarizes the entire research and provides suggestions and recommendations for future research in this area.

1.1 Problem Description

One of the realities of today's world is that many people live in regions affected by endemic drought while others face droughts on an irregular basis and therefore may be less prepared for times of water scarcity. Reccurring droughts are one of the main natural hazards and can have significant environmental and economic impacts. Compared with other natural hazards, such as floods and hurricanes, the spatial extent of droughts is usually much greater, as well the impacts of droughts are generally non-structural and difficult to quantify (*Obasi*, 1994). Also the development of droughts is slow and it is very difficult to identify the moment in which they start and finish (*Burton et al.*, 1978). From this view-point, droughts are the best example of "penetrating" natural hazards since they are usually recognized when human activities and the environment are affected.

Droughts are very complex phenomena both in terms of definition and causes (*Vicente-Serrano and Lopez-Moreno*, 2005). Nevertheless droughts are usually related to a long and sustained period in which water availability becomes scarce mainly due to an abnormal decrease in precipitation. Hydrological drought is defined as a deficit of water supply in time, in area, or both, with deficit magnitude and deficit duration taken into account (*Yevjevich*, 1967).

In the Canadian prairies, although droughts are not generally associated with catastrophic injury or death, droughts have had disastrous impacts on Canada's grain industry and on environmental and socio-economic conditions. According to the Canadian government's Discussion Paper on Drought in Western Canada (*Khandekar*, 2002), the 1997 drought in the western Prairies of Canada cost over $100 million in additional power generation costs, $20 million in unanticipated fire-fighting charges and $10 million in emergency federal and provincial drought programs, in addition to losses in tourism and costs of additional water treatment. The recent Prairie drought of 2001 was estimated to be the third most severe drought in the last 50 years and produced an estimated shortfall of $4 billion in grain revenues (*Leavitt and Chen*, 2000). The relevance of the past droughts with the future droughts is analyzed in the domain of drought frequency analysis. The basic assumption of most methods of frequency analysis is that the events observed in the past are likely to be typical of what may be expected in the future. The estimation of how often a specified event will occur is of great importance. Planning of weather related emergencies, reservoir management, pollution control, and insurance risk calculations all rely on knowledge of the frequency of drought events. Despite the high economic and social costs of droughts, and the potential savings that could be derived from better drought frequency analysis, there are few avenues presently available to estimate their frequency. Therefore, research on estimation of drought frequency, duration and

severity will provide a rigorous basis for future agricultural, insurance and resource management decisions (*Leavitt and Chen*, 2000).

1.2 Objectives

Based on the existing need for extensive research on extreme hydrological events using advanced techniques, the objectives of this thesis are developed. These objectives are designed for improvement of existing drought quantile estimation approaches reviewed in Chapter 2. The thesis is focused on developing methods in the following areas:

1. Bivariate drought frequency using copula: the two main characteristics of droughts are duration (t) and severity (m^3). For water resource planning and management the joint distribution of drought variables (i.e., bivariate analysis) can yield much more sophisticated results (*Gonzalez and Valdes*, 2003). When parametric frequency analysis is applied, two characteristics, duration and severity, may not have the same marginal distributions. By using the copula approach, each component is allowed to have its own different marginal distribution. A copula is a function which links a multivariate distribution to the one-dimensional marginal distributions. In this study the bivariate probability distribution of drought characteristics will be studied by using a suitable copula for describing the dependence between two drought characteristics.

2. Pooling groups and univariate and bivariate tests of homogeneity and discordancy-An L-moment approach: in most cases of extreme event frequency analysis, the absence of lengthy records, or any record, interferes with the reliability of statistical frequency analysis. To address this issue, the rationale of using "pooled" or "regionalized" information from multiple sites has been applied. The most common approach for pooling sites has been based on the index-event procedure

which assumes one frequency distribution for a homogeneous region (*Hosking and Wallis*, 1997). Since the index-event procedure uses more information than "at-site" analysis (which uses only data from a single catchment), there is potential for greater accuracy in the final quantile estimates. From another side, regionalization has the advantage that an ungauged site, a site with available attributes but missing or lacking in data for the variable of interest, can still be assigned to a region and extreme event quantiles can be estimated. Since the idea of regionalization has been developed, different approaches have been used for forming hydrological regions. The delineation of regions may be a complicated task. However, it is normally agreed that the formed groups have to meet the criteria of homogeneity and lack of discordancy suggested by *Hosking and Wallis* (1993) and sufficient size suggested by *Reed and Robson* (1999). Although regionalization is a very important topic in extreme hydrological event frequency analysis, there seems to be no rigorous and fast approach for this. The problem becomes more complicated since there has been very little work on tests of bivariate homogeneity and discordancy when dealing with bivariate frequency analysis approach. This study develops a comprehensive algorithm for regionalization in both univariate and bivariate analysis. A Matlab code is developed to use site characteristics and an intelligent clustering approach, called Fuzzy C-Means (FCM), to form the initial regions (clusters) and adjusts the initial formed clusters based on partial or fuzzy membership of each site to other clusters to form the final clusters that meet the criteria of homogeneity, lack of discordancy, and sufficient size.

3. Drought frequency analysis at ungauged sites using neural networks and statistics: for many engineering projects, reliable drought quantile estimation for a desired return periods is essential. The problem is that, in many cases, fre-

quency analysis needs to deal with scenarios that do not have any extreme event data at all (ungauged sites). Nonlinear regression is one of the common approaches used to find quantiles as a function of site physiographic and other characteristics (*Shu and Ouarda*, 2008). Most of these regression approaches involve a parametric regression. From another side, during the past decades there has been an emergence of the application of neural networks and other artificial intelligence approaches in function estimation and regression analysis in different areas of engineering. These relatively new techniques can provide an attractive alternative to the traditional statistical models. Artificial neural networks (ANNs) have been introduced in the domain of regional flood frequency analysis by *Shu and Burn* (2004a). For application of ANNs in the area of regional drought frequency analysis, there appears to be no work recorded in the literature. To test the functionality of nonlinear regression methods and ANNs, four methods of nonlinear regression, nonlinear regression with regionalization, Radial Basis Functions (RBFs), and Support Vector Machines (SVMs) are used for quantile estimation of droughts, specifically applied to drought records in Canadian Prairies. The first two methods are very common approaches in statistical analysis. The regionalization step in nonlinear regression with regionalization is done using the FCM clustering approach discussed in the previous objective. The two latter approaches are two strong tools, applied in other areas of science, being used here in drought quantile estimation. The four approaches are compared and analyzed.

This research is to create a comprehensive approach for analyzing the probabilities of extreme hydrological events applied to droughts. The approach will consist of a collection of procedures. It is hoped that this research can lead to the development of

approaches that can be used to estimate the probability of an extreme hydrological event at any location of interest.

1.3 Background

Storms, floods, and droughts are examples of extreme hydrological events which sometimes cause severe damage to the environment. In order to analyze the risk of occurrence of very severe events, the science of frequency analysis found its way to hydrology. According to *Chow et al.* (1988) "the objective of frequency analysis of hydrologic data is to make sense of the magnitude of extreme events and their frequency of occurrence through the use of science of probability." Major research on frequency analysis in the area of hydrology are based on the assumptions that the hydrologic data analyzed are to be Independent and Identically Distributed (IID) (*Chow et al.*, 1988). The assumption of IID data is often satisfied by selecting the maxima or minima of the variable being analyzed (e.g., the annual maximum discharge) with the expectations that successive observations of this variable will be independent. The probability $F_X(x) = Pr(X \leq x)$ of the event $X \leq x$ in any observation is called the non-exceedance probability. The non-exceedance probability $F_X(x)$ is in fact the Cumulative Distribution Function (CDF) for the actual value of X is at most x. A drought quantile is defined as the value of a drought variable with non-exceedance probability $F_X(x)$. The probability of occurrence of an event in any observation is related to the inverse of its return period (*Yue and Rasmussen*, 2002):

$$Pr(X \leq x) = 1 - \frac{1}{T} \tag{1.1}$$

where T is the return period and used to express the result from frequency analysis [years].

Another common term is the "recurrence interval" τ which is the time between occurrences of $X \leq x$. The return period of the event $X \leq x$ is the expected value of τ, $E(\tau)$, and is the average or the most probable value measured over a very large number of occurrences

$$E(\tau) = T = \frac{1}{1 - Pr(X \leq x)} \tag{1.2}$$

For example, when studying frequency analysis of droughts, if the time between the first negative exceedance and the last negative exceedance of n annual minimum river flow is 50 years and in total there are $n = 5$ negative exceedances, the average time between exceedances, or the return period, is approximately $\bar{\tau} = 50/5 = 10.0$ years. The probability of minimum discharge in this example is $Pr(X \leq x) = 1/\bar{\tau} = 1/10 = 0.1$. More clarification can be found in *Chow et al.* (1988).

The exceedance probability that x will be equalled or exceeded is given by (*Chow et al.*, 1988):

$$F'_X(X) = 1 - F_X(x) \tag{1.3}$$

1.3.1 Drought Definition

Yevjevich (1967) "defined a hydrologic drought as the deficiency in water supply on the earth's surface and used in runs as the basic concept for definition of droughts (*Tase*, 1976)." A smilar definition by *Yevjevich* (1967) is used for defining flood events which is named the "Theory of Runs". Theory of runs is a useful theory for defining both floods and droughts. Based on this definition drought is defined on the basis of differences between the processes of water supply and water demand. Drought occurs when the magnitude of a discrete series of variable X (e.g., river flow) that occurs at a given time, is smaller than some predefined arbitrary level. The demand time series is called "truncation level" and its value X_T may be defined based on

single-purpose water use for agriculture, for continuous irrigation, hydropower, water supply, low flow augmentation for quality control or a combination of various uses (*Yujica*, 1975). The periodicity of droughts can vary from a month to multi years which makes the analysis of droughts somehow difficult, therefore, based on the study various time intervals of monthly, seasonally, or annually can be selected. Also, due to seasonal variation of the streamflow, use of a variable truncation level (Figure 1.1) was suggested in *Kjeldsen et al.* (1999). In this thesis for discrete times series of streamflow, a selected arbitrary monthly variable of truncation level is assumed to represent water demand and is calculated as the average value of each month's drought severities. Based on the theory of runs, three main drought characteristics

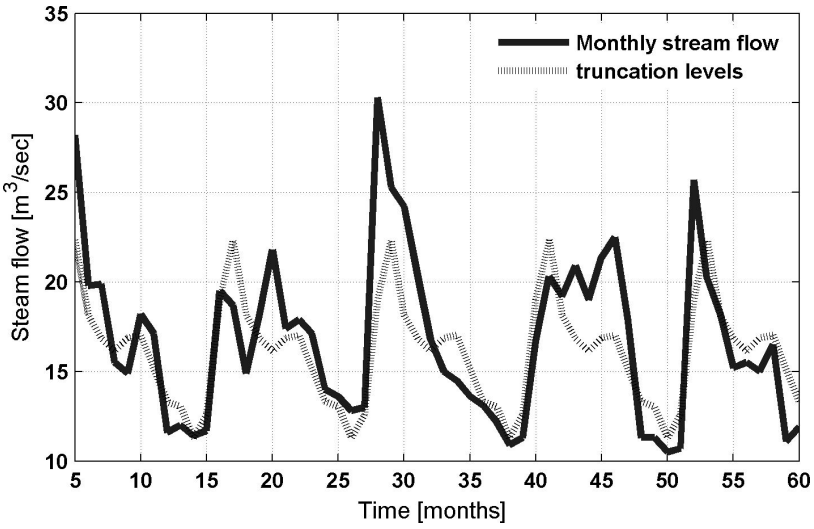


Figure 1.1: Varying truncation level

can be extracted:

1. Duration D: the length of consecutive negative deviations followed by positive deviations is defined as negative run-length; a consecutive period of time comprising the drought.

2. Severity S: the sum or integral of all negative deviations is defined as the negative run-sum or the cumulative volume of water deficit quantified as:

$$S = \sum_{i=1}^{m} (X_{C_i} - X_i) \tag{1.4}$$

where X_i is the mean flow of the ith time interval [L^3/T]; X_{C_i} is the truncation level for the ith time interval [L^3/T]; and m is the number of time intervals where $X_{C_i} \geq X_i$.

3. Magnitude M: the ratio of the negative run-sum and the negative run-length is defined as the negative run-intensity (*Yevjevich*, 1967):

$$M = \frac{S}{D} \tag{1.5}$$

Yevjevich (1967) found that the run-length properties are free of the underlying distribution of input processes. Theory of runs has been successfully applied in characterization of drought and further statistical analysis of droughts.

1.3.1.1 Method of L-moments

The method of L-moments has been widely used in regional frequency analysis to fit a distribution to a set of variables, either regional or single site. L-moments statistics are analogous to the conventional moments (mean, standard deviation, skewness, kurtosis, etc.) and were developed by *Hosking and Wallis* (1993). They have been used in a wide range of hydrological areas since they represent simple and reasonably efficient estimators for characteristics of hydrologic data. Consider a sample statistics of size n_i from a single monitoring site arranged in an ascending order, so that the ordered sample is: $x_{1:n} \leq x_{2:n} \leq ... \leq x_{n:n}$. A statistical view to an ordered

sample of certain linear combinations of the elements is that it contains information about the location, scale, and shape of the distribution from which the sample was drawn. L-moments are defined to be the expected values of these linear combinations. For convenience, the expected values of linear combinations are multiplied by scalar constants. The "linear" combinations of order statistics is emphasized in "L" in L-moments. The procedure to calculate L-moments is described below:

1. Calculate the mean or the average of the variable vector X. The mean is the first L-moment and is shown as (*Hosking and Wallis*, 1997):

$$\lambda_1 = E[x] = \mu = \beta_0 \tag{1.6}$$

2. Calculate the probability weighted moments (i.e., β_1, β_2, and β_3) by first arranging the data vector X in an ascending order and then:

$$\beta_1 = \frac{1}{n}\sum_{j=2}^{n}\frac{(j-1)}{(n-1)}x_{j:n} \tag{1.7}$$

$$\beta_2 = \frac{1}{n}\sum_{j=3}^{n}\frac{(j-1)(j-2)}{(n-1)(n-2)}x_{j:n} \tag{1.8}$$

$$\beta_3 = \frac{1}{n}\sum_{j=4}^{n}\frac{(j-1)(j-2)(j-3)}{(n-1)(n-2)(n-3)}x_{j:n} \tag{1.9}$$

3. Calculate the the second L-moment λ_2 as well as scale measures L-moments λ_3 and λ_4:

$$\lambda_2 = 2(\beta_1) - (\beta_0) \tag{1.10}$$

$$\lambda_3 = 6(\beta_2) - 6(\beta_1) + (\beta_0) \tag{1.11}$$

$$\lambda_4 = 20(\beta_3) - 30(\beta_2) + 12(\beta_1) - (\beta_0) \tag{1.12}$$

4. The second, third and fourth L-moment ratios or $L-CV$ (τ), L-skewness ratio (τ_3) and and L-kurtosis ratios (τ_4) are:

$$\tau = \frac{\lambda_2}{\lambda_1} \tag{1.13}$$

$$\tau_3 = \frac{\lambda_3}{\lambda_2} \tag{1.14}$$

$$\tau_4 = \frac{\lambda_4}{\lambda_2} \tag{1.15}$$

Normally the higher L-moments are not needed in frequency analysis but one can see how to calculate the higher moments by finding the values of β and λ. In general:

$$\beta_r = \frac{1}{n} \sum_{j=r+1}^{n} \frac{(j-1)(j-2)...(j-r)}{(n-1)(n-2)...(n-r)} x_{j:n} \tag{1.16}$$

$$\lambda_{r+1} = n^{-1} \sum_{j=1}^{n} P^{\star}_{r,k} \beta_k \tag{1.17}$$

where coefficients $P^{\star}_{r,k}$ are defined as:

$$P^{\star}_{r,k} = \frac{(-1)^{(r-k)}(r+k)!}{(k!)^2(r-k)!} \tag{1.18}$$

L-moment ratios are achieved by dividing the higher-order L-moments by the scale measure λ_2:

$$\tau_r = \frac{\lambda_r}{\lambda_2}, \qquad r = 3, 4, ... \tag{1.19}$$

In practice, the advantages of using L-moments over ordinary moments are:

- small bias and variance, especially in comparison with the method of moments (*Hosking*, 1990);
- less sensitive to outliers (*Vogel and Fennessey*, 1993);

- better identification of the parent distribution that generated a particular data sample (*Hosking*, 1990); and
- better identification of distributions of highly skewed data of the L-moments diagrams over the conventional moments diagrams (*Vogel and Fennessey*, 1993).

In *Hosking and Wallis* (1993), three statistics which are useful in regional frequency analysis were used: (1) a discordancy measure for identifying unusual sites in a region, (2) a heterogeneity measure for assessing whether a proposed region is homogeneous, and (3) a goodness of fit measure for assessing whether a candidate distribution provides an adequate fit to the data (*Hosking and Wallis*, 1993).

1.3.1.2 Test of discordancy

One of the very important stages in frequency analysis is screening the data so that gross errors and inconsistencies can be eliminated. For screening the data the discordancy measure $D(I)$ applies. The discordancy measure identifies unusual sites; those sites whose at-site sample L-moments are markedly different from those of the other sites in the data set. Discordancy is measured in terms of the L-moments of the sites' data (*Hosking and Wallis*, 1993). There is not an easy way to choose a single value of $D(I)$ that can be used as a criterion for deciding whether a site is unusual. For large regions *Hosking and Wallis* (1993) suggested $D(I) \geq 3$ as a criterion for declaring a site to be unusual or discordant.

1.3.1.3 Test of regional homogeneity

In order to determine whether the data at the different sites pooled together can be considered to be from a common regional distribution, a validation test of homogeneity has to be performed (*Hosking and Wallis*, 1997). The major assumption in a homogeneous region is that the sites' frequency distributions are identical apart

from a at-site scale factor which is the mean of the at-site data. Calculation of the homogeneity requires comparison between weighted standard deviation of the at-site sample L-CV of the sites in the region formed with statistics of a large number of simulated regions. More information on this calculation is presented in Chapter 3.

1.3.1.4 Goodness-of-fit

Assuming that the region formed is acceptably close to homogeneous, goodness of fit is a test of how well a given distribution fits the data. The distribution being tested will have location and scale parameters which can be chosen to match the regional average mean and L-CV. The goodness of fit will be judged by how well the L-skewness and L-kurtosis of the fitted distributions match the regional average L-skewness and L-kurtosis of the observed data (*Hosking and Wallis*, 1993).

1.3.1.5 Revisions to regions

Regionlaization methods enable us to define the initial groups of catchments (regions). However, it is often found that the resulting groups need to be revised due to not meeting all requirements (lack of discordancy, homogeneity, and size) for an effective region (*Hosking and Wallis*, 1993). According to *Burn and Goel* (2000) revisions to initial regions is a heuristic process in the sense that there is no set way for how to move from one stage of the process to the next. The goals of the regional revision process are to increase the homogeneity of the regions, and to ensure each region is of a sufficient size (*Burn and Goel*, 2000). Although after all considerations a region may be moderately heterogeneous, regional analysis will still yield much more accurate quantile estimates with lower standard errors than an at-site analysis (*Hosking and Wallis*, 1997; *Haan*, 2002).

Chapter 2

Literature Review

This chapter aims to review the critical points of current knowledge on drought frequency analysis. The discussion begins with a review of existing drought frequency analysis mechanisms followed by a review of pooled frequency analysis and frequency analysis of ungauged sites. Other approaches to frequency analysis such as non-stationarity of droughts and application of soft computing techniques in frequency analysis will also be reviewed.

Most literature available in the context of frequency analysis in hydrology has been written for flood flows. Some of those techniques are applicable to a wide range of cases in drought frequency analysis (*Haan*, 2002). It should also be noted that droughts have a different definition from that of low flows and thus the literature written on low flow frequency analysis is not a focus in this study. Low flow is a seasonal phenomenon and is an integral component of a flow regime in any river. Droughts, on the other hand, are understood as a penetrating event due to less than normal precipitation over any period of time (*Smakhtin*, 2001).

2.1 Path to Drought Frequency Analysis

Literature available on frequency analysis of droughts addresses a form of statistical modelling in which a set of mathematical equations describes the behavior of droughts in terms of random variables and their associated probability distributions. Drought frequency analysis can be a complicated task since it requires a series of decisions and assumptions. A summary of these considerations is illustrated in Figure **??**. The literature available in the domain of drought frequency analysis is the matter of different assumptions and choices in every step of Figure **??** and hence different paths taken by the researchers for modelling of droughts. Having different options at each step of drought frequency analysis makes modelling of drought frequency flexible, and at the same time, complicated. These choices are categorized in different levels of assumptions.

The following sections provide information and review the literature of each of the blocks in Figure **??**.

2.1.1 At-site frequency analysis

At-site analysis uses only the data from a single site. The first objective definition of droughts given by *Yevjevich* (1967) on the basis of runs theory was implemented with a single site frequency analysis. Although accurate estimation of drought frequency at sites with fairly long records is not impossible, at-site frequency analysis of sites with short time series data records, and ungauged sites (sites with no statistical records) is impossible. Since droughts can last several months or years, the historical record of one site is often too short to fully characterize droughts stochastically (*Kim et al.*, 2006). The reliable estimation of droughts requires a length of data record that is often not available. Besides, drought analysis based on data collected for a

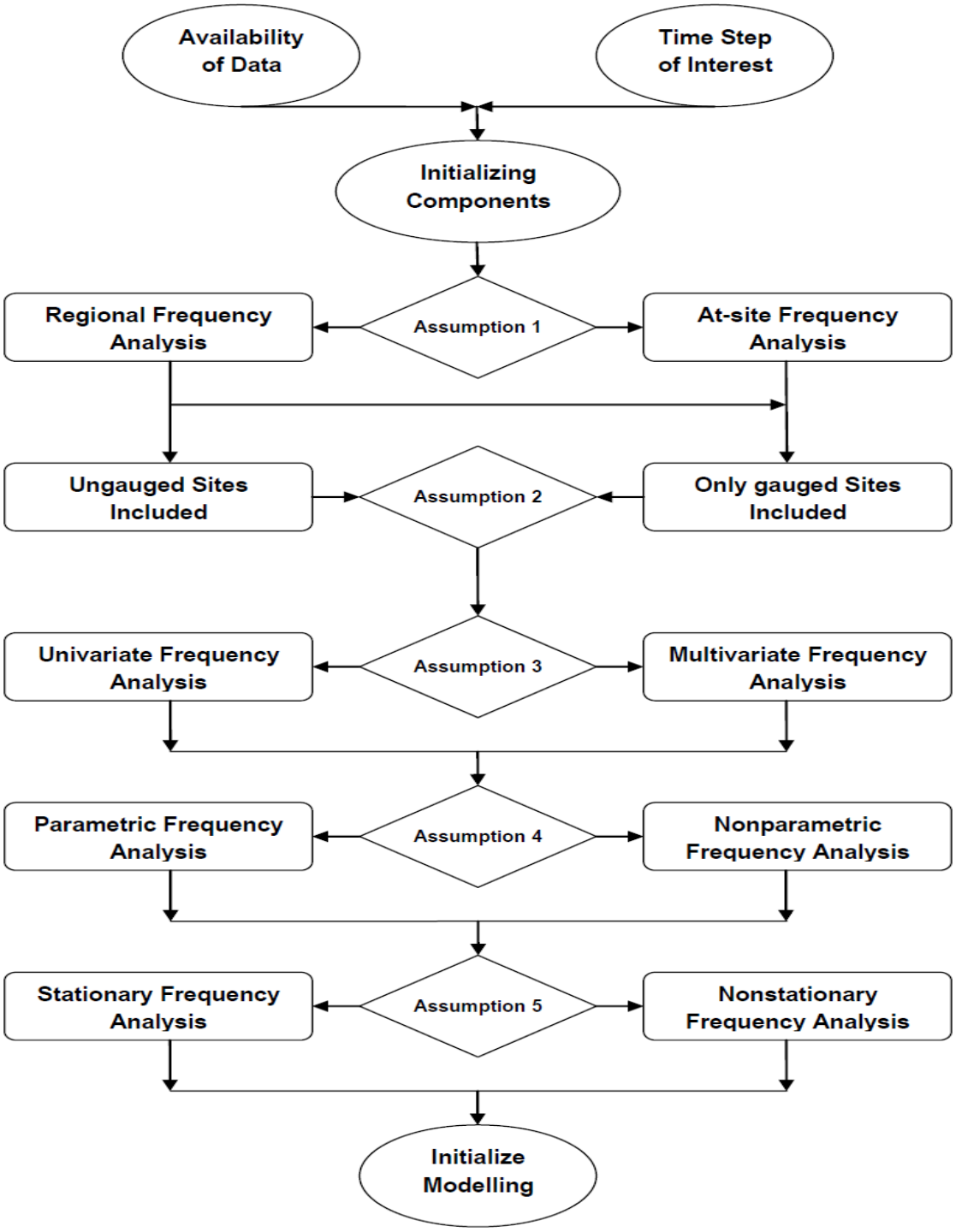


Figure 2.1: Decisions and assumptions needed to be made before initializing frequency modelling of droughts

single site brings wide sampling variations. *Hosking and Wallis* (1997) suggested by using regional analysis, such variations are expected to diminish by exploiting all the available data at multiples sites. In short, to reduce the vulnerability of agricultural production and development of large-scale multi-purpose water supply systems, at-site drought analysis is inadequate and a more comprehensive analysis at a regional scale is required (*Rossi et al.*, 1992).

2.1.2 Regional frequency analysis for droughts

It has long been accepted by many researchers that frequency analysis based on data collected from similar sites defined as regional (pooled) data is usually preferable to that developed for a single site (*Hosking and Wallis*, 1997). Therefore, regionalization is finding its importance through many researchers (*Lettenmaier et al.*, 1987; *Hosking*, 1990). Regionalization of similar catchments is based on the idea that catchments with similar climate, geology, topography, vegetation, and soils would normally have similar streamflow responses (*Smakhtin*, 2001). From a statistical point of view, a "region" is a group of sites where each site is assumed to have its data drawn from the same frequency distribution (*Hosking and Wallis*, 1997). Therefore, there is potential for greater accuracy in the final quantile estimates.

The first study concerning regional droughts was performed by *Tase* (1976) who succeeded in determining experimentally the area covered by a drought inside a fixed region, the total water deficit below the demand level, and the maximum drought intensity (*Gonzalez and Valdes*, 2003; *Tase*, 1976).

Regionalization comes at the cost of requiring a delineation of groups that are homogeneous. There seems to be no uniquely objective approach to the delineation of homogeneous regions. This is because grouping the regions should be based on the similarities in the characteristics of extreme events to be studied at different

gauging stations, but there have been controversies in defining the term "similarity" itself. Even if there exists a uniquely defined measure for similar catchments, it is not possible to use that measure for all different case studies. More traditional methods for classification of catchment are based on geographic, administrative, or physiographic boundaries (*Smakhtin*, 2001), or based on standardized flow characteristics estimated from the available observed or simulated streamflow records (*Midgley et al.*, 1994), or from maps and hydro-meteorological data (rainfall, evaporation) (*Hayes*, 1992).

Clausen and Pearson (1995) presented regional frequency analysis of annual maximum streamflow drought by using three geographical regions with different climate and physical properties in New Zealand. The annual maximum droughts were identified in terms of severity with two levels of truncation level representing the mean and 75% of the mean. Among other occasionally used methods of delineation of pooling groups *Gingras and Adamowski* (1993) and *Hayes* (1992) applied the residual analysis method. In this approach the residual pattern from a linear regression of a given design extreme event for the entire study area is examined and regions are then delineated on the basis of geographic proximity of the positive and negative residuals. Delineation of regions may be accomplished using convenient boundaries based on geographic, administrative, or physiographic considerations. However, the regions that result from using such an approach may not always appear to be "sufficiently" homogenous (*Groupe de recherche en hydrologie statistique*, 1996). *Midgley et al.* (1994) classified catchments based on standardized flow characteristics estimated from the available observed or simulated streamflow records. Regions can be delineated from maps and hydro-meteorological data such as rainfall and evaporation (*Hayes*, 1992). *Acreman and Wiltshire* (1989) used a pooling approach without fixed groups which was later developed further by *Burn* (1990a,b) into the Region of Influence (ROI) focused pooling method. *Burn and Goel* (2000) used K-means algorithm as a cluster-

ing technique for identifying groups for regional flood frequency analysis. The groups found using the clustering algorithm are subsequently revised to improve the regional characteristics. Another method of forming homogeneous regions has been canonical correlation analysis (*Ribeiro-Correa et al.*, 1995; *Ouarda et al.*, 2001).

Shu and Burn (2004b) ran an experiment on flood data and delineating homogeneous pooling groups using method of Fuzzy Expert Systems (FES) to derive an objective similarity measure between catchments. *Shu and Ouarda* (2008) used Adaptive Neuro-Fuzzy Inference Systems (ANFIS) as a mechanism for identifying the hydrological regions by generating knowledge from hydrometric station network in southern Quebec. This method requires an identification of parameters of the subtractive clustering algorithm as the clustering radius is the most important parameter that needs to be specified and is to be optimally determined though a trial and error procedure. Although significant progress has been made in recent years in regionalization, such as the ROI scheme as probably the most noteworthy one, difficulties still exist especially in defining the similarity measures and adjustment of regions. Since two of the more commonly used methods for homogeneous pooling delineation are region of influence and cluster analysis, they are reviewed here.

2.1.2.1 Region of influence (ROI)

Region of influence ROI was initially developed by (*Burn*, 1990a). ROI is based on the hydrological neighborhood determination. This is a method in which stations are included in a group on the basis of threshold values of a set of related attributes and a weighting function. In the ROI method, each site is assumed as the centre of its own region. Each site has associated with it a collection of gauged attributes that are useful for the transfer of extreme flow information. There is a need for the choice of a threshold value that functions as a cut-off for the dissimilarity measure. There

are different ways to measure the similarity of each basin with the target site; one is measured by means of a weighted Euclidean distance in the M-dimensional attribute space defined by a set of N physiographic and climatic indexes which are considered to influence the frequency behavior of the extreme flows of the basin. The distance measure employed has the following expression (*Burn*, 1990a):

$$D_{i,j} = [\sum_{m=1}^{N} W_m(X_{m,i} - X_{m,j})^2] \tag{2.1}$$

where $D_{i,j}$ = the Euclidean distance from site i to the site j, $X_{m,i}$ = the standardized value of the m^{th} pooling variable (catchment attribute) for site i, W_m = a weight reflecting the relative importance of the m^{th} attribute and N = the total number of pooling variables.

The equation above allows for the calculation of a dissimilarity index for any pair of sites. Catchments with higher similarity with the target site have a lower $D_{i,j}$ value and enter the pooling group first. Since different attributes have different units the standardization of the attributes is necessary. There are several methods available for data standardization.

2.1.2.2 K-means clustering

The K-means algorithm can be applied to form clusters based on attributes into K partitions (*Changa et al.*, 2008; *Burn and Goel*, 2000). This comprises grouping of pooling sites using the clustering algorithm (outlined below) and later modifying the formed clusters using the homogeneity test. The algorithm assumes the attributes

are from a vector space. The objective is to achieve a minimized total intra-cluster variance (distance) or squared error function D_v:

$$D_v = \sum_{k=1}^{K} \sum_{x_j \in S_k} |x_j - c_k|^2 \tag{2.2}$$

where c_k is the the centroid or mean point of all points in cluster k; S_k is the set of points in the kth cluster; and x_j is the standardized value for attribute j from site i. The K-means algorithm starts by throwing random seeds as initial centroids and making an initial set of K groups, either at random or using some heuristic approach (Figure 2.2 (a)). It then calculates the mean point, or centroid, of each set. The next step involves creating a new partition by associating each point with the closest centroid (Figure 2.2(b)). Then, the centroids are recalculated for the new clusters (Figure 2.2(c)). The algorithm is repeated by alternate application of these two steps until convergence (Figure 2.2(d)). This is obtained when the points no longer switch clusters (or alternatively when the centroids are no longer changed) (*Changa et al.*, 2008).

This algorithm has a drawback in terms of performance; there is no guarantee of finding a global optimum and the quality of the final solution depends on the initial set of clusters and may, in practice, be much poorer than the global optimum. Since the algorithm is extremely fast, a common method is to run the algorithm several times and return the best clustering found.

2.1.2.3 Frequency analysis for ungauged sites

Most drought frequency analysis methods require adequate observed streamflow records which can only be provided for gauged catchments. An ungauged site is a site where no data have been observed (*Hosking and Wallis*, 1993). Most recent literature suggests that frequency analysis at an ungauged site can be done using the regional

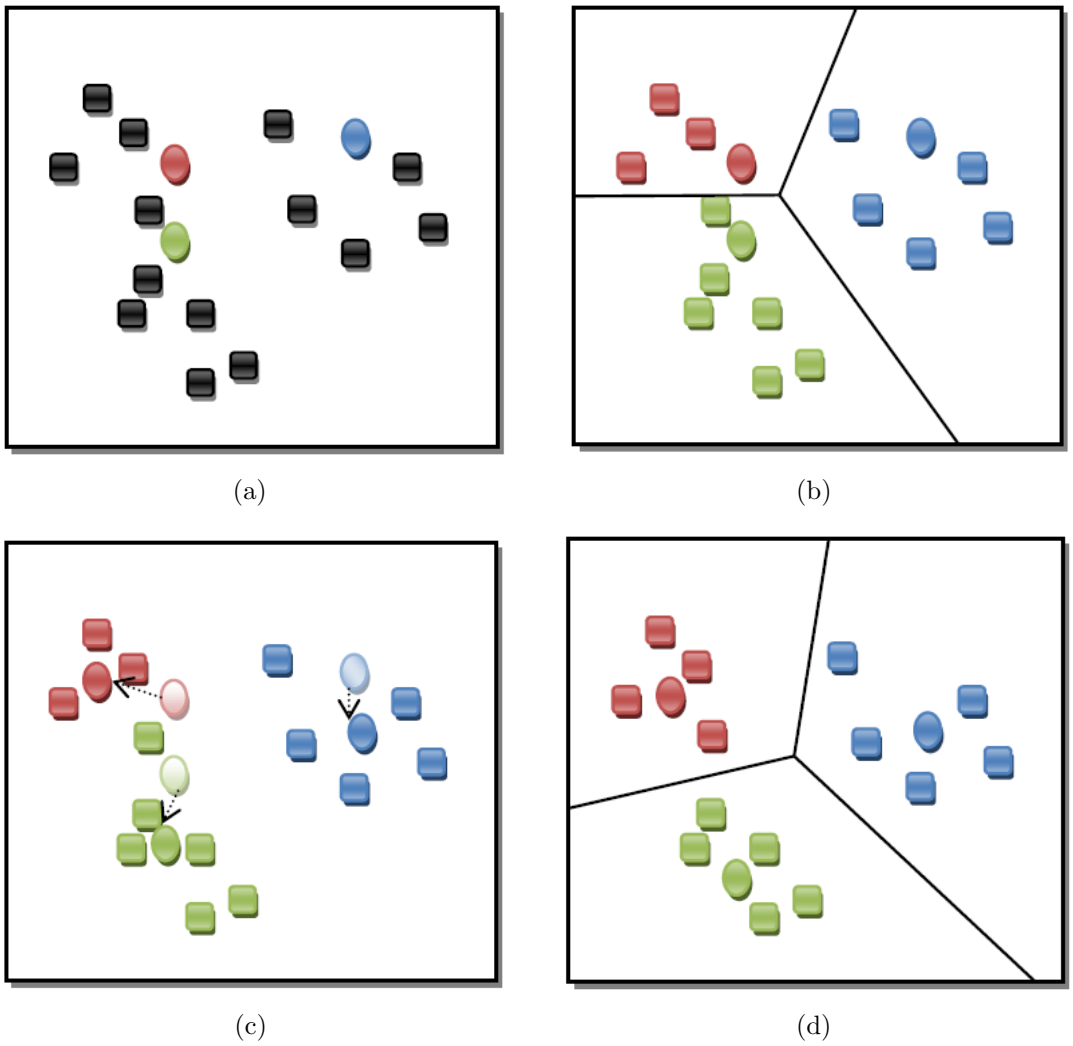


Figure 2.2: The objective of K-means clustering is to minimize the sum of squares of intra-cluster variance between data and the corresponding cluster centroid after assigning initial seeds to the data set

frequency analysis approach (*Hosking and Wallis*, 1997). An ungauged site can be assigned to one of the regions identified for the gauged sites, using the ungauged site's characteristics. For an ungauged location, information from hydrologically similar gauged catchments was used to characterize the flood regime (*Burn and Goel*, 2000). Regionalization was done using a clustering algorithm as a starting point. The result of regionalization can be used for estimating extreme flow quantiles for

gauged or ungauged sites. *Shu and Ouarda* (2008) used ANFIS for flood quantile estimation at ungauged sites. In the first steps, regionalization was achieved and then using a jackknife cross-validation procedure for each catchment of the study area, its flood records were temporarily removed from the database, and thus assumed to be ungauged. The training was done without including the ungagued site. Regional estimates can be tested using the calibrated model for the ungauged sites.

Estimating the parent distribution of an ungauged site can be achieved by regionalization and parametric approach. Then there remains only the problem of estimating the index event of ungauged site, which is usually the mean μ of the at-site frequency distribution at ungauged sites. This can be done by regarding μ as being a function of site characteristics. The relationship between μ and site characteristics by using data from the gauged sites can be calibrated and used for estimating μ of the ungauged site (*Hosking and Wallis*, 1993).

2.1.3 Univariate frequency analysis

The classical approach to the drought problem began with the evaluation of the instantaneously smallest value by means of the theory of extremes (*Gumbel*, 1958). This approach does not reveal anything about the drought duration. In earlier studies, in order to satisfy the assumptions of IID data, hydrologic data were carefully selected which in practice was often achieved by selecting the annual maximum or minimum of the variable being analyzed (e.g, annual minimum discharge during the year) with the expectation that successive observations of this variable from year to year will be independent (*Chow et al.*, 1988). *Tase* (1976) preferred exclusively the univariate experimental methods such as Monte Carlo or sample generation since application of analytical methods in the investigation of area-deficit-intensity characteristics of drought faced many difficulties. A univariate drought frequency analysis does not

fully characterize the drought potential due to the existence of correlation among the drought characteristics, namely severity and duration.

2.1.4 Bivariate frequency analysis

Bivariate frequency analysis can be used to consider the occurrence and effect of two drought characteristics simultaneously. Bivariate analysis of drought is finding more interest in recent years, however, most of the previous work on drought frequency was in dealing with univariate analysis of drought and little work has been done on bivariate characterization of drought.

Hisdal and Tallaksen (2003) introduced a method to calculate the probability of a specific area to be affected by a drought of a given severity. *Sen* (1980) derived a joint and marginal PDF of regional drought/flood descriptors for simple cases on the basis of random fields and probability theory. Other researchers have studied joint distribution of drought severity and duration using the conditional distribution of drought severity given drought duration and its distribution (*Gonzalez and Valdes*, 2003; *Shiau and Shen*, 2001). *Beersma and Buishand* (2004) derived joint probability of annual maximum precipitation deficit and discharge deficit. Three theoretical distributions of bivariate normal, bivariate Gumbel and a logistic Gumbel dependence structure were used to join the standardized transformed precipitation deficit and discharge deficit. The theoretical distributions were compared with an empirical bivariate distribution obtained with a re-sampling model. The re-sampling was performed to simulate values of precipitation, evaporation and discharge. *Hisdal and Tallaksen* (2003) produced drought severity-area-frequency curves using the probability distribution functions of the area covered by the drought deficit volumes. These curves showed the estimation of the probability of an area with a drought of a given severity, and thereby return periods could be assigned to historical drought events.

The last comparisons of drought characteristics showed that streamflow droughts are less homogeneous over the region, less frequent and last for longer time periods than precipitation droughts. *Yue and Rasmussen* (2002) did some valuable work on explaining some useful concepts in bivariate frequency analysis. *Kim et al.* (2003) examined a methodology for estimating the return periods of droughts using a non-parametric kernel estimator. The kernel estimator was developed for both univariate and bivariate frequency analysis. According to them, bivariate analysis showed a shorter return period for the severe droughts occurring during 1990s for the Conchos River Basin in Mexico. Since drought severity and drought duration exhibit significant correlation, a bivariate distribution is used to model the drought duration and severity jointly by *Shiau and Saralees* (2007). In parametric analysis, the biggest problem of studying drought severity and duration jointly is that drought severity and duration do not often follow the same distribution. Therefore, a "copula" is applied to form the bivariate distribution on data from the Yellow River in China (*Shiau and Saralees*, 2007).

Song and Singh (2009) modelled the joint probability distribution of periodic hydrologic data using meta-elliptical copulas, and monthly precipitation data from a gauging station in Texas, US, was used to illustrate parameter estimation. *Shiau and Modarres* (2009) developed a probabilistic approach to establish a drought severity-duration-frequency (SDF) relationship. They used rainfall data from two rain gauges in Iran and the copula approach was used for bivariate analysis (*Shiau and Modarres*, 2009). *Poulin et al.* (2009) compared three nonparametric estimators of the tail-dependence coefficient by simulations with seven families of copulas. *Poulin et al.* (2009) showed the importance of taking into account the tail dependence in the context of bivariate frequency analysis based on copulas for risk estimation. *Kim et al.*

(2006) presented a semi-nonparametric model and the nonparametric bivariate frequency analysis for characterizing droughts in the Conchos River Basin.

It is commonly agreed that since both severity and duration play an important role in drought characterization and management, the bivariate return periods estimated in theses studies would be useful for both design and management of water resources (*Kim et al.*, 2006; *Shiau and Saralees*, 2007; *Shiau and Modarres*, 2009).

2.1.5 Parametric frequency analysis

The parametric approach for frequency analysis is based on the premise that observations of hydrologic variables follow specified distributions. A probability distribution is a function representing the probability of occurrence of a random variable (*Chow et al.*, 1988). Fitting a distribution to a set of hydrologic data can generate a great deal of probabilistic information about the entire population. Fitting distributions can be accomplished by the method of moments, the method of L-moments, or the method of maximum likelihood. Drought characteristics commonly fit one of Gamma, Pearson Type-III, Generalized Pareto, log-normal or Wakeby distribution.

Burn et al. (2004) did a univariate analysis of drought for the data from the Athabasca River in Alberta. After reconstructing missing drought data, *Burn et al.* (2004) used them as a source of historical data for estimating drought severity quantiles. The drought quantiles were then fitted to a log-normal distribution for frequency analysis. *Shiau and Saralees* (2007) did a bivariate assessment to investigate the hydrological droughts of the Yellow River in northern China. The two major variables of drought, duration and severity, were fitted to different distributions and then a copula was used to assess the joint distribution of drought events.

2.1.6 Nonparametric frequency analysis

In contrast with parametric methods for estimating the density functions which assume that samples come from a population with a given PDF, nonparametric methods are distribution free. In general, nonparametric procedures for frequency analysis are becoming more accepted in hydrological practice. Studies on drought frequency analysis show that there is not a universally accepted parametric distribution for drought variables and results are sometimes strongly biased for high and low quantiles (*Kim et al.*, 2006). Nonparametric function estimations have advantages in that they always reproduce the attributes represented by samples. *Kim et al.* (2003) used a nonparametric kernel estimator for univariate and bivariate behaviors of drought return periods. *Kim et al.* (2006) studied a multivariate kernel estimator for bivariate drought characterization using the Palmer Drought Severity Index (PDSI) on droughts in the Conchos River Basin, Mexico. *Haghighatjou et al.* (2008) also asserts that parametric methods, although having been used successfully in some cases, are not fitting the observed data very well, or they divert from extreme tails. *Haghighatjou et al.* (2008) used both parametric and nonparametric approaches for frequency analysis of monthly precipitation in five locations from five cities in Iran. For the nonparametric approach, they used a kernel function with 4 different methods for finding the optimum smoothing parameter. However, the kernel method is not efficient in extrapolating a distribution function beyond an available record length. In work by *Adamowski and Feluch* (1990), this problem was investigated by using a new mixture distribution model for inclusion of historical data into the analysis. Then the nonparametric kernel approach was used for frequency analysis of floods based on the reconstructed historical data. *Ouarda and Shu* (2009) introduced Artifical Neural Networks (ANNs) to obtain improved regional low-flow estimates at ungauged sites in the province of Quebec, Canada. Each ANN was trained using the

Levenberg-Marquardt algorithm. The bootstrap aggregation approach was used to generate individual networks in the ensemble. The jackknife validation procedure was used to evaluate the performance of the proposed models. *Shu and Ouarda* (2008) developed a methodology for using ANFIS for flood quantile estimation at ungauged sites with identification ability of fuzzy models and the learning capability of ANNs. The proposed approach was applied to 151 catchments in the province of Quebec, Canada. Results showed that the ANFIS approach had a much better generalization capability than the Non-Linear Regression (NLR) and Non-Linear Regression with Regionalization (NLR-R) approaches and was more comparable to the ANN approach.

2.1.7 Stationary drought frequency analysis

In most frequency analysis literature the assumption is that the hydrologic system producing extreme events (e.g., a drought system) is stochastic, space-independent, and time-independent (*Chow et al.*, 1988). In other words, most literature on frequency analysis assumes that the parameters of the time series distribution have not changed over time. All work reviewed in this literature are based on the assumption that the hydrological system is stationary.

2.1.8 Non-stationary drought frequency analysis

Reoccurring droughts are considered to be a main natural hazard that can have significant environmental and economical impacts. During the last several years many hydrological studies have identified significant trends in the flow time series and therefore drought events. Are droughts becoming longer, or more severe, and happening more frequently? To answer this question a trend analysis on droughts should be done. A time series whose distribution parameters have changed over time is called

non-stationary. There are different sources causing non-stationarity in hydrological records such as forest fires, El Niño, land use changes, or climate change (*Cunderlik and Burn*, 2003). When significant non-stationarity is identified in the flow time series, it means that the parameters describing the location, scale and shape properties of the drought series change over time. Therefore, the standard parametric methods which are time independent under stationary conditions cannot be applied for drought frequency analysis. *Sadri et al.* (2009) did a trend analysis on rainfall data in Denmark. *Cunderlik and Burn* (2003) proposed a second order non-stationary approach to pooled flood frequency analysis, where non-stationarity was assumed only in the first two moments of the time series. Doing a trend analysis on rainfall data, *Wood* (1987) considered some evidence that the weather in the UK is becoming more variable with a tendency for drier summers and wetter autumns. By stating that this pattern has been observed only over the past 10 years, *Wood* (1987) suggested that engineering hydrologists should consider using paleo-hydrological data to improve the estimates of flood and drought severity.

A comprehensive literature on the effects of climate change on non-stationarity of low flows and extreme hydrological events including drought has been discussed in *Smakhtin* (2001). Also, a study on the impact of land-use, climatic change, and groundwater abstractions on streamflow droughts using four different physically based models operating with daily and monthly time steps was discussed by *Smakhtin* (2001). He discovered that both duration and deficits are increasing in most of the catchments with lower precipitation and higher storage capacity; the drought duration is increasing substantially (*Smakhtin*, 2001).

Studies on non-stationarity modelling due to climate change impacts on streamflow are normally performed in two distinct directions. They are either through the analysis of available historical flow records or by investigation of the effects of various

possible climate change scenarios on streamflow by means of physically based hydrological models (*Smakhtin*, 2001). *Vorosmarty et al.* (2000) combined a global runoff model, a global climate model, and population projections to compare the relative effects on water availability of projected climate change due to global warming and population growth in the year 2025. The model forecasted an overall reduction of global runoff of 6%, resulting in a 4% increase in water stress due to climate change alone. However, it was noted that the risk of recurring droughts with greater magnitudes due to population growth and economic development can be larger in the future as well.

Overall, very few literature has reflected the impact of climate change on drought frequency analysis. Non-stationarity is one of the realities in drought frequency and without considering it the strength of stochastic methods and intelligent learning on drought frequency remains unrevealed (*Smakhtin*, 2001).

2.2 Summary

In summary drought frequency analysis can be challenging when dealing with sites that have short record lengths or are ungauged. However, there has not been an easy or quick way to pool the similar sites together and adjust the initially formed regions so that they meet the requirements of effective hydrological regions based on index event criteria. Most adjustments of regions have been based on subjective judgment thus far. From another perspective, when carrying on a bivariate regional frequency analysis for droughts it is important to jointly consider both variables of severity and duration in each step including a bivariate test of homogeneity and discordancy. This topic has also been untouched in the case of droughts. There is also a need to provide a more straightforward procedure of bivariate frequency analysis of droughts

at sites with short record lengths or for ungauged sites. This means application of L-comoments statistics and copula in the process. Frequency analysis of an extreme hydrological event such as drought is most important at the tails of distributions. Work has to be done to compare different methods of drought frequency analysis in quantile estimation of ungauged sites at quantiles closer to tail of the distributions. In this work, neural networks and machine learning methods are introduced to examine different approaches for drought frequency analysis and compare the results with non-linear regression and nonlinear regression with regionalization methods. The reason that neural networks were selected to be examined is that very few ideas have been implemented using soft computing and intelligent techniques on drought frequency analysis. Both parametric and nonparametric approaches will be studied for drought frequency analysis.

Chapter 3

A Fuzzy C-Means Approach for Regionalization

One of the problems with drought frequency analysis is that, in most cases, the absence of lengthy records limits the reliability of statistical estimates. To address this issue, "pooled" or "regionalized" information from multiple sites is often used (*Burn et al.*, 1997). Regional frequency analysis uses data from a number of measuring sites to produce regions. From another side, drought is a multivariate phenomena whose two main variables are severity and duration. Therefore, correcting the initial regions formed for achieving effective regions (that are not including discordant sites and are homogeneous) should be based on bivariate L-moments criteria. Bivariate L-moments are matrices with L-comoment elements defined in this Chapter.

3.1 Background

Regional drought frequency analysis attempts to collect similar sites in one region in order to overcome the shortage of observed data. *Hosking and Wallis* (1997) recommended an "L-moment" statistics approach for judging the closeness of observed samples to suggested distributions. The L-moments are strong tools for univariate discordancy and homogeneity tests and have several theoretical advantages, including being able to characterize a wider range of distributions, to consider the correlation between variables, and, when estimated from a sample, to be robust to the presence of outliers. However, univariate L-moments calculate the discordancy and the homogeneity statistics based on only severity or duration of observed data and not based on severity and duration jointly. As a result, the final region formed is not necessarily homogenous and not discordant for both variables. To overcome this problem *Serfling and Xiao* (2007) developed multivariate L-moments for defining the joint statistical properties of multivariate phenomena. Bivariate L-comoment analysis as an extension of the univariate discordancy statistic and homogeneity test was presented by *Chebana and Ouarda* (2007).

From a statistical point of view, a "region" is a group of sites each of which is assumed to have data drawn from the same frequency distribution (*Hosking and Wallis*, 1997). According to the "index event" method one of the most important criteria for assessing whether a site can be included in a region is a heterogeneity measure. However, in most cases, satisfying this criterion is a subjective, often challenging and time consuming task. Multiple revisions to a region are often unavoidable. To address this issue the idea of using Fuzzy C-Means (FCM) clustering algorithm is applied. Fuzzy C-Means (FCM) is a method of clustering that allows one station to belong to two or more regions. Using the FCM algorithm, regions can be developed faster and easier. Clustering enables us to define the initial groups of catchments (regions). However,

it is often found that the resulting groups do not meet the requirements of lack of discordancy, homogeneity, and size for an effective region. Therefore, revisions to initial clusters are inevitable. The goals of the regional revision process are to remove discordant sites, increase the homogeneity of the regions, and to ensure each region is of a sufficient size. The theory of both univariate and bivariate discordancy and heterogeneity tests are reviewed. In order to accept that a region is sufficiently large, the guideline of $5T$ is used where T is return period (*Jakob et al.*, 1999). The rest of this chapter is organized as follows. In Section 3.1 theoretical background for FCM clustering algorithm and both univariate and bivariate discordancy and homogeneity tests are explained. Section 3.2 explains a case study for regionalization using both univariate and bivariate homogeneity approach. Results of regionalization are presented in Section 3.3 and conclusions and summary are in Section 3.4. Finally, the algorithms for clustering with univariate and bivariate homogeneity approaches are presented in Appendix A.

3.1.1 Fuzzy C-Means Clustering

The FCM algorithm is a modification of the K-means algorithm. This algorithm minimizes intra-cluster variance (*Ayvaza et al.*, 2007). It comprises the grouping of sites using the clustering algorithm (outlined below). The algorithm assumes the attributes are from a vector space. The objective is to achieve a minimized total intra-cluster variance (distance or squared error) function D_v

$$D_v = \sum_{k=1}^{K} \sum_{x_j \in S_k} |x_j - c_k|^2 \tag{3.1}$$

where K is the total number of clusters. The FCM algorithm starts by making an initial set of k groups, either at random or using some heuristic procedure. It then

calculates the mean point, or centroid, of each set. The next step is construction of a new partition by associating each point with the closest centroid. Then the centroids are recalculated for the new clusters and the algorithm is repeated by alternate application of these two steps until convergence. Like K-means, this algorithm minimizes intra-cluster variance (*Ayvaza et al.*, 2007). This is obtained when the points no longer switch clusters (or alternatively when the centroids are no longer changed) (*Burn and Goel*, 2000). In contrast to the K-means algorithm, which assigns each site to only one cluster, partial membership is permitted in FCM, meaning that each point has a degree of membership in each of the clusters. Thus points on the edge of a cluster may be in that cluster to a lesser degree than points in the centre of a cluster.

The degree of belonging of site i in the kth cluster is equal to the inverse of the distance of site i to the centroid of that cluster

$$b_k(i) = \frac{1}{d(\text{centre}_k, i)} \tag{3.2}$$

where $b_k(i)$ is the degree of belonging of site i in the kth cluster; and $d(\text{centre}_k, i)$ is the distance of site i to the centroid of that cluster k. Each station is assigned to the cluster to which it has the largest membership value. The coefficients are normalized and "fuzzified" with a real parameter so that the sum of membership of one site of interest to all different clusters is unity (*Ayvaza et al.*, 2007).

$$\forall i(\sum_{k=1}^{K} U_k(i) = 1) \tag{3.3}$$

where $U_k(i)$ is the normalized coefficient of site i in the kth cluster ($\in$ [01]).

3.1.1.1 Adjustment of clusters formed

FCM clustering method enables the definition of initial groups of catchments (regions). However, it is often found that the resulting groups do not meet the requirements for an effective region (*Hosking and Wallis*, 1993). *Hosking and Wallis* (1993) and *Jakob et al.* (1999) indicate that an effective region should satisfy the following criteria

1. Discordancy $D(i)$: For each cluster, the first check is the discordancy of each site as a member of that cluster; *Hosking and Wallis* (1993) suggested that if $D(I) \geq 3$ it is too large and that site is grossly discordant with the group as a whole and should be moved from the cluster into another possible host cluster.

2. Homogeneity H_1: Test of homogeneity is a natural way to know whether the between site dispersion of the sample L-moments for the group of sites under consideration is larger than would be expected of a homogeneous region. H_1 is the standardized test value for the group L-CV and shows the homogeneity of the cluster. The cluster is strongly homogeneous if $0 \leq H_1 < 1$, acceptably homogeneous if $1 \leq H_1 < 2$ and heterogeneous if $H_1 \geq 2$. In this study $H_1 \leq 2$ was considered to show homogeneity.

3. Size of the region J: *Jakob et al.* (1999) indicated that a region ideally should contain $5T$ station-years of data to provide an effective estimate for an event with a return period of T years. For $T = 100$ years, sum of number of drought events in one cluster should be at minimum $N = 500$.

The goal of the regional revision process is to remove discordant site(s) from the clusters and find a home cluster for the removed site(s), to make sure the regions formed are homogeneous, and to ensure each region is of a sufficient size. There are several techniques to achieve this including (*Hosking and Wallis*, 1997):

- Move a site or a few sites
- Delete a site or a few sites
- Subdivide a region
- Break up the region
- Merge a region with another or others
- Merge two or more regions
- Obtain more data and redefine groups

FCM is a great tool to achieve the corrected final clusters since it calculates the degree of membership of each site into each cluster. This extra piece of information that FCM provides reduces the amount of subjective judgment and complexity of deciding which site(s) can be moved from/to a region or different regions.

3.1.1.2 Test of univariate discordancy

After each cluster is formed, the first assessment is the discordancy measure of each site i as a member of that cluster among a set of N sites; *Hosking and Wallis* (1993) suggested that if $D(i) \geq 3$, that site is grossly discordant with the group as a whole and should be moved from the cluster into another possible host cluster. Let $u_i = [t^{(i)} t_3^{(i)} t_4^{(i)}]$ be a vector containing the L-CV, L-skewness and L-kurtosis values for a site i. The value $\bar{u}$ is the unweighted group average (*Hosking and Wallis*, 1997)

$$\bar{u} = N^{-1} \sum_{i=1}^{N} u_i, \quad i = 1, ..., N. \tag{3.4}$$

The discordancy measure for site i is defined as

$$D_i = \frac{1}{3} N (u_i - \bar{u})^T S^{-1} (u_i - \bar{u}) \tag{3.5}$$

where S is the matrix of sums of squares and cross-products

$$S = \sum_{i=1}^{N} (u_i - \bar{u})(u_i - \bar{u})^T \tag{3.6}$$

Discordancy can be illustrated heuristically: in a two dimensional space a group of sites will yield a cloud of L-CV versus L-skewness. Any point that is far from the centre of this cloud is flagged as discordant (*Hosking and Wallis*, 1993).

3.1.1.3 Test of univariate heterogeneity

H_1 is the standardized test value for the group L-CV and shows the homogeneity of the cluster. Estimation of the degree of heterogeneity in a group of sites is an assessment of whether the between-site variations in sample L-moments is what would be expected for a homogeneous region. Based on *Hosking and Wallis* (1993), all sites in a homogeneous region have the same population L-moments, however, their sample L-moments will be different. Test of homogeneity is a natural way to know whether the between site dispersion of the sample L-moments for the group of sites under consideration is larger than would be expected of a homogeneous region. A simple measure of the dispersion of the sample L-moment is the standard deviation of the at-site L-CVs. The reason to concentrate on L-CV is that the between site variation in L-CV has a much larger effect (than variation of the other L-moments) on the variance of the estimates of the quantiles ($Q_i(F)$) (*Hosking and Wallis*, 1993).

If the weighted standard deviation of the at-site sample L-CV is

$$V = \frac{\sum_{i=1}^{N} n_i (t^{(i)} - \bar{t})^2}{\sum_{i=1}^{N} n_i} \tag{3.7}$$

where $t^{(i)}$ is the sample L-CV at site i; $\bar{t}$ is the group average L-moment ratios, with sites weighted proportionally to their record length; and V is the weighted standard deviation of the at-site sample $L - CV$.

The heterogeneity measure is calculated as

$$H_1 = \frac{(V - \mu_v)}{\sigma_v} \tag{3.8}$$

where μ_v and σ_v are the mean and the standard deviation of a large number N_{sim} of simulated regions (using Monte Carlo simulation) from the kappa distribution. For detailed information refer to *Hosking and Wallis* (1997).

3.1.2 Bivariate L-moments

If $X^{(j)}$ is a random variable with distribution F_j, for two random variables of $j = 1, 2$ multivariate L-moments are matrices Λ_r with L-comoment elements defined by

$$\lambda_{k[ij]} = Cov(X^{(i)}, P^{\star}_{k-1}(F_j(X^j))), \quad i, j = 1, 2 \quad and \quad k = 2, 3, ... \tag{3.9}$$

where k is the order moments ≥ 1; and $P^{\star}_{k-1}$ is the shifted Legendre polynomial. For example, the kth L-comoment of $X^{(1)}$ with respect to $X^{(2)}$ is (*Chebana and Ouarda*, 2007)

$$\lambda_{k[12]} = Cov(X^{(1)}, P^{\star}_{k-1}(F_2(X^2))) \tag{3.10}$$

Analogously, the first L-comoment elements are

$$\lambda_{2[12]} = 2Cov(X^{(1)}, F_2(X^{(2)})) \tag{3.11}$$

$$\lambda_{3[12]} = 6Cov(X^{(1)}, (F_2(X^{(2)}) - 1/2)^2) \tag{3.12}$$

$$\lambda_{4[12]} = Cov(X^{(1)}, 20(F_2(X^{(2)}) - 1/2)^3 - 3(F_2(X^{(2)}) - 1/2) + 1) \tag{3.13}$$

which are the L-coCV, L-coskewness and L-cokurtosis, respectively. For $k = 2$ the L-comoment coefficient is given by

$$\tau_{2[12]} = \frac{\lambda_{2[12]}}{\lambda_1^{(1)}} \tag{3.14}$$

and for $k \geq 2$ the L-comoment coefficients are

$$\tau_{k[12]} = \frac{\lambda_{k[12]}}{\lambda_2^{(1)}} \tag{3.15}$$

The matrix of the L-comoment coefficients for $k = 2$ is written as (*Chebana and Ouarda*, 2007)

$$\Lambda_2^\star = \begin{pmatrix} \tau_{2[11]} & \tau_{2[12]} \\ \tau_{2[21]} & \tau_{2[22]} \end{pmatrix} \tag{3.16}$$

and in general

$$\Lambda_k^\star = (\tau_{k[ij]})_{i,j=1,2} = \begin{pmatrix} \tau_{k[11]} & \tau_{k[12]} \\ \tau_{k[21]} & \tau_{k[22]} \end{pmatrix} \tag{3.17}$$

According to *Chebana and Ouarda* (2007), the L-comoments are similar in structure and behaviour to the univariate L-moments and capture their attractive properties. The univariate L-moments are explained in Chapter 1. More detailed information

on bivariate and multivariate L-comoments has been addressed in *Serfling and Xiao* (2007)

3.1.2.1 Test of bivariate homogeneity

The logic of bivariate homogeneity is the same as in univariate homogeneity described by *Hosking and Wallis* (1993). *Chebana and Ouarda* (2007) described the statistic $V_{\|.\|}$ as

$$V_{\|.\|} = ((\sum_{i=1}^{N} n_i)^{-1} \sum_{i=1}^{N} n_i \|\Lambda_2^{\star(i)} - \bar{\Lambda_2^{\star}}\|^2)^{1/2} \tag{3.18}$$

where $\|.\|$ is the norm of matrix V; and $\Lambda_2^{\star(i)}$ is the L-covariation coefficient matrix for site i with record length n_i, $i = 1, ..., N$.

$$\bar{\Lambda_2^{\star}} = (\sum_{i=1}^{N} n_i)^{-1} \sum_{i=1}^{N} n_i \Lambda_2^{\star(i)} \tag{3.19}$$

$V_{\|.\|}$ reduces to the V statistic of *Hosking and Wallis* (1993) when handling only one variable. Similarly to the univariate case the statistic that measures the heterogeneity of a set of sites is given by (*Chebana and Ouarda*, 2007)

$$H_{\|.\|} = \frac{V_{\|.\|} - \mu_{Vsim}}{\sigma_{Vsim}} \tag{3.20}$$

where μ_{Vsim} is the mean of the N_{sim} values of $V_{\|.\|}$ of simulated regions; and σ_{Vsim} is the standard deviation of the N_{sim} values of $V_{\|.\|}$ of simulated regions. The heterogeneity criteria in bivariate analysis is also similar to that in univariate analysis as in *Hosking and Wallis* (1993), meaning that depending on the value of $H_{\|.\|}$ a decision concerning the homogeneity of the observed region can be taken. In this case, a region of sites is homogeneous if $H_{\|.\|} < 1$, acceptably homogenous if $1 \leq H_{\|.\|} < 2$ and definitely heterogeneous if $H_{\|.\|} \geq 2$. In this work, the bivariate heterogeneity measure considers

only the L-CV measure of variation. Other measures described in *Hosking and Wallis* (1993) can also be considered for the extension by following the same procedure (*Chebana and Ouarda*, 2007).

3.1.2.2 Test of bivariate discordancy

The discordancy test proposed by *Hosking and Wallis* (1993) was extended to its multivariate framework by *Chebana and Ouarda* (2007). The discordancy measure of site i among a set of N sites is a preliminary step in evaluating effective regions. According to *Chebana and Ouarda* (2007), if a matrix of $U_i^t = [\Lambda_2^{\star(i)}\Lambda_3^{\star(i)}\Lambda_4^{\star(i)}]$ is considered for each site i, the following matrix D_i is defined by

$$D_i = \frac{1}{3}(U_i - \bar{U})^T S^{-1}(U_i - \bar{U}) \tag{3.21}$$

where

$$S = (N-1)^{-1}\sum_{i=1} N(U_i - \bar{U})(U_i - \bar{U})^T \tag{3.22}$$

$$\bar{U} = N^{-1}\sum_{i=1}^{N} U - i \tag{3.23}$$

where $\Lambda_2^{\star(i)}$, $\Lambda_3^{\star(i)}$ and $\Lambda_4^{\star(i)}$ are defined as matrices in Equation 3.17. It is possible to use a norm $\|D_i\|$ of the matrix D_i. Several matrix norms have been presented as examples in *Chebana and Ouarda* (2007). Using a norm transforms a matrix from multidimensional space to the real line and has the advantage of defining an intuitive distance in a vector space and reducing exactly to the usual univariate case (*Chebana and Ouarda*, 2007). A site i is discordant with respect to the considered set of sites if $\|D_i\|$ exceeds a critical value of 3. This value is accepted from *Hosking and Wallis* (1993) as an extension of univariate discordancy.

3.2 Case Study

The methodology explained in this study is applied to archived hydrological records of unregulated flow monitoring sites for rivers in the Canadian prairies and nearby areas in the provinces of Alberta, Saskatchewan, and Manitoba. The monthly records of 59 sites (22 sites from Alberta, 18 from Saskatchewan, and 19 from Manitoba) were selected from the "Archived Hydrometric Data" Website (*Water Survey of Canada*, 2006). These sites are all natural sites meaning that there has been minimal human related interference with the flow regime. The record lengths of flows for these stations vary from 15 to 88 years. The major step in delineation of pooling groups is the definition of similar regions based on certain attributes. Among the attributes considered are hydrological, climatic (weather regimes), and physiographic (basin) characteristics. Using archived hydrological and meteorological data and GIS maps and information acquisition, nine characteristics or attributes were extracted:

1. Latitude of gauging station
2. Longitude of gauging station
3. Drainage area $[km^2]$
4. Mean catchment elevation $[m]$
5. Mean annual catchment precipitation $[mm/yr]$
6. Mean daily maximum temperature $[^{\circ}C]$
7. Mean daily minimum temperature $[^{\circ}C]$
8. Mean catchment annual evapotranspiration $[mm/yr]$
9. Mean catchment run-off $[mm/yr]$

The truncation level of each month was assigned as the average flow of that month over the entire period of record. The drought events (i.e., pairs of duration and severity for each event) at each of the 59 sites were extracted using a code written in MATLAB. *Hosking and Wallis* (1997) suggested that a site should have ≥ 20 historic events in order to have contributed into statistical analysis of frequency correctly, therefore, any site with < 20 drought events was removed from the analysis process. Therefore, the number of sites for clustering analysis was reduced to 36. Figure 3.1 shows the location of sites originally selected and the remaining 36 sites.

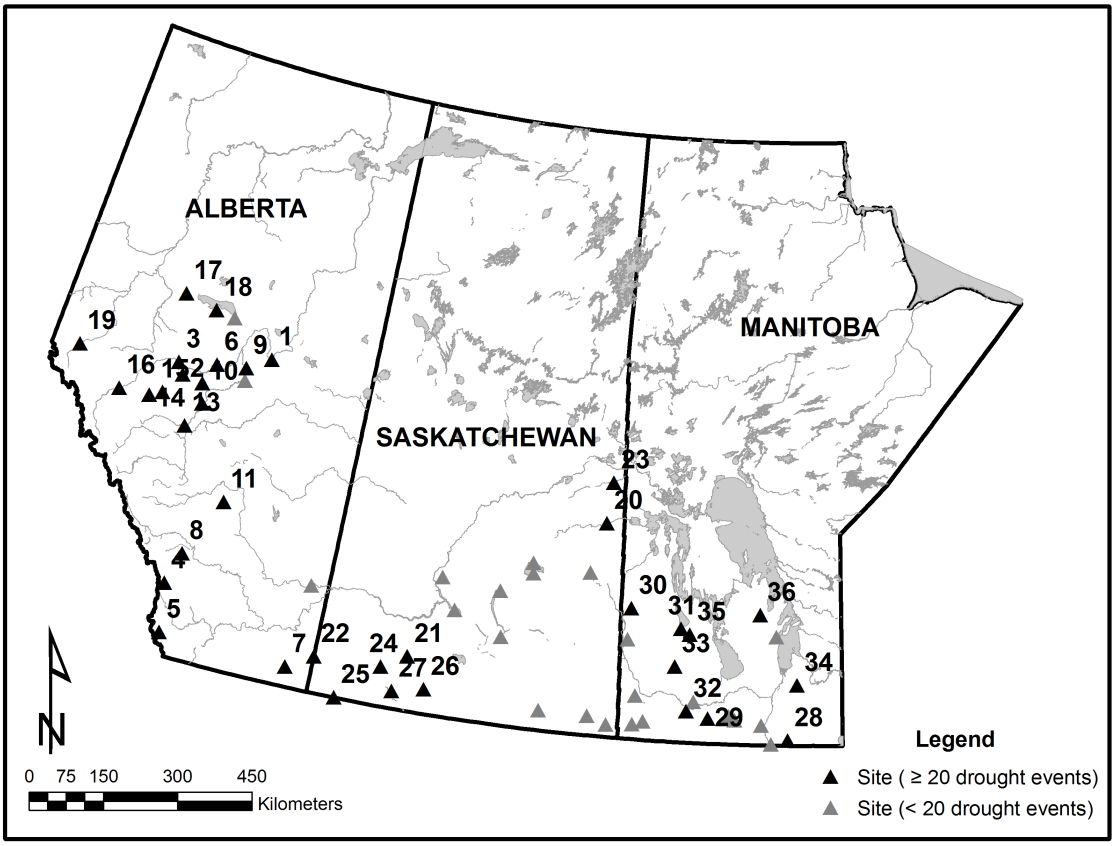


Figure 3.1: The location of the 36 natural sites on the Prairie Provinces plus sites with less than 20 drought events

3.2.1 Trend analysis results

Changes in streamflow patterns can be seen as evidence of climate change (*Cunderlik and Burn*, 2003). Therefore, an essential part of drought frequency analysis is studying any trends in streamflow. The Mann-Kendall (MK) nonparametric test was used for at-site and regional analysis to detect statistically significant trends. Having a collection of 36 hypothesis tests, the at-site significance level p_i, $i = 1, ..., 36$ of each site was computed.

The severity data had no tied values whereas duration data of each site had some tied values so variance had to be corrected for tied data. For a 5% significance two-sided test, all trends in duration were decreasing trends and minimal increasing trends were detected for severity at only a few sites. Therefore, assuming stationarity is conservative. Also, regional MK test revealed no significant trends at 5% significance level. Therefore the null hypothesis that there is no trend was accepted and according to p-values it was concluded that there was no evidence to reject the null hypothesis. Valuable information on how to perform the MK test for both at-site and regional scales is available in *Douglas et al.* (2000) and *Burn and Hag Elnur* (2002).

3.2.2 Formation of initial clusters and adjustments in univariate analysis

In this study, an FCM algorithm was designed to take inputs of return period (T), number of desired clusters, and sites's characteristics (as row vectors) as inputs and return clusters which are meeting the three criteria of homogeneity, lack of discordancy, and sufficient size. The algorithm is presented in Appendix A. Based on the algorithm, a model was developed using MATLAB code.

In this study, a return period of $T = 100$ years, a number of clusters equivalent to three and a matrix of 36×9 of site characteristics was given to the clustering model as inputs. Choice of having three clusters for 36 sites was rather a subjective decision. In different literature, there is no assumption that distinct number of clusters for certain number of sites satisfy the homogeneity condition. In other words, there is no "correct" number of clusters. However, the aim is to choose the number of clusters within which at-site frequency distributions vary so little with the site characteristics that regional frequency analysis is preferable to at-site analysis (*Hosking and Wallis*, 1997). Therefore, the aim is to seek a balance between using regions that are too small or two large. As a rule of thumb, methods that tend to form clusters of roughly equal size should give good results (*Hosking and Wallis*, 1997). The FCM algorithm developed for this study is flexible to make different number clusters depending on the input.

Clusters formed using FCM need not be final (*Hosking and Wallis*, 1997). Clusters should be adjusted in order to meet the three requirements of lack of discordancy, homogeneity, and size (*Hosking and Wallis*, 1993). For each cluster, the first check is the discordancy of each site as a member of that cluster; if the $D(I) \geq 3$, that site should be removed from the cluster and moved to the next cluster with which it has the highest membership. The aim is to get the other clusters to adopt the discordant site. Moving of discordant sites continues until that site is no longer discordant. Sites that get introduced in clusters and stay discordant at any stage including the last stage, are permanently removed from the group of sites. In the next step, the total number of drought events of all sites (i.e., $\sum n_i$) of each cluster should be calculated; for $T = 100$ years, sum of drought events in one cluster should be at least 500. If this criteria is already met, we can move on to the next level, otherwise, site(s) in other clusters can join our candidate cluster. We choose the site(s) which have the second

highest membership into the candidate cluster after their own cluster to compensate for lack of number of drought events in our target cluster. Note that one site can contribute to two clusters or more at the same time and this is an advantage of fuzzy clustering. Similarly to reach effective size of each cluster, adjustments for meeting the homogeneity can be done. More sites can be added to the target cluster until the homogeneity criteria ($i.e., H < 2$) is met. Partial membership of sites into clusters in FCM make the decision of which site to remove or add to a cluster very easy.

3.2.3 Formation of initial clusters and adjustments in bivariate analysis

The algorithm for initial formation and correction of clusters in bivariate homogeneity and discordancy test is the same as univariate algorithm described briefly in the previous section. The difference is that instead of reading the values of univariate discordancy and homogeneity, the algorithm reads the bivariate discordancy and homogeneity from two MATLAB functions written by *Chebana and Ouarda* (2007). The correction of clusters is then done based on the new readings of bivariate discordancy and homogeneity, accordingly.

3.3 Results

3.3.1 Regionalization results

Table 3.1 shows a summary of three clusters formed by FCM algorithm. Note that since formation of clusters is based on sites' characteristics not sites' statistic, the initial clusters formed have the same site members regardless of whether the focus is on severity or duration. However, discordancy ($D_{i,D}$ and $D_{i,S}$) and homogeneity

values for the two variables of duration and severity are different. Discordant sites and heterogeneous clusters are accented by a $\star$ sign.

Table 3.1: Initial clusters (univariate analysis)

Cluster 1				Cluster 2				Cluster 3			
Site	$D_{i,D}$	$D_{i,S}$	$\sum n_i$	Site	$D_{i,D}$	$D_{i,S}$	$\sum n_i$	Site	$D_{i,D}$	$D_{i,S}$	$\sum n_i$
4	0.8	1.08	48	20	0.82	1.06	54	1	0.25	0.73	104
7	0.47	0.83	44	23	1.22	0.05	50	2	0.24	4.21 **	22
8	0.14	1.16	38	28	0.84	2.31	38	3	0.39	2.79	54
11	1.1	2.34	20	29	1.48	1.88	20	5	0.61	1.3	23
21	1.82	1.11	27	30	0.76	0.35	25	6	2.69	0.66	28
22	2.64	0.55	21	31	1.82	1.64	72	9	0.27	0.27	67
24	2.11	1.34	20	32	0.56	0.67	35	10	1.63	0.71	23
25	0.2	0.78	31	33	0.83	0.67	70	12	0.29	0.25	105
26	0.19	0.38	24	34	0.86	0.65	63	13	3.43 *	0.56	25
27	0.53	0.43	33	35	1.37	1.53	29	14	1.47	0.65	87
				36	0.44	0.2	47	15	0.63	0.24	85
								16	0.07	0.29	93
								17	1.29	1.73	61
								18	0.43	0.46	63
								19	1.31	0.16	83
Sum			306				503				923
Homogeneity	4.42**	1.94*			4.60**	2.29**			9.86**	7.62**	

Revisions to initial clusters are performed using an algorithm developed in MATLAB and the principle criteria described in Section 3.1.1.1. Table 3.2 and Table 3.3 provide the information of modelling after all sites have been adjusted.

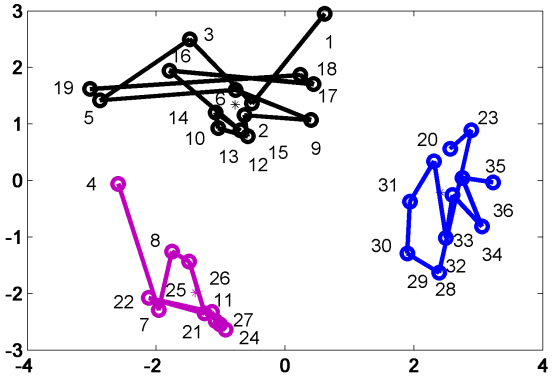
In Figure 3.2(a) the initial clusters and their sites are presented graphically. Since each site is a vector with nine attributes, or in other words, each site is a nine dimensional vector, it is not possible to display the clustering procedure graphically on a two dimensional plane. Therefore, in order to display the functionality of FCM, a projection technique needs to be applied. A Principal Component Analysis (PCA) technique is used which can also be coded as a command line in MATLAB (*Samania et al.*, 2007). Therefore, it should be noted that the distance between any two stations,

Table 3.2: Duration (final clusters)

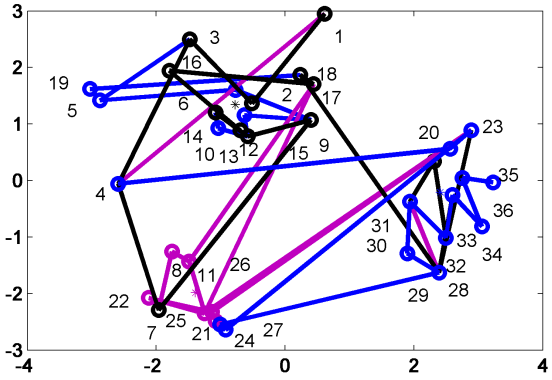
Cluster 1			Cluster 2			Cluster 3		
Site	$D_{i,D}$	$\sum n_i$	Site	$D_{i,D}$	$\sum n_i$	Site	$D_{i,D}$	$\sum n_i$
1	0.21	104	4	0.13	48	1	0.08	104
4	0.52	48	20	0.97	54	2	2.65	22
7	0.36	44	23	1.3	50	3	1.8	54
8	0.25	38	26	0.66	24	4	1.18	48
11	2.65	20	27	0.48	33	7	0.6	44
17	0.25	61	28	1.64	38	9	0.13	67
21	2.19	27	29	2.1	20	12	0.73	105
24	2.28	20	30	1.06	25	14	0.83	87
25	0.2	31	32	0.63	35	15	0.87	85
26	0.31	24	33	0.99	70	16	1.31	93
27	0.49	33	34	0.75	63	17	0.3	61
23	0.55	50	35	1.83	29	18	0.59	63
28	1.31	38	36	0.46	47	28	2.21	38
30	2.43	25				33	0.63	70
						34	1.1	63
Sum		563			536			1004
Homogeneity	1.30*			1.23*			1.88*	

as well as between any station and the centroid of the cluster, is representative of the intra-cluster variance or the squared error and is not into a real scale. Figure 3.2(b) and (c) show the final adjusted clusters for duration and severity, respectively.

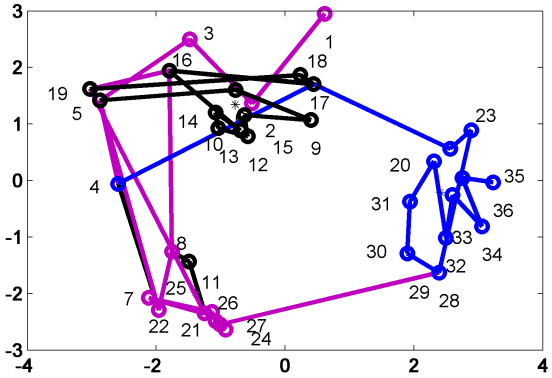
An algorithm developed in MATLAB is designed to result in bivariate homogeneous clusters as outputs. The values of bivariate discordancy and homogeneity are the criteria for adjusting clusters. As it can be noted, for the two variables duration and severity, there is only one output as final clusters. Table 3.4 has the result of initial clusters formed. The final clusters in bivariate clustering are presented in Table 3.5. Figure 3.3 is the mapping of final clusters formed using bivariate homogeneity analysis.



(a)



(b)



(c)

Figure 3.2: (a) Initial 3 clusters formed by FCM using sites' characteristics, (b) Final 3 clusters corrected for duration. Sites 5, 6, 10, 13, 19, 22, and 31 do not have a home cluster, (c) Final 3 clusters corrected for severity. Sites 1, 2, 3, and 11 do not have a home cluster

Table 3.3: Severity (final clusters)

Cluster 1			Cluster 2			Cluster 3		
Site	$D_{i,S}$	$\sum n_i$	Site	$D_{i,S}$	$\sum n_i$	Site	$D_{i,S}$	$\sum n_i$
4	0.83	48	4	0.46	48	5	0.98	23
5	0.75	23	17	1.52	61	6	2.5	28
7	0.42	44	20	1.1	54	9	0.73	67
8	1.35	38	23	0.06	50	10	1.05	23
16	0.11	93	28	1.55	38	12	0.15	105
19	0.11	83	29	2.1	20	13	1.94	25
21	1.39	27	30	0.47	25	14	0.52	87
22	2.96	21	31	1.79	72	15	0.26	85
24	0.89	20	32	0.68	35	16	0.22	93
25	0.76	31	33	0.64	70	17	2.13	61
26	0.31	24	34	0.64	63	18	0.53	63
27	0.89	33	35	1.82	29	19	0.97	83
28	2.23	38	36	0.16	47			
Sum		523			612			743
Homogeneity	0.31			1.64*			1.00*	

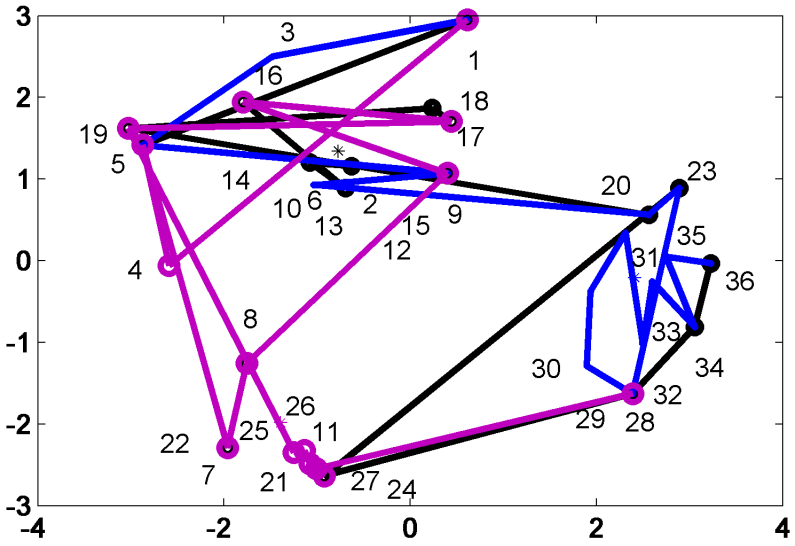


Figure 3.3: Final 3 clusters corrected using bivariate homogeneity. Sites 2, 6, 11, 12, and 22 do not have a home cluster

Table 3.4: Initial clusters (Bivariate analysis)

Cluster 1			Cluster 2			Cluster 3		
Site	$\|D_i\|_2$	$\sum n_i$	Site	$\|D_i\|_2$	$\sum n_i$	Site	$\|D_i\|_2$	$\sum n_i$
4	2.1282	48	20	0.7754	54	1	0.9038	104
7	0.609	44	23	1.1482	50	2	3.63**	22
8	2.3937	38	28	1.565	38	3	2.2271	54
11	2.4719	20	29	1.683	20	5	0.6036	23
21	0.894	27	30	1.6901	25	6	2.6261	28
22	2.5168	21	31	2.2381	72	9	0.6877	67
24	0.857	20	32	0.5915	35	10	0.391	23
25	0.5488	31	33	1.0295	70	12	1.9148	105
26	0.2444	24	34	1.3148	63	13	2.7166	25
27	0.7123	33	35	1.8326	29	14	0.7346	87
			36	1.151	47	15	0.7224	85
						16	0.1991	93
						17	0.7067	61
						18	0.3885	63
						19	0.4783	83
Sum		306			456			923
Homogeneity	-0.22			1.52*			7.93**	

3.4 Comparison and Summary

This Chapter covered both univariate and bivariate L-moment homogeneity test developed by *Hosking and Wallis* (1993) and later by *Chebana and Ouarda* (2007). The same group of sites and the same number of clusters are used for both univariate and bivariate analysis. Initial clusters formed in both univariate and bivariate approaches were the same. However, the final clusters, after all revisions from both methods, were not the same. The major reason for this difference is the fact that univariate homogeneity and discordancy criteria choose only one variable at a time to analyze and a site which can be recognized as discordant or heterogeneous when analyzing duration data may not necessarily be discordant or heterogeneous when analyzing its severity variable data. The possible solution to this problem is either think of

Table 3.5: Final clusters (Bivariate analysis)

Cluster 1			Cluster 2			Cluster 3		
Site	$\|D_i\|_2$	$\sum n_i$	Site	$\|D_i\|_2$	$\sum n_i$	Site	$\|D_i\|_2$	$\sum n_i$
1	2.0869	104	1	1.2504	104	1	2.601	104
4	2.5677	48	3	2.935	54	5	2.1607	23
5	2.4595	23	5	2.6952	23	7	0.7166	44
7	0.8429	44	9	0.671	67	8	1.9698	38
8	1.8808	38	13	2.989	25	9	0.5518	67
9	0.5178	67	20	0.7426	54	10	2.6504	23
16	0.5377	93	23	0.3389	50	14	1.378	87
17	0.6845	61	28	0.9402	38	15	0.8055	85
19	1.3979	83	29	2.3127	20	16	0.5625	93
21	1.353	27	30	0.9379	25	17	1.2959	61
24	2.2327	20	31	2.3556	72	18	0.2637	63
25	2.1404	31	32	0.5759	35	19	1.1548	83
26	0.5979	24	33	1.0748	70	20	2.063	54
27	1.0638	33	34	1.1004	63	23	0.6383	50
28	0.9545	38	35	2.3701	29	26	2.673	24
			36	0.6351	47	28	1.5035	50
						34	1.522	63
						36	1.1531	47
Sum		734			776			1059
Homogeneity	1.34*			1.79*			1.51*	

drought as a univariate phenomenon or apply the bivariate L-comoment approach to recognize joint heterogeneity and joint discordancy indexes for both variables, severity and duration. This can be seen by looking at the results: Comparing Table 3.2 with 3.3 it can be seen that when studying duration, other than sites 5, 6, 10, 13, 19, 22, and 31 the remaining 29 sites are included in at least one cluster. When severity is the variable of interest, sites 1, 2, 3 and 11 ended up having no home cluster(s) and had to be deleted. This is due to the fact that these sites have been recognized as discordant or they increased the heterogeneity in the clusters they initially were included. They then got removed from those clusters and moved into the next cluster with which each had the next highest membership. Removing site and

moving them down to the other clusters continued until each site could find a home cluster. Sites that ended up discordant or increased the heterogeneity in all clusters, were removed from all clusters permanently. Another observation of this study in the domain of regionalization, when doing bivariate frequency analysis of drought, is that it is rather impossible to use univariate homogeneity and discordancy tests for adjusting the initial hydrological regions formed. Using bivariate L-comoment homogeneity and discordancy tests for adjusting initial clusters resolves the issue of getting two different sets of clusters for each drought variable. Looking at Table 3.5, there is only one group of final clusters. Sites 2, 6, 11, 12, and 22 do not appear in any of the final clusters based on the adjustment criteria regardless of drought variable but since there is no site of interest in the assumption of study, deleting of some sites is not considerable here. If it happens that one of the deleted sites is the site of interest, there are several ways to find a home cluster for that site including: changing the number of clusters formed (the program developed in MATLAB for FCM clustering gets the number of desired clusters as input from the user and is capable to simply run the experiment on a different number of clusters) and/or obtaining more data and redefine groups. Bivariate homogeneity and discordancy tests are the multivariate version of L-moments approach developed by *Hosking and Wallis* (1993) and can be effectively used to model drought events described by their duration and severity. The proposed procedure is also easy to use and implement. The model presented in this study is a useful tool for illustrating the advantages of FCM clustering and bivariate homogeneity and discordancy tests in regional drought frequency analysis. It should be noted that the performance of the proposed approach can be influenced by several factors such as the size of the region $\sum n_i$, each sites' record lengths n_i, and the degree of regional heterogeneity. In summary, univariate tests can give a false indication of the regions in bivariate drought frequency analysis.

Chapter 4

Copula-based Pooled Frequency Analysis of Droughts in the Canadian Prairies

One of the difficulties of drought frequency analysis is calculating the joint return period of drought based on the two correlated variables of droughts: duration and severity. In most drought frequency analysis literature, there is an assumption that the two variables of severity and duration are from the same distributions mostly normal distribution. In practice, that is not the case. In this Chapter bivariate drought frequency analysis is applied with applying a technique called "copula". Copulas are functions that connect multivariate probability distributions to their one-dimensional marginal probability distribution while still capture the essential features of dependence and correlation among the random variables.

4.1 Introduction

The most significant fact about drought is its dynamic and multi-attribute nature. *Tase* (1976) used experimental methods such as Monte Carlo or sample generation since application of analytical methods in the bivariate characteristics of drought faced many difficulties. *Sen* (1980) derived a joint and marginal PDF of regional drought/flood descriptors for simple cases on the basis of random fields and probability theory. Other researchers have studied joint distribution of drought severity and duration using the conditional distribution of drought severity given drought duration and its distribution (*Gonzalez and Valdes*, 2003; *Shiau and Shen*, 2001). In these studies, significant correlation relationships are not revealed by separate consideration of correlated characteristics. *Hisdal and Tallaksen* (2003) produced drought severity-duration-frequency (SDF) curves using the probability distribution function approach of the area covered by the drought deficit volumes. In practice, drought SDF curves (analogous to rainfall IDF curves) are derived empirically meaning that no analytic approaches for such multivariate curves have been proposed. Simultaneous assessment of the multi-attributes of droughts can yield much more sophisticated results in evaluating the risk of droughts. Much of the previous work on drought frequency dealt with univariate analysis. This chapter aims to investigate joint distribution of drought quantiles in terms of copula. Bivariate frequency analysis of drought using copula, although still recent, is becoming more popular. The reason is that in reality, the two characteristics of drought (severity and duration) are correlated and may not have the same marginal distributions. The main objectives of this chapter are

1. Study of the bivariate probability distribution of drought variables by using a suitable copula to describe the dependence between the two drought characteristics.

2. Development of a probabilistic approach for drought bivariate severity-duration CDF and return period.

3. Development of a case study based on droughts of the Canadian Prairie Provinces.

In this study, the drought data and clusters evaluated in Chapter 3 are used. The summary of drought data are presented in Table 4.1. The location of stream monitoring sites are approximately between $120° - 97°$W and $49° - 56°$N.

Table 4.1: Statistics summary of the study data

Variable	Min	Mean	Max	STD
Drainage area $[km^2]$	111	5183	74600	12769
Mean elevation $[m]$	222	760	1879	407
Mean annual precipitation $[mm/yr]$	337	463	600	69
Mean daily max temperature $[°C]$	6	9	12	2
Mean daily min temperature $[°C]$	-5	-4	-2	1
Mean annual evapotranspiration $[mm/yr]$	16	219	369	105
Mean run-off $[mm/yr]$	13	178	1260	242

4.2 Methodology

The steps involved with the subsequent sections of this chapter to achieve copula-based drought frequency analysis are as follows

1. Fitting candidate distributions to both drought variables (i.e., duration and severity) on a regional basis.

2. Calculating copula parameters and the best fitting copula based on Q-Q plots for candidate sites.

3. Calculating copula-based joint return period of given ranges of severity and duration and presenting the 3D mesh of severity-duration-joint return period of candidate sites.

These steps are explained in further detail in the next sections.

4.2.1 Copulas

Copulas are functions which link joint probability distributions to their one-dimensional marginal distributions (*Singh and Zhang*, 2007). Using the copula approach, each component is allowed to have its own different marginal distribution. This characteristic gives the copula approach a high level of flexibility for modelling compared with regular bivariate analysis. The theoretical basis of a copula was first introduced by *Sklar* (1959), who used it to derive the joint distribution of random variables having non-normal marginal distributions. According to the Sklar's theorem, for a bivarate case, copula exists as function C that binds the margins F_X and F_Y to give joint distribution F_{XY} (*Sklar*, 1959). Sklar's theorem can be stated as follows

$$F_{XY}(x, y) = C(F_X(x), F_Y(y)) \tag{4.1}$$

Equation 4.1 shows that a copula can describe a multivariate distribution in terms of a univariate distribution. If marginal distributions F_X and F_Y are continuous, the copula function C is unique. Otherwise, the copula C is unique on the range of values of the marginal distributions. Determining C involves estimating the marginal distribution of each variable separately and the dependence function. These two steps enable the derivation of the joint probability parameter regardless of the dependence between different marginal distributions of the variables.

Several families of copulas have been widely used in risk analysis, financial domain and actuarial science. In the area of hydrology, the one-parameter Archimedean copula family is more applicable. Archimedean copulas are easy to construct; they include a large variety of copula families most of which can be applied whether the correlation

between variables is positive or negative. *de Michele and Salvadori* (2003) used copulas to model different combinations of rainfall depth and duration. *de Michele et al.* (2005) used copulas to model the dependence structure between flood peak and flood volume to check the adequacy of a spillway. Archimedean copula is an important family of copula; if U and V are uniformly distributed random variables defined as $U = F_X(X)$ and $V = F_Y(Y)$, then the one parameter Archimedean copula, denoted C_θ, has cumulative density function (*Sklar*, 1959)

$$C_\theta(u, v) = \phi^{-1}\{\phi(u) + \phi(v)\}, \quad 0 < u, v < 1 \tag{4.2}$$

where θ is the parameter hidden in the generating function ϕ; $\phi(\cdot)$ is the copula generating function; u is the specific value of U; and v is the specific value of V. $\phi(\cdot)$ is the copula generator that is a convex decreasing function satisfying $\phi(1) = 0$; and $\phi^{-1}(\cdot) = 0$ when $v \geq \phi(0)$ (*Singh and Zhang*, 2007).

4.2.2 Obtaining Kendall's τ and copula's parameter

The first step in determining the copula for a given data set is to find the degree of correspondence between two variables and the significance of this correspondence. This is calculated by a nonparametric statistic, namely τ rank correlation coefficient (*Zhang and Singh*, 2006). It assumes that for a random sample of bivariate observations of size n : $(x_1, y_1), (x_2, y_2), \cdots (x_n, y_n)$ the underlying distribution function $H_{X,Y}(x, y)$ has an associated Archimedean copula C_θ which can also be regarded as an alternative expression of the joint CDF. Kendall's τ is the rank correlation coefficient and can be estimated from the observations as

$$\tau = \binom{n}{2}^{-1} \sum_{1 \leq i < j \leq n} sgn[(x_i - x_j)(y_i - y_j)] \tag{4.3}$$

where n is the number of observations, $i, j = 1, 2, ..., n$ and

$$sgn(\theta) = \begin{cases} 1 & : \quad \theta > 0 \\ 0 & : \quad \theta = 0 \\ -1 & : \quad \theta < 0 \end{cases} \tag{4.4}$$

The two pairs (x_i, y_i) and (x_j, y_j) are said to be concordant if $(x_i - x_j)(y_j - y_i) > 0$, and discordant if $(x_i - x_j)(y_i - y_j) < 0$. Another explanation of τ is

$$\tau = \frac{2}{n(n-1)} (n_{cp} - n_{dp}) \tag{4.5}$$

where n_{cp} is the total number of concordant pairs, n_{dp} is the total number of discordant pairs, and n is the number of observations.

The parameter θ can be determined for each site with a candidate copula from the calculated Kendall's τ. The following sections explain briefly the procedure for obtaining θ parameter. More extensive reading on Archimedean copulas is available from *Zhang and Singh* (2006).

4.2.2.1 Gumbel-Hougaard copula family

The generating function $|\phi(t)|$ for this family is expressed as (*Zhang and Singh*, 2006)

$$\phi(t) = (-\ln t)^{\theta} \tag{4.6}$$

where $t = u$ or v varying from 0 to 1, and θ is a parameter of the generating function. Thus, Equation (4.2) can be written as

$$\begin{aligned} C_\theta(u,v) &= C_\theta[F_X(x), F_Y(y)] = H_{X,Y}(x,y) \\ &= \exp\{-[(-\ln u)^\theta + (-\ln v)^\theta]^{1/\theta}\}, \quad \theta \in [1,\infty) \end{aligned} \tag{4.7}$$

where $H_{X,Y}(x,y)$ is the joint probability distribution function of two random variables X and Y. Kendall's coefficient of correlation τ between X and Y is $\tau = 1-\theta^{-1}$. Note that the relationship between Kendall's τ and the generating function shows that for the Gumbel copula, only the positive correlation structure of the bivariate data can be analyzed (*Zhang and Singh*, 2007).

4.2.2.2 Clayton copula family

For this family the generation function is written as (*Zhang and Singh*, 2006)

$$\phi(t) = t^{-\theta} - 1 \tag{4.8}$$

Thus, Equation (4.2) can be expressed as

$$C_\theta(u,v) = C_\theta[F_X(x), F_Y(y)] = H_{X,Y}(x,y) = [u^{-\theta} + v^{-\theta} - 1]^{(-1/\theta)}\}, \quad \theta \geq 0 \tag{4.9}$$

The correlation for this copula is

$$\tau = \frac{\theta}{\theta + 2} \tag{4.10}$$

Similar to Gumbel copula, the Clayton copula is only suitable for positively correlated random variables.

4.2.2.3 Frank copula family

The generating function of this family is expressed as

$$\phi(t) = \ln\left[\frac{e^{\theta t} - 1}{e^{\theta} - 1}\right] \tag{4.11}$$

Then Equation 4.2 can be written as

$$\begin{aligned} C_\theta(u, v) &= C_\theta[F_X(x), F_Y(y)] = H_{X,Y}(x, y) \\ &= \frac{1}{\theta}\ln[1 + \frac{[\exp(\theta u) - 1][\exp(\theta v) - 1]}{\exp(\theta) - 1}]\}, \quad \theta \neq 0 \end{aligned} \tag{4.12}$$

The correlation for Frank copula is

$$\tau = 1 - \frac{4}{\theta}[D_1(-\theta) - 1] \tag{4.13}$$

where D_1 is the first order Debye function D_k which can be defined for both positive and negative arguments (*Zhang and Singh*, 2007).

4.2.3 Identification of the preferred copula

When using different copulas, the question of which copula should be used to obtain joint distributions of variables needs to be answered. This question was addressed by *Genest and Rivest* (1993) who described a procedure for identification of the best fitting copula using a Q-Q plot procedure. A Q-Q plot is a plot of the quantiles of nonparametric copula versus the quantiles of parametric copula. The best matching copula should have its parametric and nonparametric statistics fitting on line in a Q-Q plot. Producing a Q-Q plot involves the following steps

1. Construct a nonparametric estimate of the distribution function called $K_N(z)$, where z is a specific value of $Z = Z(x, y)$, by obtaining z_i= [number of (x_j, y_j) such that $x_j < x_i$ and $y_j < y_i$]$/(N-1)$ for $i = 1, 2, \ldots, N$.

2. Construct a parametric estimate of K using the equations above with z obtained from step 1.

3. Define an intermediate random variable $Z = Z(x, y)$ which has a distribution function $K(z) = P(Z \leq z)$, when z is specific value of Z. This distribution function is related to the generating functions of the Archimedean copula, determined earlier

$$k(z) = \frac{\phi(z)}{\phi'(z)} \tag{4.14}$$

Where ϕ' = derivative of ϕ with respect to z. Once displaying $K_N(z)$ versus $K(z)$, if the samples do come from the same distribution, the plot will be linear.

4.2.4 Bivariate return period

In bivariate frequency analysis the physical meaning of the marginal CDF (also called the non-exceedance probability) and the return period remain the same as those in the univariate analysis (*Yue and Rasmussen*, 2002). Bivariate events can be described using concepts such as conditional probability distributions, conditional return periods, and joint return periods (*Yue and Rasmussen*, 2002). The joint return period of an event (X, Y) having joint cumulative distribution $F(x, y)$ can be defined as

$$T(x, y) = \frac{1}{1 - F(x, y)} \tag{4.15}$$

where $F(x, y) = \Pr[X \leq x, Y \leq y]$.

The joint return period for bivariate drought can be calculated by multiplying the above $T(x, y)$ relation by cycle (λ) of each site which is the average duration of dry period in years (i.e., total record length t divided by total number of drought events n) (*Burn and DeWit*, 1996). $f(x, y)$ is the joint PDF of two continuous random variables X and Y, and $f_Y(y)$ is the marginal PDF of Y (*Yue and Rasmussen*, 2002).

4.3 Results

4.3.1 Fitting Candidate Distributions to the Pooled Drought Variables

In parametric approach, the aim is to fit a single frequency distribution to the homogeneous region. An index drought procedure can be used to estimate the dispersion and shape of at-site data based on regional averaging, while the mean is estimated from at-site data (*Hosking and Wallis*, 1997). The candidate distributions are fitted by the method of L-moments. The goodness-of-fit statistic Z^{DIST} is computed for each of the homogeneous regions according to the procedure described in *Hosking and Wallis* (1993) for each of the candidate distributions. The distributions which give an acceptable fit have a $|Z^{DIST}| \leq 1.64$. The parameters obtained for a region should be scaled appropriately at any candidate site to estimate quantiles of the at-site frequency distributions. The methodology explained in this chapter is applied to the three clusters formed and tested using tests of bivariate homogeneity and discordancy. Based on the conventional L-moment approach of *Hosking and Wallis* (1997), the best fitting marginal distribution for the drought severity and duration were calculated from observed drought data. Since copula is a joint distribution function of the marginal univariate distribution functions, the univariate CDFs of drought severity and duration were fitted from the observed data. The results show that the

drought severity and duration can be fitted best to the Wakeby, Generalized Pareto (GP), and Pearson type III distributions. One site from each of the three clusters is selected to continue with bivariate and copula analysis. These sites are Denniel Creek Near Val Marie in Saskatchewan from cluster 1 (site 27), Little Saskatchewan River Near Minnedosa in Manitoba from cluster 2 (site 33), and Athabasca River at Hinton in Alberta from cluster 3 (site 16). In the following sections, these sites have been addressed based on their numbers. The marginal distribution parameters, which are assigned to a region, were rescaled for the candidate sites using the index event method. Table 4.2 shows the best fit distribution for each pooled region and the regional parameters. Table 4.3 shows the summary of statistics of candidate sites to be modelled.

Table 4.2: Candidate sites and distribution parameters

Cluster	Site	Variable	Distribution and Parameters
1	27	Duration [month] Severity $[*10^6 m^3]$	PIII: $\mu = 1.00, \sigma = 1.09, \gamma = 2.69$ Wakeby: $\xi = -0.115, \alpha = 0.00, \beta = 0.00,$ $\gamma = 0.92, \delta = 0.173$
2	33	Duration [month] Severity $[*10^6 m^3]$	GP: $\xi = 0.13, \alpha = 0.57, \kappa = -0.35$ PIII: $\mu = 1.00, \sigma = 1.48, \gamma = 3.01$
3	16	Duration [month] Severity $[*10^6 m^3]$	PIII: $\mu = 1.00, \sigma = 1.24, \gamma = 2.80$ PIII: $\mu = 1.00, \sigma = 1.35, \gamma = 2.97$

Table 4.3: Statistics of variables selected for each selected site

Site	Cycle [yr]	Location	Variable	Min	Mean	Max	STD
27	1.97	49.31N 107.7W	Duration [month]	1	5.42	22	6.03
			Severity $[*10^6 m^3]$	0.03	2.91	12.47	3.69
33	0.93	50.36N 99.91W	Duration [month]	1	5.26	36	6.35
			Severity $[*10^6 m^3]$	0.18	27.98	216.08	39.15
16	0.49	53.42N 117.57W	Duration [month]	1	3.25	14	2.87
			Severity $[*10^6 m^3]$	0.26	221.68	1185.07	278.98

4.3.2 Identification of dependence of variables, copula and determination of its parameter

The value of Kendall's τ and the parameter θ of each of the candidate Archimedean copula was calculated for one candidate site from each of the three clusters. The results are shown in Table 4.4.

Table 4.4: Copula parameter (θ) values of candidate sites and copulas

Site	τ	Clayton	Gumbel	Frank
27	0.56	2.57	2.29	7.01
33	0.61	3.19	2.59	8.33
16	0.49	1.94	1.97	5.60

4.3.3 Q-Q plots

The next step is to identify the most appropriate copula among the candidate copula families for each site of interest. For each of the candidate family of copula, the Q-Q plot of each candidate site is shown in Figures 4.1, 4.2, and 4.3. Superimposed on the plot is a robust linear regression fit of the order statistics of the two samples. This line is extrapolated out to the ends of the sample to help evaluate the linearity of the data.

Based on the results of the Q-Q plots, the copula families Gumbel, Gumbel, and Clayton were selected as the most appropriate copula for selected sites of interest of each cluster, respectively. This choice of best fitting copulas can be confirmed by checking Table 4.4. A lower copula parameter suggests a better fit. The generating function $\phi(t)$ with $t = u$ or v, value of a uniformly distributed variable varying from 0 to 1 for each copula can now be written using the explanations in subsections 4.2.2.1, 4.2.2.2, and 4.2.2.3. Table 4.5 shows the summary of copula analysis and generating function for each site.

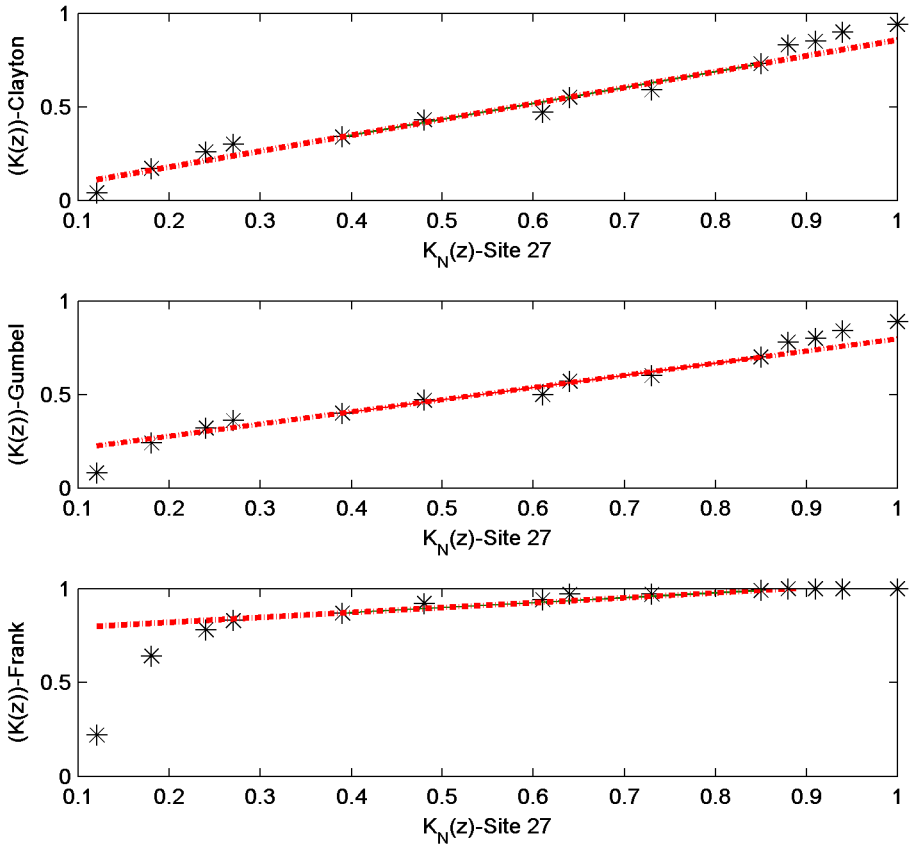


Figure 4.1: Q-Q plots of site 27

Table 4.5: Summary of copula analysis for three candidate sites

Site	Copula family	θ	$\phi(t)$
27	Gumbel	2.29	$(-\ln t)^{2.29}$
33	Gumbel	2.59	$(-\ln t)^{2.59}$
16	Clayton	1.94	$\ln \frac{1-1.94(1-t)}{t}$

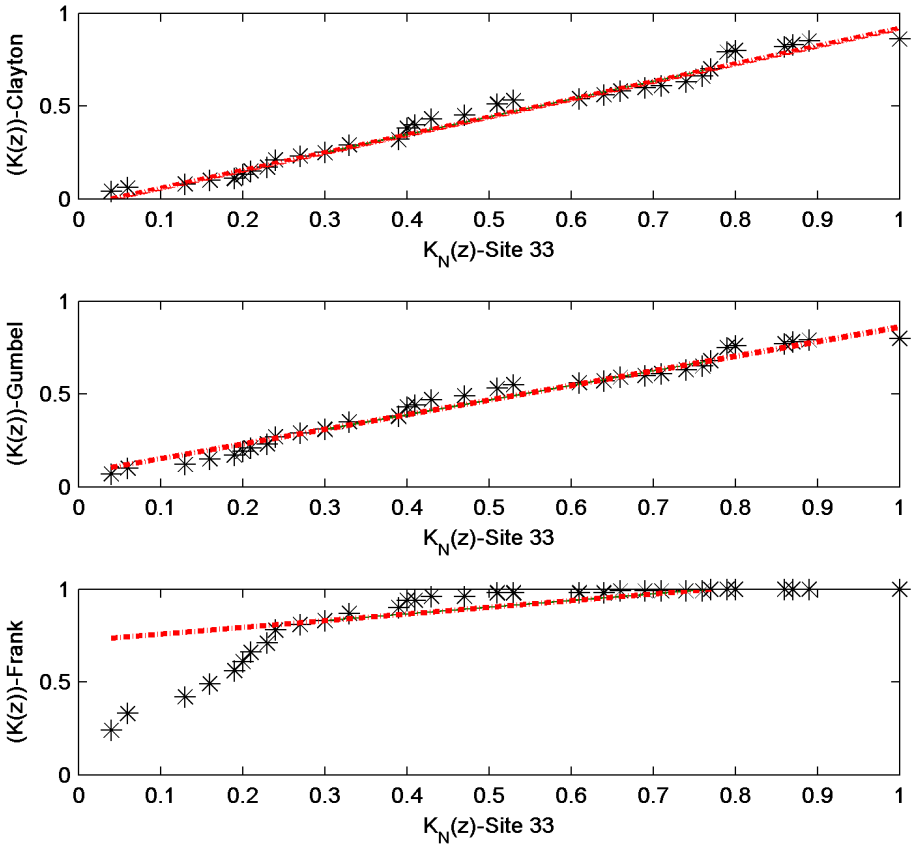


Figure 4.2: Q-Q plots of site 33

4.3.4 Determination of the joint probability distribution and joint return period based on Copula

The CDF and the joint return period $T(x, y)$ of candidate sites are shown in Figures 4.4, 4.5, and 4.6. The same figures also show joint return period contours of 5, 10, 25, 50, and 100 years for these sites. Note that when looking at the figures scaling of the severity and duration is different for each site. The upper end of each axes is the variable's quantile corresponding to 0.999 CDF of the best fitting marginal distribution. The joint return period was computed for selected return periods of 5,

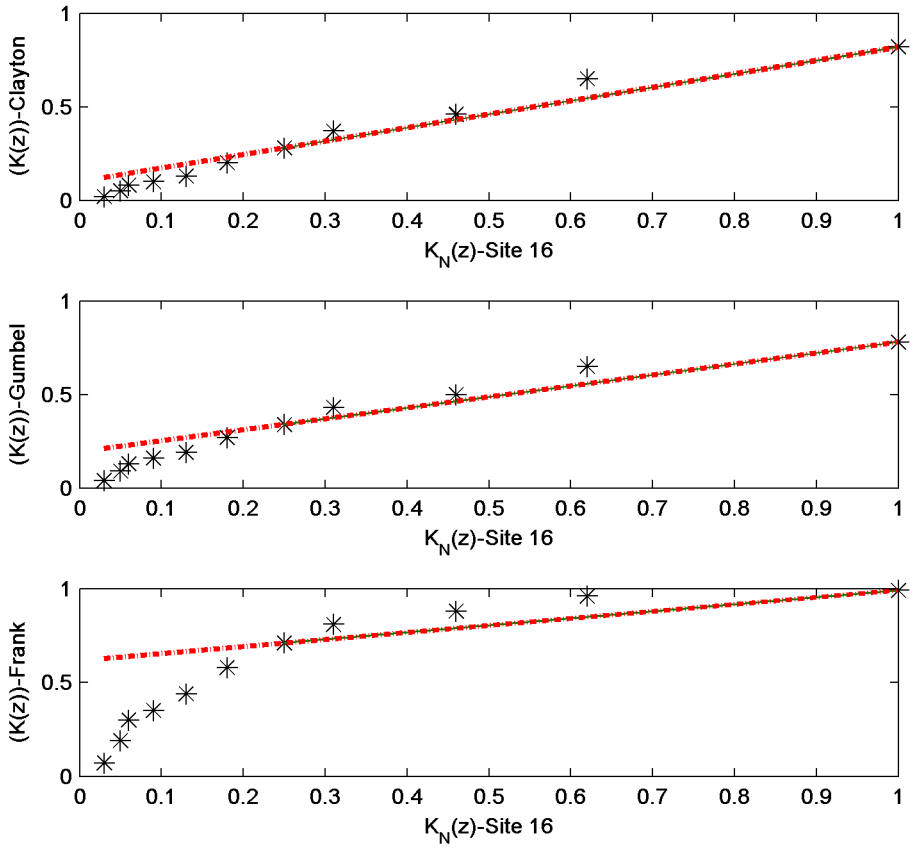
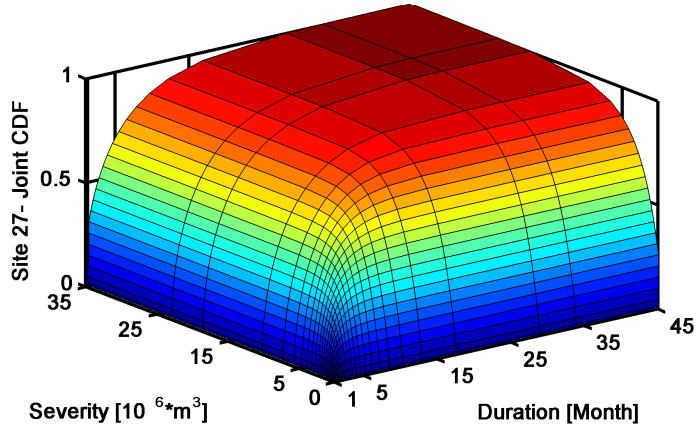


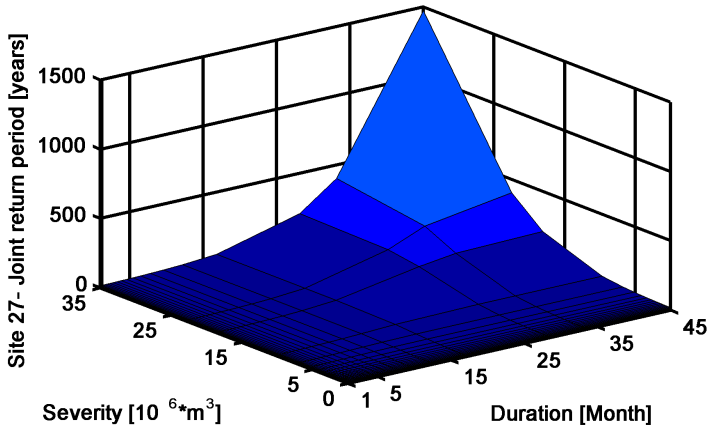
Figure 4.3: Q-Q plots of site site 16

10, 25, 50, and 100 years to show a better view of the results. These contours should be approaching zero in very higher end of the tail of distributions (higher than 0.999).

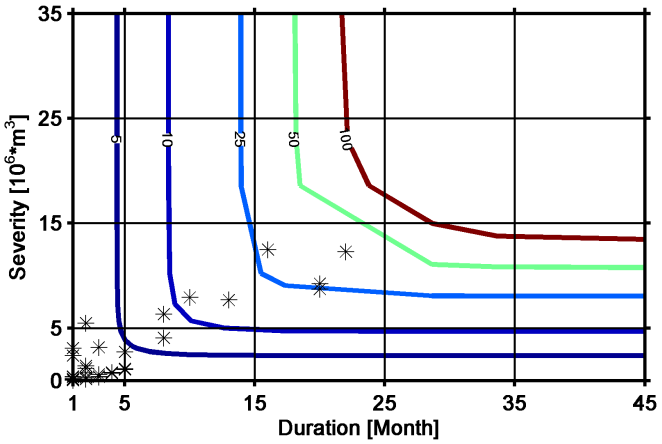
The results from contour plots (Figures 4.4(c), 4.5(c), and 4.6(c)) show that for a given return period, there may be several historical events with different combinations of severity and duration. Although it is generally perceived that the return period to be adapted for design purposes should be with regard to the worst case scenario in terms of historical events, Figure 4.4(c) is a good example that the longer drought is not necessarily the most severe one. Site 27 has a 22-month duration drought as the longest drought event, however, this event corresponds to a severity of $12.27 \times 10^6 m^3$



(a)

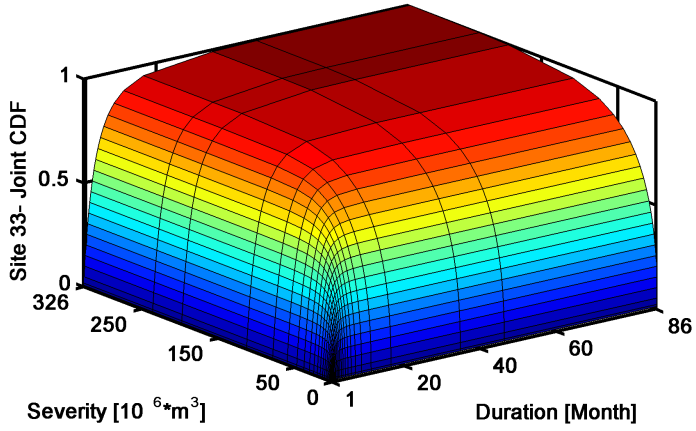


(b)

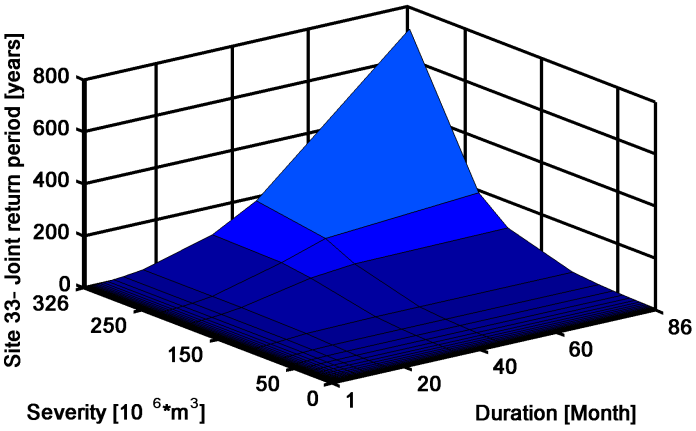


(c)

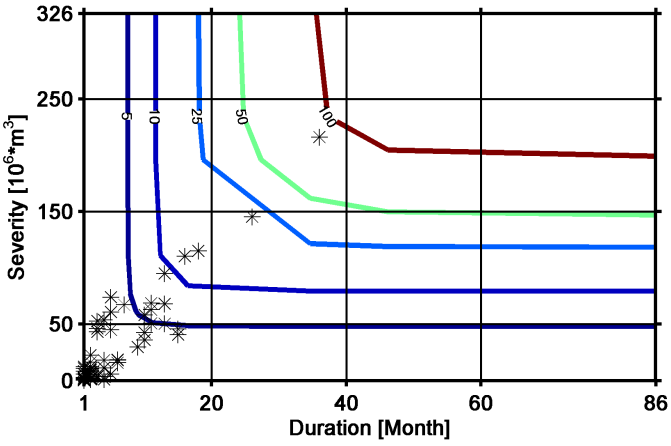
Figure 4.4: (a) Joint CDF, (b) joint return period, and (c) contour plot (stars represent observed events) of site 27 from cluster 1



(a)

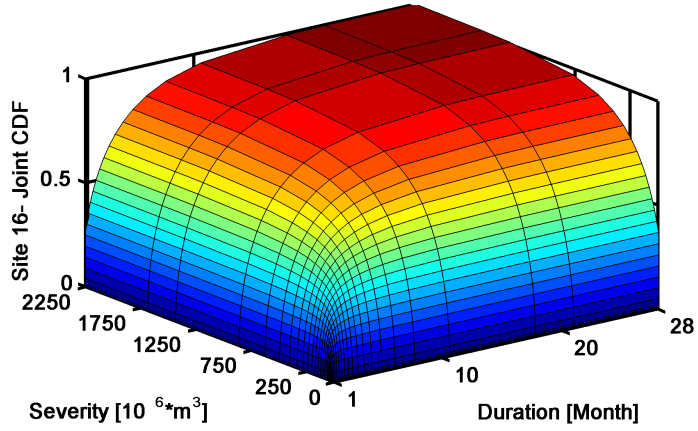


(b)

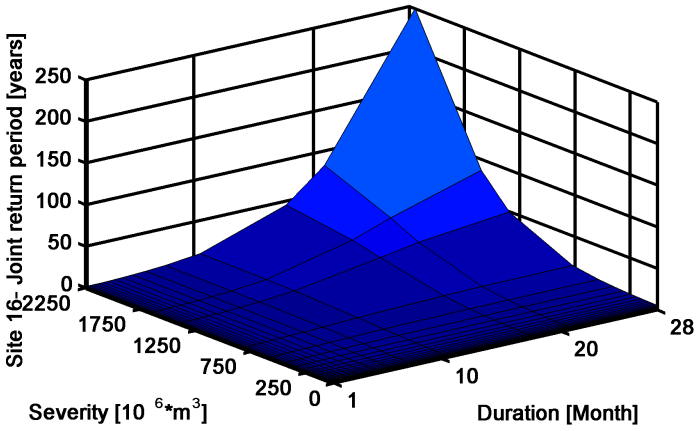


(c)

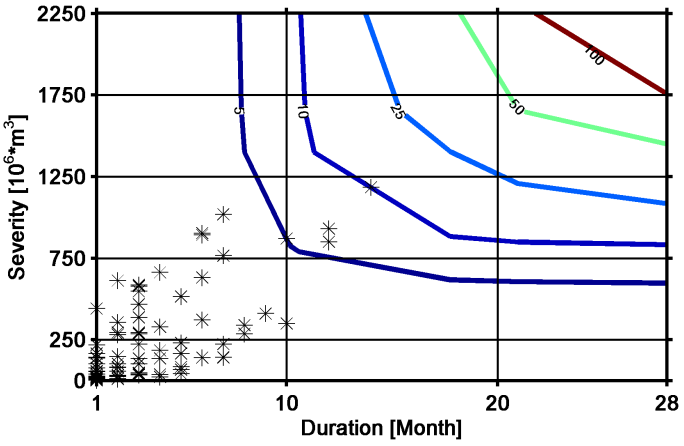
Figure 4.5: (a) Joint CDF, (b) joint return period, and (c) contour plot (stars represent observed events) of site 33 from cluster 2



(a)



(b)



(c)

Figure 4.6: (a) Joint CDF, (b) joint return period, and (c) contour plot (stars represent observed events) of site 16 from cluster 3

which is not the most severe observed drought. The duration corresponding to this most severe drought is 16 months. In general for all sites, the contours are to show that for a desired return period varying droughts can occur with respect to their severity and duration. Comparing the contour plots of the selected sites with each other, for the same return period and duration, drought severity is the most at site 16 and the least at site 27. Site 16 is a catchment in western Alberta which receives about 480 mm/yr mean annual precipitation and the catchment can experience shorter but more severe droughts. Site 27 is a catchment in southern Saskatchewan receiving about 388 mm/yr mean annual precipitation and the catchment can experience longer but less severe droughts. The more severe droughts occur in the humid regions due to highly fluctuating rainfall.

4.4 Conclusions

This chapter studied application of Gumbel, Clayton, and Frank families copula in bivariate drought frequency analysis. In each of selected sites, each of duration and severity variables were fitted to their best fitting parametric marginal distributions. The the joint CDF of the two variables given the best fitting copula parameter was constructed using a piece of Matlab code. The joint return period graphs and contours of selected return periods can be calculated from the joint CDF graphs' data. It should be noted that in general, drought frequency analysis results obtained by statistical analysis are inherently uncertain, since we can rarely be sure what the "correct" model is. However, when using regional L-moment and L-comoment algorithms, the data that satisfy all the assumptions and that underlie the index-event procedure including goodness-of-fit and when applying copula which allows fitting two different marginal distributions for each variable and takes care of the statistical dependance between

observations, the results are considered consistent. The copula method relaxes the restrictions of traditional bivariate frequency analysis by allowing each variable to fit different distribution other than the normal distribution. Bivariate contours for selected return periods showed that for any return period of interest, a range of droughts defined by pairs of severity and duration have been observed in the historical data.

Chapter 5

Nonparametric methods for Pooled Drought Frequency Analysis at Ungauged Sites

For many engineering projects, reliable at-site drought quantile estimation for desired return periods is essential. The problem is that in many cases there is a lack of at-site lengthy or reliable hydrological data, or there can be no observed data available (ungauged sites). To overcome this problem, *Hosking and Wallis* (1997) suggested a linear regression approach to relate a drought quantile of interest to a vector of catchment's physiographic, climatic and geomorphologic characteristics. The linear regression approach has been successfully applied in many cases, but has some disadvantages such as not fitting to the observed data very well or diverting from tails particularly for skewed data (*Haghighatjou et al.*, 2008). One of the new practices in the domain of drought frequency analysis in this work is the study of soft computing and heuristic techniques such as neural networks and machine

learning algorithms in quantile estimation of desired return period for an ungauged site. The performance of Radial Basis Function and Support Vector Machines is compared to the results from traditional statistical method of nonlinear regression. As well, the effect of regionalization on nonlinear regression is studied and compared with the other methods.

5.1 Introduction

In this Chapter, four approaches for estimating drought quantiles at ungauged sites at desired return periods are proposed. Two of these approaches include Radial Basis Function (RBF) and Support Vector Machines (SVMs) to provide intelligent non-parametric drought quantile estimation. The other two approaches are statistical methods of Nonlinear Regression (NLR) and Nonlinear Regression with Regionalization (NLR-R).

During the past decades there has been an emergence of applications of neural networks and other artificial intelligent systems in function estimation and regression analysis in different areas of engineering. These relatively new techniques provide an attractive alternative to the traditional statistical models such as linear regression. The capability of dealing with imprecision gives artificial intelligence great potential for hydrological analysis and water resources decision making (*Shu and Ouarda*, 2008). Artificial neural networks (ANNs) have been introduced in the domain of regional flood frequency analysis by *Shu and Burn* (2004a). To the author's knowledge, no work has addressed the application of ANNs in the area of regional drought frequency analysis.

Another powerful tool for solving problems in nonlinear classification and function estimation are SVMs. SVMs have led to many other recent developments in kernel

based learning methods in general and have been introduced within the context of statistical learning theory and structural risk minimization.

This chapter is organized as follows: in Section 5.2 a general introduction to RBF is presented. In Section 5.3 a description of SVMs for regression is reviewed. Statistical regression methods are reviewed in Section 5.4. Section 5.5 contains details related to the implementation of this study. In Section 5.6 the functionalities of all four methods are compared, and finally Section 5.7 provides the summary and conclusions of this work.

5.2 Radial Basis Functions (RBFs)

Application of RBF became popular in the mid 1980's due to their exact interpolation of a set of data points in a high-dimensional space (*Ghodsi and Schuurmans*, 2003). A radial basis network represents a special category of the Feed Forward (FF) neural network for stochastic approximation. The technique provides an interpolating function which passes through every data point. Advantages of RBF over Multilayer Perceptron (MLP) neural networks are:

- RBF trains faster;
- Each basis function can have its own width parameter ν_j;
- RBF is not suffering from local minima in the same way as Multi-Layer Perceptrons, at the same time it does not guarantee finding a global optimum;
- Suitable parameters can be chosen for the units of hidden layer

Disadvantages of RBF are:

- Selecting the appropriate number of basis functions;

- Change of input data changes the number of basis functions;
- Requires good coverage of the input space by radial basis functions.

A Radial Basis function is a real-valued function whose value depends only on the distance from the origin, so that $\phi(x) = \phi(\|x\|)$; or alternatively on the distance from some other point μ, called a center, so that $\phi(x, \mu) = \phi(\|x - \mu\|)$.

If a mapping from a n-dimensional input space $\vec{x_i}$, $(i = 1, ..., n)$ to a one-dimensional target value y, is desired where the input set consists of M input vectors, with corresponding targets y_j, $(j = 1, ..., M)$, an exact interpolation is achieved by introducing a set of M basis functions, one for each data vector, and then setting the weights for the linear combination of basis functions (*Ghodsi and Schuurmans*, 2003). There are several forms of basis function on RBF such as Gaussian, triangular, and univariate but the most common one is the Gaussian with the general form

$$\Phi_j(\vec{x}) = e^{-(\frac{\|x-\mu_j\|^2}{2\nu_j^2})} \tag{5.1}$$

where μ_j is the center of basis function; Φ_j is the Gaussian basis function; and ν_j is the bandwidth parameter and controls the smoothness of the interpolating function. The form for the RBF neural network mapping is

$$y_j(\vec{x}) = \sum_{(j=1)}^{M} w_j \Phi_j(x) + w_0 \tag{5.2}$$

where y_j is the output vector; w_j is the weighting vector; and w_0 is the bias parameter. The bias w_0 can be absorbed into the final summation by including an extra bias

function Φ_0 whose activation is set to 1. Therefore, the RBF, after absorbing the bias parameters into the weights, can be written in matrix notion as

$$Y = W\Phi \tag{5.3}$$

where Y is the matrix of output values; and W is the matrix of second-layer weights to be estimated. The basic architecture of RBF network structure consists of an input layer, a single hidden layer with a radial activation function and an output layer. Basis functions act like hidden units. Figure 5.1 shows a schematic representation of an RBF network.

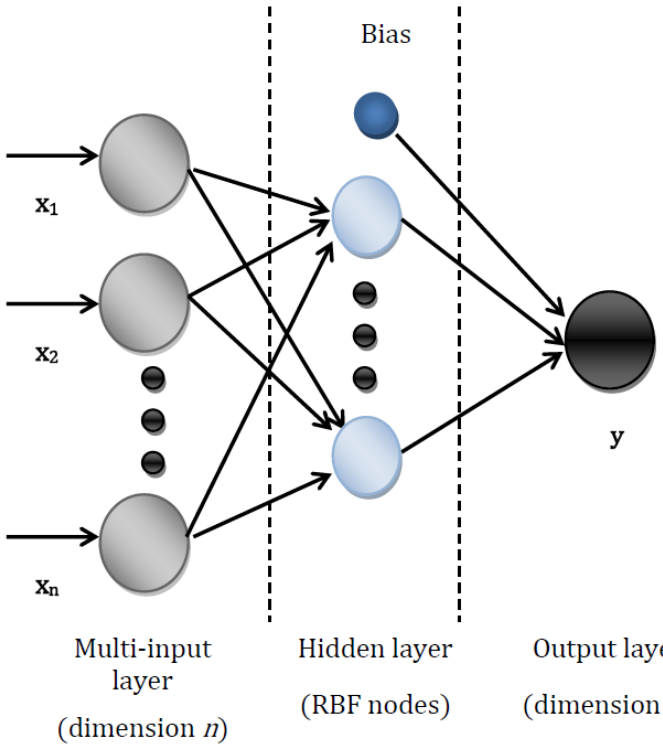


Figure 5.1: A graphical architecture of RBF network. An extra bias function whose outputs is fixed at 1 severs as the bias for each output unit

The centers and bandwidths can be determined during the training process. The approach includes modelling the input distribution as a Gaussian mixture and then

estimate the center and the width parameters of the Gaussian mixture components via Estimation-Maximization (EM) algorithm which is an unsupervised learning algorithm (*Bishop*, 1995). Equation (5.3) is a classical least squares estimation problem. In order to minimize $\|Y = W\Phi\|^2$, W must satisfy (*Karray and De Silva*, 2004)

$$W = (\Phi^T\Phi)^{-1}\Phi^T Y \tag{5.4}$$

In summary, training RBF networks proceeds through two steps:

1. The first step determines the first layer of weights in which the basis function parameters μ_j and ν_j are selected.

2. In the second step the basis functions are kept fixed while the second-layer weights are estimated via linear least squares

Thus, the first stage is an unsupervised method which is relatively fast, and the second stage requires the solution of a linear problem, which is also fast (*Ghodsi and Schuurmans*, 2003).

5.2.1 Overfitting and underfitting

One of the critical issues in using RBF networks is selecting the appropriate number of basis functions that show good performance on both training and testing data. As mentioned, a set of M training data vectors can be modelled exactly with M RBFs. Although such a model follows the training data perfectly, the model cannot represent features of unseen testing data. For an optimal training performance of the network, the hidden layer nodes should be optimized (*Karray and De Silva*, 2004). To achieve a sufficient number of basis functions, the difference between the training error (err) and the generalization (testing) error (Err) of hidden neurons must be

minimized. In practice, it is often observed that up to a certain point, the model error on testing data tend to decrease as the training error decreases. However, if one attempts to decrease the training error too far by increasing model complexity and number of hidden neurons, the testing error often can increase dramatically (*Ghodsi and Schuurmans*, 2003). The reason is that after a certain point, the model starts to overfit the training set which means that the model starts losing generality. For the case in which a new data point has been introduced to the trained model, the training error is an estimate of the expectation of the squared error on the training data, $E(\hat{y} - y)^2$

$$err = \sum_{j=1}^{N} (y - \hat{y})^2 \tag{5.5}$$

where N is the total number of training data sets; y is the target space; and $\hat{y}$ is the estimated target value. Generalization error is an estimate of mean squared error

$$Err = (\hat{f} - f)^2 \tag{5.6}$$

where $\hat{f}$ is the estimated model and f is the true model, and both are single values. This shows that err and Err do not demonstrate a linear relationship meaning that, a smaller training error does not necessarily result in a smaller testing error (*Ghodsi and Schuurmans*, 2003).

5.3 Support Vector Machine Regression (SVR)

One of the most recognized intelligent algorithms in machine learning is the Support Vector Machine (SVM) invented by Vladimir Vapnik and his coworkers in 1995 for tackling separation of two series of data points based on supervised learning (*Khan and Coulibaly*, 2006). *Vapnik* (2006) developed Support Vector Machine Regression

(SVM-R or SVR) from SVMs concept to overcome the shortcoming of neural networks. Advantages of Support Vector Machine - Regression (SVR) are:

- SVRs guarantee a global solution;
- Initial conditions do not change for different training data sizes.

In order to discuss the theory of SVR, an explanation on SVM theory is necessary to be reviewed.

5.3.1 Linear Support Vector Machine

Similar to other neural networks and fuzzy systems, SVMs are typical nonparametric classifiers, meaning that no primary knowledge is assumed for tackling the pattern classification problem. These systems acquire knowledge for classifying input data into one of the given classes through training using input-output pairs. The optimization technique in the SVMs consists of solving a linearly constrained solvable quadratic optimization problem which is guaranteed to find a unique, optimal, and global minimum for the error surface. SVR still contains all the main features that characterize maximum margin algorithm of SVMs. The simplest case of SVMs deal with linear machines on separable data. Nonlinear SVMs trained on non-separable data result in a very similar quadratic programming problem. Suppose that we want to classify some data points into two classes of positive and negative data points. This separation idea is based on hyperplane classifier or linear separability. For l training data points we label the data (*Vapnik*, 2006)

$$\{\mathbf{x}_i, y_i\}, \quad i = 1, ..., l, \quad y_i \in \{-1, 1\}, \mathbf{x}_i \in \Re^n \tag{5.7}$$

If there is a "separating hyperplane" which separates the positive from the negative examples, the points $\mathbf{x}$ which lie on the hyperplane satisfy

$$\mathbf{w}.\mathbf{x} + b = 0 \tag{5.8}$$

where $\mathbf{w}$ is the normal to the hyperplane; $|b|/(w)$ is the perpendicular distance from the hyperplane to the origin; and $\|\mathbf{w}\|$ is the Euclidean norm of $\mathbf{w}$. For the linearly separable case, the optimization involves finding a separating hyperplane with the largest margin (d_+, d_-). To formulate this, suppose that all the training data satisfy the following constraints

$$\mathbf{x}.\mathbf{w} + b \geq +1, \quad for \quad y_i = +1 \tag{5.9}$$

$$\mathbf{x}.\mathbf{w} + b \leq -1, \quad for \quad y_i = -1 \tag{5.10}$$

The points for which the equality in Equations (5.9) and (5.10) hold lie on the hyperplane $H_1 : \mathbf{x}.\mathbf{w} + b = 1$ with normal $\mathbf{w}$ and perpendicular distance from the origin $|1 - b|/\|\mathbf{w}\|$ and hyperplane $H_2 : \mathbf{x}.\mathbf{w} + b = -1$ with normal again $\mathbf{w}$ and perpendicular distance from the origin $|-1 - b|/\|\mathbf{w}\|$. Hence $d_+ = d_- = 1/\|\mathbf{w}\|$ and the margin is simply $2/\|\mathbf{w}\|$ (*Burges*, 1998). Thus the optimization function can be written as

$$minimize \quad \|\mathbf{w}\|^2 \tag{5.11}$$

Subject to

$$y_i(\mathbf{x}.\mathbf{w} + b) - 1 \geq 0 \quad \forall i \tag{5.12}$$

The solution for a typical two dimensional case which is linearly separable is shown in Figure 5.2. Those training points for which the equality in Equation (5.12) holds (those that lie on the hyperplanes H_1 and H_2), and whose removal would change the

solution found, are called support vectors and are indicated in Figure 5.2 (*Burges*, 1998). This is a convex, quadratic programming problem, in a convex set. Using La-

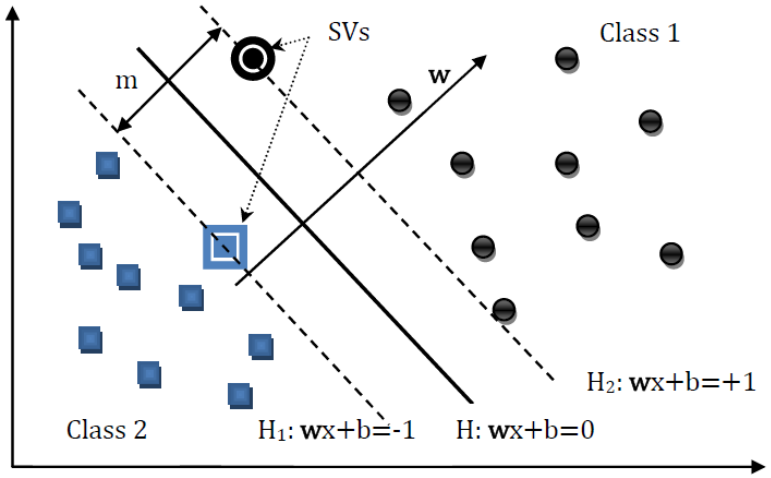


Figure 5.2: Linearly separating hyperplane for the separable case. Theoretically, the best hyperplane is to maximize the margin m. Support vectors are emphasized.

grange multipliers optimization technique, the minimization problem can be rewritten as minimizing L_p

$$L_p \equiv \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{l} \alpha_i y_i (\mathbf{w}.\mathbf{x}_i + b) + \sum_{i=1}^{l} \alpha_i \tag{5.13}$$

where α_i is the positive Lagrange multipliers ($i = 1, 2, ..., l$); $\mathbf{w}$ and b are to be minimized. The superiority of SVMs come from this specific formulation of a convex objective function with constraints. Since the function is solved using Lagrange multipliers, it guarantees the following:

- A global optimal solution exists that will be found;
- The result is a general solution avoiding overtraining;
- The solution is sparse and only a limited set of training points contribute to this solution.

5.3.1.1 Nonlinear Support Vector Machines

Solving the optimization problem of SVMs can look complicated when one tries to solve the above methods for the case where the decision function is not a linear function of the data. In fact, in many cases the surface separating the two classes is not linear. According to *Burges* (1998), a rather old "kernel trick" function can be used to accomplish this in a straightforward way. If the data were first mapped to some other n-dimensional Euclidean space H, using a mapping which we will call ϕ

$$\Phi : \Re^n \longmapsto H \tag{5.14}$$

then the training algorithm would only depend on the data through dot products in H, i.e., on functions of the form $\Phi(\mathbf{x}_i).\Phi(\mathbf{x}_j)$. Assuming that there is a kernel function K such that

$$K(\mathbf{x}_i, \mathbf{x}_j) = \Phi(\mathbf{x}_i).\Phi(\mathbf{x}_j) \tag{5.15}$$

we would only need to use K in the training algorithm, and would never need to explicitly even know what Φ is. One example of kernel function that can be used in the above equation is the radial basis Gaussian (RBF kernel)

$$K(\mathbf{x}_i, \mathbf{x}_j) = e^{-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\nu^2}} \tag{5.16}$$

where ν is the spread or variance parameter of Gaussian function. Since H is infinite dimensional, it would not be very easy to work with Φ explicitly, unless one replaces $\mathbf{x}_i.\mathbf{x}_j$ by $K(\mathbf{x}_i.\mathbf{x}_j)$ everywhere in the training algorithm. It must be noted that the only way that data appears in the training algorithm of Lagrange multipliers problem is in the form of dot products $\mathbf{x}_i.\mathbf{x}_j$.

5.3.1.2 Generalization for Support Vector Machine Regression (SVR)

In $\varepsilon - SV$ regression our goal is to find a function $f(x)$ that has at most ε deviation from the actual targets y_i for all the training data, and at the same time as flat as possible *Smola and Schölkopf* (2003). ε is the "soft margin" meaning that we do not care if the target estimate has an error less that ε. Analogously to the "soft margin" one can introduce slack variables ξ_i, ξ_i^* to cope with otherwise infeasible constraints of the optimization problem. The formulation stated by Vapnik (*Vapnik*, 1995):

$$minimize \quad R = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}(\xi_i + \xi_i^*) \tag{5.17}$$

$$s.t. = \begin{cases} y_i(w, x_i) - b \leq \varepsilon + \xi_i \\ (w, x_i) + b - y_i \leq \varepsilon + \xi_i^* \\ \xi_i + \xi_i^* \geq 0 \end{cases} \tag{5.18}$$

where C is the positive constant; l is the number of training data sets; ξ_i is the slack variables as well as ξ_i^*; and ε is the bias. Figure 5.3 illustrates the situation graphically. Deviations are penalized in a linear fashion. Similar to SVMs, the final

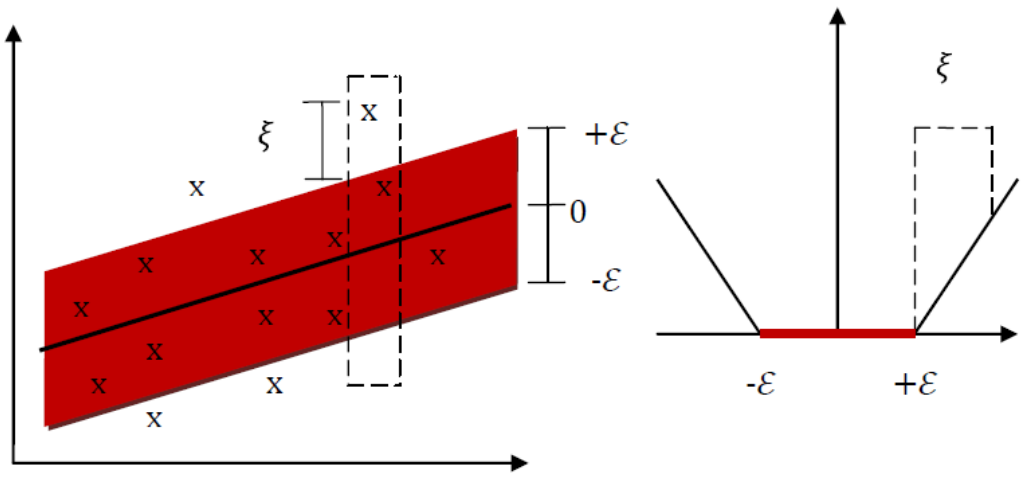


Figure 5.3: Soft margin loss setting for a linear SVM Regression.

goal is to minimize the norm $\|w\|^2$ and b. All the considerations of the previous sections hold, since still there is a linear separation but in a different space (*Smola and Schölkopf*, 2003).

One of the advantages of SVMs, and SVR as the part of it, is that it can be used to avoid difficulties of using linear functions in the high dimensional feature space and optimization problem is transformed into dual convex quadratic programmes. In regression case the loss function is used to penalize errors that are greater than a threshold ξ. Such loss functions usually lead to the sparse representation of the decision rule, giving significant algorithmic and representational advantages (*Burges*, 1998). A good tutorial on detailed calculations of SVR can be found in *Smola and Schölkopf* (2003).

5.4 Nonlinear Regression

Nonlinear regression is one of the common approaches used for obtaining regional estimates (widely used for flood quantile estimation) (*Shu and Ouarda*, 2008). In nonlinear regression, quantiles can be found as a function of site physiographical and other characteristics (*Shu and Ouarda*, 2008). Therefore, besides studying RBF and SVR models, in this chapter two other approaches are studied for quantile estimation. These include: nonlinear regression and nonlinear regression with regionalization. Results from RBF and SVR will be compared to results obtained from nonlinear regression approaches.

In nonlinear regression approach, the relationship between the drought quantile S_T and the catchment characteristics are assumed to be the power form function which has the following form (*Shu and Ouarda*, 2008)

$$S_T = a x_1^{\theta_1} x_2^{\theta_2} x_3^{\theta_3} ... x_n^{\theta_n} \tag{5.19}$$

where S_T is the model quantile; a is the multiplicative error term; x_i is the model characteristics; and n is the number of catchment characteristics. Solving Equation (5.19) can be achieved using an affine regression technique which requires linearizing the power-form model by a natural logarithmic transformation of quantile data. We tackle the problem by calculating the least squares fit of y_i on matrix of $\mathbf{x}$ and by solving the affine model (a linear model plus a constant)

$$y = \mathbf{x}\beta + \varepsilon \tag{5.20}$$

for β where y is the $1 \times N$ vector of natural log of observations; $\mathbf{x}$ is the $N \times l$ matrix of regressors; β is the $l \times 1$ vector of parameters; and ε is the $N \times 1$ vector of random disturbances ($\varepsilon : \mathbf{N}(0, \sigma^2\mathbf{I})$). Using an affine regression approach a value of $\varepsilon = 1$ is added to logarithmic n-dimensional input vector of $\mathbf{x}_{l\times 1}$. The natural log of drought quantile for each site then can be calculated as dot product of $y_i = (x_i \times \beta)$ where y_i is the natural log of output.

5.4.1 Nonlinear regression with regionalization

The regionalization step, the Fuzzy C-Means (FCM) clustering algorithm, is first used to identify the hydrological clusters described in Chapter 3, then the normalized weights of sites in final corrected clusters based on severity variable were used in the NLR method described to obtain drought quantile estimates (Table 3.3). Using FCM, each site in the corrected cluster has a normalized membership weight $w \in [0 \quad 1]$ value while taking a zero membership into any other cluster which it does not belong to after clusters are corrected. Then the NLR method described in previous section

was used to obtain the quantile estimates. The objective function is sum of weighted squared residuals

$$f(\mathbf{x}_0) = \sum_{k=1}^{K} (\mathbf{x}_0^\top \cdot \beta_k) \cdot w_{0,k}^\top \tag{5.21}$$

where k is the cluster number; $\mathbf{x}_0$ is the $1 \times l$ input target vector; β_k is the $l \times 1$ regression coefficient in the presence of a weight factor w; and $w_{0,k}$ is the normalized weight value of input target vector into cluster k.

5.5 Application

5.5.1 Study Area

The introduced regression methods were applied to the same hydrometric station network presented in Chapter 4. Six types of characteristics, physiographical, meteorological, and hydrological were selected as input vectors of each site including

- Drainage area (DA) [km^2]
- Mean elevation (ME) [m]
- Mean annual precipitation (MAP) [mm/yr]
- Mean daily maximum temperature (MDMT) [$^\circ C$]
- Mean annual evapotranspiration (MAET) [mm/yr]
- Mean runoff (MRO) [mm/yr]

Summary of statistics of these characteristics is shown in Table 4.1. The map of Canada showing the selected 36 hydrologic sites is presented in Figure 3.1. For each of the 36 sites the most appropriate statistical distribution was identified using a parametric approach to the historical records and the equivalent at-site drought

quantiles for the number of different return periods were obtained. Table 5.1 shows the summary of severity statistics of droughts for three selected return periods of 5, 10, and 50 years. Figure 5.4 shows the scatter plots between the quantiles and the

Table 5.1: Statistics of the study data

Severity for assigned return period	Min	Mean	Max	STD
S5 $[10^6.m^3]$	0.56	79.10	583.2	131.10
S10 $[10^6.m^3]$	0.78	120.20	811.95	190.40
S50 $[10^6.m^3]$	1.26	222.00	1351.50	339.50

variables described above.

5.5.2 Evaluation method

The evaluation method was adapted from *Shu and Ouarda* (2008). The performance of each drought frequency analysis model is evaluated based on the following indices

$$NASH = 1 - \frac{\sum_{i=1}^{n}(q_i - \hat{q}_i)^2}{\sum_{i=1}^{n}(q_i - \bar{q})^2} \tag{5.22}$$

The NASH criterion provides overall assessment of the quality of estimation. Models with NASH values close to 0.8 are generally acceptable, while models with NASH values close to 1 are deemed to produce near perfect estimation (*Shu and Ouarda*, 2008). The Root Mean Squared Error (RMSE)

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(q_i - \hat{q}_i)^2} \tag{5.23}$$

and the mean BIAS:

$$BIAS = \frac{1}{N}\sum_{i=1}^{N}(q_i - \hat{q}_i) \tag{5.24}$$

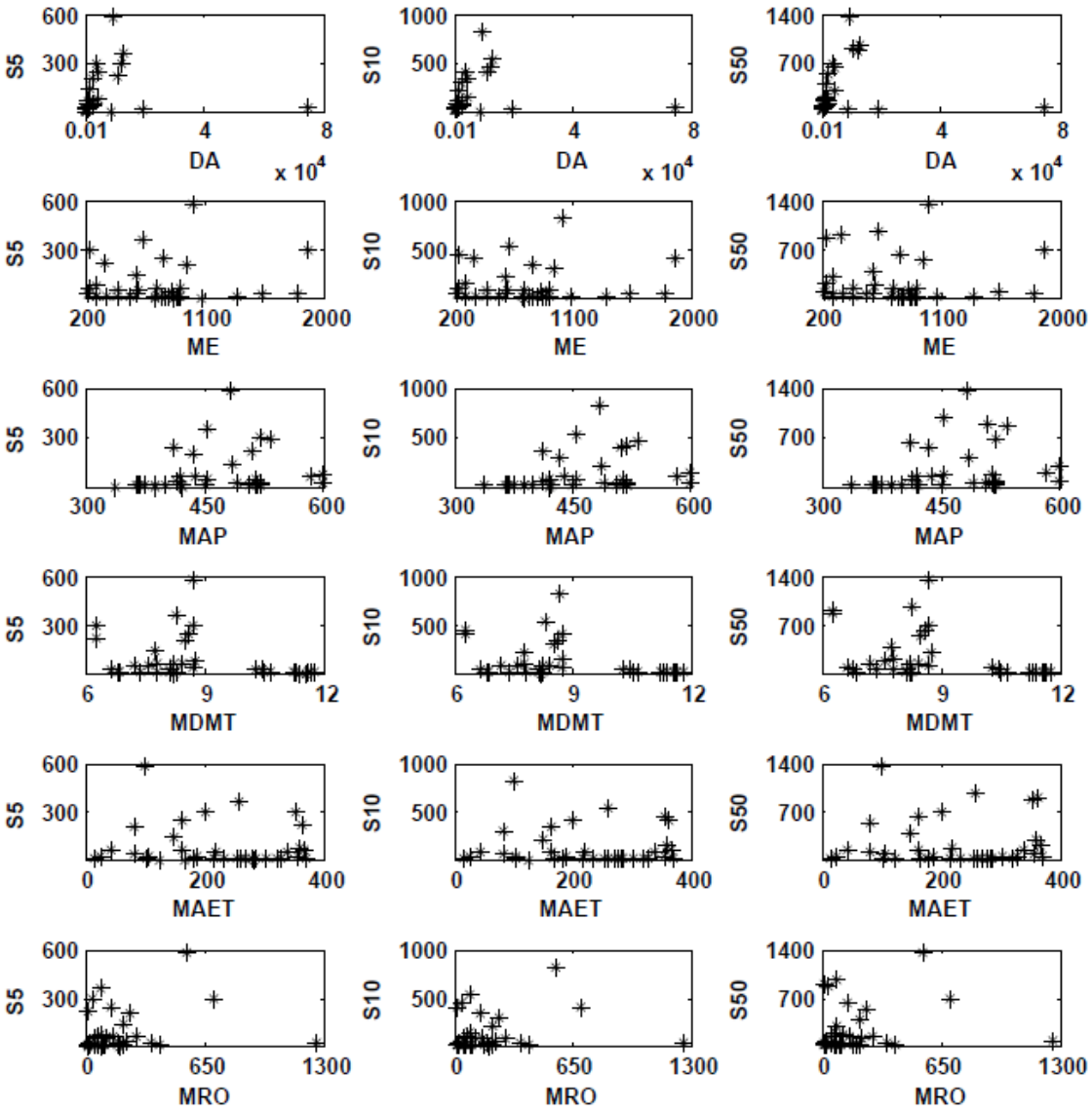


Figure 5.4: Scatter plots of site characteristics and drought severity quantiles. Unit of severity is $10^6.m^3$. DA: Drainage Area; ME: Mean Elevation; MAP: Mean Annual Precipitation; MDMT: Mean Daily Maximum Temperature; MAET: Mean Annual Evapotranspiration; and MRO: Mean Run off.

where N is the total number of sites used in modelling; q_i is the at-site estimation for site i; $\hat{q}_i$ is the quantile estimation obtained from modelling; and $\bar{q}$ is the mean of at-site estimation.

5.5.3 Experiment design

To assess the model performance on quantile estimation for desired return periods, a cross-validation (leave one out) procedure was used. In this procedure, for each catchment in the study area, its drought records were temporarily removed from the database, thus it was assumed to be "ungauged". Then each regional drought frequency analysis model was calibrated using data of the remaining sites. The estimated quntiles were obtained using the calibrated model. They were then compared against their corresponding at-site values.

The training input vectors of site characteristics were transformed into natural logarithmic scale. This was done to make training of neural network easier. The output quantiles were also transformed into natural logarithmic scale but only for NLR and NLR-R methods to avoid getting negative quantiles as testing outputs. Obviously, the output in testing NLR and NLR-R were then transformed exponentially.

When applying RBF networks, the center and spread of radial basis functions were evaluated for a mixture of Gaussians using the EM algorithm. To achieve optimum architecture for the network, the training and testing were performed for a set of number of hidden neurons varying from 1 to 16. A number of neurons in the hidden layer which reduces the difference between the MSE of training and the MSE of testing was selected for the network design.

For function approximation using SVR, a type RBF kernel was selected. To minimize the training error a grid search of $exp(C)$ and $exp(\sigma^2)$ is used within an

assigned range. The pair (2.3410 1.6841) are the optimal result in a 10×10 search shown in Figure 5.5.

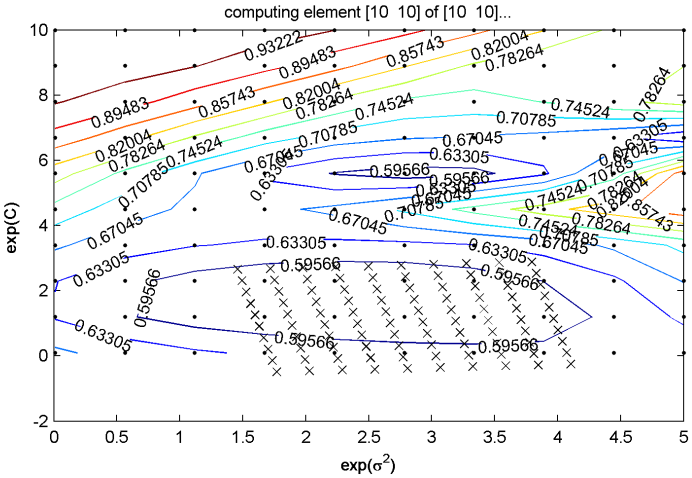


Figure 5.5: Minimizing the training error in SVR by a grid search

.

The three methods of RBF, SVMs, and NLR treat the entire study area as one hydrological region. Only NLR-R approach reflects the effect of regionlization in quantile estimation.

5.6 Results and Discussion

Figure 5.6 shows a sample for optimizing the number of hidden nodes among 1 to 16 nodes. In this figure, once the hidden nodes exceed number 8, the overfitting of the training data set begins. In this study, a model with 8, 8, and 9 hidden neurons for quantiles of S5, S10, and S50, respectively were found as the optimum number of hidden neurons. This shows the fact that design of RBF networks has to change with the change in the input data. It should be noted that one of the qualities of RBF is that the algorithm trains over time and since it is not designed to find the global optimum, every time it might give a different output. So running the test several times to see a consistency of certain output is important. The results regression of

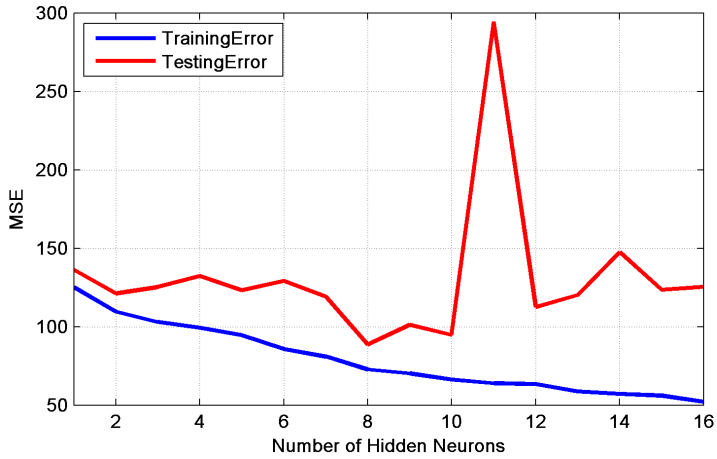


Figure 5.6: The best number of hidden nodes is when the MSE between the training error and testing error is minimized

Table 5.2: Regression results using cross-validation

	Variable	RBF	SVM-R	NLR	NLR-R
NASH	S5	0.530	0.146	-0.558	0.422
	S10	0.500	0.172	-0.550	0.437
	S50	0.477	0.207	-0.542	0.391
RMSE [$\times 10^6$ m^3]	S5	88.595	119.445	161.344	98.243
	S10	132.782	170.799	233.758	140.819
	S50	242.040	297.947	415.593	261.253
BIAS [$\times 10^6$ m^3]	S5	-1.172	-2.879	21.709	31.135
	S10	8.875	-4.263	33.042	54.644
	S50	13.632	-7.693	62.417	114.526

quantiles using RBF, SVR, NLR, and NLR-R approaches are presented in form of the indices suggested in *Shu and Ouarda* (2008) in Table 5.2.

The NASH value is a better indicator of the precision of models in function estimation. RBF model shows the best output while NLR-R shows similar results while the results of NLR are totally unacceptable.

RMSE shows the prediction accuracy of a model in an absolute scale (*Shu and Ouarda*, 2008) so the better values should be closer to zero. The RMSE values

computed by RBF and NLR-R are the lowest meaning that they have the highest accuracy of quantile estimation.

The magnitude of systematic overestimation or underestimation of a model is evaluated using the BIAS index. The results indicate that in general SVR tends to overestimate the targets especially at the lower quantiles. The result for BIAS suggest that SVR has a tendency to present reliable estimation. The predicted values (outputs) of all four models for drought severity quanties (S5, S10, S50) are in Figures 5.7 through 5.10. The regression coefficient R^2 is shown on each figure representing the linear correlation between observed values and predicted values.

In general, regression models are known to have a good descriptive interpolation ability but a limited predictive capacity (extrapolation). To compare the extrapolation ability of the models, the catchment located at the outmost part of the sample can be considered. We can observe that only RBF can have an acceptable prediction quality while all other three methods tend to underestimate the quantiles for all three return periods. Therefore, among all the four models RBF showed the closest extrapolation ability.

Although RBF appeared to have a better response over the other methods, the overall function approximation qualities of RBF, SVMs, and NLR-R may be improved. Two suggested approaches to improve the results are: (1) increase the number of training input; (2) decrease the dimensionality of training vectors. The lower RMSE values of NLR-R indicate that regionalization improves the model performance.

5.7 Conclusions and Summary

In this chapter, a new set of work based on using RBF and SVR and NLR-R for drought quantile estimation at ungauged sites was introduced and studied. Both RBF

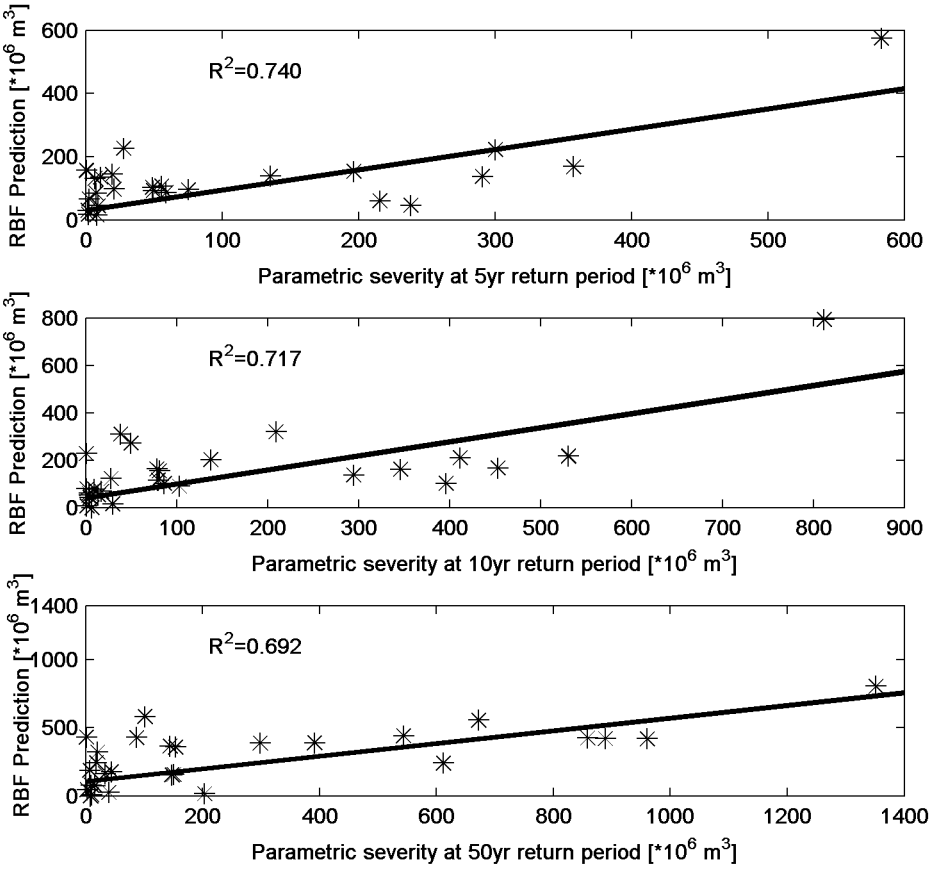


Figure 5.7: Cross validation estimation using radial basis function

and NLR-R approaches showed better estimations comparing to SVR and NLR. The results of RBF, NLR-R, and SVR approaches can be improved either by increasing the training input data or decreasing the dimensionality of each site (i.e., 6 in this set of analysis). Use of intelligent algorithms requires a long series of training data. This experiment shows that 36 sites and their statistical information can not be sufficient for getting satisfactory prediction in all cases. However, the new approaches still did better than the traditional NLR regression method. The two algorithms RBF and SVMs are strong in terms of initialization and unlike ANNs which may require several rounds of random selection, the initialization of a RBF and SVR networks

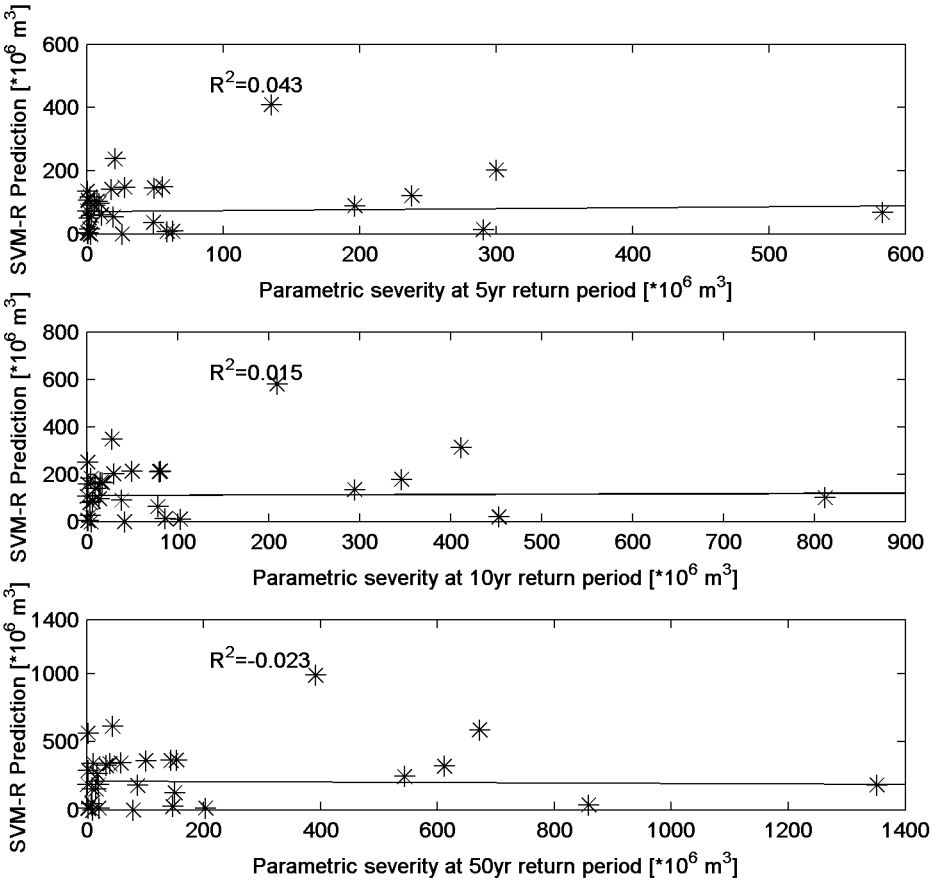


Figure 5.8: Cross validation estimation using support vector regression

can be performed using the one pass subtractive clustering algorithm, and a bivariate grid search, respectively. Also regionalization is shown to be worthwhile.

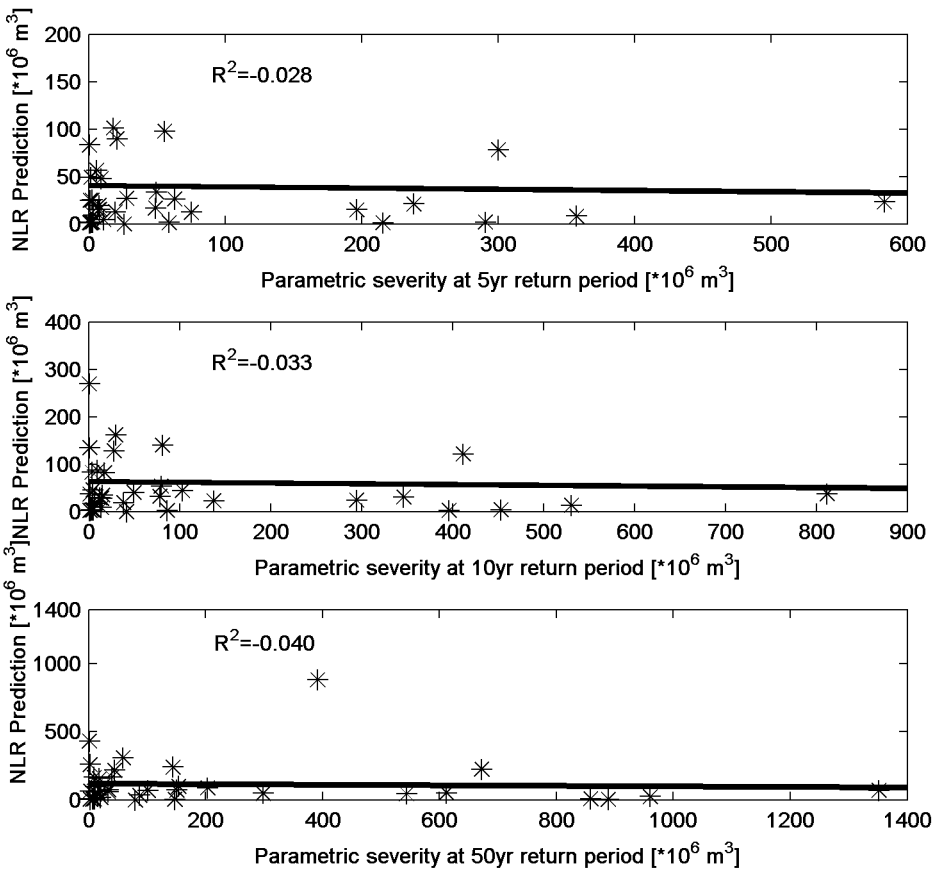


Figure 5.9: Cross validation estimation using nonlinear regression

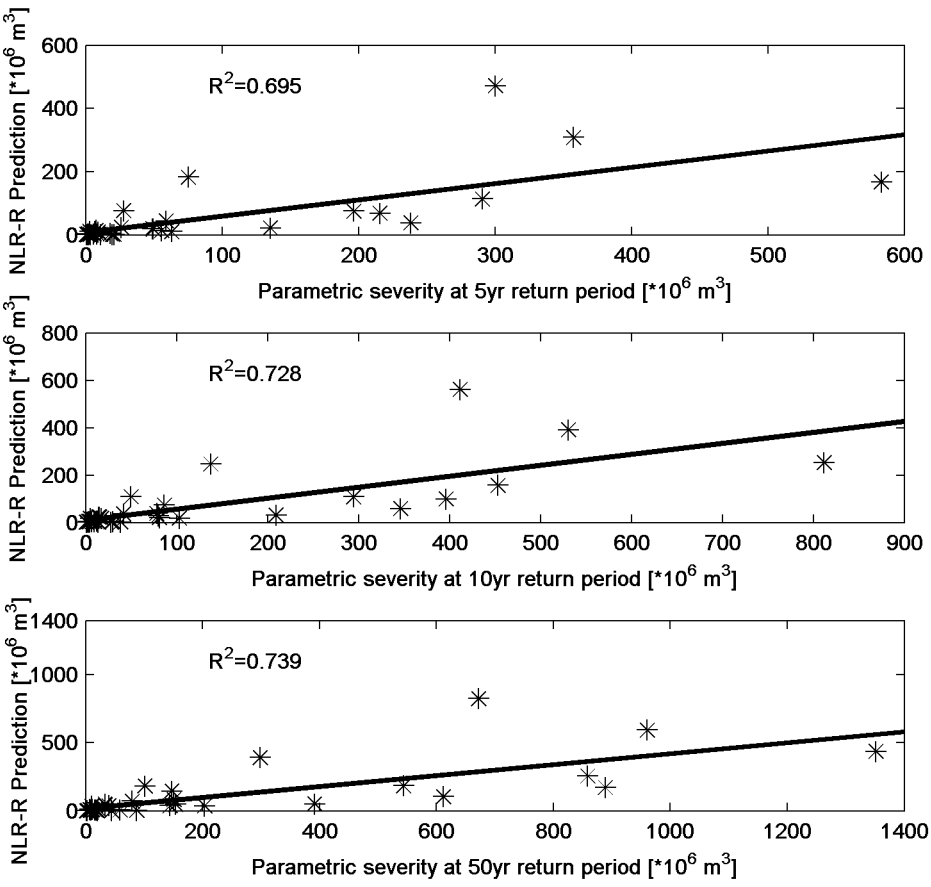


Figure 5.10: Cross validation estimation using nonlinear regression with regionalization

Chapter 6

Conclusions

This thesis covered several issues related to drought frequency analysis and developed different approaches to solve the frequency analysis issues.

Regionalization of sites for drought frequency analysis can result in different final regions depending on the application of a univariate or bivariate homogeneity approach. The major reason for this difference is the fact that univariate homogeneity and discordancy criteria choose only one variable at a time for analysis. A site which can be recognized as discordant or heterogeneous based on L-moments of the duration data may not necessarily be discordant or heterogeneous when analyzing L-moments of the severity data. The possible solution to this problem is to consider drought as a univariate phenomenon or apply the bivariate L-comoment approach to recognize joint heterogeneity and joint discordancy indexes for two variables of severity and duration. Adjustment of the clusters that are originally formed using FCM algorithm might have the consequence that some sites are not included in any final clusters as their inclusion in any set of clusters will either recognize them as discordant or

increase the heterogeneity. This might not be a problem in principle but the problem arises when the site of interest is the site that can not find a final home. The solution to this can be in increasing the number of total desired clusters. As the number of clusters increases to an acceptable extent, the chance that all selected sites find a final home increases. The cluster's heterogeneity is also expected to decrease with an increase of the number of clusters. The model presented in this study is a useful tool for illustrating the advantages of FCM clustering and bivariate homogeneity and discordancy tests in regional drought frequency analysis. In summary the advantages of using soft clustering algorithm of FCM are:

1. The algorithm classifies the object automatically based only on the criteria (i.e. minimum distance to the centroid). The learning process is unsupervised learning.

2. The convergence period and the calculation time are extremely short.

3. Partial membership in FCM algorithm provides a good asset for evaluation of each catchment.

The performance of FCM approach can be influenced by several factors such as the size of the region, the site's record length, and the degree of regional heterogeneity. In any case, univariate tests can give a false indication of the regions in bivariate drought frequency analysis. The bivariate homogeneity and discordancy approach is a bivariate version of L-moment approach and can be effectively used to model drought events described by their duration and severity.

Each drought event characterized by drought severity and duration was separately modelled using a parametric probability distribution function. A copula function was employed to link the fitted models and to construct a joint distribution function of drought severity and duration. Bivariate drought frequency distributions can be

developed using a copula method without assuming the two variables fit the same form of marginal distributions. Results from bivariate copula contour plots show that although it is generally perceived that the return period to be chosen for decision purposes should be with regard to the worst case scenario in terms of historical events, this study shows a good example that a longer drought is not necessarily the most severe one. The contours show that for a desired return period, varying droughts can occur with respect to their severity and duration. Comparing the contour plots of the selected sites with each other, for the same return period and duration, drought severity is the greatest in a catchment with the highest amount of mean annual precipitation. This shows that more severe droughts occur in the humid regions due to highly fluctuating rainfall. Also, longer droughts in terms of duration occur in regions with lower mean annual precipitation which causes accumulated water resource deficits.

The methodologies for using NLR-R and RBF for nonparametric drought quantile estimation at ungauged sites was introduced and studied. Both RBF and NLR-R approaches showed better estimations comparing to SVR and NLR. The results of RBF and SVR can be improved if the dimensionality and input vectors are adjusted compared with the number of training sets of inputs. Regionalization of sites was shown to have a great influence on improving the result of regression.

The contributions of this thesis in the domain of drought frequency analysis can be summarized as followed:

- An algorithm was developed in Matlab using fuzzy membership qualities that identifies homogeneous regions, thereby speeding-up a process which is considered both difficult and which requires the greatest amount of subjective judgement

- The Regional drought frequency analysis algorithm developed has the flexibility to accept differing numbers of regions and different return periods as inputs
- Development of an approach for drought frequency analysis that is statistically efficient and reasonably straight forward to implement
- Application of bivariate L-comoments in creating effective regions for bivariate drought frequency analysis
- Development of a method in which bivariate regional frequency analysis of droughts can be improved through the use of bivariate L-comoments and copulas
- Demonstrate the importance and application of various soft computing techniques in drought frequency analysis, including:
  - Radial Basis Functions
  - Support Vector Machine Regression, and
  - Nonlinear regression with FCMs regionalization

It is hoped that this material provides a comprehensive review, a routine, and a source of reference for bivariate drought frequency analysis for researchers in any sector of the world that are interested in looking into the issues of drought frequency analysis using stochastic and soft computing techniques.

6.1 Future Work

The algorithm developed in this study for performing a FCM clustering and then adjusting the initial clusters to create final clusters to meet various hydrological constraints is a unique approach for regional frequency analysis and one of the contri-

butions of this study. The current algorithm is capable of determining a clustering of sites in most cases, but more research is required for the algorithm to be capable of determining the final clusters which satisfy constraints of homogeneity, sufficient size, and lack of discordancy for *all* possible input data scenarios.

Developing models on trivariate copulas for drought frequency analysis which include severity, duration, and magnitude as an extension of bivariate copulas and comparing the results from trivariate frequency analysis with bivariate frequency of drought is a subject of interest for potential future research. In addition, combining a model for bivariate frequency analysis using copula with a physically based model to improve the final clusters would be an interesting research topic.

RBF and SVMs have been applied for the first time to drought quantile estimation within this thesis, and has been shown to be very useful. Applying nonparametric approaches for a larger set of data in order to evaluate a better quantile estimation would be very useful in giving a better idea of which learning algorithms and statistical approaches provide a better response.

Investigation of drought using both physically based hydrology models and stochastic hydrology methods for regional drought frequency analysis can produce a more comprehensive analysis of drought frequency. This approach can be applied to the effects of climate change on nonstationarity and drought to develop an advanced drought frequency model that can accommodate climate change factors.

References

Acreman, M. C., and S. E. Wiltshire (1989), The regions are dead: Long live the regions. methods for identifying and dispensing with regions for flood frequency analysis, in *FRIENDS in Hydrology*, edited by L. Roald, K. Nordseth, and K. Hassel, pp. 175–188, IAHS Publl.187, England.

Adamowski, K., and W. Feluch (1990), Nonparametric flood frequency analysis with historical information, *Journal of Hydraulic Enginering*, *116*(8), 1035–1047.

Ayvaza, M. T., H. Karahana, and M. M. Aral (2007), Aquifer parameter and zone structure estimation using kernel-based fuzzy c-means clustering and genetic algorithm, *Journal of Hydrology*, *343*(3-4), 240.

Beersma, J. J., and T. A. Buishand (2004), Joint probability of precipitation and discharge deficits in the netherlands, *Water Resources Research*, *40*(12).

Bishop, C. (1995), *Neural networks for pattern recognition*, Oxford University Press, Oxford.

Burges, C. (1998), A tutorial on support vector machnices for pattern recognition, *Data Mining and Knowledge Discovery*, *2*, 121–167.

Burn, D. H. (1990a), Evaluation of regional flood frequency analysis with a region of influence approach, *Water Resources Research*, *26*(10), 2257–2265.

Burn, D. H. (1990b), An appraisal of the 'regiona of influence' approach to flood frequency analysis, *Hydrological Science*, *2*(35), 149–165.

Burn, D. H., and W. J. DeWit (1996), Spatial characterization of drought events using synthtic hydrology, *Canadian Journal of Civil Engineering*, *23*, 1231–1240.

Burn, D. H., and N. K. Goel (2000), The formation of groups for regional flood frequency analysis, *Hydrological Sciences Journal*, *45*(1), 97–112.

Burn, D. H., and M. A. Hag Elnur (2002), Detection of hydrologic trends and varaiability, *Journal of Hdyrology*, *255*, 107–122.

Burn, D. H., Z. Zrinji, and M. Kowalchuk (1997), Regionalization of catchments for regional flood requency analysis, *Journal of Hydrologic Engineering*, *2*(2), 76–82.

Burn, D. H., J. Wychreschuk, and D. V. Bonin (2004), An integrated approach to the estimation of streamflow drought quantiles, *Hydrological Sciences Journal*, *49*(6), 1011–1024.

Burton, I., R. W. Kates, and G. F. White (1978), *The environment as hazard*, 240 pp., Oxford University Press, New York.

Changa, F.-J., M.-J. Tsaia, W.-P. Tsaia, and E. E. Herricksb (2008), Assessing the ecological hydrology of natural flow conditions in Taiwan, *Journal of Hydrology*, *354*(1-4), 75–89.

Chebana, F., and T. Ouarda (2007), Multivariate L-moment homogeneity test, *Water Resources Research*, *43*(W08406), 1–14.

Chow, V. T., D. R. Maidment, and L. W. Mays (1988), *Applied Hydrology*, Water resources and environmental engineering, 572 pp., McGraw-Hill.

Clausen, B., and C. P. Pearson (1995), Regional frequency analysis of annual maximum streamflow drought, *Journal of Hydrology*, *173*, 111–130.

Cunderlik, J. M., and D. H. Burn (2003), Non-stationary pooled frequency analysis, *Journal of Hydrology*, *276*, 210–223.

de Michele, C., and G. Salvadori (2003), A generalized pareto intensity-duration model of storm rainfall exploiting 2-copulas, *Journal of Geophysical Research*, *108*(D2).

de Michele, C., G. Salvadori, M. Canossi, A. Petaccia, and R. Rosso (2005), Bivariate statistical approach to check adequacy of dam spillway, *Journal of Hydrologic Engineering*, *10*(1), 50–57.

Douglas, E. M., R. M. Vogel, and C. N. Kroll (2000), Trends in floods and low flows in the United States: impact of spatial correlation, *Journal of Hydrology*, *240*, 90–105.

Genest, C., and L. Rivest (1993), Statistical inference procedures for bivariate Archimedean copulas, *Journal of American Statistical Association*, *88*, 1034–1043.

Ghodsi, A., and D. Schuurmans (2003), Automatic basis selection techniques for RBF networks, *Journal of Neural Networks*, *16*(5-6), 809–816.

Gingras, D., and K. Adamowski (1993), Homogeneous region delineation based on annual flood generation mechanisms, *Hydrological Sciences Journal*, *38*(2), 103–121.

Gonzalez, J., and J. B. Valdes (2003), Bivariate drought recurrence analysis using tree ring reconstructions, *Journal of Hydrologic Engineering*, *8*(5), 247–258.

Groupe de recherche en hydrologie statistique, G. (1996), Presentation and review of some methods for regional flood frequency analysis, *Journal of Hydrology*, *186*, 63–84.

Gumbel, E. J. (1958), Statistical theory of floods and droughts, *Journal of Institutional Water Engineering*, *12*(3), 157–184.

Haan, C. T. (2002), *Statistical Methods in Hydrology*, 496 pp., Iowa State Press, Ames, IA.

Haghighatjou, P., A. Akhoond-Ali, A. Behnia, and R. Chinipardaz (2008), Parametric and nonparametric frequency analysis of monthly precipitation in Iran, *Journal of Applied Sciences*, *8*(18), 3242–3248.

Hayes, D. C. (1992), Low-flow characteristics of streams in Virginia, *Open-File Report 89-586*, United States Geologic Survey.

Hisdal, H., and L. M. Tallaksen (2003), Estimation of regional meteorological and hydrological drought characteristics: a case study for Denmark, *Journal of Hydrology*, *281*, 230–247.

Hosking, J. (1990), L-moments: Analysis and estimation of distributions using linear combinations of order statistics, *Journal of Royal Statistical Society*, pp. 105–124.

Hosking, J. R. M., and J. R. Wallis (1993), Some statistics useful in regional frequency analysis, *Water Resources Research*, *29*(2), 271–281.

Hosking, J. R. M., and J. R. Wallis (1997), *Regional Frequency Analysis: An Approach Based on L-Moments*, 224 pp., Cambridge University Press, United States of America.

Jakob, D., D. W. Reed, and A. Robson (1999), *Statistical procedures for flood frequency estimation, Flood Estimation*, vol. 3, chap. 16, p. 153 Ű 180, Institute of Hydrology, Wallingford, UK.

Karray, F. O., and C. De Silva (2004), *Soft Computing And Intelligent Systems Design*, 560 pp., Pearson Education Limited, England.

Khan, M. S., and P. Coulibaly (2006), Application of support vector machine in lake water level prediction, *Jounral of Hydrologic Engineering*, *11*, 199–205.

Khandekar, M. L. (2002), Trends and changes in extreme weather events: An assessment with a focus on alberta and canadian prairies, *Tech. Rep. ISBN 0-7785-2428-0*, Alberta Environment.

Kidson, R., and K. Richards (2005), Flood frequency analysis: assumptions and alternatives, *Progress in Physical Geography*, *29*(3), 392–410.

Kim, T., J. B. Valdes, and C. Yoo (2003), A nonparametric approach for estimating return periods of droughts in arid regions, *Journal of Hydrologic Engineering*, *8*(5), 237–246.

Kim, T.-W., J. B. Valdes, and J. Aparicio (2006), Spatial characteriszation of droughts in the Conchos River basin based on bivariate frequency analysis, *Water International*, *31*(1), 50–58.

Kjeldsen, T. R., A. Lundorf, and D. Rosbjerg (1999), Regional partial duration series modelling of hydrological droughts in Zimbabwean rivers using a two-component exponential distribution, in *IAHS-AISH publication ISSN 0144-7815 CODEN IAPUEP*, vol. 222, pp. 145–153.

Leavitt, P., and G. Chen (2000), Sustainable agriculture in Western Canada: Planning for drought using the past, accessed September, 2009.

Lettenmaier, D. P., J. R. Wallis, and E. F. Wood (1987), Effect of regional heterogeneity on flood frequency estimation, *Water Resources Research*, *23*(2), 313–323.

Midgley, D. C., W. V. Pitman, and B. J. Middleton (1994), Surface water resources of South Africa 1990, *Tech. Rep. 298/5.1/94*, Pretoria, South Africa.

Obasi, G. O. P. (1994), WMO's role in the international decade for natural disaster reduction, *Bull. Amer. Meteor. Soc.*, *75*, 1655–1661.

Ouarda, T. B. M. J., and C. Shu (2009), Regional low-flow frequency analysis using single and ensemeble artifical neural networks, *Water Resources Research*, *45*, 1–16.

Ouarda, T. B. M. J., C. Girard, G. S. Cavadias, and B. Bobee (2001), Regional flood frequency estimation with canonical correlation analysis, *Journal of Hydrology*, *254*, 157–173.

Poulin, A., D. Hauard, A.-C. Favre, and S. Pugin (2009), Importance of tail dependence in bivariate frequency analysis, *Jounral of Hydrologic Engineering*, pp. 394–403.

Reed, D. W., and A. J. Robson (1999), *Flood Estimation Handbook*, vol. 3, Institute of Hydrology, Wallingford, UK.

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

Rossi, G., M. Benedini, G. Tsakiris, and S. Giakoumakis (1992), On regional drought estimation and analysis, *Water Resources Management*, *6*, 249–277.

Sadri, S., H. Madsen, P. S. Mikkelsen, and D. H. Burn (2009), Analysis of extreme rainfall treands in Denmark, IAHR/CSCE 33rd Annual conference, Vancouver, BC.

Samania, N., M. Gohari-Moghadam, and A. A. Safavi (2007), A simple neural network model for the determination of aquifer parameters, *Journal of Hydrology*, *340*(1-2), 1–11.

Sen, Z. (1980), Regional drought and flood frequency analysis: Theoretical consideration, *Journal of Hydrology*, *46*, 265–279.

Serfling, R., and P. Xiao (2007), A contribution to multivariate L-moments: L-commoment matrices, *Journal of Multivariate Analysis*, *98*, 1765–1781.

Shiau, J. T., and R. Modarres (2009), Copula-based drought severity-duration-frequency analysis in Iran, *Journal of Meteorologica Applications*, *16*, 481–489.

Shiau, J.-T., and H. W. Shen (2001), Resource analysis of hydrologic droughts of different severity, *Journal of Water Resources Planning Mannagement*, *127*(1), 30–40.

Shiau, S., Jenq-Tzong; Feng, and N. Saralees (2007), Assessment of hydrological droughts for the Yellow River, China, using copulas, *Hydrological Processes*, *21*, 2157–2163, doi:10.1002/hyp.6400.

Shu, C., and D. H. Burn (2004a), Artificial neural network ensembles and their application in pooled flood frequency analysis, *Water Resources Research*, *40*, 1–10.

Shu, C., and D. H. Burn (2004b), Homogenous pooling group delineation for flood frequency analysis using a fuzzy expert system with genetic enhancement, *Journal of Hydrology*, *291*, 132–149.

Shu, C., and T. Ouarda (2008), Regional flood frequency analysis at ungauged sites using the adaptive neuro-fuzzy inference system, *Journal of Hydrology*, *349*, 31–43.

Singh, V. P., and L. Zhang (2007), IDF curves using Frank Archimedean copula, *Journal of Hydrologic Engineering*, *6*, 651–662.

Sklar, A. (1959), *Fonctions de répartition à n dimensions et leurs marges*, vol. 8, pp. 229–231, Publications de Institut de Statistique Universite de Paris, Paris.

Smakhtin, V. U. (2001), Low flow hydrology: a review, *Journal of Hydrology*, *240*, 147–186.

Smola, A. J., and B. Schölkopf (2003), A tutorial on support vector regression, *Tech. rep.*, RSISE, Australian National University, Cancerra, 0200, Australia.

Song, S., and V. P. Singh (2009), Meta-elliptical copulas for drought frequency analysis of periodic hydraulic data, *Stochastic Environmental Research and Risk Assessment*, *24*(3), 425–444.

Tase, N. (1976), *Area-deficit-intensity characteristics of droughts*, vol. 87, Colorado State University, Fort Collins, Colorado.

Vapnik, V. (1995), *The Nautre of Statisical Learning Theory*, Springer, New York.

Vapnik, V. (2006), *Estimation of dependencies based on empirical data*, Springer Science + Media Inc., U.S.A.

Vicente-Serrano, S. M., and J. I. Lopez-Moreno (2005), Hydrological response to different time scales of climatological drought: an evaluation of the standardized precipitation index in a mountainous mediterranean basin, *Hydrology and Earth System Sciences*, *9*, 523–533.

Vogel, R. M., and N. M. Fennessey (1993), L-moment diagrams should replace product moment diagrams, *Water Resources Research*, *29*(6), 1745–1752.

Vorosmarty, C. V., P. Green, J. Salisbury, and R. B. Lammers (2000), The vulnerability of global water resources: Major impacts from climate change or human development?, *Science*, (289), 284–288.

Water Survey of Canada (2006), Archived hydrometric data.

Wood, T. R. (1987), *Present-day hydrology of the River Severn. Paleohydrology in Practice: A River Basin Analysis*, pp. 79–97, Wiley, New York.

Yevjevich, V. (1967), An objective approach to definitions and investigations of continental hydrologic droughts, *Paper 23*, Colorado State University.

Yue, S., and P. Rasmussen (2002), Bivariate frequency analysis: discussion of some useful concepts in hydrological application, *Hydrological Processes*, *16*(16), 2881–2898.

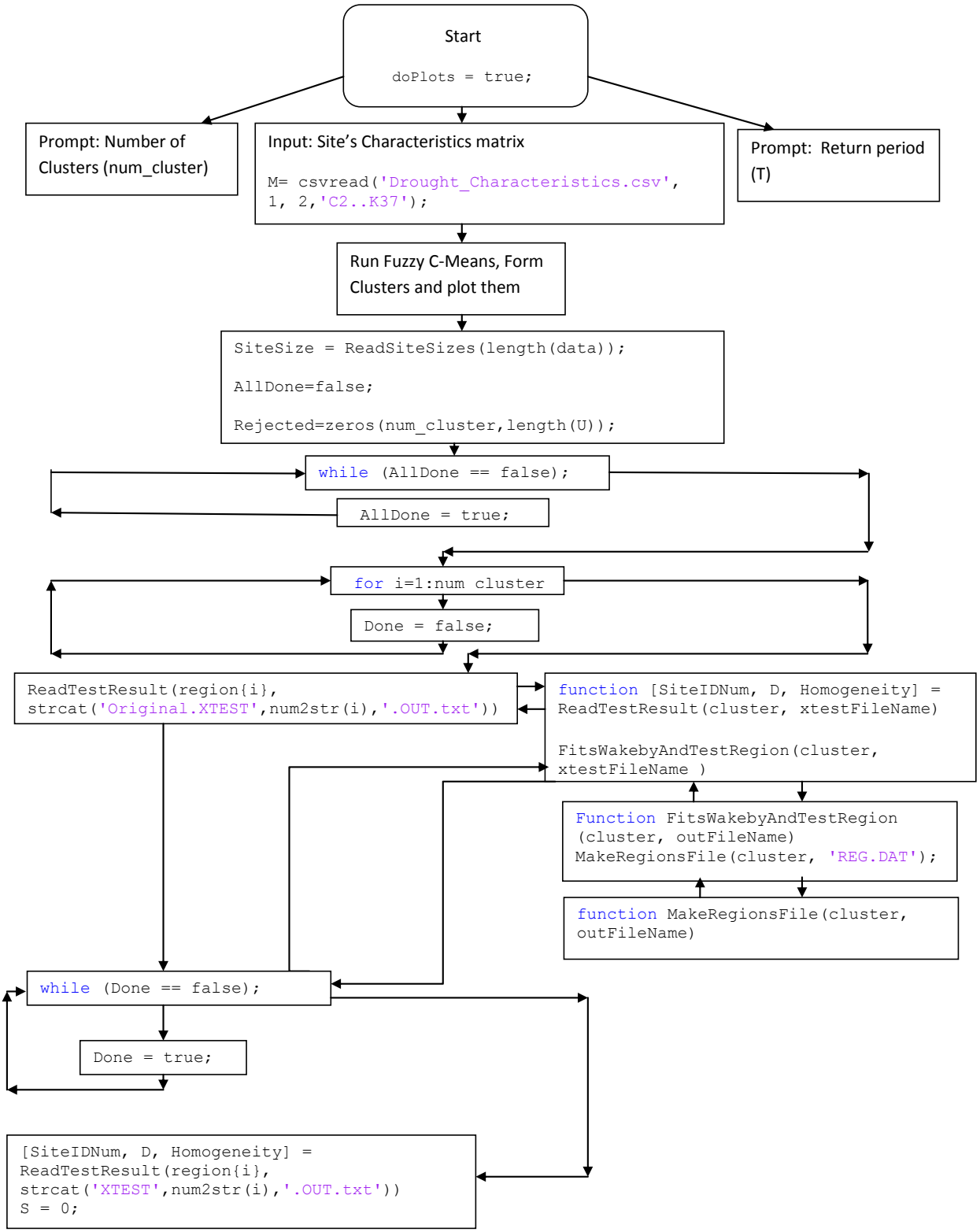
Yujica, Y. (1975), Analysis of drought characterisitics by the theory of runs, *Tech. Rep. 80*, Colorado State University, Fort Collins, Colorado.

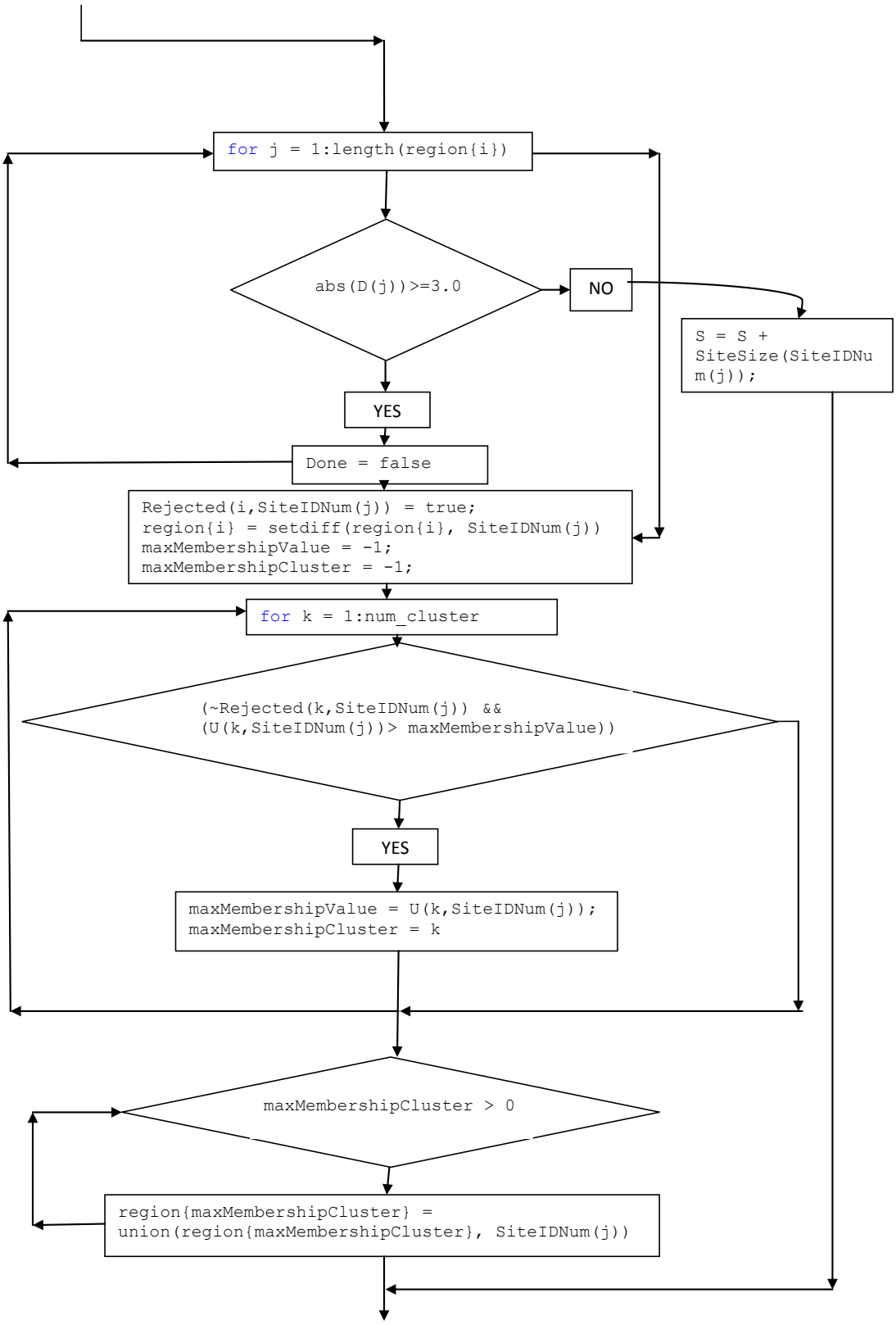
Zhang, L., and V. P. Singh (2006), Bivariate flood frequency analysis using the copula method, *Journal of Hydrologic Engineering*, *11*(2), 150–160.

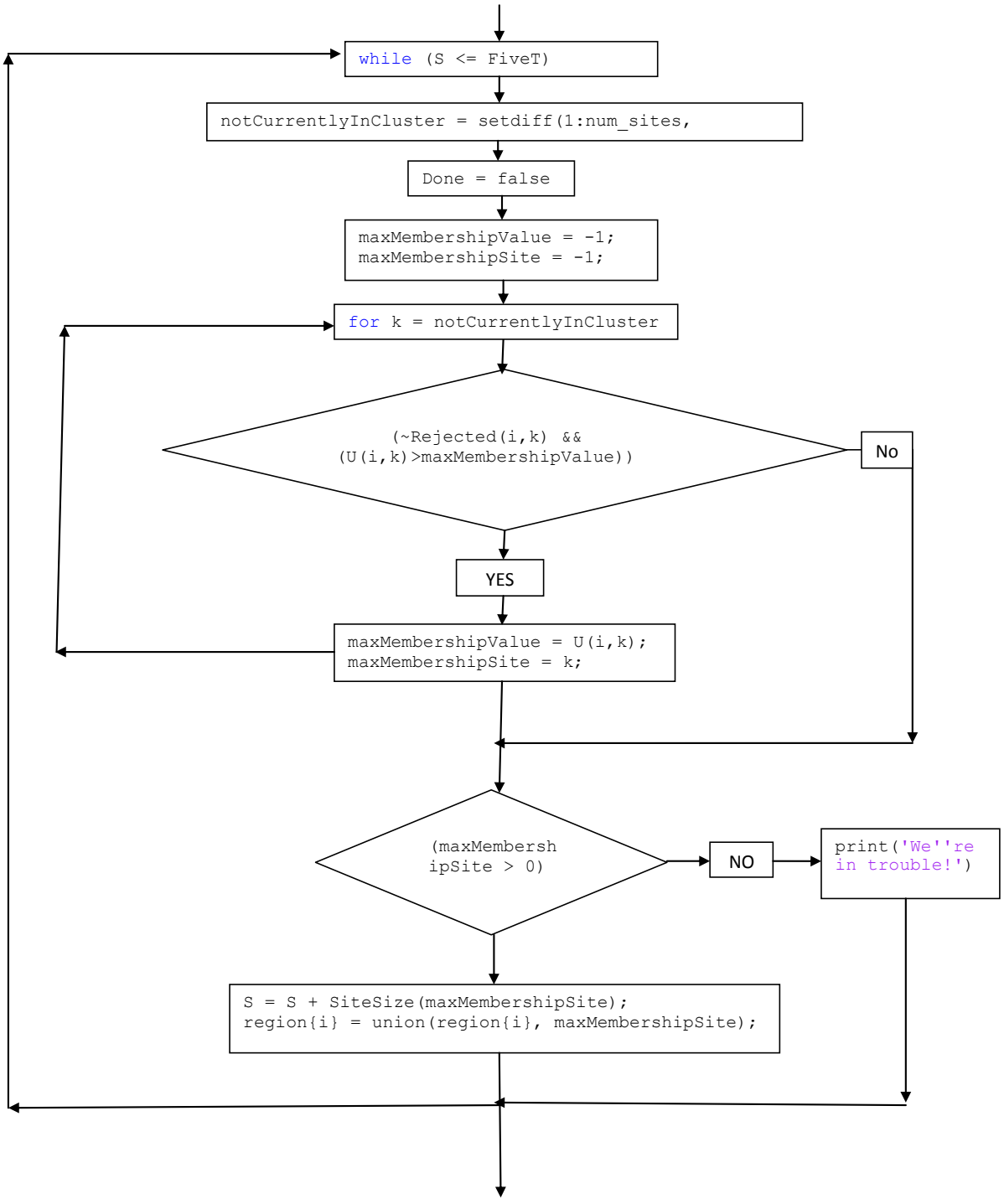
Zhang, L., and V. P. Singh (2007), Bivariate rainfall frequency distributions using Archimedean copulas, *Journal of Hydrology*, *332*, 93–109.

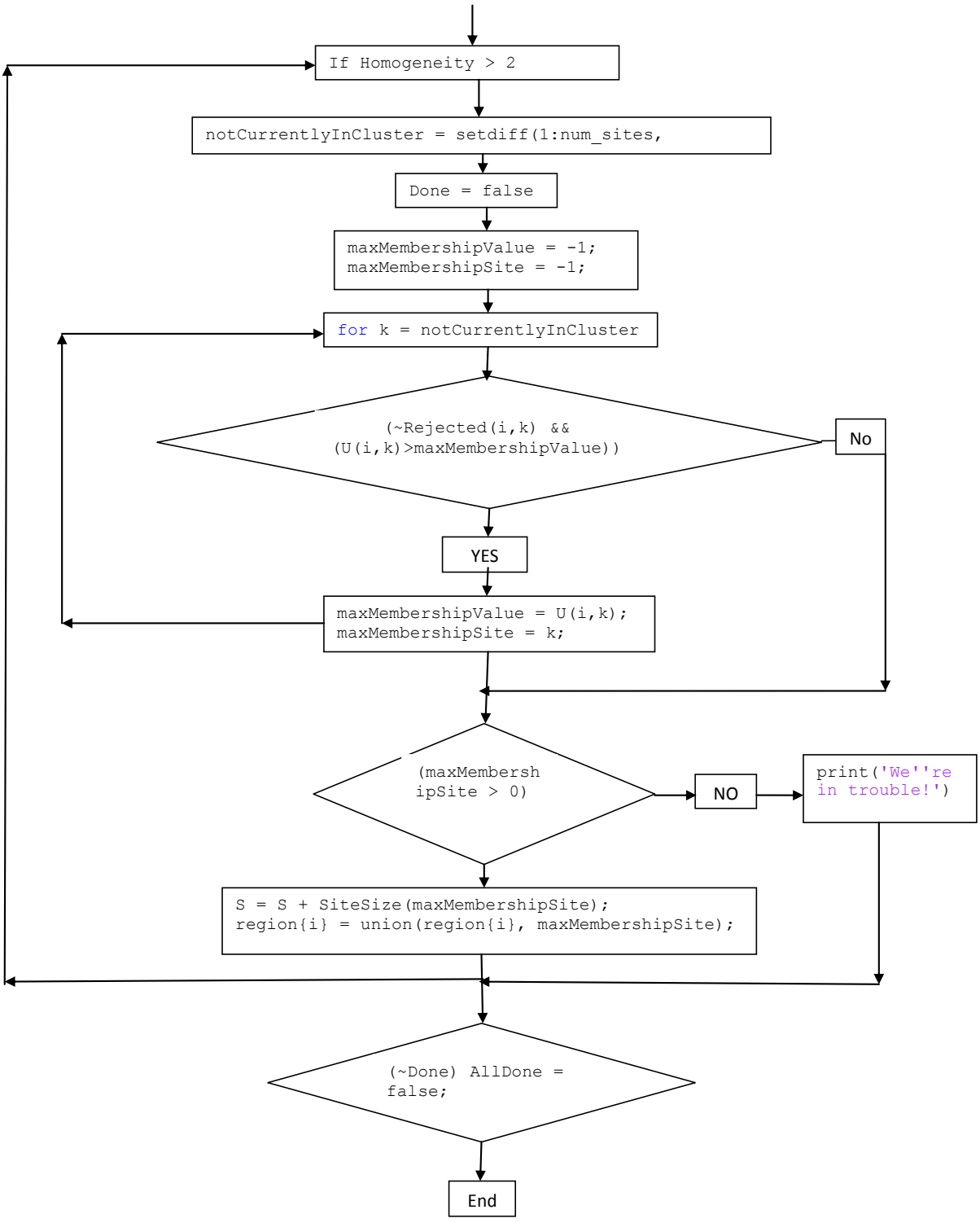
Appendix A

Model Flowchart


Start
doPlots = true;
Prompt: Number of Clusters (num_cluster)
Input: Site's Characteristics matrix
M= csvread('Drought_Characteristics.csv', 1, 2,'C2..K37');
Prompt: Return period (T)
Run Fuzzy C-Means, Form Clusters and plot them
SiteSize = ReadSiteSizes(length(data));
AllDone=false;
Rejected=zeros(num_cluster,length(U));
while (AllDone == false);
AllDone = true;
for i=1:num cluster
Done = false;
ReadTestResult(region{i}, strcat('Original.XTEST',num2str(i),'.OUT.txt'))
function [SiteIDNum, D, Homogeneity] = ReadTestResult(cluster, xtestFileName)
FitsWakebyAndTestRegion(cluster, xtestFileName)
Function FitsWakebyAndTestRegion (cluster, outFileName)
MakeRegionsFile(cluster, 'REG.DAT');
function MakeRegionsFile(cluster, outFileName)
while (Done == false);
Done = true;
[SiteIDNum, D, Homogeneity] = ReadTestResult(region{i}, strcat('XTEST',num2str(i),'.OUT.txt'))
S = 0;


for j = 1:length(region{i})
abs(D(j))>=3.0
NO
S = S + SiteSize(SiteIDNum(j));
YES
Done = false
Rejected(i,SiteIDNum(j)) = true;
region{i} = setdiff(region{i}, SiteIDNum(j))
maxMembershipValue = -1;
maxMembershipCluster = -1;
for k = 1:num_cluster
(~Rejected(k,SiteIDNum(j)) &&
(U(k,SiteIDNum(j))> maxMembershipValue))
YES
maxMembershipValue = U(k,SiteIDNum(j));
maxMembershipCluster = k
maxMembershipCluster > 0
region{maxMembershipCluster} =
union(region{maxMembershipCluster}, SiteIDNum(j))


while (S <= FiveT)
notCurrentlyInCluster = setdiff(1:num_sites,
Done = false
maxMembershipValue = -1;
maxMembershipSite = -1;
for k = notCurrentlyInCluster
(~Rejected(i,k) &&
(U(i,k)>maxMembershipValue))
No
YES
maxMembershipValue = U(i,k);
maxMembershipSite = k;
(maxMembersh
ipSite > 0)
NO
print('We''re
in trouble!')
S = S + SiteSize(maxMembershipSite);
region{i} = union(region{i}, maxMembershipSite);


If Homogeneity > 2
notCurrentlyInCluster = setdiff(1:num_sites,
Done = false
maxMembershipValue = -1;
maxMembershipSite = -1;
for k = notCurrentlyInCluster
(~Rejected(i,k) &&
(U(i,k)>maxMembershipValue))
No
YES
maxMembershipValue = U(i,k);
maxMembershipSite = k;
(maxMembersh
ipSite > 0)
NO
print('We''re
in trouble!')
S = S + SiteSize(maxMembershipSite);
region{i} = union(region{i}, maxMembershipSite);
(~Done) AllDone =
false;
End