WATERSHED REGIONALIZATION FOR REGIONAL FLOOD FREQUENCY ANALYSIS,

IN THE RIFT VALLEY LAKE BASIN, ETHIOPIA

MSc. THESIS

BY

ABDISA SIME KEBEBEW

March, 2021

ACEWM/Addis Ababa University

WATERSHED REGIONALIZATION FOR REGIONAL FLOOD FREQUENCY ANALYSIS,

IN THE RIFT VALLEY LAKE BASIN, ETHIOPIA

MSc. THESIS

BY

ABDISA SIME KEBEBEW

A THESIS SUBMITTED TO THE AFRICA CENTRE OF EXCELLENCE FOR WATER MANAGEMENT, ADDIS ABABA UNIVERSITY, IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTERS OF SCIENCE IN WATER MANAGEMENT, SPECIALIZATION IN HYDROLOGY AND WATER RESOURCES

March, 2021

ACEWM/Addis Ababa University

AFRICA CENTRE OF EXCELLENT FOR WATER MANAGEMENT

ADDIS ABABA UNIVERSITY

THESIS APPROVAL SHEET

APPROVED BY BOARD OF EXAMINERS

This is to certify that we the undersigned, have examined this MSc. Thesis titled "**Watershed Regionalization for Regional Flood Frequency Analysis, In the Rift Valley Lake Basin, Ethiopia**" and that in our opinion; it is fully adequate in scope and quality, as an MSc research thesis for the degree of Master of Science in Water Management, specialization in Hydrology and Water Resources.

Advisor: Adane Abebe (Dr. Ing)

Signature ---------- Date ----------

External Examiner: Tekalegn Ayele (PhD)

Signature ---------- Date ----------

Internal Examiner: Getachew Tegegne (PhD)

Signature ---------- Date ----------

Chairperson: Nigus Gabbiye (PhD)

Signature ---------- Date ----------

Declaration and Copyright

I, ABDISA SIME declare that the content of this thesis is my own original work with the exception of such quotations or references which have been credited to their authors or sources, and that this thesis has not been previously submitted to this or any other University for a degree award.

Signature……

Date……

Abdisa Sime Kebebew

Email: abdisime10@gmail.com

ACKNOWLEDGEMENT

First and foremost, I want to thank the almighty God and His Mother, Saint Virgin Mary, for giving me strength and courage during all my works.

I would like to express my whole hearted gratitude to my advisor Dr. Ing. Adane Abebe for his major support in supervising whole part of my thesis work, His critical and constructive comments and encouragements since the early stage of the thesis work through providing priceless and untapped knowledge. Without him this research would not have been realized.

I would like to give my appreciation to Africa Centre of Excellence for Water Management for providing me research sponsorship and all the needed assistance.

I would like also to thank (MoWIE); Hydrology department, National Meteorological Agency (NMA) for their help by providing necessary data to conduct this research work.

My special heartfelt gratitude also goes to my whole family and friends for their affection and encouragement.

ABSTRACT

The use of regional information to predict magnitude of flow both at site and ungauged catchments are useful for planning and management of water resources. The main objective of this study was to regionalize watersheds in the Rift Valley Lake Basin (RVLB) and flood frequency analysis for the delineated homogeneous regions. In regionalization of the watersheds; Physiographic, drainage, meteorological, soil cover, land-use pattern characteristics and geographical location attributes have been used. Cluster analysis was done by Hierarchical clustering to obtain number of clusters, and final clustering by K-mean method. Accordingly four regions have been identified and checked using homogeneity tests. Using goodness of fit tests (Chi-square test, Kolmogorov–Smirnov, and Anderson–Darling), the best fit distribution models have been selected. Generalized extreme value distribution is the best fit for region I, Log-normal (2P) is selected for region II, Wakeby distribution is found to be the best for region III, and Generalized pareto is chosen for region IV. For the selected distributions efficient parameter estimation technique was selected by performing standard error analysis. Thus, method of moment (MOM) is the one with the lowest error so, selected for region I, and maximum likelihood (ML) method is found the most efficient method for the regions II to IV. For each region unique regional frequency curve is developed with standardized annual maximum flow series (AM), which is a crucial to estimate flood quantile in ungauged areas of the basin. Regional regression model was developed for all region except for region I which consists only one gauged catchment based on their R^2 values. Accordingly 0.82, 0.83, and 0.79 of R^2 values respectively for all the three regions. For checking performance of the model, validation of regional model was carried out by computing the relative errors, over five gauged watersheds that is representative for each region considering as pseudo ungauged. The relative errors between observed and estimated mean annual maximum flows resulted all regional model performs good having maximum of 10.6% of relative error. So, for any current and future water resources developments in the area, the developed regional model can be applied.

Keywords; Rift Valley Lake Basin, Regionalization, Cluster analysis, Flood frequency, Distribution models, Parameter estimations

TABLE OF CONTENT

ABBREVIATIONS

AM	Annual Maximum Series
CV	Coefficient of Variation
CC	Site to Site Coefficient of Variation
COV	Coefficient of variation
CCA	Canonical Correlation Analysis
CDF	Cumulative Distribution Function
DEM	Digital Elevation Model
FFA	Flood Frequency Analysis
FAO	Food and Agricultural organization
ML	Maximum Likelihood
MAF	Mean Annual Flood
MAR	Mean Annual Rainfall
MOM	Method of Moment
MOR	Method of Residuals
MoWIE	Ministry of Water, Irrigation and Electricity
NMA	National Meteorological Agency
PD	Partial Duration Series
PDF	Probability Density Function
PWM	Probability Weighted Moment
RFFA	Regional Flood Frequency Analysis
RVLB	Rift Valley Lake Basin
ROI	Region of Influence
TS	Time Series

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

1.1 Background

Flood from its inherent behavior is destructive. Therefore, it is important to evaluate the magnitude and frequency of these hydrological events to mitigate the potential loss of human life, the effect on socio-economics and the degradation of ecology by either regulating or managing using different structures like dams, barrages, culverts, levees and bridges (Rao and Srinivas, 2007). In statistical hydrology, flood frequency analysis practice has been used to relate magnitude to flood incidence frequency using different distribution models, different parameter estimation methods and quantile estimation corresponding to the appropriate return period or likelihood of excess (Hussen and Wagesho, 2016).

The availability of appropriate historical time series data is important in flood frequency analysis (FFA) to estimate accurate flood quantile (Sine and Moges, 2013). In developing countries like Ethiopia hydrological data may be of short duration, missed or totally unavailable. In such case, quantile estimated depending on small samples can be unreasonable or physically unrealistic, particularly for distribution with three or more parameters (Rao and Hamed, 2000).

Watershed regionalization is a technique of grouping hydrologically analogous catchments into homogeneous regions on the basis that statistical characteristics of extreme flows at multiple locations in a region are identical (Cunnane, 1989). Cluster analysis are the most widely used to group catchment that are hydrologically alike (Isik et al., 2008). Regional flood frequency analysis (RFFA) is a method that is used to estimate flood quantile for a required site using information combined from a group of sites in a region whose catchments are comparable to the target site interms flood generating mechanism (Basu and Srinivas, 2016).

Now a days the use of regional information to calculate flood magnitude at a site with small or no measured data has become increasingly important because many water resource projects are found in areas where observed flow data are either inadequate (short duration) or missing (Hussen and Wagesho, 2016; Demissie, 2008; Yirefu, 2010). Most importantly, RFFA can be used at an ungauged location, where data from similar measured locations is used to help characterize the extreme flow regime at the ungauged site (Burn, 1997).

1.2 Statement of the Problem

It is known that floods from their intrinsic behavior causes unforeseen harm to socio-economic, human life and ecological well-being. Most of the rivers in Ethiopia have either of short duration historical time series data or totally no record (ungauged), and it is very difficult to control or regulate floods in this situation by using only at-site information. This situation becomes worsen when it comes to Rift Valley Lake Basin which incorporates a network of rivers; many examples can be given; such as Segen River in Segen watershed, Shafe River in Abaya Chamo sub-basin, Woji River in Lake Ziway sub-basin and number of others are ungauged, and in considering the gauged rivers, the duration of record is not adequate. So, study of regional flood frequency analysis should have given much consideration, even though little attention was paid.

Among the studies conducted in the area, only few scientific researches have been focused on the use of regional information to quantify at site magnitude of flood. Gebeyehu, (1989) studied flood frequency analysis (FFA) for the whole country and delineated homogeneous regions depending on monthly rainfall form and geographical vicinity; but geographical closeness does not necessarily guarantee hydrological homogeneity especially in this area where, topographic and hydro-meteorological variation is very high. Hussen and Wagesho, (2016) have investigated regional flood frequency for Abaya Chamo sub-basin within rift valley lake basin (RVLB) based on at-site L-moment statistics, and the limitation of this study is, regional homogeneity analysis was done by the same L-moment statistics that was used for regionalizing the watersheds, in this approach the regions may seem homogeneous but in reality it is questionable.

As a result of flash flood or over flows of the lakes due to inflows from the rivers, many farmers especially those who lives along the bank of the rivers and lakes has been affected and still suffering from the lack of information about the coming floods, over flow of Abaya lake in 2020 can be mentioned as one example. Therefore, flood frequency analysis is a core issue that needs a carful investigation in Rift Valley Lake Basin. This in turn shows developing regional growth curves for the basin is addressing a central question with much help in ungauged areas. Thus, this study can address the aforementioned flood problems and research gaps in the study area and be able to become a baseline information for next studies by initiating the researchers to further study on related topics.

1.3 Objective of the Study

General objective

The main objective of this study is to regionalize watersheds for analysis of regional flood frequency, in the Rift Valley Lake Basin, Ethiopia.

Specific objectives

- ✓ To identify hydrologically homogeneous region in the study area.
- ✓ To establish regional flood frequency curves for the delineated homogeneous regions.
- ✓ To develop method of calculating flood quantiles for the ungauged areas.

1.4 Research Questions

- ✓ Among the catchments in the study area, which regions are hydrologically similar?
- ✓ How does the regional growth curves are developed and what are the values of standardized regional quantiles that corresponds to a given return period?
- ✓ How does flood quantile be estimated for ungauged catchments in the study area?

1.5 Significance of the Study

Flood is known to cause disasters so, it should be either controlled or managed. Accordingly this study is useful to;

- ✓ Allow Water Resource planners, decision makers and any concerned body to understand the expected magnitude of flood in both gauged and ungauged catchments in the study area.
- ✓ Facilitate researchers to further study on Watershed regionalization related topics to come up with more homogenous region to arrive at relatively more reliable quantile estimates.
- ✓ Add values on the previous publications to bring these best approach to the development partners, local experts, researchers and other professionals to incorporate them into their water resources sustainable development and research programs.

1.6 Scope and Limitation of the Study

This study is prioritized to investigate regionalization of watersheds in the Rift Valley Lake Basin for analysis of flood frequency of the area. The research therefore focuses on hard cluster analysis, particularly K-mean method to regionalize watersheds and it did not consider soft clustering. The landuse land cover data used as regionalization attributes is developed in 2008 by MoWIE so, changes in landuse land cover is not taken into account in this study. Non-stationarity of stream flow data was not checked and considered one of the limitation of this study. Goodness of fit tests were applied to fit distributions models and among parameter estimation methods; MOM, PWM, and ML methods was compared for the development of regional growth curve. The developed Regional regression model was validated using five gauged watersheds each representing the region clustered in by considering as pseudo ungauged.

2 LITRATURE REVIEW

2.1 Flood Frequency Models

Flood frequency analysis is used to estimate magnitude of flood quantile Q_T, at any required river site and this extreme hydrologic event is expressed interms of its return period T. Here, the assignment is to obtain information from recorded stream flow data to obtain the relationship between Q and T and different models can be assigned for this job (Cunnane, 1989; Rao and Hamed, 2000; Sine and Moges, 2013). Accordingly the models are:

- ✓ Annual maximum series (AM)
- ✓ Partial duration series (PD) or peaks over a threshold (POT) model and
- ✓ Time series model (TS)

Only the peak flow in a record year is considered in the annual maximum (AM) flow sequence (Rao and Hamed, 2000). In the partial duration (PD) models all peaks above a some fixed value are used for the analysis (Rao and Hamed, 2000; Sine and Moges, 2013). In the time series (TS) model Hosking and Wallis,(1993); Rao and Hamed, (2000), the flow hydrograph is considered a time series in which the flow hydrograph is considered to be a time series in which, at evenly spaced intervals of time, the flows are represented by a series of ordinates. Hydrological time series analysis usually uses time intervals of days, months and years, but in flood frequency analysis (FFA), different studies recommend days to be used and so does for this research (Cunnane, 1989; Hosking and Wallis, 1997).

2.1.1 Statistical efficiency of estimates of Q_T by each model

According to Cunnane, (1988; 1989) denoting the estimates of Q_T obtained by AM method as Q_T and that acquired from the same hydrometric record by the PD method as $\bar{Q}_T$, it is usually seen that these two estimates are unequal. Moreover the sampling variance of Q_T is not equal to that of $\bar{Q}_T$, i.e. var $(Q_T) \neq$ var $(\bar{Q}_T)$ and from statistical point of view the method which has smallest sampling variance enjoys an advantage.

Under certain condition Cunnane, (1973; 1989) examined the relative values of var (Q_T) and var $(\bar{Q}_T)$ and found that var $(Q_T) <$ var $(\bar{Q}_T)$ provided that $\lambda < 1.65$ where, λ is the mean number of peaks per years included in the PD series. For $\lambda > 1.65$ the opposite was true.

This shows that the AM method is statistically more efficient than the PD method when λ is small but less efficient when λ is large. In many practical situations the assumptions of the PD model may not be valid if λ is increased to high level and the other thing about the PD model is that, the observation may not be independent and this can affect assumption of independence of flood peaks for statistical analysis (Cunnane, 1988; 1989; Rao and Hamed, 2000; Demissie, 2008; Sine and Moges, 2013; Nejc et al., 2013). So, for these reasons AM model is most commonly applied model in flood frequency analysis and selected for this study as well.

2.2 Over View of Regionalization

Regional flood frequency analysis is often used to improve the estimation of flooding probabilities at location that have data record lengths that are short relative to the return period of interest (Burn, 1990; Isik et al., 2008). So, annual extreme flow data from a number of sites can be used to compensate for an inadequate temporal representation of extreme flows at a given location. Regional flood frequency analysis can therefore be employed at gauged locations, where the intent is to extend the available at-site information (at least to 5T peak flows, where T is denoted as target return period), or at ungauged locations, where information from similar gauged sites are used to assist with the characterization of the extreme flow regime at the ungauged site (Burn, 2000; Rao and Srinivas, 2007).

One of the initial steps in the regional flood frequency analysis involves identifying homogeneous regions characterized by similar hydrologic response (Basu and Srinivas, 2014). A region can be considered to contain a group of sites from which extreme flow information can be combined for improving the estimation of extreme flow quantiles at any site in the region (Burn et al., 1997). According to Burn, (2000) the characteristics that the regions should possess to ensure effective information transfer and therefore efficient estimation of extreme flow quantiles include;

- ✓ Catchment should be hydrologically homogeneous, i.e. extreme flow information that is transferred from the region is similar to the extreme flow information at that site (Hosking and Wallis, 1997).
- ✓ The region should be identifiable, i.e. a regional home can be readily determined for a new catchment which may be ungauged (Nathan and Mcmahon, 1990).

- ✓ The regions should be sufficiently large, i.e. larger regions imply that more extreme flow information is incorporated in to the estimation of extreme flow quantiles, thus improving the estimates, provided that the extreme flow information is sufficiently similar to the target site (Rao and Hamed, 1997).

2.2.1 Regionalization Methods

In RFFA regions are mostly chosen contain geographically nearby watersheds based on political, administrative, or physiographic boundaries (Rao and Srinivas, 2007). However, regions formed in this manner will not necessarily be homogeneous interms of hydrological response, given the potentially large amount of spatial variability in the physiographic or hydrological characteristics in this type of region (Tasker, 1983; Burn, 2000). A lot of regional flood frequency analysis (RFFA) approaches have been suggested based on at site similarities by analyzing catchment attributes such as physiographic characteristics, geographical location, and on-site flood statistics to address this problem (Rao and Srinivas, 2007).

The available approaches include the method of residuals, the canonical correlation analysis, the region of influence (ROI) approach and its extensions, the hierarchical approach and its extension to ROI framework and the cluster analysis. Regions are developed using the positive and negative signs of residuals derived from a regional regression model, comparing flood quantities at each measured site to watershed characteristics in the Method of Residuals (MOR) approach to RFFA (Thomas and Benson, 1975). This technique delineates catchments in arbitrary manner and the regions are most of the time arranged to be coincident with recognized hydrologic boundaries or administrative areas. Therefore, it is possible that the regions identified by this method would include watersheds whose flood frequency characteristics may not be comparable (Acreman and Sinslair, 1985; Wiltshire, 1986).

According to Cavadias, (1990) in the canonical correlation analysis (CCA) drainage basins are represented as points in the spaces of pairs of uncorrelated flood-related canonical variables and pairs of uncorrelated basin-related canonical variables to examine similarity in the corresponding point patterns in these spaces.

If the point patterns are sufficiently similar, regions are formed in the space of the flood-related canonical variables. The problem with this method to regionalization is that similarity in point patterns may not be found (Bob and Rasmussen, 1995).

In the region of influence (ROI) approach of Burn, (1990), each site is permitted to have its own region. A target site's in ROI approach consists of those sites in the study region whose distance in a weighted multi-dimensional attribute space to the target site does not exceed the threshold value selected. Each site could be weighted according to its similarity to the required area in the estimation of a regional growth curve (Rao and Srinivas, 2007). The choice and weighting of features and sites is one of the difficulties where no strict mathematical solution is presented, and also as number of features available for the analysis rises it will be difficult to arrive at realistic quantile estimate (Bob and Rasmussen, 1995; Cunderlik and Burn, 2006).

Gabriele and Nigel, (1991) proposed hierarchical approach to RFFA which explicitly accounts for spatial variability in different flood characteristics. Later on Zolt and Burn, (1996) integrated the idea of hierarchical approach in to the ROI framework by defining a set of ROIs for each site as opposed to a distinct ROI.

Cluster analysis is the common name of a number of multivariate statistical processes that are useful for grouping data points (representing watersheds) into clusters (regions) using the watershed-related attributes (Rao and Srinivas, 2007; Basu and Srinivas, 2016; Hailegeorgis and Alfredsen, 2017). Regionalization attributes within a cluster shall be similar as possible, and those in different clusters should be dissimilar as possible (Cunnane, 1988; Isik and Singh, 2008).

The methods reviewed clearly demonstrate that each method of regionalization has its own strengths and constraints. Subjectivity in the attribute selection, weight’s, setting threshold values, characteristic distance measures, and others makes each methods to have no well-established criteria to select one method over the others. Clustering method is a technique suited to effectively use any available data, especially in effectively identifying patterns in both big and small dataset considered advantageous (Rao and Srinivas, 2007).

2.2.2 Homogeneity Tests of the Proposed Regions

A measure of discordance is useful for classifying sites with gross errors in their information or those that are grossly discordant from the other area as a whole. In regional flood frequency analysis, discordance measure explained by Hosking and Wallis, (1997) is commonly used by hydrologists. Areas are considered as points in the three-dimensional space of the sample L-moment ratios to estimate discordance values for areas in a region (L-CV, L-Skewness, and L-Kurtosis). The region's Centroid is considered to reflect the average sample L-moment ratio of the areas in the region. Any point that is far away from the centroid of the region is considered as discordant (Rao and Srinivas, 2007).

Cunnane, (1989) have used the values of mean coefficient of variation (CV) and the site-to-site coefficient of variation (CC) of both convention and L-moment of the proposed region to investigate hydrological homogeneity of the catchments. Many researchers suggests that the higher the values of COV and CC, the lower the performance of the index flood method for the considered region. This is because of the dominance of the flood quintile estimation variance by the variance of the at-site sample mean. Hence for better performance of the index flood method, CC should be kept low (Lettenmaier, 1985; Cunnane, 1989; Sine and Moges, 2013).

Acreman and Sinslair, (1985) used clustering algorithm to delineate homogeneous regions depending on the watershed characteristics and then using a likelihood ratio test to check whether an estimated GEV distribution for a degree of dissimilarity. Wiltshire, (1986a) used CV based statistics that depend on region size, record length and choice of distribution.

Hosking and Wallis, (1993) proposed heterogeneity measure based on L-moment ratios. In a homogeneous region all sites are expected to have the same population L-moment ratios. However, their sample L-moment ratios (LMRs: L-coefficient of variation (L-CV), L-skewness and L-kurtosis) may be different due to sampling variability. The regional homogeneity tests are developed to scrutinize whether the between-site dispersion of the sample LMRs for the group of sites under consideration is larger than the dispersion expected in a homogeneous region (Rao and Srinivas, 2007).

2.2.3 Revisions and Regional Modifications

If the regions derived from clustering method are not statistically homogeneous, to enhance their homogeneity, they should be revised and modified. This stage of regionalization is reasonable because when generated on the basis of a collection of attributes that is not comprehensive, the regions are not necessarily expected to be homogeneous (Rao and Srinivas, 2007).

The possibilities recommended by Hosking and Wallis, (1997) for modifying the regions formed by clustering algorithm include; (i) removing one or more catchments from the dataset; (ii) relocating one or more catchments from one region to another; (iii) splitting a region to form more new regions; (iv) letting a catchment to be shared by two or more regions; (v) dissolving regions by transferring their catchments to other regions; (vi) merging a region; (vii) merging the regions and form new groups; and (viii) collecting more data and forming regions. Of these, the first three alternatives are useful for reducing a region's heterogeneity measure, while alternatives (iv) to (vii) help guarantee that each region is sufficiently big (Burn, 2000; Rao and Srinivas, 2007).

2.3 Flood Frequency Distributions

Many flood frequency distributions have been recommended for modeling flood flows, but none has been universally accepted because they have different governing processes and underlying characteristics (Sine and Moges, 2013). The choice of distribution for the AM time series has been the topic of interest and influenced by the factors like whether the distributions would be; Widely accepted, Simple and convenient to apply, Consistent, flexible or robust (low sensibility to outliers), Theoretically well based and Documented in the guide (Cunnane, 1988; 1989; Rao and Hamed, 2000; Yirefu, 2010).

Cunnane, (1989) listed the commonly used statistical distributions for AM series models;

- ✓ **Normal and related distributions**; Normal distribution, Log normal two parameter distribution, Log normal three parameter distribution.
- ✓ **The Gamma distributions**; Exponential distribution, Two parameter Gamma distribution, Pearson three distribution, Log person two distribution.
- ✓ **Extreme value distributions**; Generalized extreme value distribution, Extreme value type I distribution, Extreme value type II distribution, Weibull distribution.

- ✓ **Wake-by distributions**; Five parameter wake-by distribution, Four parameter wake-by distribution, Generalized Pareto distribution.
- ✓ **Logistic distributions**; Log-logistic distribution, Generalized logistic distribution.

Table 2-1: Mathematical Expression of Statistical Distributions (Cunnane, 1989)

Name of Distribution	**Probability Density Function of *x***	**Variate and Parameter ranges**
Normal distribution (N)	$F(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{1}{2\sigma^2}(x-\mu)^2}$	$-\infty < x < \infty$ $\mu \; and \; \sigma \; are \; parameters$
Two parameter Lognormal distribution (LN2)	$F(x) = \frac{1}{x\sigma_y\sqrt{2\pi}} \exp\{\frac{[-logx - \mu_y]^2}{2\sigma_y^2}\}$	$0 < x$ $\mu_y \; and \; \sigma_y$
Three parameter Lognormal distribution (LN3)	$F(x) = \frac{1}{(X-a)\sigma_y\sqrt{2\pi}} \exp\{-\frac{1}{2\sigma_y^2}[logx - a) - \mu_y]^2\}$	$a < x$ $\mu_y, \sigma_y, \; and \; a$
Exponential distribution	$F(x) = \frac{1}{a} \exp(-\frac{x-\varepsilon}{a})$	$\varepsilon < x < \infty$ $(i.e, \beta = 1$ $a \; and \; \varepsilon$
Two parameter Gamma distribution (G2)	$F(x) = \frac{1}{a^\beta \Gamma(\beta)} x^{(\beta-1)} e^{-(x/a)}$	$0 < x < \infty$ $(i.e, \varepsilon = 0$ $a \; and \; \beta$
Pearson-III Distribution (P-III)	$F(x) = \frac{1}{a\Gamma(\beta)} (\frac{x-\gamma}{a})^{\beta-1} e^{-(\frac{x-\gamma}{a})}$	$\gamma < x < \infty$ $\beta, a, and \; \gamma$
Log Pearson-III Distribution (LP-III)	$F(x) = \frac{1}{ax\Gamma(\beta)} [\frac{\log(x)-\gamma}{a}]^{\beta-1} e^{-\{\frac{\log(x)-\gamma}{a}\}}$	$1 < \beta$ $0 < 1/a$ $a, \beta, and \; \gamma$

Generalized Extreme value distribution (GEV)	$F(x) = \frac{1}{a}[1 - k\left(\frac{x-\mu}{a}\right)]^{1/k - 1} e^{-[1-k\left(\frac{x-\mu}{a}\right)]^{1/k}}$	$a > 0$ $\mu + \frac{a}{k} \leq x \leq \infty, if\ k > 0$ $-\infty < x \leq \mu + \frac{a}{k}, if\ k > 0$
Extreme value type-I distribution (EV-I)	$F(x) = \frac{1}{a}\exp[-\left(\frac{x-\beta}{a}\right) - e^{-\left(\frac{x-\beta}{a}\right)}]$	$-\infty < x < \infty$ $a > 0$
Five parameter Wakeby distribution (WAK-5)	$x = m + a[1 - (1-F)^b] - [1 - (1-F)^{-d}]$ $Where\ F = F(x)$	Analytically explained only in inverse form
Four parameter Wakeby distribution (WAK-4)	$F(x) = \frac{a}{\beta}[1 - (1-F)^{\beta}] - \frac{\gamma}{\delta}[1 - (1F)^{-\delta}]$	$m\ or\ \varepsilon = 0$
Generalized pareto distribution	$F(x) = \frac{1}{a}[1 - \frac{k}{a}(x-\varepsilon)]^{1/k - 1}$	$\varepsilon \leq x < \infty, for\ k \leq 0$ $\varepsilon \leq x \leq \varepsilon + {}^{a}/_{k}, for\ k > 0$
Logistic distribution	$F(x) = [1 + e^{\left(\frac{x-m}{a}\right)}]^{-1}$	$-\infty < x < \infty$
Generalized Logistic distribution	$F(x) = \frac{1}{a}[1 - k\left(\frac{x-\varepsilon}{a}\right)]^{(\frac{1}{k}-1)}[1 + \{1 - k(\frac{x-\varepsilon}{a})\}^{1/k}]^{-2}$	$\varepsilon + \frac{a}{k} \leq x < \infty, for\ k < 0$ $-\infty < x \leq \varepsilon + \frac{a}{k}, for\ k > 0$

2.4 Summary of Parameter Estimation Techniques

There are a multitude of methods for estimating parameters of hydrologic frequency models. Some of the methods used in hydrology includes; (1) Method of Moments (Rao and Hamed, 2000; Wooldridge, 2001), (2) Probability Weighted Moments (Greenwood, 1979), (3) Maximum Likelihood estimation (Myung, 2003), (4) Least square methods (Strejc, 1980), and (5) Method of Maximum Entropy (Singh, 1998). From those methods above the widely used techniques like;

Method of Moments, Probability Weighted Moment and Maximum Likelihood Estimation have been reviewed.

The method of moments (MOM) is a relatively easy technique and is more commonly applied. For numerical processes involved in ML estimation, it can also be used to obtain starting values. Parameter estimation by MOM is known to be biased and inefficient especially with three-parameter distribution but it is more preferable for two parameter distribution types (Rao and Hamed, 2000; Wooldridge, 2001).

Greenwood, (1979) introduced the method of probability weighted moments (PWM) and showed its usefulness in deriving explicit expressions for parameters of distributions whose inverse forms $X = X(F)$ can be explicitly defined. Hosking and Wallis, (1986) developed the theory of probability weighted moments and applied to estimate parameters of several distributions. According to (Greenwood, 1979; Rao and Hamed, 2000) PWM gives parameter estimations equivalent to the ML estimations, however in some cases the estimation processes are much less complex and the computations are easier. Indeed in small samples PWM may be as efficient as ML. With a suitable choice of distribution PWM estimation also contributes to robustness and is attractive from that point of view.

The method of maximum likelihood (ML) estimation is widely accepted as one of the most powerful parameter estimation methods. Asymptotically, ML parameter estimates are unbiased, minimum variance, and normally distributed, while in some cases these properties hold for small samples (Rao and Hamed, 2000). Yapo et al., (1996) assessed ML parameter estimation method for conceptual rainfall-runoff models and evaluated the influence of data variability and length on model reliability. Kitanidis and Lane, (1985) applied the ML method to estimate hydrologic spatial processes. The equations are usually complex that can only be solved by numerical techniques (Cunnane, 1989).

2.5 Index Flood Analysis for Ungauged Catchments

The concept underlying the index flood method is that the distributions of flood at different sites in a region are the same except for a scale or index flood parameter which reflects the size, rainfall, and runoff characteristics of each watershed (David, 1993). Here, mean is employed as the index flood. The index flood parameter reflects the important physiographic and meteorological

characteristics of a watershed (Rao and Srinivas, 2007). The difficulty of estimating p^{th} quantile x_p is then reduced to estimation of the mean for a site μ_x, and the ratio x_p/μ_x of the p^{th} quantile to the mean. The average can be calculated from the record available at a site, even if that record is short. The stated ratio is calculated by using regional information and these normalized regional flood distributions can be shown on the Growth curve (David, 1993).

The index flood method is also found to be an accurate and reproducible to be used at ungauged sites. In the derivation of index flood equation, empirical methods based on simple or multiple regression can be used (Stedinger and Lu, 1995); these methods are often used to relate the index flood to physical characteristics of river basins such as climate, geomorphology, slope, and land use. Comparison of some those methods have been investigated by Grover et al., (2002) who analyzed the use of method OLS (Ordinary Least Squares), and the WLS method (Weighted Least Squares).

3 MATERIALS AND METHODS

3.1 Description of Study Area

3.1.1 Location and Topography

The Rift Valley lakes basin, is part of the Great East African Rift System which is geographically located between 36.459° to 39.407°E and 4.367° to 8.421°N. Totally the basin has 52,739 km^2 of area and surrounded in the north by Awash River basin, in the west by Omo-Gibe River basin, in southeast by Genale-Dawa River basin, and in the east by WabiSheble River basin (figure 3.1). The basin has a wider elevation gradient ranging from 343 m (in the valley floor) to above 4200 m above mean sea level in the eastern and western part of the basin.

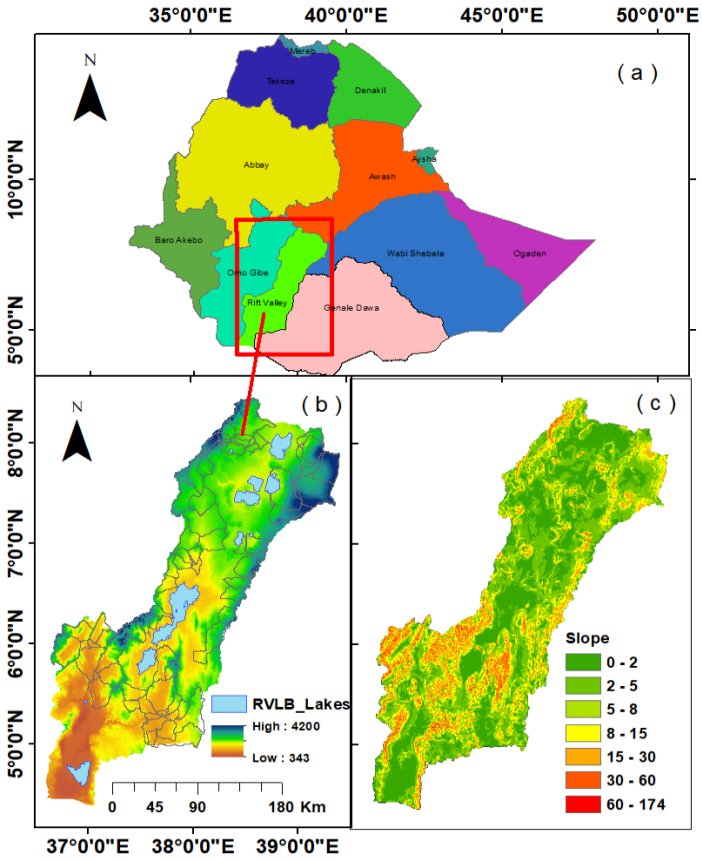

Figure 3-1: Study area description map; (a) Ethiopian basins map, (b) Elevation map, (c) Slope map of RVLB

The catchments within the basin is well characterized in a range from gentle to steep slopes (see table 3.1) according to the terrain slope classification system of Food and agriculture organization (FAO, 1985).

Table 3-1: Slope classification of the study area (%)

Classification	**Flat slope**	**Gentle slope**	**slopping**	**Moderate steep**	**Steep**	**Very steep**	**Extremely steep**
Slope (%)	0-2	2-5	5-8	8-15	15-30	30-60	>60
Weito	9.49	19.32	14.29	20.88	25.57	10.13	0.31
Upper Gelana	10.12	33.82	19.93	23.72	11.46	0.95	-
Hammassa	15.13	48.19	20.54	11.37	4.20	0.56	-
Badessa	5.01	13.58	14.92	31.12	31.87	3.49	-
Gidabo	19.62	32.06	17.42	21.35	8.63	0.91	-
Kola	15.28	51.49	13.89	14.38	4.95	-	-
Bilate	35.02	42.25	11.85	6.90	3.46	0.52	-
Tikur wuha	17.16	37.75	18.50	14.44	10.33	1.82	-
Ferfuro	12.47	42.90	24.96	16.14	3.46	0.07	-
Gedemso	18.67	44.82	15.58	15.66	5.26	-	-
Horakelo	16.23	34.47	23.67	18.98	5.96	0.69	-
Ketar	15.77	38.40	21.77	17.61	6.01	0.43	-
Chiufa	3.40	22.15	30.67	31.52	11.33	0.94	-
Meki	18.26	25.77	19.73	18.41	14.17	3.59	0.07
Rinzaf	4.92	23.09	18.15	21.84	19.97	11.80	0.23
Gombora	17.02	48.43	28.30	5.25	1.00	-	-
Hare	0.93	7.68	11.30	20.93	41.97	16.82	0.35
Kulfo	0.78	4.01	7.60	19.96	42.27	24.29	1.08
Guder	13.89	46.12	17.47	17.58	4.94	-	-
Shafe	8.11	11.37	12.68	25.12	33.33	9.28	0.12
Baso	1.81	11.84	22.57	26.94	26.38	10.06	0.40
Sagan	22.91	30.29	15.57	16.25	11.27	3.58	0.13
Bisare	28.34	41.74	19.29	6.93	3.36	0.34	-
Woji	42.88	27.33	8.71	10.97	6.89	3.17	0.06

From terrain slope map of the catchments from Arch GIS, most of the area (30.79%) of the catchments are classified as gentle slope, 17.89% and 18.09% of the catchments area are classified as slopping and moderate steep slope respectively. The areas that are characterized under flat slope shared 14.72% and only 4.31% and 0.11% of the area are very steep and extremely steep slope.

According to the traditional system there are five climatic zones; Wurch (cold climate more than 3000 m. altitude), Dega (temperate like climate highlands with 2500-3000 m. altitude), Woina Dega (warm at 1500-2500 m. altitude), kola (hot and arid type, 500-1500 m. altitude), and Berha (hot and hyper arid type less than 500 m. altitude) climate (NMA, 2001). Table 3.2, describes those five climatic zones of the study area.

Table 3-2: Traditional climate zone classification of the study area (%)

Classification	**%Berha**	**%Kola**	**%Woina Dega**	**%Dega**	**%Wurch**
Elevation (M)	0-500	500-1500	1500-2500	2500-3000	>3000
Weito	0.002	62.819	29.514	5.700	1.966
Upper Gelana	0.048	16.928	77.790	5.233	-
Hammassa	0.019	20.364	78.454	1.163	-
Badessa	0.014	0.041	81.516	18.429	-
Gidabo	-	0.009	75.608	22.295	2.088
Kola	-	0.015	83.585	16.123	0.277
Bilate	0.001	0.425	89.924	7.382	2.268
Tikur wuha	-	0.011	77.930	22.060	-
Ferfuro	-	0.014	87.214	12.772	-
Gedemso	-	0.008	37.630	42.095	20.267
Horakelo	0.015	0.943	60.454	27.087	11.502
Ketar	-	0.002	22.469	45.568	31.962
Chiufa	-	0.008	63.267	25.078	11.647
Meki	-	0.004	75.046	19.048	5.902
Rinzaf	-	0.027	60.352	18.372	21.249
Gombora	-	-	0.037	99.963	-
Hare	0.088	6.050	26.717	51.336	15.809
Kulfo	0.011	8.214	52.289	25.867	13.619
Guder	-	0.054	46.392	53.554	-
Shafe	0.024	14.477	63.226	22.103	0.170
Baso	0.053	9.469	47.959	41.824	0.694
Sagan	0.001	64.896	32.102	2.940	0.061
Bisare	0.006	33.754	66.207	0.032	-
Woji	-	0.004	85.573	7.574	6.849

Based on this classification as shown in table 3.2, majority of the catchments (59.22%) are Wiona Dega, about 24.73% of the area are characterized under Dega. The rest 9.94%, 6.1% and 0.01% falls in to Kola, Wurch and Berha respectively traditionally.

3.1.2 Hydro-Meteorology

The rainfall distribution of the study area is bimodal with the short rainy season ('Belg') extending from March to May and the long rainy season ('Kiremt') from June to September and up to October in the southern part (Gebrehiwot et al., 2019). Since the basin is characterized by a wider altitudinal gradient (table 3.2), it exhibits various hydro climatic classes having different amount of rainfall and temperature (Ulsido and Alemu, 2014; Alemu et al., 2018). Accordingly average annual rainfall varies approximately 786 mm to 1413 mm with the mean of approximately 1223 mm.

RVLB has a diverse network of streams, among those the main rivers flowing into different lake are; Meki and Ketar rivers flowing in to Lake Ziway; Harokelo and Gedemso flowing into Lake Langano; Tikur Wuha River draining into Lake Hawassa; Bilate, Gelana, Gidabo, Hamassa, Baso and Hare drains Lake Abaya; Kulfo flows to Lake chamo; and Weito river which rises from Guge mountains flows into Lake Chew Bahir along with Segen river which rises from Delo mountain that joins Weito river. The irrigation development in the basin is growing (10%) of estimated irrigation potential but, the lakes has not much used due to salinity and other water quality problems (Yirga et al., 2019; Shumet and Kassa, 2016).

3.1.3 Land use and Soil type

The land use condition in the Rift Valley Lake Basin includes; intensively cultivated, moderately cultivated, Shurbland, forest, alpine vegetation, water body, marshland, exposed surface, urban area, grassland, woodland, and riparian vegetation according to landuse map of the area. It is estimated that 67% is intensively cultivated, 14.24% is moderately cultivated, 6.94% is Shurbland, 4.15% is forest, 2.21% is alpine vegetation, 1.83% is water body, 1.27% is marshland, 0.69% is exposed surface, 0.58% is urban area, 0.48% is grassland, 0.43% is woodland, 0.17% is riparian vegetation (figure 3.2).

The soil type of the basin is diverse and interms of the textural component it contains; fine and medium, medium and coarse, coarse, medium, and fine. Statistically fine and medium is 75.79%, medium and course is 10.49%, course is 7.91%, medium is 2.99%, and fine is 2.82% of the study area (figure 3.2).

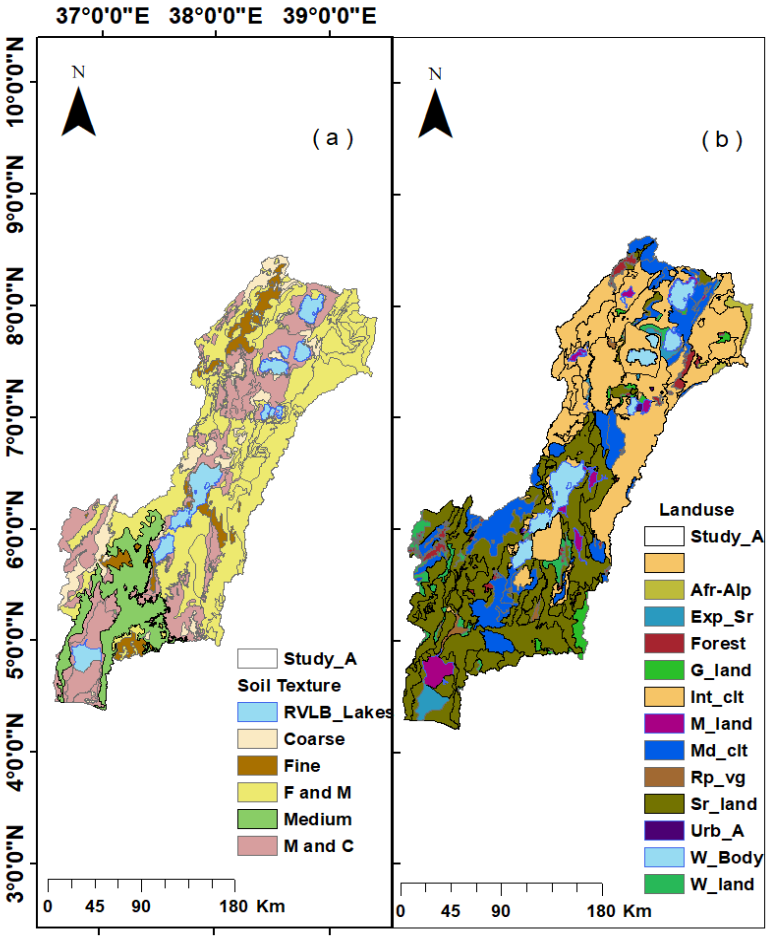

Figure 3-2: Hydrological Soil type (a) and Landuse (b) of the study area

3.2 Data Collection

3.2.1 Time series Data

Stream flow data

To carry out this study hydrological data is collected from the basin development authority in the Ministry of Water, Irrigation and Electricity (MoWIE). In Rift valley lake basin there are about 35 gauging station. Out of 35 gauging stations, 6 gauging stations are used to measure water level of the lakes found in the basin and 1 gauging station (forty spring) is under influence of ground water, so 28 gauging stations are considered for data quality check and after checking quality of each data 23 gauging stations are selected for the analysis. The stations have various length of record with

fair distribution over the basin. Hydrological gauging stations with their respective coordinates, elevation, and years of record are shown in table 3.3 below.

Table 3-3: Location and record period of gauging stations

SN	Gauging Stations	Stn.no	Long.	Lat.	Elevation (m)	Record period
1	Weito	83002	37.43	5.72	673	1985-2007
2	Upper Gelana	82029	38.18	6.15	1857	1989-2006
3	Hammassa_Wajjifo	82030	37.82	6.57	1636	1985-2007
4	Hammassa_Humbo	82036	37.75	6.67	1308	1985-2006
5	Badessa	82034	38.30	6.38	1568	1990-2006
6	Gidabo_Measso	82044	38.43	6.43	2644	1997-2015
7	Kola	82028	38.40	6.63	1861	1990-2006
8	Gidabo_Apposto	82016	38.38	6.75	2596	1985-2006
9	Bilate_Tena	82005	38.13	6.93	1566	1970-2015
10	Tikurwuha_Awasabridge	82012	38.48	7.08	1704	1985-2002
11	Tikurwuha_Dato	82027	38.50	7.10	1700	1985-2006
12	Bilate_Alabakulitto	82008	38.07	7.28	1743	1992-2007
13	Ferfuro	81026	38.12	7.73	2009	1980-2015
14	Gedemso	81016	38.87	7.47	2180	1980-2010
15	Horakelo	81005	38.70	7.67	1606	1989-2007
16	Ketar_Fete	81011	39.05	7.78	2435	1992-2007
17	Chiufa	81025	39.07	7.98	1806	1985-2007
18	Ketar_Abura	81019	39.05	8.07	1780	1995-2010
19	Meki	81018	38.80	8.18	1666	1995-2011
20	Rinzaf	82032	38.37	8.12	2074	1990-2009
21	Gombora	82040	37.92	7.58	2315	1990-2005
22	Hare	82019	37.60	6.07	1186	1990-2007
23	Kulfo	82020	37.53	6.03	1438	1998-2013

Meteorological data

Rainfall data are collected from National Meteorological Agency (NMA) for the purpose of using as one of meteorological attributes used for characterizing flow regime of the area and as a parameter in developing regional regression equation. The meteorological stations used to calculate areal precipitation of the catchments are listed in table 3.4, with their respective type, location, and record period.

Table 3-4: Location, type and observation period of meteorological stations

SN	Met. Stations	Type	Location		Elevation	Record period
			Long.	Lat.		
1	Beto	Ordinary	37.09	6.05	1082	1987-2014
2	Gerese	Ordinary	37.30	5.92	2329	1988-2011
3	Jinka	Principal	36.55	5.77	1373	1987-2014
4	Key Afer	Ordinary	36.73	5.52	1597	1987-2011
5	Morka	Ordinary	37.31	6.42	1221	1989-2012
6	Konso	Principal	37.43	5.33	1431	1987-2014
7	Wolaita	Principal	37.73	6.81	1854	1987-2011
8	Danna 2	Principal	37.55	6.59	1303	1990-2011
9	Danna 1	Principal	37.57	6.64	1279	1994-2014
10	Dilla	Principal	38.30	6.37	1579	1988-2011
11	Yirga Chefe	Ordinary	38.20	6.15	1856	1987-2013
12	Hosana	Principal	37.85	7.57	2307	1987-2014
13	Bilate	Principal	38.08	6.82	1361	1987-2011
14	Awassa	Synoptic	38.48	7.07	1694	1987-2014
15	Butajira plc.st	Ordinary	38.37	8.15	2000	1987-2006
16	Kofele	Principal	38.80	7.07	2620	2001-2014
17	Gidolle	Ordinary	37.37	5.63	2087	2008-2014
18	Bulbula	Ordinary	38.67	7.76	1645	1987-2014
19	Chelelektu	Ordinary	38.13	6.05	1732	1987-2014
20	Dagaga	Ordinary	38.84	7.43	2067	1989-2013
21	Sagure	Ordinary	39.15	7.77	2480	1987-2013
22	Kulumsa	Principal	39.16	8.01	2211	1987-2013
23	Ziway	Principal	38.70	7.93	1640	1987-2014
24	Arba Minch	Synoptic	37.56	6.06	1207	1987-2014
25	Mirab Abaya	Principal	37.78	6.30	1260	1991-2011
26	Burji	Principal	37.87	5.48	1815	1997-2012
27	Konso	Principal	37.43	5.33	1431	1987-2014
28	Bako Gazer	Ordinary	36.57	5.90	1645	2006-2011
29	Teltelle	Ordinary	37.38	5.06	1449	1997-2011
30	Werabe	Principal	38.19	7.85	2057	2006-2014
31	Bui	Principal	38.55	8.33	2054	1987-2014
32	Amaro Kelo	Ordinary	37.90	5.84	1659	1987-2009

3.2.2 Spatial Data

Land use land cover map

Land use land cover map was collected from GIS department in the ministry of water, irrigation and energy (MoWIE) that was developed in 2008 and used as land use attribute in regionalization.

Soil map

The soil map of the study area was obtained from MoWIE of Ethiopia which was compiled as Soil and Terrain Database of East Africa with the scale of 1:1,000,000 produced by Food and Agricultural Organization (FAO) in 2006 to use as soil characteristics in regionalizing watersheds.

Topographic data

The Digital Elevation Model (DEM) 30*30 m of Rift Valley Lake Basin was also collected from MoWIE of Ethiopia to calculate topographic characteristics of the catchments.

3.3 Data Analysis

In order to obtain reliable quantile estimates the hydro-meteorological data used in the analysis has to be checked of possible errors or inconsistency and for any indication that contravene basic statistical assumptions of FFA. Therefore filling of missing data, checking consistency and independency of the observed data has to be given due attention.

3.3.1 Meteorological Data Analysis

Filling of missed rainfall data

Time series data can be missed due to interruption of measurements caused by human induced or natural factors and these data should be reasonably filled before the analysis. Missed rainfall data was calculated from the observations of close by stations to missed station. There are a number of methods to compute the missed value like; arithmetic mean method, normal ratio method and inverse distance weighting method (equation 3.1). From those methods described inverse distance weightage method is used in most scientific researches and best in areas that topography varies rapidly (Lu and David, 2008) and so applied in this study.

$$P_x = \frac{\sum_{i=1}^{n} P_i / {d_i}^2}{\sum_{i=1}^{n} 1 / {d_i}^2} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (Equation\ 3.1)$$

Where, P_x is calculated precipitation at station x, P_i precipitation values of neighboring stations, d_i distance of stations, n is number of stations.

Consistency check

A time series rainfall data should be consistence to have a reliable areal representation. Double mass curve method is often used to test the consistency of rainfall record. Accumulated rainfall at the gauge station whose record is in doubt is plotted as ordinate versus the average concurrent accumulated average rainfall of nearby stations whose rainfall data are reliable (Kohler, 1949). If a record of meteorological data is consistence with time, statistical characteristics of the record like mean, variance, and higher-order moments will not change in time.

When a significant change in trend of the curve is observed, indicates that rainfall data is inconsistent at that station and it should be corrected by using (equation 3.2).

$$P_{cx} = P_x * \frac{M_c}{M} \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (Equation\ 3.2)$$

Where, P_{cx} is the corrected precipitation, P_x is observed precipitation at a time period, M_c is correct slope of the double mass curve, M is the original slope of the double mass curve.

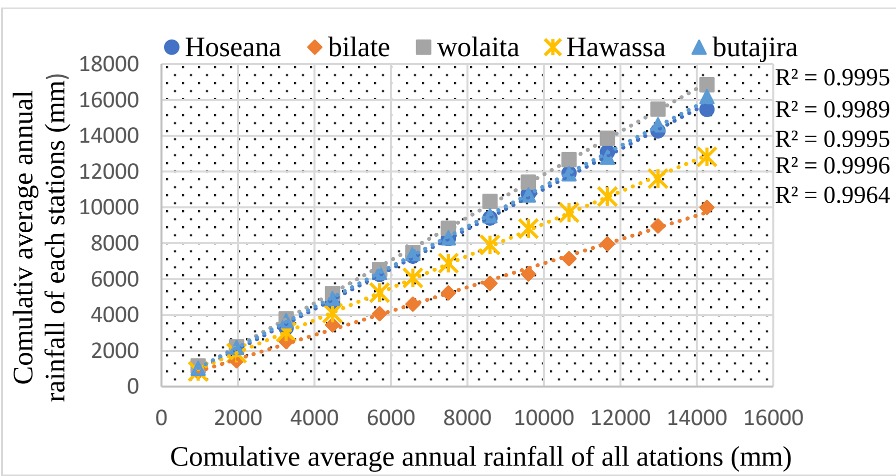

Figure 3-3: Plot of Double mass curve for Bilate Tena catchment

Figure 3.3, shows double mass curve analysis for Bilate Tena catchment, as seen the curves did not show break down in the trend or the lines are fairly smooth. Wolaita, Butajira and Hoseana stations recorded relatively higher annual rainfall than Bilate and Hawassa station as shown in the

curve. The double mass curve analysis was computed for all catchments in the study area and the lines have not gone under significant change in slope, so are fairly smooth.

Average areal rainfall over the catchment

Estimation of average rainfall over the catchment area from rainfall measurements made at a few measurement stations in that area has an important steps in many hydrological applications like management of surface water resources or flood forecasting (Bastin et al., 1984). The most common methods used to calculate average areal precipitation are; arithmetic mean method, isohyetal and Thiessen polygon method.

For this study Thiessen polygon method was used because this method assign relative weights to the gauges in computing areal average (see equation 3.3).

$$\bar{P} = \frac{1}{A}\sum_{i=1}^{n} A_i * P_i\,, \qquad A = \sum_{i=1}^{n} A_i \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (Equation\ 3.3)$$

Where; A_i, is area of polygon i in the watershed (km^2), P_i, rainfall amount in polygon i (mm), $\bar{P}$ is average areal rainfall (mm). Thiessen polygon of stations were developed using Arc GIS 10.4 software for this research. For the case of Bilate Tena catchment; 54.57% of the area was covered by Hoseana gauging station, 24.5% goes to Bilate station, 14.04% to Hawassa station and the rest 4.51% and 2.38% of area are represented by Butajira and Wolaita stations respectively.

Table 3-5: Mean annual rainfall and stations areal coverage of Bilate Tena catchment

Rainfall gauging stations	**Area coverage (km^2)**	**Rainfall (mm)**	**% of area covered**
Hoseana	2028.74	1203.53	54.57
Bilate	910.86	804.86	24.50
Wolaita	88.34	1284.40	2.38
Hawassaa	521.90	970.86	14.04
Butajira	167.83	1333.66	4.51

The developed Thiessen polygon for Bilate Tena catchment is shown in figure 3.4 below.

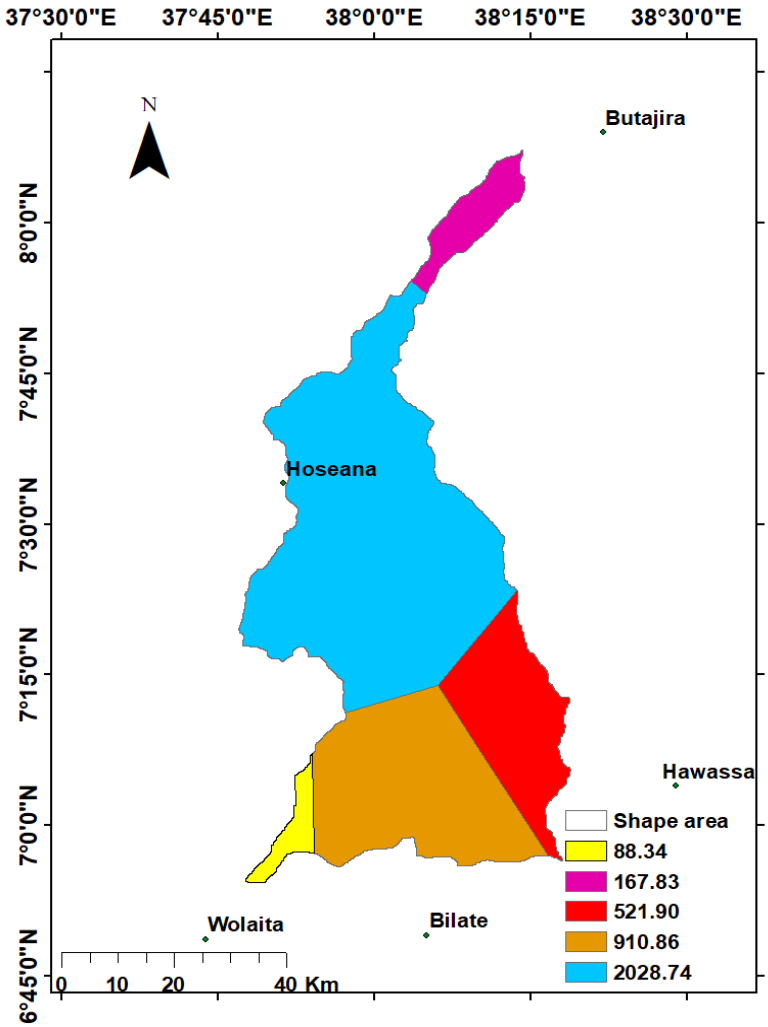

Figure 3-4: Thiessen polygon of Bilate Tena catchment

For all other catchments the above described procedures have been applied.

3.3.2 Stream flow Data Analysis

Filling of missed data

Failure of any gauging stations or absence of observer causes a break in instantaneous daily observed flow. Based on the visual observation the stream flow records of the gauging stations have short to high breaks in observed flow. So, filling those values in scientific and reasonable technique is not an option to arrive at reliable quantile estimate. For this study missed stream flow data were computed using regression analysis having R squared value near to unity and arithmetic mean method.

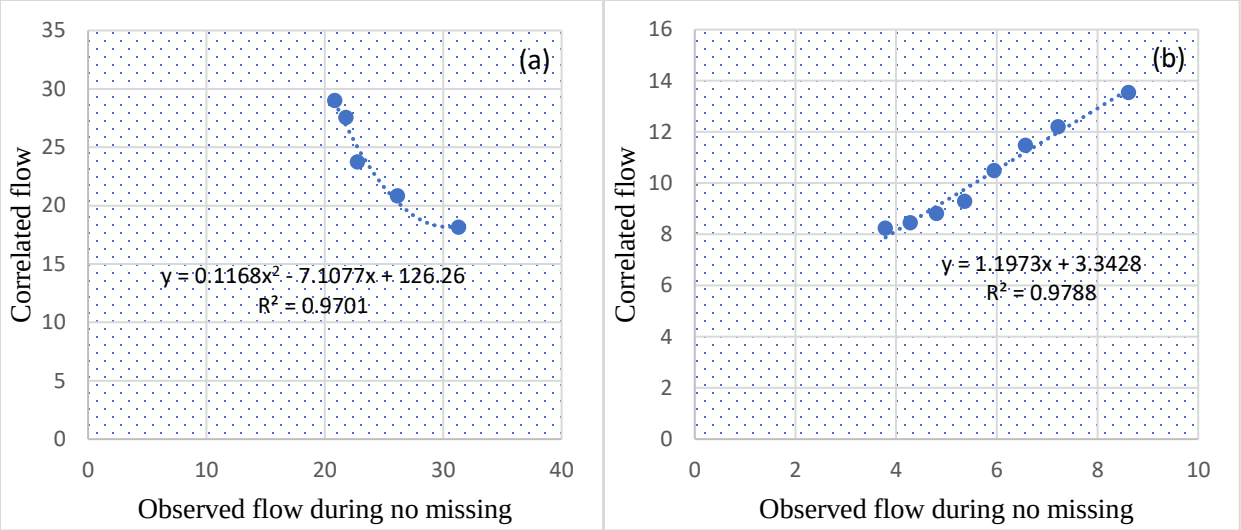

Figure 3-5: Regression equation developed to fill missed stream flows in; (a) Meki, and (b) Keter Fete

Independence test

It is usually assumed that all the peak magnitudes in the AM series are mutually independent in the statistical sense. Independence of the events in the data series may bias the result of frequency analysis (Wolfowitz, 1942; Haddad and Moravej, 1943).

Given a sample of size N, the Hallin and Puri, (1995) (W-W) test is used to test the independence of a data and to search for patterns in the data. For data $x_1, x_2, \dots, x_n$ the statistic R is computed as follows;

$$R = \sum_{i=1}^{N-1} x_i x_{i+1} + x_1 x_N \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (Equation\ 3.4)$$

R follows a normal distribution with mean and variance when the elements in sample are independent (see equations 3.5 and 3.6).

$$\bar{R} = \frac{(s_1^2 - s_2)}{N-1} \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (Equation\ 3.5)$$

$$var(R) = \frac{s_2^2 - s_4}{N-1} - \bar{R}^2 + \frac{(s_1^4 - 4s_1 s_3 + s_2^2 - 2s_4)}{(N-1)(N-2)} \dots\dots\dots\dots\dots\dots\dots (Equation\ 3.6)$$

Where, $s_r = Nm'_r$ and m'_r is the r[th] moment of the sample about the origin.

The statistic $u = {(R - \bar{R})}/{(var(R))^{1/2}}$ is approximately normally distributed with mean zero and variance unity and is used to test the hypothesis of independence at significance level, by comparing the statistic u with the standard normal variate $u_{\alpha/2}$ corresponding to a probability of exceedence $\alpha/2$ (Rao and Hamed, 2000). Accordingly the independence test has been done for each hydrological gauging stations using Wald- Wolfowitz (W-W) see (table 3.6).

Table 3-6: Summary of independence test

St.Name	W-W stastic	Critical test stastic	Remark
Weito	0.949	1.96	Independent
Gato	2.343	1.96	Dependent
Upper Gelana	-0.259	1.96	Independent
Hamassa Wajifo	0.751	1.96	Independent
Hamassa Humbo	0.205	1.96	Independent
Badessa	-0.936	1.96	Independent
Gidabo Measso	-0.318	1.96	Independent
Kola Aletawondo	0.402	1.96	Independent
Gidabo Apposto	1.497	1.96	Independent
Bilate Tena	0.539	1.96	Independent
Tikurwuha Hawassabridge	-2.131	1.96	Independent
Tikurwuha Dato	-0.839	1.96	Independent
Wosha	2.856	1.96	Dependent
Melka Oda	2.681	1.96	Dependent
Bilate Alabakulito	1.917	1.96	Independent
Ferfuro	0.005	1.96	Independent
Gedemso Langano	0.191	1.96	Independent
Horakelo Langano	1.8	1.96	Independent
Ketar Fete	-2.478	1.96	Independent
Kerersitu Adamitulu	0.817	1.96	Independent
Chiufa	1.17	1.96	Independent
Ketar Abura	-0.001	1.96	Independent
Meki	1.059	1.96	Independent
Rinzaf	0.946	1.96	Independent
Guder	2.381	1.96	Independent
Gombora Hoseana	-0.708	1.96	Independent
Hare	0.339	1.96	Independent
Kulfo	0.987	1.96	Independent

According to the test value *u* three stations; Gato Nr. Gidolle (82015), Wosha Nr. Wondo-Genet (82023), Melka odda at Melka odda town (82003) goes beyond the critical value at 5% significance level $u_{0.05} = 1.96$, so they are rejected from the analysis. The hypothesis of independence and stationarity has been confirmed for the remaining stations.

3.4 Delineation of Homogeneous Regions

As the main objective of regionalization is grouping of the watersheds with required homogeneity in a sense of flood generating mechanism, delineation of hydrologically homogeneous region is the most important step.

3.4.1 Attributes used in Regionalization

Geographically contiguous regions based on geographical, political, administrative, or physiographic boundaries have been used for a long time in hydrology for RFFA. However, this practice is not acceptable, because the delineated regions do not guarantee hydrological homogeneity (Rao and Srinivas, 2007).

At-site flood statistics have also been used as attributes for regionalization in the past (Stedinger and Tasker, 1985). A drawback in using flood statistics as attributes to form regions is that the resulting regions may appear homogeneous but are not necessarily effective for RFFA, because of formation of the regions and testing of regional homogeneity is done with the same flood statistics (Amiri et al., 2018).

So, attributes used for regionalization of watersheds in the RVLB includes:

a. **Physiographic characteristics;** drainage area A (km^2) and length of longest stream in the sub-basin Lm (km).
b. **Drainage characteristics;** stream density of the sub-basin S_d (/ km).
c. **Geographical location attributes;** latitude and longitude of gauging stations.
d. **Meteorological characteristics;** mean annual precipitation (mm)
e. **Soil cover characteristics;** percentages of fine and medium, coarse, medium and coarse, fine, and medium interms of texture.
f. **Land use pattern;** percentages of intensively cultivated, moderately cultivated, Shurbland, forest, alpine vegetation, water body, marshland, exposed surface, urban area, grassland, woodland, and riparian vegetation.

The descriptive static's of feature vectors (also referred to as 'data point', 'data vector') is descried in the flowing table.

Table 3-7: Descriptive statics of the attributes

	N	Minimum	Maximum	Mean	Std. Deviation
Area (km^2)	29	73.72	5190.64	1045.14	1364.12
Length of main river (km)	29	21.14	105.60	46.45	22.84
Minimum Elevation	29	585.00	2275.00	1521.03	389.34
Maximum Elevation	29	2629.00	4162.00	3242.79	473.91
Mean Elevation	29	1393.02	2847.02	2149.07	342.57
Gravelius index (Kg)	29	1.40	2.47	1.88	0.31
Stream Density (Sd)	29	.03	.31	0.12	0.06
Average annual PPn. (mm)	29	582.26	1413.03	1088.09	228.23
Latitude	29	5.24	8.18	6.97	0.81
longitude	29	37.43	39.07	38.23	0.49
% Fine and Medium	29	16.85	100.00	72.62	27.10
% Coarse	29	0.00	57.51	8.68	16.69
% Fine	29	0.00	26.22	4.03	7.73
% Medium and Course	29	0.00	44.95	11.85	14.89
% Medium	29	0.00	46.83	2.82	9.79
% Moderately cultivated	29	0.00	87.92	14.87	23.38
% Intensively cultivated	29	0.00	100.00	62.24	35.19
% Grass land	29	0.00	10.83	0.69	2.15
% Riparian vegetation	29	0.00	2.32	0.20	0.61
% Shurbland	29	0.00	86.87	13.72	23.56
% Exposed surface	29	0.00	4.19	0.39	0.99
% Forest	29	0.00	30.74	3.44	7.24
% Urban area	29	0.00	9.89	0.46	1.86
% Alpine vegetation	29	0.00	15.44	1.60	4.04
% Water body	29	0.00	13.33	0.85	2.97
% Woodland	29	0.00	6.62	0.33	1.28
% Marshland	29	0.00	15.74	1.20	3.22

3.4.2 Catchment Clustering

The aim of clustering is to classify catchment groups in such a way that the catchments within a cluster are comparable, while the catchments in different clusters are different. The first step is to specify characteristic distance for the similarity or dissimilarity of the variables' (Isik and Singh, 2008).

In the area where there are climatological and topographical variations, hybrid (hierarchical and k-mean) cluster analysis is best to cluster dataset. The hierarchical clustering processes are not affected by initialization and local minima but, initial guesses influence the k-mean clustering processes (like number of clusters). In the sense that feature vectors can transfer from one cluster to another to reduce the objective function, the k-mean clustering procedures are dynamic. In comparison, in hierarchical clustering procedures, the feature vectors committed to a cluster in the early stages do not transfer to another. So, for these study catchments are preliminary grouped using hierarchical method then using number of clusters from agglomeration schedule then apply to k-means finally using SPSS software.

Hierarchical clustering

There are several hierarchical clustering techniques, such as single linkage, total linkage, centroid, Ward's minimum variance, and the average distance method. In this study Wards method was used to obtain optimum number of clusters because it was identified by different researchers as the method that overtakes other algorithms in terms of separation to give relatively dense clusters within small within group variance (Unal et al., 2003; Murtagh, 2014). The distinctions made at each successive stage of study are represented by a two dimensional diagram known as a dendrogram or tree diagram.

In Ward's method, estimate of the distance amongst clusters is computed using analysis of variance at each stage (see equation 3.7). For any two clusters at any point, it minimizes the total of squared deviations. This technique minimizes the heterogeneity of clusters, i.e., by maximizing intra-group homogeneity, clusters are formed. The intra-group sum of squares of the deviations of values from the average of the clusters is understood as the indicator of homogeneity of clusters, and is Ward's criterion. The criteria for linking clusters is based on the assumption that Ward's criterion is minimally increased at each point of clustering. Ward's approach tends to remove small clusters and generate clusters of about the same size (Kent and Lioyd, 1975).

$$W = \sum_{k=1}^{p} \sum_{i=1}^{N} \sum_{j=1}^{M} \left(x_{kij} - \bar{x}_{kj}\right)^2 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (Equation\ 3.7)$$

$$\bar{x}_{kj} = \frac{1}{N} \sum_{i=1}^{N} x_{kij} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (Equation\ 3.8)$$

Where, W is total within cluster error sum of squares; P represents the number of clusters; N represents the number of gauging stations in each cluster; M denotes the number of variables (Months) in elements; $x_{kij} = j^{th}$ variable at the i^{th} element in the k^{th} cluster; and $\bar{x}_{kj}$ = average value of the i^{th} elements at the j^{th} variable in the k^{th} cluster. Squared Euclidean distance is used as a resemblance degree for clustering.

$$d_{ij} = \sum_{k=1}^{n} (x_{ik} - x_{jk})^2 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (Equation\ 3.9)$$

K-mean clustering

K-mean clustering was implemented in this study to finally classify catchments in a homogeneous clusters after number of cluster is fixed by hierarchical clustering. This approach makes it possible to transfer elements from one cluster to another, a reallocation that is not available in hierarchical methods, with the intention of minimizing the variance within the cluster and maximizing the variance between the clusters (Ralamboundrainy, 1995; Nageshwar and Carol, 1990). As a potential enhancement of hierarchical techniques, the k-means approach was used in this study. K-mean algorithm is provided in equation 3.10 shown below.

$$J = \sum_{j=1}^{n} \sum_{i=1}^{n} \left\| X_i^{(j)} - C_j \right\|^2 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (Equation\ 3.10)$$

Where, $\left\| X_i^{((j)} - C_j \right\|^2$ is the selected distance measure between a data point $X_i^{(j)}$ and the cluster center C_j. The basic algorithm follows the following steps for the specified number of clusters k:

a. Split the data into initial k clusters,
b. The means or centroids of the k clusters are computed,
c. Calculate the distance of each centroid for each case and allocate it to the closest centroid,
d. Repeat Steps 2 and 3 until no catchments are reallocated.

3.4.3 Regional Homogeneity Analysis

Catchments that are grouped based on k-mean method again has to be tested statistically to further know if, they are really homogenous. Firstly, the sites was checked for possible discordance then using site to site coefficient of variations (CC) statistical homogeneity of each region has been further evaluated.

Discordancy Test

The aim of this test is to identify the sites that are grossly discordant from the group as a whole (see equation 3.11).

$$D_i = \frac{1}{3} N (U_i - \overline{U})^T A^{-1} (U_i - \overline{U}) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (Equation\ 3.11)$$

Where, D_i is the discordancy amount for catchment i, N is the number of catchments in the region, U_i is a vector containing the $L-CV$, $L-skweness$, and $L-kurtosis$ for a catchment i, $\overline{U}$, is the regional mean for U_i, and A is expressed as;

$$A = \sum_{j=1}^{N} (U_j - \overline{U})(U_j - \overline{U})^T \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (Equation\ 3.12)$$

To declare site i to be discordant, D_i of that specific site has to pass certain critical value that depends on the number of sites in the region suggested by (Hosking and Wallis, 1997). The critical value of D_i for proposed number of sites is given in the (table 3.8).

Table 3-8: Critical values for the discordancy statistic D_i

Number of sites in a region	Critical value	Number of sites in a region	Critical value
5	1.333	10	2.491
6	1.648	11	2.632
7	1.917	12	2.757
8	2.140	13	2.869
9	2.329	14	2.971
		≥15	3

CV- Based Homogeneity Test

Cunnane, (1989) suggests the procedures described below to show regional homogeneity of the proposed stations. For each catchment in a region compute mean ($\bar{Q}$), standard deviation (σ) and coefficient of variation (CV).

$$\bar{Q}_i = \frac{\sum_{i=1}^{n} Q_i}{n_i} \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots (Equation\ 3.13)$$

$$\sigma_i = \sqrt{\frac{\sum_{i=1}^{n}(Q_i - \bar{Q}_i)^2}{n-1}} \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots (Equation\ 3.14)$$

$$CV_i = \frac{\sigma_i}{\bar{Q}_i} \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots (Equation\ 3.15)$$

Where, Q_i is the flow rate of station i

($\bar{Q}$), The mean flow for sites i

(σ_i), Standard deviation of Q_i for site i

(CV_i), Coefficient of variation for site i

For each region, using the statistics above, computed the regional mean, CV and finally the resulting CC is calculated using the equation shown below;

$$\overline{CV}_i = \frac{\sum_{i=1}^{n} CV_i}{n_i} \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots (Equation\ 3.16)$$

$$\sigma_{cv} = \sqrt{\frac{\sum_{i=1}^{n}(CV_i - \overline{CV}_i)^2}{n-1}} \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots (Equation\ 3.17)$$

$$CC = \frac{\sigma_{cv}}{\bar{Q}_i} \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots (Equation\ 3.18)$$

Where, n, is the number of site in a region

$\overline{CV}$, Mean coefficient of variation

σ_{cv}, Standard deviation of at-site CV values

The station in a region confirmed to be homogenous if, $CC < 0.3$

LCV- Based Homogeneity Test

LCV- based homogeneity measure is more accurate and effective way of testing the homogeneity of the site (station) when compared with that of the CV-based homogeneity test (Lettenmaier, 1985; Demissie, 2008). Hosking and Wallis, (1986) gives unbiased estimators of M_{10k} and M_{1j0} as;

$$M_{10k} = \frac{1}{n}\frac{\sum_{i=1}^{n}\binom{n-i}{k}}{\binom{n-1}{k}}X_i, \quad k = 0,1,2,\dots,n-1 \dots\dots\dots\dots\dots\dots\dots\dots (Equation\ 3.19)$$

$$M_{1j0} = \frac{1}{n}\frac{\sum_{i=1}^{n}\binom{i-1}{j}}{\binom{n-1}{j}}X_i, \quad J = 0,1,2,\dots,n-1 \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (Equation\ 3.20)$$

Where, i is rank of observed stream flow data in ascending order.

The first few moments are;

$$L_1 = M_{100} \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (Equation\ (3.21)$$

$$L_2 = M_{100} - 2 * M_{101} \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (Equation\ 3.22)$$

$$L_3 = M_{100} - 6 * M_{101} + 6 * M_{102} \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (Equation\ 3.23)$$

$$L_4 = M_{100} - 12 * M_{101} + 30 * M_{102} - 20 * M_{103} \dots\dots\dots\dots\dots\dots (Equation\ 3.24)$$

Like the conventional moments L-moment can be used to specify and summarize probability distribution. In particular L_1, the first L-moment, is the mean of the statistical distribution and identical to the first conventional moment, and L_2 is a linear measure of spread or dispersion similar to standard deviation (Demissie, 2008; Yirefu, 2010). L-moment ratio, which are analogous to conventional moment ratio are expressed by Rao and Hamed, (2000) as;

$$\tau = \frac{L_1}{L_2}, measure\ of\ scale\ and\ dispersion\ (L_{cv}) \dots\dots\dots\dots\dots\dots (Equation\ 3.25)$$

$$\tau_3 = \frac{L_3}{L_2}, measure\ of\ skewness\ (L_{cs}) \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (Equation\ 3.26)$$

$$\tau_4 = \frac{L_4}{L_2}, measure\ of\ kurtosis\ (L_{ck}) \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (Equation\ 3.27)$$

Using the above procedural formula we have;

$$L_{cv} = \frac{L_2}{L_1} \text{ , and } \bar{L}_{cv} = \frac{\sum_{i=1}^{n} L_{cv}}{n} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (Equation\ 3.28)$$

$$\sigma_{L_{cv}} = \sqrt{\frac{\sum_{i=1}^{n}(L_{cv} - \bar{L}_{cv})^2}{n-1}} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (Equation\ 3.29)$$

$$CC = \frac{\sigma_{L_{cv}}}{\bar{L}_{cv}} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (Equation\ 3.30)$$

The stations in a region confirmed to be homogeneous if, $CC < 0.3$.

3.5 Statistical Distribution, Parameter and Standard Error Estimation Methods

3.5.1 Methods of selecting distribution models

Goodness-of-fit test measurements was used to select best fit distribution from candidate distribution using software which is called Easy Fit available at *http://www.mathwave.com/easyfit-distribution -fitting.html*. Using this software goodness of fit tests; Chi-Square, Kolmogorov-Smirnov and Anderson Darling tests are done and the best fit distribution from the candidate is displayed automatically.

Kolmogorov–Smirnov (K-S)

The Kolmogorov–Smirnov (K–S) test, the highest class of statistics on empirical distribution functions (EDF) is focused on the highest change between the hypothetical and empirical distributions (Frank, 2012). The aim of this test is to relate the empirical cumulative frequency $S_n(x)$ with the cumulative distribution function of an assumed theoretical distribution $F_x(x)$ the highest variance among $S_n(x)$ and $F_x(x)$ is the K-S test statistic. For a model size n, the data are reordered in ascending order where $x_1 < x_2 < \cdots < x_1$ and the K-S statistic is evaluated for each value;

$$S_n(x) = 0; if\ x < x_1 = {}^{k}/_{n}\,; if\ x_k\langle x_{k+1} = 1; if\ x\rangle x_n \ldots\ldots\ldots (Equation\ 3.31)$$

$$D_n = max|F_x(x) - S_n(x)| \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (Equation\ 3.32)$$

$$P(D_n \leq D_n^{\alpha}) = 1 - \alpha \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (Equation\ 3.33)$$

Where D_n^{α} the critical value, α is the highest level, and k is the descending order of the data set.

Anderson–Darling (A–D)

The Anderson–Darling (A–D) is accurate at the tails of the partial differential functions. For comparatively high extremes, to select the best-fit distribution for annual daily peak discharges, the A-D test may be more suitable. In the A–D test measure, the quadratic class of the empirical distribution function test statistic, is described as A^2 as follows;

$$A^2 = -\sum_{i=1}^{n}\left[(2i-1)lnF_x(x_i) + ln\frac{[1-F_x(x_{n+1-i})]}{n}\right] - n \ldots\ldots\ldots \quad (Equation\ 3.34)$$

Where, $F_x(x_i)$ is the CDF of the expected distribution at x_i for, $i = 1,2,\ldots,n$. The annual daily peak-flow data are arranged in ascending order, as $x_1 < x_2 < \cdots x_n$. The hypothetical and experimental cumulative distribution functions at the tails of the PDF are both comparatively smooth in the K-S test, which means that the A-D test gives the tails more precision (Sinclair and Spurr, 2012).

Chi-squared (C–S)

Chi-squared (C–S) test relates the theoretical probability distribution functions (PDF) with the empirical probability distribution functions (see Equation 3.35).

$$x^2 = \sum_{i=1}^{k}\frac{(Q_{i-}E_i)}{E_i} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (Equation\ 3.35)$$

Where x^2 is the test statistic, Q_i is observed frequency, and E_i is expected frequency explained as;

$$E_i = F(x_2) - F(x_2) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (Equation\ 3.36)$$

Where, F is the cumulative distribution function of theoretical probability distribution functions, x_1 and x_2 are the lower and upper limits (Kim, 1993).

3.5.2 Parameter Estimation Methods

Once appropriate distribution has fitted to a region, a parameter estimation method of low sampling variance and easily computable has to be done. For this study most commonly used methods; Method of Moment (MOM), Probability Weighted Moment (PWM) and Maximum Likelihood (ML) was compared based on their standard error values.

Method of Moment (MOM)

Estimates of the parameters of a probability distribution function are obtained in the MOM by equating the moments of the sample with the moments of the probability distribution function. For a distribution with n parameters, $\alpha_1, \alpha_2, \dots, \alpha_n$ which are to be estimated, the first n sample moments are set equal to the corresponding population moments that are given in terms of unknown parameters. These n equations are then solved simultaneously for the unknown parameters $\alpha_1, \alpha_2, \dots, \alpha_n$. So, following this general concept parameters have been estimated for the fitted distributions using detailed formulas on (Rao and Hamed, 2000).

Probability Weighted Moment (PWM)

In this process, parameter value can be estimated, as in the case of MOM, by equating distribution moments with the corresponding sample moments. For a distribution with n parameters, $\emptyset_1, \emptyset_2, \dots, \emptyset_n$ are to be estimated, the first n sample moments are set equal to the corresponding population moments. The resulting equations are the solved simultaneously for the unknown parameters $\emptyset_1, \emptyset_2, \dots, \emptyset_n$. Like that of MOM in computing the parameters using PWM the detail formulas on Rao and Hamed, (2000) have been applied for this study.

Maximum Likelihood (ML)

Estimating by the ML technique involves selecting parameter estimates that give the observations the highest likelihood of occurrence (Rao and Hamed, 2000). For a distribution with a probability density function (PDF) given by $f(x)$ and parameters $\alpha_1, \alpha_2, \dots, \alpha_n$ the likelihood function is expressed as the joint pdf of the observations conditional on the given values of the parameters $\alpha_1, \alpha_2, \dots, \alpha_n$ as in (equation 3.37).

$$L(\alpha_1, \alpha_2, \dots, \alpha_n) = \prod_{i=1}^{k} f(x_i;\ \alpha_1, \alpha_2, \dots, \alpha_n) \dots\dots\dots\dots\dots\dots\dots\dots\dots (Equation\ 3.37)$$

The value of $\alpha_1, \alpha_2, \dots, \alpha_n$ that maximizes the likelihood function are calculated by partial differentiation with respect to $\alpha_1, \alpha_2, \dots, \alpha_n$ and setting this partial derivatives equal to zero as shown in equation 3.38 the resulting set of equations are then solved simultaneously to obtain the values of $\alpha_1, \alpha_2, \dots, \alpha_n$.

$$\frac{\partial \ln L(\alpha_1, \alpha_2, \dots, \alpha_n)}{\partial \alpha_i} = 0\ ;\ \ i = 1,2, \dots n \dots\dots\dots\dots\dots\dots\dots\dots\dots (Equation\ 3.38)$$

The equations are usually complex that can only be solved by numerical techniques but, in this study it was computed using Easy Fit software. Easy Fit software is a data analyzing program that automates the process of fitting statistical distribution to a data set. In this software different tests such as Kolmogorov Simrinov, Anderson Darling and Chi-squared have been applied.

3.5.3 Standard Error

The parameters estimated by those methods have their own measure of variability which is standard error of estimates S_T as shown in (equation 3.39).

$$S_T = \sqrt{E\{x_T - E(x_T)\}^2} \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (Equation\ 3.39)$$

The standard error of estimate depends generally on the method of parameter estimation. Each method gives a different value of standard error estimates. So, the most efficient method is the one which gives small standard error of estimates. Therefore in this study standard errors of fitted distributions have been calculated for all the three parameter estimation methods mentioned above using the detail formulas explained in (Rao and Hamed, 2000).

3.6 Regional Frequency Curve

Index flood estimation at a given site is the methodology used to assess the average value of the local maximum annual peak distribution, provided the basin lies in a homogenous region (Bocchiola et al., 2003). In this case, regional quantile estimates Q_T at a given site for a given return period T can be obtained as in (equation 3.40) where, q_T is the quantile estimate from the regional distribution for the given return period, and μ_i is the mean flow at the site (Rao and Hamed, 2000).

$$Q_T = \mu_i q_T \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (Equation\ 3.40)$$

Regional distribution parameters were obtained using the regional weighted average of dimensionless moments obtained by the use of dimensionless rescaled data $q_{ij} = {Q_{ij}}/{\mu_i}$. Accordingly standardized regional quantile (${Q_T}/{\bar{Q}}$) versus required return period was drawn and one can get quantile estimate of specific site of any return period in a region by just multiplying rescaled regional quantile with mean flood of that site.

3.7 Regional Regression Model

In this study regression analysis was applied to predict mean annual flood (MAF) for ungauged catchments in the study area. Prior to the development of regression equation correlation matrix between pairs of catchment characteristics have been identified. Regional regression is built as a multiple regression representing the relationship between mean annual flood (dependent variable) and morpho-climatic parameters (independent variables) like; catchment area (A), length of the main river (Lm), gravelius index (Kg), stream density (Sd) and mean annual rainfall (MAR).

$$\bar{Q} = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + \cdots + \beta_{p-1} * x_{p-1} \dots\dots\dots\dots\dots\dots (Equation\ 3.41)$$

Where; $\bar{Q}$ is index flood, x_i is morpho-climatic characteristics of the catchment and β_i is regression coefficients.

3.8 Validation of Regression Equations

Validation of regional regression model was assessed using mean annual flood (MAF) of five verification catchments excluded from regression model development (i.e Tikurwuha at Dato, Bilate at Tena, Ketar at Abura, Hamassa at Wajifo and Hare) considered as pseudo ungauged catchments. Independent morpho-climatic characteristics of those catchments have been applied in regression equation to calculate predicted quantile to analyze the relative error as shown in the (equation 3.42).

$$RE = \left|\frac{Q_P - Q_O}{Q_O + 1}\right| * 100 \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (Equation\ 3.42)$$

Where, Q_P and Q_O are predicted and observed quantile flows respectively, RE is the computed relative error.

3.9 Materials Used

- ✓ Arc GIS 10.4; for catchment delineation and catchment characteristics extraction
- ✓ IBM SPSS 20; were used for cluster analysis and for developing regional regression model
- ✓ RStudio 3.4 "nsRFA" package; were used for L-moment and regional homogeneity analysis
- ✓ Easy Fit 5.5; were used for fitting statistical distribution model

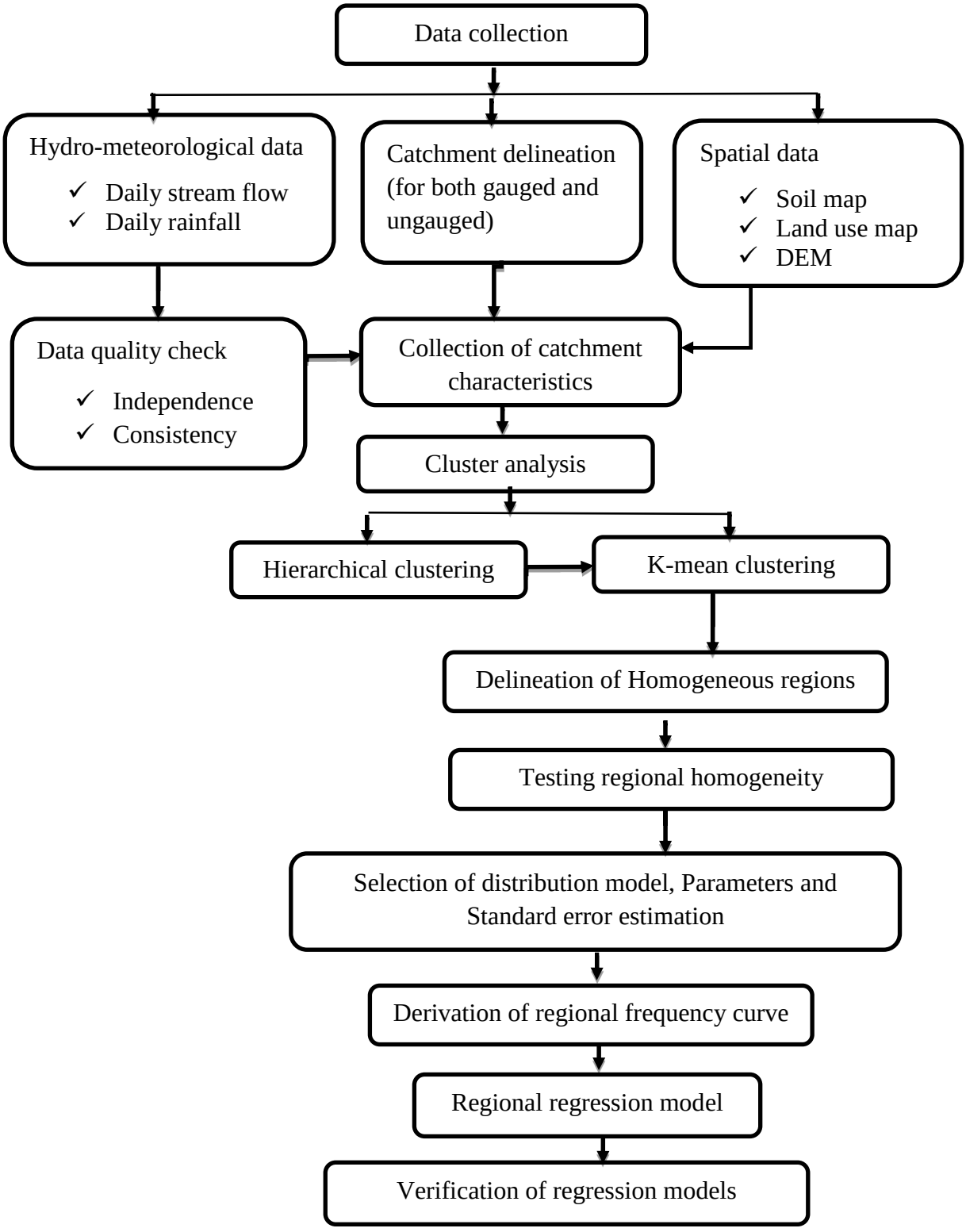

Figure 3-6: Conceptual frame works for the study

4 RESULTS AND DISCUSSION

In this study homogeneous regions were identified using cluster analysis and tested for hydrological homogeneity. Using goodness of fit tests, distribution models have been selected and parameters estimated for the fitted distributions, which are efficient based on standard error analysis. Unique regional frequency curve and regional regression model was developed and tested for validity to calculate quantiles at ungauged catchments as presented in this section.

4.1 Cluster Analysis

4.1.1 Hierarchical Clustering

The main purpose of hierarchical clustering in this study is to get optimum number of cluster and to have initial cluster memberships of the catchments. Then, using outputs from hierarchical method the K-means approach have been computed to obtain possible homogeneous groups after several iterations. Using the proximity matrix squared Euclidean distance was calculated between all pairs of attributes. The screen plot of distances (coefficients column) against stage using agglomeration schedule (Annex-E) is used to decide number of clusters as shown in (figure 4.1).

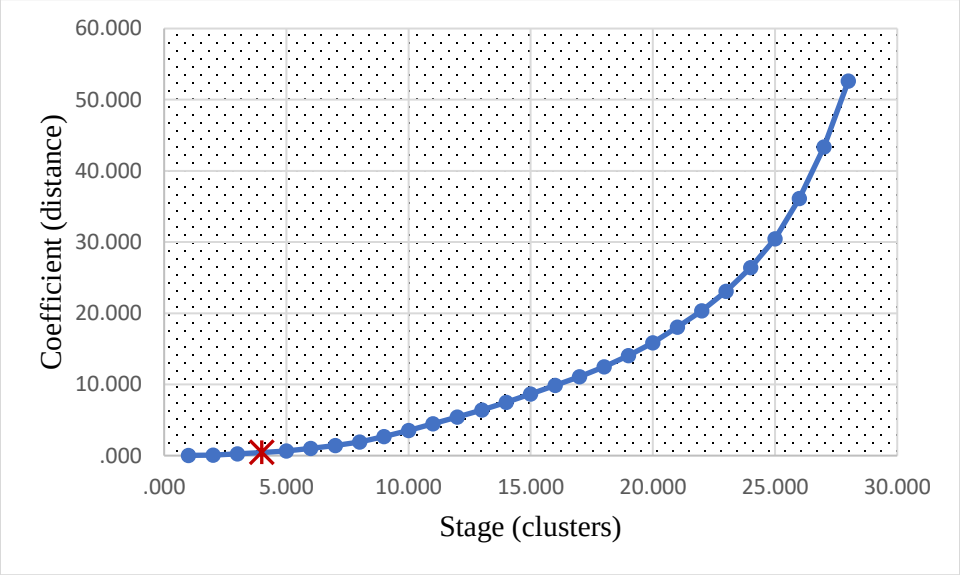

Figure 4-1: Scree plot of coefficients (distance) by stage (number of clusters)

When there is a big change between the coefficients of two succeeding stages the trend in the screen plot altered and elbow occurs.

This indicates that the heterogeneity of the clusters being combined is growing and that it would be ideal to stop the clustering process before the clusters become too different. In figure 4.1, there is a greater difference in the coefficients between stages 4 and 5. With a change of about 0.225, this is the first clear escalation from visual inspection of screen plot and agglomeration schedule. Therefore cluster has been chosen to stop at 4th stage suggesting four cluster is optimal to form hydrologically homogeneous regions in the Rift Valley Lake Basin.

Figure 4.2, shows the dendrogram produced by SPSS with an added crosses horizontal lines specifying number of clusters and the numbers on the y-axis indicating the lists of catchment numbers prior to initial cluster membership.

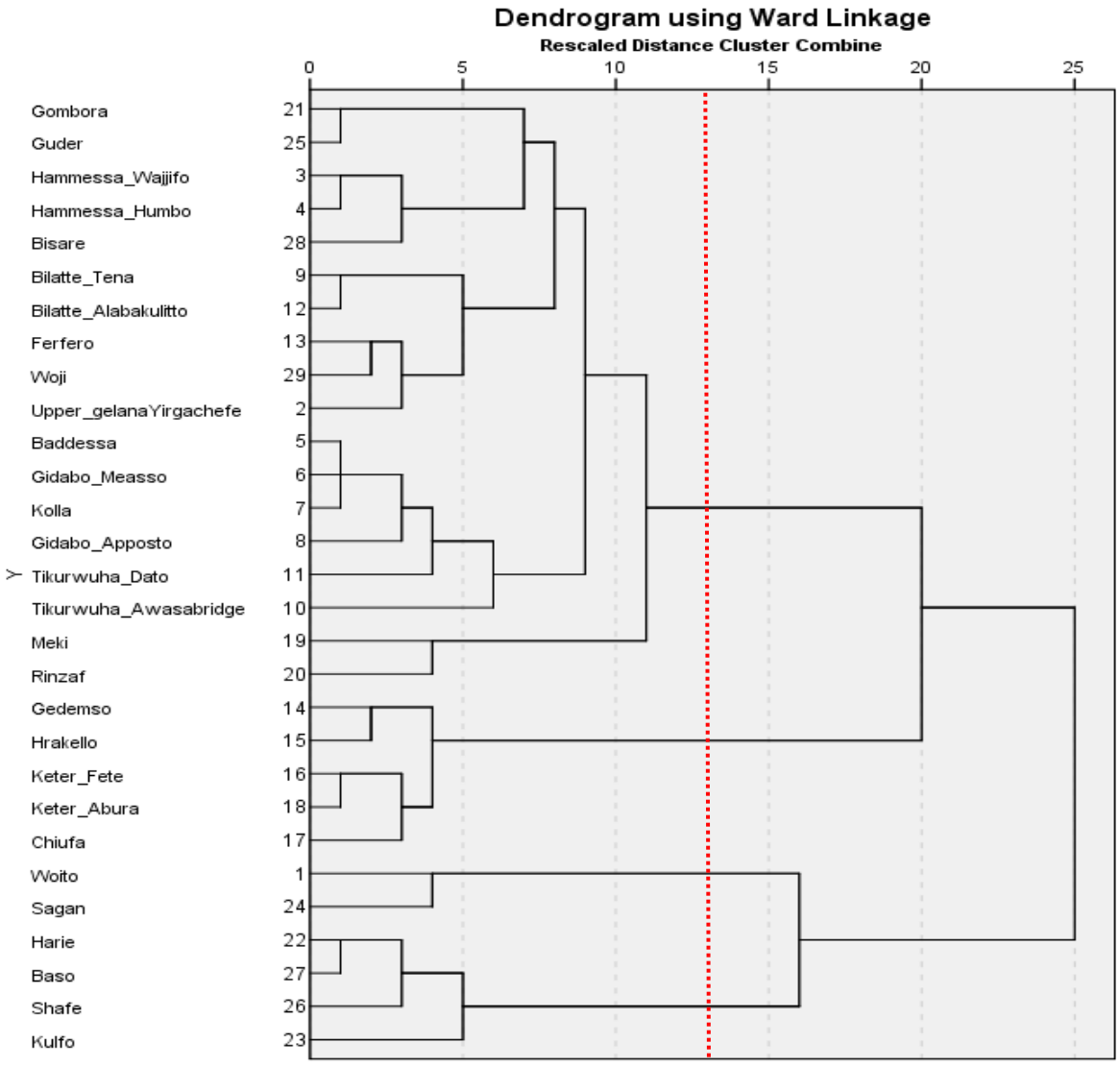

Figure 4-2: Dendrogram using Ward method.

This plot indicates the distances of clusters at each cluster number defined by intersecting the branches using Wards method as initial cluster. From the screen plot it has been concluded to stop the cluster examination after the 4th stage. This decision is reflected in the dendrogram where the last three vertical lines (representing the last three stages in the agglomeration schedule) were cut from the cluster solution. By stopping the clustering at this point, four clusters are revealed within the dataset as the cut-off line crosses four horizontal lines.

4.1.2 K-means Clustering

The final cluster membership of each stations have been iterated using k-means method in such a way that minimizes within-cluster variance and maximizes between-cluster variance. Accordingly the clustering was performed with four cluster solution obtained from hierarchical clustering output. The final cluster center using standardized catchment characteristics is shown in the (table 4.1) the negative sign indicating the position with respect to cluster center.

Table 4-1: Final cluster center

Catchment characteristics	**Cluster**			
	1	2	3	4
Area (km^2)	2.79684	-.57663	.36024	-.51597
Length of the main channel (L_m)	-.01956	-.09753	.58230	-.51012
Minimum elevation	-2.04200	.80090	.54108	-.69331
Maximum elevation	.05952	-.81437	1.07533	-.51718
Mean elevation	-2.12557	.24906	.60312	-.35235
Gravelius index (K_g)	.95036	.01982	-.00550	-.19845
Stream density (S_d)	-1.10303	1.03121	-.51935	.01811
Average annual PPn	.18447	.10890	-.23349	.12036
Latitude	-1.85195	.28577	.96222	-.79187
Longitude	-1.49026	.02356	.93033	-.64877
%Fine and Medium	-1.24962	.40309	-.03485	.00261
%Coarse	.00350	-.49286	.27585	.06846
%Fine	.03148	-.18066	.42799	-.30783
%Medium and Coarse	1.27302	.10179	-.27881	-.04704
%Medium	1.49286	-.28775	-.28775	.19060
%Moderately cultivated	.08114	-.60045	-.24673	.65082
%Intensively cultivated	-1.72997	.87484	.24717	-.51356
%Grassland	2.20057	-.21641	.03203	-.32065
%Riparian vegetation	2.91505	-.32991	-.32991	-.02216

%Shurbland	2.20630	-.58128	-.49097	.45661
%Exposed surface	-.39772	-.39772	.75567	-.39772
%Forest	.13186	-.47387	.69568	-.39035
%Urban area	-.24606	.67556	-.17762	-.24606
%Alpine vegetation	-.39600	-.39600	.75241	-.39600
%Water body	-.28556	.35446	.09097	-.28198
%Woodland	3.16571	-.25862	-.19349	-.25862
%Marshland	-.37249	.32551	.07817	-.23153

The cluster membership is obtained using k-means method after five iterations that are more similar interms of proposed similarity measure. Table 4.2, shows the groups of catchments clustered together using k-means algorithm.

Table 4-2: Cluster membership

Case Number	St.name	Cluster	Distance
1	Weito	1	3.577
2	Upper Gelana	4	4.321
3	Hammassa Wajifo	4	3.864
4	Hammassa Humbo	4	3.533
5	Badessa	4	3.313
6	Gidabo Measso	4	3.080
7	Kola	2	2.643
8	Gidabo Apposto	2	3.726
9	Bilate Tena	3	4.579
10	Tikurwuha_Hawasabridge	2	6.294
11	Tikurwuha Dato	2	4.965
12	Bilate Alabakulito	3	3.809
13	Ferfuro	2	2.910
14	Gedemso	3	3.033
15	Horakelo	3	4.066
16	Ketar Fete	3	4.720
17	Chiufa	3	4.731
18	Ketar Abura	3	3.595
19	Meki	3	5.232
20	Rinzaf	3	4.937
21	Gombora	2	2.788
22	Hare	4	4.080
23	Kulfo	4	5.346

24	Sagan	1	3.577
25	Guder	2	3.468
26	Shafe	4	4.140
27	Baso	4	3.016
28	Bisare	4	3.538
29	Woji	3	3.874

From table 4.2, two catchments are grouped in cluster one, seven catchments in cluster two, ten catchments in cluster three and the rest ten in cluster four. The hydrological homogeneity of each cluster is done in the following session to test the output gained here from SPSS.

4.2 Regional Homogeneity Test Outputs

Once a set of initial regions are identified, it is necessary to access the regions hydrological similarity. The tests applied in this study are; measure of discordance and dispersion analysis (Cv-based homogeneity and LCv-based homogeneity measure).

4.2.1 Discordance Measure

Based on the preliminary definition sketch for discordancy Ferfuro catchment is found to be discordant from region two as the average regional L-moments of this region is far from the average value. Chiufa and Rinzaf catchments are also stations that are discordant from region three and also for the region four Gidabo near Measso as shown in (figures 4.3 and 4.4).

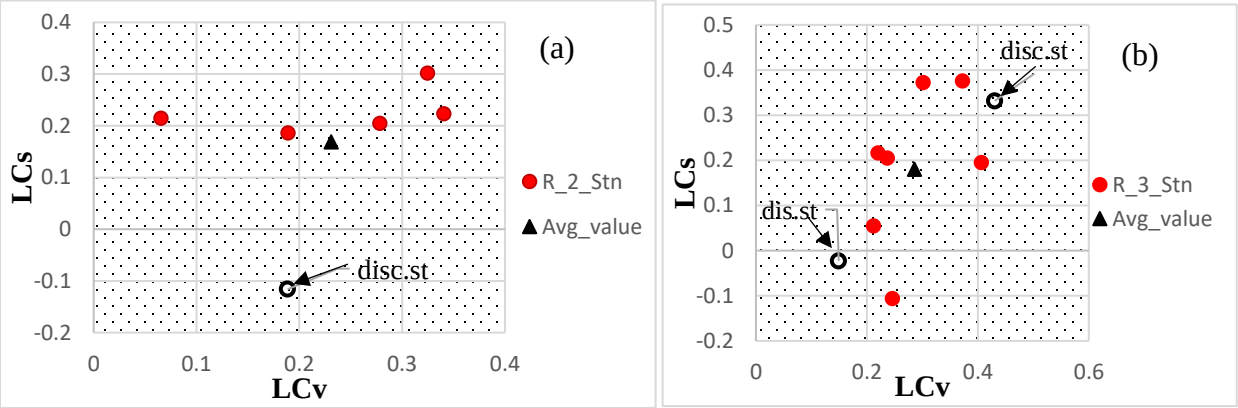

Figure 4-3: (a) LCv versus LCs of AM flow for region one, (b) LCv versus LCs of maximum AM for region two

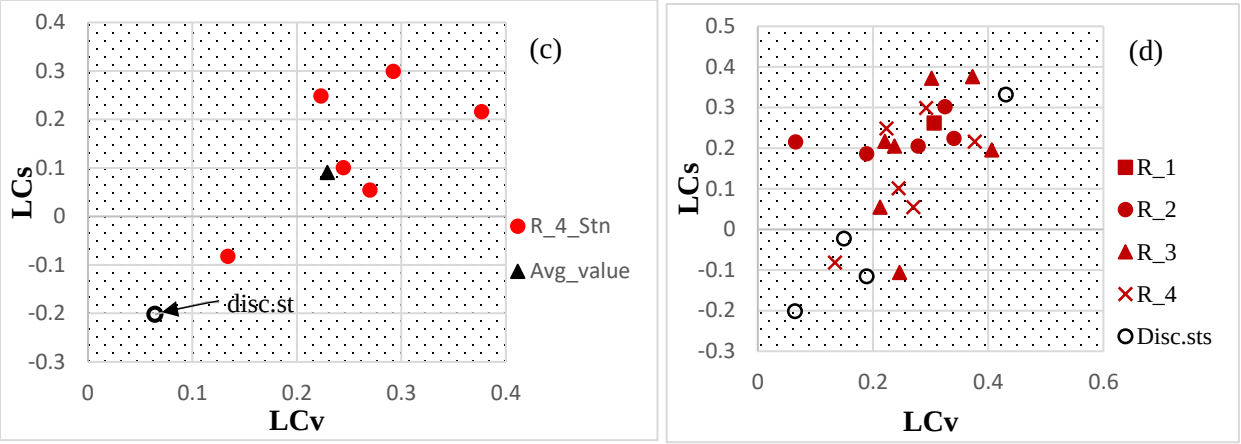

Figure 4-4: (c) LCv versus LCs of AM for region three, (d) LCv versus LCs of AM within all regions

The discordance value of each station within a corresponding region is provided in table 4.3, for region two, three and four respectively. Region one consists only one gauged station, thus cosidered as independent homogeneous region or at site frequency curve is developed independently if it fits to unique distribution. Where, τ is measure of scale and dispersion, τ_3 ,measure of skewness and τ_4 is measure of kurtosis.

Table 4-3: Discordance test result for region two

St. Name	**τ**	**τ_3**	**τ_4**	**Discordancy Measure**	**Remark**
Kola	0.340	0.223	0.077	0.45	Homogeneous
Gidabo Aposto	0.278	0.205	0.169	0.87	Homogeneous
Tikurwuha_Hawassabridge	0.189	0.186	-0.025	1.22	Homogeneous
Tikurwuha Dato	0.065	0.215	0.289	1.21	Homogeneous
Gombora	0.325	0.302	0.244	1.25	Homogeneous

Table 4-4: Discordance test result for region three

St. Name	τ	τ_3	τ_4	Discordancy Measure	Remark
Bilate Tena	0.237	0.205	0.105	0.34	Homogeneous
Bilate Alabakulito	0.212	0.054	-0.063	1.27	Homogeneous
Gedemso	0.373	0.376	0.093	0.93	Homogeneous
Horakelo	0.406	0.195	0.059	1.28	Homogeneous
Ketar Fete	0.220	0.216	0.298	1.24	Homogeneous
Ketar Abura	0.246	-0.106	-0.009	1.34	Homogeneous
Meki	0.302	0.372	0.260	0.60	Homogeneous

Table 4-5: Discordance test result for region four

St. Name	τ	τ_3	τ_4	Discordancy Measure	Remark
Upper Gelana	0.270	0.054	-0.028	1.23	Homogeneous
Hamassa Wajifo	0.134	-0.082	0.306	1.62	Homogeneous
Hamassa Humbo	0.377	0.216	0.025	1.37	Homogeneous
Badessa	0.292	0.299	0.075	0.68	Homogeneous
Hare	0.223	0.248	0.074	0.96	Homogeneous
Kulfo	0.244	0.101	0.064	0.14	Homogeneous

Number of stations in region two is five and the critical discordance statistics is $D_i < 1.33$ so, stations within this region have meet the requirement and therefore are not discordant. Region three which contains seven stations the critical value is $D_i < 1.917$ and no stations have passed this value according to (table 4.4). Finally for region four critical value is $D_i < 1.648$ because it contains six stations and gain no station have passed the threshold value.

4.2.2 CC-Based Homogeneity Test

In this study site to site coefficient of variation of both conventional moment and L-moment are used. Table 4.6, shows CC values of both conventional and L-moments.

Table 4-6: Cv and LCv based homogeneity measures for the regions in RVLB

	St. Name	Cv in con. moment	Cv in L moment	Conclusion
Region one	Weito	0.560	0.307	Homogeneous
Region two	Kola	0.597	0.340	Homogeneous
	Gidabo Apposto	0.500	0.278	
	Tikurwuha Hawasabridge	0.329	0.189	
	Tikurwuha Dato	0.327	0.189	
	Gombora	0.590	0.325	
	Mean	0.468	0.264	
	Standard deviation	0.120	0.065	
	CC	**0.256**	**0.245**	
Region three	Bilate Tena	0.420	0.237	Homogeneous
	Bilate Alabakulito	0.360	0.212	
	Gedemso	0.678	0.373	
	Horakelo	0.703	0.406	
	Ketar Fete	0.401	0.220	
	Ketar Abura	0.416	0.246	
	Meki	0.569	0.302	
	Mean	0.507	0.285	
	Standard deviation	0.131	0.072	
	CC	**0.259**	**0.251**	
Region four	Upper Gelana	0.462	0.270	Homogeneous
	Hamassa Wajifo	0.242	0.134	
	Hamessa Humbo	0.661	0.377	
	Badessa	0.527	0.292	
	Hare	0.396	0.223	
	Kulfo	0.419	0.244	
	Mean	0.451	0.257	
	Standard deviation	0.128	0.073	
	CC	**0.283**	**0.286**	

From the above computed values all regions satisfy the criteria of CC being < 0.3 in both conventional moments and L-moment. Site to site coefficient of variation as expressed in (table 4.6) is 0.256 and 0.245 both conventional and L-moments respectively for region two. For region

three 0.259 and 0.251 respectively that shows homogeneity within the region. The final region four has also 0.283 and 0.286 still satisfy the criteria, therefore all regions are hydrologically homogeneous. Figure 4.5, shows the delineated homogeneous regions of Rift Valley Lake Basin (RVLB) from ArcMap.

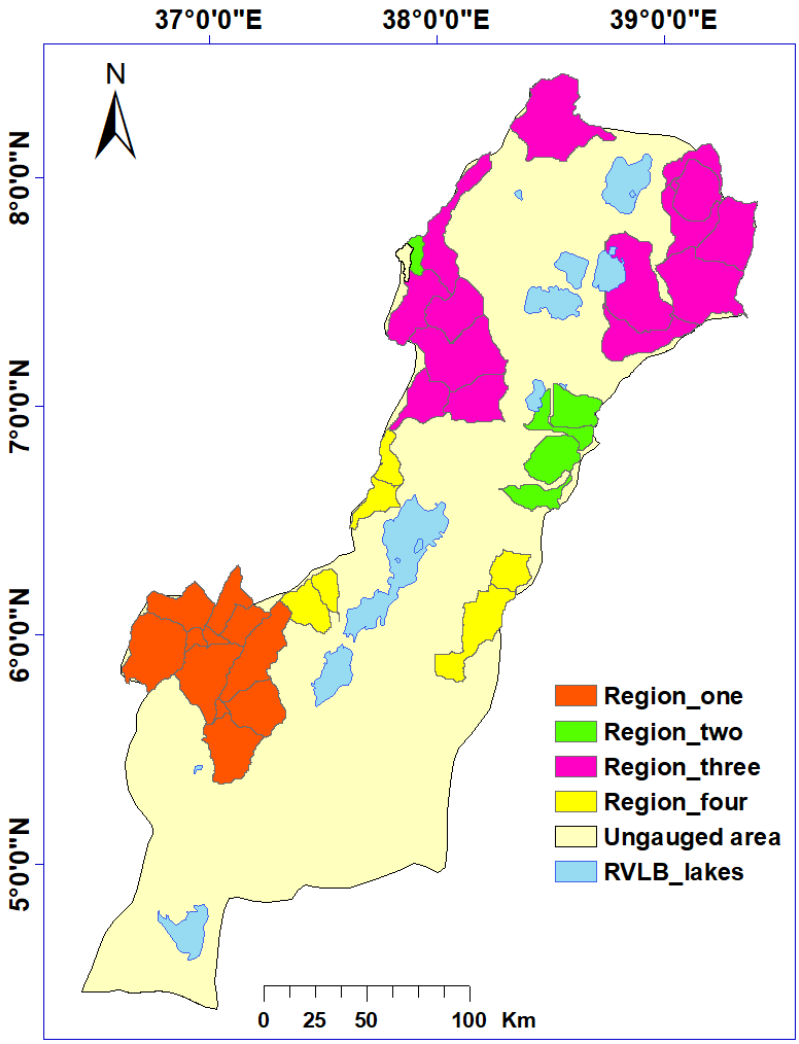

Figure 4-5: Delineated homogeneous regions in RVLB

Accordingly the study area is grouped into four gauged clusters and satisfied regional homogeneity criteria; the first consisting only Weito catchment; the second cluster contains; Kola, Gidabo Apposto, Tikurwuha at Hawassa bridge, Tikurwuha at Dato village, and Gombora catchments. In the third region; Bilate Tena, Bilate Alabakulito, Gedemso, Horakelo, Ketar Fete, Ketar Abura and Meki catchments. Finally region four contains; Upper Gelana, Hamassa Wajifo, Hamassa Humbo,

Badessa, Hare and Kulfo homogeneous catchments. Regional flood frequency is computed for each region.

4.3 Selection of Distribution Models, Parameters and Standard Error Estimates

4.3.1 Selection of Regional Distribution

In this study the selection of best fit statistical distribution is computed using regional goodness of fit test from the most commonly used distribution models. Using Easy Fit software all goodness of fit tests such as; Chi-square, Kolmogorov Smirnov and Anderson Darling tests are computed and best fit regional distribution is displayed as shown in the following tables.

Table 4-7: Goodness of fit summary for region one

		Kolmogorov Smrinov		Anderson Darling		Chi-squared		Weighted
Sn	Distribution	Statistic	Rank	Statistic	Rank	Statistic	Rank	Rank
1	Gen. Extreme Value (GEV)	0.1177	2	0.1761	1	0.6425	3	2
2	Gen. Logistic (GL)	0.1255	4	0.1931	4	0.5703	2	3.3
3	Gen. Pareto (GP)	0.1239	3	3.9414	5	Not available	-	4
4	Lognormal (LN 2P)	0.117	1	0.1761	2	0.7422	4	2.3
5	Uniform	0.2017	6	4.4844	6	Not available	-	6
6	Wakeby	0.1258	5	0.1917	3	0.5694	1	3

For region one; according to Kolmogorov Smirnov test, null hypothesis is accepted at 10% significance level that the data is fit to Generalized extreme value distribution because Statistic (S) $0.11767 < S_{critical} = 0.27851$. Using Anderson Darling test statistic (S) is $0.17608 < S_{critical} = 1.9286$ so null hypothesis is accepted to fit GEV at 10% significance level. Finally using Chi-squared test Statistic (S) $0.6425 < S_{critical} = 2.7055$ again accepted at the same significance level, therefore generalized extreme value (GEV) is the best fit distribution for this region with respect to this and weighted rank.

Table 4-8: Goodness of fit summary for region two

Sn	Distributions	Kolmogorov Smrinov		Anderson Darling		Chi-squared		Weighted Rank
		Statistic	Rank	Statistic	Rank	Statistic	Rank	
1	Gen. Extreme Value (GEV)	0.11922	4	2.2307	2	21.119	3	3
2	Gen. Logistic (GL)	0.12831	5	2.5242	3	19.811	2	3.3
3	Gen. Pareto (GP)	0.1173	2	5.4833	4	Not available	-	3
4	Lognormal (LN 2P)	0.11526	1	1.8931	1	19.321	1	1
5	Uniform	0.26275	6	28.066	6	Not available	-	6
6	Wakeby	0.1173	3	5.4833	5	Not available	-	4

For region two; Using Kolmogorov Smirnov test statistic (S) $0.11526 < S_{critical} = 0.12923$, and using Anderson Darling test statistic (S) $1.8931 < S_{critical} = 1.9286$, at 10% significance level therefore null hypothesis is accepted that the data is fit to Lognormal distribution. But using Chi-squared (S)$19.321 > S_{critical} = 10.645$ so null hypothesis is rejected. Even though null hypothesis is rejected using Chi-square method considering the ranks and other two methods Lognormal is the selected distribution for this region.

Table 4-9: Goodness of fit summary for region thee

Sn	Distributions	Kolmogorov Smrinov		Anderson Darling		Chi-squared		Weighted Rank
		Statistic	Rank	Statistic	Rank	Statistic	Rank	
1	Gen. Extreme Value (GEV)	0.0667	3	0.5774	2	4.2271	2	2.3
2	Gen. Logistic (GL)	0.0704	4	0.8149	3	4.5789	3	3.3
3	Gen. Pareto (GP)	0.057	2	8.2564	5	Not available	-	3.5
4	Lognormal (LN 2P)	0.1538	6	4.5295	4	32.608	4	5
5	Uniform	0.1103	5	30.381	6	Not available	-	5.5
6	Wakeby	0.047	1	0.4503	1	2.3433	1	1

For region three; Using Kolmogorov Smirnov statistic (S) $0.04704 < S_{critical} = 0.12051$, using Anderson Darling statistic (S) $0.45027 < S_{critical} = 1.9286$, and using Chi-squared test statistic (S) $2.3422 < S_{critical} = 10.645$ at 10% significance level using all methods thus, null hypothesis is accepted to fit the data to Wakeby distribution.

Table 4-10: Goodness of fit summary for region four

Sn	Distribution	Kolmogorov Smrinov		Anderson Darling		Chi-squared		Weighted Rank
		Statistic	Rank	Statistic	Rank	Statistic	Rank	
1	Gen. Extreme Value (GEV)	0.0881	3	1.1201	3	6.1176	3	3
2	Gen. Logistic (GL)	0.104	5	1.5475	5	10.715	5	5
3	Gen. Pareto (GP)	0.0531	1	0.3725	1	3.9862	1	1
4	Lognormal (LN 2P)	0.1024	4	1.1224	4	9.1797	4	4
5	Uniform	0.1583	6	24.937	6	Not available	-	6
6	Wakeby	0.0531	2	0.3725	2	3.9862	2	2

For region four; Using Kolmogorov Smirnov test statistic (S) $0.05309 < S_{critical} = 0.11935$, using Anderson Darling test statistic (S) $0.37249 < S_{critical} = 1.9286$, and finally using Chi-squared test statistic (S) $3.9862 < S_{critical} = 10.645$ so null hypothesis is accepted at 10% significance level that the data is fit to Generalized pareto distribution.

4.3.2 Parameter Estimation

The methods used in this study to estimate parameters for selected regional distribution are; method of moment (MOM), probability weighted moment (PWM) and maximum likelihood (ML) as those methods are widely used by many researchers. Table 4.11, shows the summary of parameter estimates by those three methods expressed from standardized flood data for each homogeneous regions.

Table 4-11: Summary of parameters for each region

Region	Selected distribution	Parameter using MOM		Parameter using PWM		Parameter using ML	
		Parameter	Value	Parameter	Value	Parameter	Value
I	GEV	κ	0.0129	κ	-0.1389	κ	0.1382
		α	0.4431	α	0.3814	α	0.3817
		μ	0.7498	μ	0.7196	μ	0.7195
II	LN (2P)	σ_γ	0.2705	σ_γ	0.4114	σ_γ	0.9072
		μ_γ	-0.0366	μ_γ	-0.0846	μ_γ	-0.4284
III	Wakeby	Not available		Not available		α	1.7957
						β	0.8237
						γ	0.0209
						δ	0.7114
						ζ	-0.0596
IV	GP	ε	0.2054	ε	1.0565	ε	0.0766
		κ	1.0517	κ	1.6585	κ	-0.2307
		α	1.6304	α	0.2051	α	1.1364

Even though parameters have been calculated with all three methods for GEV, LN (2P) and GP distributions, only ML methods were computed for region four using Wakeby distribution. For Wakeby distribution MOM is not feasible and PWM has certain assumptions like all parameters should be positive but, in this study when PWM is used it violates the basic assumptions so, using this method didn't succeed and ML method becomes the only choice here.

4.3.3 Standard Error Estimation

In order to obtain a reliable quantile estimates assessment of possible errors has to be done among the available parameter estimation methods. Table 4.12, shows the standard error estimates of parameters corresponding to a required return period. SEE is usually given in a unit of flow series fitted to a distribution model and in this case unit of standardized flow.

Table 4-12: Standard error estimates

T	SEE region one			SEE region two		SEE region four		
	MOM	**PWM**	**ML**	**MOM**	**ML**	**MOM**	**PWM**	**ML**
2	0.625	1.692	0.788	1.497	0.903	0.664	0.612	0.083
5	0.866	3.157	1.498	2.059	1.248	1.117	0.767	0.201
10	1.023	4.154	1.960	2.613	1.577	1.462	0.993	0.297
25	1.220	5.424	2.522	3.456	2.177	1.673	1.114	0.430
50	1.364	6.363	2.919	4.129	2.771	1.732	1.143	0.535
100	1.506	7.281	3.297	4.809	3.491	1.756	1.153	0.641
500	1.830	9.287	4.105	6.366	5.674	1.770	1.157	0.880
1000	1.967	10.067	4.425	7.020	6.848	1.771	1.158	0.974

For generalized extreme value of region one according to SEE estimates among all proposed parameter MOM is the one with the lowest standard error so, GEV with MOM parameter estimation method is selected to drive frequency curve for this region. For region two of LN (2P) SEE is calculated analytically using MOM and ML method only because there is no analytical method to calculate for PWM. Using numerical simulation that is approximate and comparing with the analytical method is not reasonable so, ML is selected since it has lowest SEE compared to MOM and this region LN (2P) with ML parameter estimation is used to drive frequency curve. Region three with Wakeby parent distribution has no mathematical expression for calculating SEE estimates, eventhougth there is numerical approximation to calculate SEE for this method, because of the main objective of calculating SEE here is to compare among the proposed methods, SEE has not approximated for Wakeby distribution. Therefore ML were chosen by default for growth curve derivation in region three. For region four of GP distribution ML has lowest SEE compared to MOM and PWM at the corresponding return periods so, GP with ML is chosen to drive frequency curve for region four. Finally the selected distribution models with the robust parameter estimates along with regional L-moment values are displayed in (table 4.13).

Table 4-13: Selected distribution and Parameter estimation method

Region	Average regional L-moments			Selected distribution	Selected PEM
	τ	τ_3	τ_4		
I	0.30583	0.26188	0.20148	GEV	MOM
II	0.22886	0.22170	0.15655	LN (2P)	ML
III	0.27431	0.18354	0.10688	Wakeby	ML
IV	0.25755	0.14002	0.08704	GP	ML

4.4 Derivation of Regional Frequency Curve

From the results described previously in (table 4.14) using the chosen distribution and parameter estimation method standardized annual quantile have been computed for each region. The standardized regional quantile estimates $Q_t/\bar{Q}$ of each region at the proposed return period is described in the following table.

Table 4-14: Standardized quantile values of homogeneous regions

T	Region I	Region II	Region III	Region IV
2	1.278*$\bar{Q}$	0.586*$\bar{Q}$	0.908*$\bar{Q}$	0.931*$\bar{Q}$
5	1.769*$\bar{Q}$	0.803*$\bar{Q}$	1.605*$\bar{Q}$	2.291*$\bar{Q}$
10	2.090*$\bar{Q}$	0.980*$\bar{Q}$	1.917*$\bar{Q}$	3.530*$\bar{Q}$
25	2.491*$\bar{Q}$	1.262*$\bar{Q}$	2.233*$\bar{Q}$	5.502*$\bar{Q}$
50	2.785*$\bar{Q}$	1.511*$\bar{Q}$	2.491*$\bar{Q}$	7.297*$\bar{Q}$
100	3.075*$\bar{Q}$	1.788*$\bar{Q}$	2.844*$\bar{Q}$	9.403*$\bar{Q}$
500	3.735*$\bar{Q}$	2.538*$\bar{Q}$	4.640*$\bar{Q}$	15.810*$\bar{Q}$
1000	4.014*$\bar{Q}$	2.907*$\bar{Q}$	6.307*$\bar{Q}$	19.393*$\bar{Q}$

Table 4.14, can be used to calculate quantile of any site in a specific region by simply multiplying standardized regional quantile with index flood or mean flood at a particular site in the region at the required return period that can be used for any water resource projects. For each region regional flood frequency curve of standardized quantiles versus return period is developed to estimate standardized flows alternatively as described in the following figures.

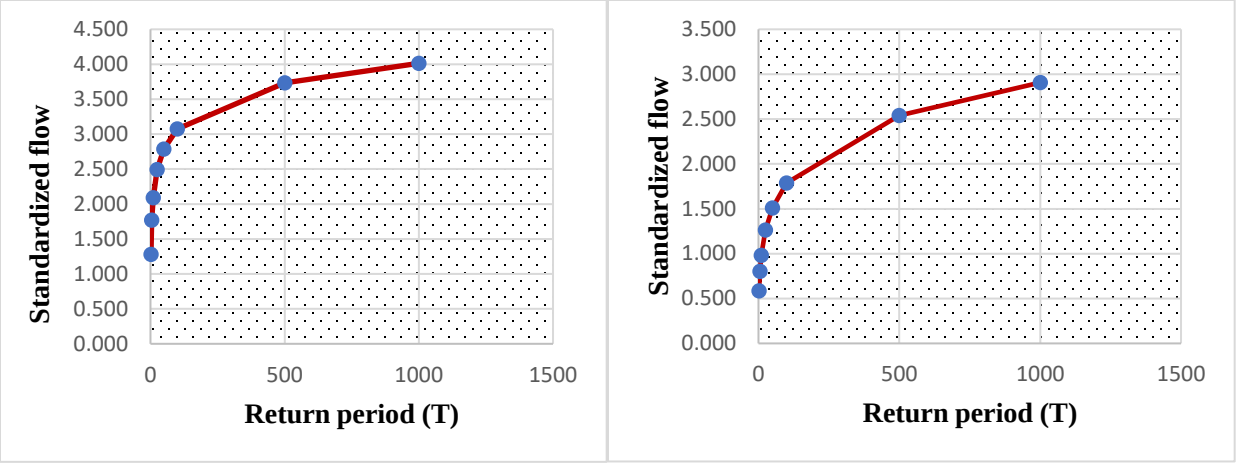

Figure 4-6: (a) Regional growth curve for region one, (b) Regional growth curve for region two

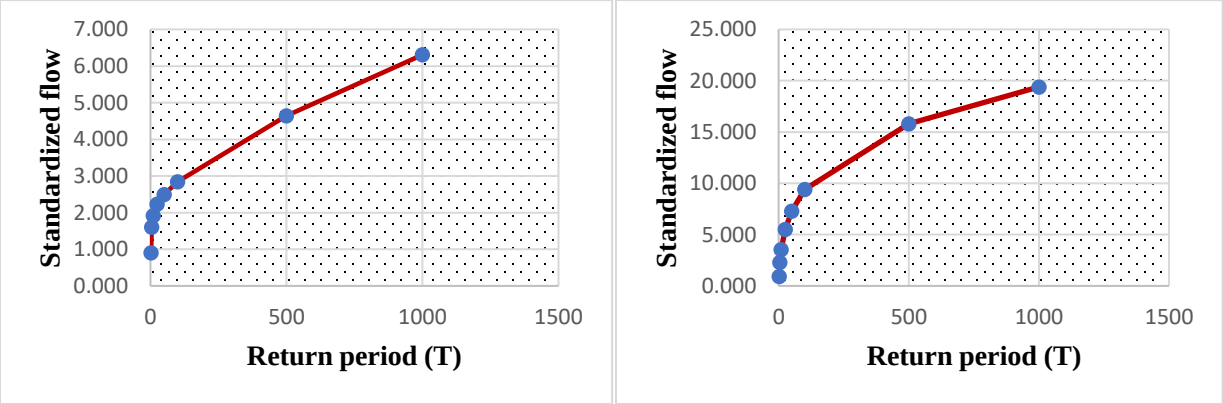

Figure 4-7: (c) Regional growth curve for region three, (d) Regional growth curve for region four

At any site within the region once mean annual flow is known the flood quantile for different return periods can be computed from the above curves.

4.5 Regional Regression Equations

As one of the aim this study is to predict mean annual flow for ungauged areas, Using regression models, the mean annual flow can be determined from easily available data such as, catchment features and meteorological variables. The R^2 values of regression model between mean annual flood and physiographic or meteorological characteristics of the catchments have been carried out in specific region. Before conducting regression model a preliminary correlation analysis were done between index flood and catchment characteristics of specific regions in the study area.

As described in (table 4.15) for region two; the highest R^2 value were obtained when length of the main river (L_m) and mean annual rainfall (MAR) are combined. For region three; gravelius index (K_g), stream density (S_d) and mean annual rainfall (MAR) are highly correlated to mean annual flood; finally for region four area (A) and length of the main river (L_m) were used to implement regression analysis. Region one consists only one catchment, since performing regression analysis for single catchment is not possible, in this study regression analysis for region one has not computed.

Table 4-15: Derived regression equation for regions in RVLB

Region	**Regression equation**	R^2
II	$\bar{Q} = 0.41 * L_m + 0.036 * MAR - 42.989$	0.82
III	$\bar{Q} = -1888.335 * S_d - 0.212 * MAR + 114.849 * K_g + 248.848$	0.83
IV	$\bar{Q} = 0.142 * A - 1.049 * L_m + 13.757$	0.79

As shown in (table 4.15) the following catchment physiographic and meteorological parameters are significant in predicting mean annual flow of certain catchment; length of the main river (L_m), mean annual rainfall (MAR), stream density (S_d), gravelius index (K_g) and area of catchment (A) according to their correlation in the respective region. To determine index flood of any ungauged area in the region, inserting those parameters and mean annual flood of that area can easily be computed from regression equation developed. Once index flood is calculated maximum annual flood corresponding to a required return period can be found by multiplying index flood value by regional standardized flood quantile of respective return periods within a region. So, using this regional flood frequency analysis relatively better quantile estimates can be computed even though there is no recorded flood data.

4.6 Validation of Regional Models

To validate the performance of regional models certain gauged catchments were considered ungauged and did not used in development of regression equation. The pseudo ungauged catchments selected are; Tikurwuha Dato from region two, Bilate Tena and Ketar Abura from region three; Hamassa Wajifo and Hare catchments.

As presented in (table 4.16) the observed mean annual flood and estimated mean annual flood from regional models are compared.

Table 4-16: Comparison of observed and estimated mean annual maximum flow from regional model

Regions	Stations	Observed mean flow	From regional model	Relative error
II	Tikurwuha Dato	4.725	5.319	0.104
III	Bilate Tena	141.081	133.658	0.052
	Keter Abura	98.664	109.259	0.106
IV	Hamassa Wajifo	30.838	32.304	0.046
	Hare	7.020	6.545	0.059

From (table 4.16) for region two the estimated flow is greater than the observed one but the relative error is 10.4% which is acceptable. In region three the model performs good estimates having relative error of 5% and 10.6% respectively from two representative stations. When it comes to region four the model performs very good estimates having 4.6% and 5.9% of errors from the two stations represented. This shows the model performs well in estimating index flood in ungauged catchments in the study area.

5 COCLUSSIONS AND RECOMMENDATIONS

5.1 Conclusions

In this study, regionalization of catchments within the Rift Valley Lake Basin and subsequently regional flood frequency analysis of both gauged and ungauged regions have been carried out. In the delineation of homogeneous regions cluster analysis have been applied, using Hierarchical method in selecting optimum number of the clusters and K-mean method to classify catchments after several iterations using selected attributes. Accordingly four regions have been identified and delineated. Region one consisting two catchments, region two consists seven, region three ten and region four ten catchments respectively including ungauged catchments. All four regions have satisfied homogeneity tests (discordance measure, Cv and LCv based homogeneity tests) except; Ferfuro from region two, Chiufa and Rinzaf from region three and Gidabo near Measso from region four catchments was discordant.

Choices of the best fit distribution for each homogeneous regions of RVLB have been conducted using goodness of fit analysis. For regions one to four, the distributions that are found to be suitable are; Generalized extreme value, Log-normal (two parameters), Wakeby and Generalized pareto distributions respectively. MOM, PWM, and ML methods have been compared using standard error tests, accordingly for region one MOM found to be suitable, ML is found the efficient and robust method for estimation of parameters for region two to four. For each regions regional frequency curves have been developed with standardized annual maximum flow series, that are useful to compute the quantile of ungauged areas within the basin. The estimates of mean annual maximum flow for ungauged catchments can be found from the regression equation developed for each regions except region one which contains only one gauged stations. The validity of regression equation have also checked using five pseudo ungauged catchments. Therefore the result of the present study provides useful information to support water resource developments of RVLB, where a lot of ungauged sites exists and flood is an issue.

5.2 Recommendations

The main recommendation points based on the methods and results of this study are presented as follows.

- ✓ In order to get a reliable estimates of regional quantile and best performing regional regression model at ungauged catchments more recent hydrological time series data should be included.
- ✓ The mean annual rainfall over the catchment was calculated using Thiessen polygon but this method has a limitation especially where altitudinal variation is very high, so the coming studies should consider other methods.
- ✓ In this study K-mean clustering was used for delineating homogeneous regions using selected catchment characteristics; the future study should have to consider other regionalization approaches including soft clustering.
- ✓ More parameters which are believed to affect mean annual maximum flows in ungauged catchments have to be assessed and incorporated to modify regression equation.
- ✓ Validation of regional models were carried out by using five gauged catchments considering as pseudo ungauged; therefore the coming researches should add more validation catchments to enhance the performance of regression model.

REFERENCES

Acreman, M. C., & Sinslair, C. D. (1985). Classification of Drainage Basins according to their physical Characteristics; an application for flood frequency, analysis in Scotland. *Hydrologic Engineering*, *84*(3–4), 365–380.

Alemu, O. A., Bayou, C., & Assefa, M. M. (2018). Soil Erosion Modelling and Risk Assessment in Data Scarce Rift Valley Lake Regions, Ethiopia. *Water Science Journal*, *10*(11), 1–17. https://doi.org/10.3390/w10111684

Amiri, B. J., Baheri, B., Fohrer, N., & Adamowski, J. (2018). Regionalization of Flood Magnitudes using Watersheds' Ecological Attributes. *Geocarto International*, *35*(9), 917–933. https://doi.org/10.1080/10106049.2018.1552321

Bastin, G., Lorent, B., & Gevers, M. (1984). Optimal Estimation of the Average Areal Rainfall and Optimal Selection of Rain Gauge Locations. *Water Resources Research*, *20*(4), 463–470. https://doi.org//10.1029/WR020i004p00463

Basu, B., & Srinivas, V. V. (2014). Regional flood frequency analysis using kernel-based fuzzy clustering approach. *Water Resources Research*, *50*(4), 3295–3316. https://doi.org/10.1002/2012WR012828.Received

Basu, B., & Srinivas, V. V. (2016). Regional Flood Frequency Analysis Using Entropy-Based Clustering Approach. *Journal of Hydrologic Engineering*, *21*(8), 1–12. https://doi.org/10.1061/(ASCE)HE.1943-5584.0001351.

Bob, B., & Rasmussen, P. (1995). *Recent advances in flood frequency analysis L-moments Annual flood series*. U.S. National Report to International Union of Geodesy and Geophysics, University of Quebec, Quebec.

Bocchiola, D., Michele, C., & Rosso, R. (2003). Review of recent advances in index flood estimation. *Hydrology and Earth System Sciences*, *7*(3), 283–296.

Burn, D. H. (1990). Evaluation of Regional Flood Frequency Analysis With a Region of Influence Approach. *Water Resources Research*, *26*(10), 2257–2265.

Burn, D. H. (1997). Catchment similarity for regional flood frequency analysis using seasonality

measures. *Journal of Hydrology*, *202*(97), 212–230. https://doi.org//10.1016/S0022-1694(97)00068-1

Burn, D. H. (2000). The formation of groups for regional flood frequency analysis. *Hydrological Science Journal*, *45*(1), 97–112. https://doi.org/10.1080/02626660009492308

Burn, D. H., Zrinji, Z., & Kowalchuk, M. (1997). Regionalization of catchments for regional flood frequency analysis. *Journal of Hydrologic Engineering*, *2*(2), 76–82. https://doi.org//10.1061/(ASCE)1084-0699(1997)2:2(76)

Cavadias, G. S. (1990). *The canonical correlation approach to regional flood estimation*. Paper Presented to the Annual Conference of the Canadian Society of Civil Engineering, McGill University, Montreal, Canada.

Cunderlik, J. M., & Burn, D. H. (2006). Switching the pooling similarity distances : Mahalanobis for Euclidean. *Water Resources Research*, *42*, 1–10. https://doi.org/10.1029/2005WR004245

Cunnane, C. (1973). A particular comparison of annual maxima and partial duration series methods of flood frequency prediction. *Journal of Hydrology*, *18*(3–4), 257–271. https://doi.org//10.1016/0022-1694(73)90051-6

Cunnane, C. (1988). Methods and Merits of Regional flood frequency Analysis. *Journal of Hydrology*, *100*(1–3), 269–290.

Cunnane, C. (1989). *Statistical distributions for flood frequency analysis.* WMO Rep., World Meteorological Organization, Geneva, Switzerland.

David, R, M. (1993). *Handbook of Hydrology*. University of Texas, Austin,Texas: McGRAW-HILL,INC.

Demissie, M. (2008). *Regional Flood Frequency Analysis for Upper Awash Sub- Basin (Upstream of Koka)*. MSc. thesis, Civil and Enviromental Engineering, Addis Ababa University.

Frank, M. (2012). The Kolmogorov-Smirnov Test for Goodness of Fit. *Journal of American Statistical Association*, *46*(253), 68–79.

Gabriele, S., & Nigel, A. (1991). A Hierarchical Approach to Regional Flood Frequency Analysis. *Water Resources Research*, *27*(6), 1281–1289.

Gebeyehu A. (1989). *Regional flood frequency analysis*. PHD. thesis, Royal Institute of Technology, Stockholm University, Sweden.

Gebrehiwot, B., Gessesse, B., & Melgani, F. (2019). Characterizing the spatiotemporal distribution of meteorological drought as a response to climate variability : The case of rift valley lakes basin of Ethiopia. *Weather and Climate Extremes*, *26*(April), 100237. https://doi.org/10.1016/j.wace.2019.100237

Greenwood, J. A. (1979). Probability Weighted Moments: Definition and Relation to Parameters of Several Distributions Expressible in Inverse Form. *Water Resources Research*, *15*(5), 1049–1054. https://doi.org//10.1029/WR015i005p01049

Grover, P. L., Burn, D. H., & Cunderlik, J. M. (2002). A comparison of index flood estimation procedures for ungauged catchments. *Canadian Journal of Civil Engineering*, *29*(5), 734–741. https://doi.org/10.1139/l02-065

Haddad, O. B., & Moravej, M. (1943). Discussion of " Trend , Independence , Stationarity , and Homogeneity Tests on Maximum Rainfall Series of Standard Durations Recorded in Turkey ." *Journal of Hydrologic Engineering*, *20*(10), 1–3. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000973

Hailegeorgis, T. T., & Alfredsen, K. (2017). Regional flood frequency analysis and prediction in ungauged basins including estimation of major uncertainties for. *Journal of Hydrology*, *9*(5), 104–126. https://doi.org/10.1016/j.ejrh.2016.11.004

Hallin, M., & Puri, M. L. (1995). A multivariate Wald-Wolfowitz rank test against serial dependence. *The Canadian Journal of Statistics*, *23*(1), 55–65. https://doi.org//10.2307/3315547

Hosking, J. R. M., & Wallis, J. R. (1986). The Value of Historical Data in Flood Frequency Analysis. *Water Resources Research*, *22*(11), 1606–1612. https://doi.org//10.1029/WR022i011p01606

Hosking, J. R. M., & Wallis, J. R. (1993). Some Statistics Useful in Regional Frequency

Analysis. *Water Resources Research*, *29*(2), 271–281. https://doi.org//10.1029/92WR01980

Hosking, J. R. M., & Wallis, J. R. (1997). *Regional frequency analysis — An approach based on L-moment*. Cambridge University Press: New York.

Hussen, B., & Wagesho, N. (2016). Regional Flood Frequency Analysis for Abaya – Chamo Sub-Basin, Rift Valley Basin Ethiopia. *Journal of Resources Development and Management*, *24*, 15–28.

Isik, S., Singh, V. P., & Asce, F. (2008). Hydrologic Regionalization of Watersheds in Turkey. *Journal of Hydrologic Engineering*, *13*(9), 824–834. https://doi.org//10.1061/(ASCE)1084-0699(2008)13:9(824)

Kent, K., & Lioyd, F. (1975). A Monte Carlo Comparison of Six Clustering Procedures. *International Biometric Society*, *31*(3), 777–783. https://doi.org//10.2307/2529565

Kim, J. H. (1993). Chi-Square Goodness of Fit Tests for Randomly Censored Data. *The Annals of Statistics*, *21*(3), 1621–1639.

Kitanidis, P. K., & Lane, R. W. (1985). Maximum likelihood parameter estimation of hydrologic Spatial Processes by the Gauss-Newton Method. *Journal of Hydrology*, *79*(1–2), 53–71. https://doi.org/10.1016/0022-1694(85)90181-7

Kohler, M. A. (1949). On the Use of Double-Mass Analysis for Testing the Consistency of Meteorological Records and for Making Required Adjestments. *American Meteorological Society*, *30*(5), 471–476. https://doi.org//10.1175/1520-0477-30.5.188

Lettenmaier, D, P. (1985). Testing Flood Frequency Estimation Method Using a Regional Flood Generation Model. *Water Resources Research*, *21*(12), 1903–1914. https://doi.org/10.1029/WR021i012p01903

Lu, G., & David, W. (2008). An adaptive inverse-distance weighting spatial interpolation technique. *Journal of Computer and Geoscience*, *34*(9), 1044–1055. https://doi.org/10.1016/j.cageo.2007.07.010

Murtagh, F. (2014). Ward ’ s Hierarchical Agglomerative Clustering Method : Which Algorithms Implement Ward ’ s Criterion ? *Journal of Classification*, *31*(8), 274–295.

https://doi.org//10.1007/s00357-014-9161-z

Myung, I. J. (2003). Tutorial on Maximum Likelihood Estimation. *Journal of Mathematical Psychology*, *47*(1), 90–100. https://doi.org//10.1016/S0022-2496(02)00028-7

Nageshwar, B., & Carol, O. (1990). Comparision of Method of Residuals and Cluster Analysis for Flood Regionalizations. *Journal of Water Resources Planning and Management*, *115*(6), 793–808. https://doi.org//10.1061/(ASCE)0733-9496(1989)115:6(793)

Nathan, R, J., & Mcmahon, T, A. (1990). Identification of homogeneous regions for the purpuse of regionalization. *Journal of Hydrology*, *121*(4), 217–238.

Nejc, B., Mitja, B., & Mojca, S. (2013). Comparison between the peaks over threshold method and the annual maximum method for flood frequency analysis. *Hydrological Science Journal*, *59*(5), 37–48. https://doi.org/10.1080/02626667.2013.831174

Ralamboundrainy, H. (1995). A conceptual version of the K-means algorithm. *Pattern Recognition Letters*, *16*(11), 1147–1157. https://doi.org//10.1016/0167-8655(95)00075-R

Rao, A, R., & Hamed, K, H. (2000). *Flood Frequency Analysis*. CRC Press: Boca Raton, Florida USA.

Rao, A. R., & Hamed, K. H. (1997). Regional flood frequency analysis of wabash river flood data by L-moments. *Journal of Hydrologic Engineering*, *2*(4), 169–179.

Rao, A. R., & Srinivas, V. V. (2007). *Regionalization of Watersheds an Approach Based on Cluster Analysis* (V.P. Singh, Ed.). Bangalore, India: Springer Science+Business Media B.V.

Shumet, A. G., & Kassa, T. M. (2016). Assessing the Impact of Existing and Future Water Demand on Economic and Environmental Aspects (Case Study from Rift Valley Lake Basin : Meki-Ziway Sub Basin), Ethiopia. *International Journal of Waste Resources*, *6*(2), 2–14. https://doi.org/10.4172/2252-5211.1000223

Sinclair, D., & Spurr, B. (2012). Approximations to the Distribution Function of the Anderson-Darling Test Statistic. *Journal of American Statistical Association*, *83*(404), 1990–1991. https://doi.org//10.1080/01621459.1988.10478720

Sine, A., & Moges, S. (2013). Basin Regionalization for the Purpose of Water Resource Development in a Limited Data Situation: Case of Blue Nile River Basin, Ethiopia. *Journal of Hydrologic Engineering*, *18*(10), 2–10. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000730

Singh, Vijay P. (1998). *Entropy - Based Parameter Estimation In Hydrology*. Louisiana State University, Baton Rouge U.S.A.

Stedinger, J, & Lu, L. H. (1995). Appraisal of regional and index flood quantile estimators. *Stochastic Hydrology and Hydraulics*, *9*(1), 49–75. https://doi.org/10.1007/BF01581758

Stedinger, Jery, & Tasker, G. (1985). Regional Hydrologic AnalysisOrdinary , Weighted , and Generalized Least Squares Compared. *Water Resources Research*, *21*(9), 1421–1432. https://doi.org/10.1029/WR021i009p01421

Strejc, V. (1980). Least squares parameter estimation. *Czechoslovak Academy of Science*, *16*(5), 535–550. https://doi.org/10.1016/0005-1098(80)90077-1

Tasker, D, D. (1983). Comparing methods of hydrologic regionalization. *American Water Resources Association*, *18*(6), 965–970.

Thomas, D. M., & Benson, M. A. (1975). *Generalization of Streamflow Characteristics From Drainage-Basin Characteristics*. US Government Printing Office, Washington, DC.

Ulsido, M. D., & Alemu, E. (2014). Irrigation Water Management in Small Scale Irrigation Schemes : the Case of the Ethiopian Rift Valley Lake Basin. *Environmental Research, Engineering and Management*, *67*(1), 5–15. https://doi.org//10.5755/j01.erem.67.1.6240

Unal, Y., Kindap, T., & Karaca, M. (2003). Redefining the Climate Zones of Turkey using Cluster Analysis. *International Journal of Climatology*, *23*(9), 1045–1055. https://doi.org//10.1002/joc.910

Wiltshire, S. E. (1986a). Regional flood frequency analysis i: Homogeneity statistics. *Hydrological Sciences Journal*, *31*(3), 321–333. https://doi.org//10.1080/02626668609491051

Wiltshire, S. E. (1986b). Regional flood frequency analysis II : Multivariate classification of

drainage basins in Britain. *Hydrological Science Journal*, *31*(3), 335–346. https://doi.org/10.1080/02626668609491052

Wolfowitz, J. (1942). Additive Partition Functions and a Class of Statistical Hypotheses. *Mathematical Statistics*, *13*(3), 247–279. https://doi.org//stable/2235939

Wooldridge, J. M. (2001). Applications of Generalized Method of Moments Estimation. *Journal of Economic Perispectives*, *15*(4), 87–100. https://doi.org/10.1257/jep.15.4.87

Yapo, P. O., Gupta, H. V., & Sorooshian, S. (1996). Automatic calibration of conceptual rainfall-runoff models: Sensitivity to calibration data. *Journal of Hydrology*, *181*(1–4), 23–48. https://doi.org/10.1016/0022-1694(95)02918-4

Yirefu, S. M. (2010). *Regional Flood Frequency Analysis, Upstream of Awash with the confluence of Kesem river*. MSc. thesis, Civil and Enviromental Engineering, Addis Ababa University.

Yirga, H., Urge, M., Goetsch, A. L., & Tolera, A. (2019). Quality of Water from Rift Valley Lakes of Ethiopia for Livestock Drinking. *East African Journal of Veterinary and Animal Sciences*, *3*(1), 9–16.

Zolt, Z., & Burn, D, H. (1996). Regional Flood Frequency with Hierarchical Region of Influence. *Journal of Water Resources Planning and Management*, *122*(4), 245–252.

ANNEXES

ANNEX-A: Catchment characteristics (Physiographic and drainage)

St.name	Area (km^2)	Lm (km)	Elmin (m)	Elmax (m)	Elmean (m)	KG	Sd (/km)
Weito	4530.12	40.265	585	3421	1448.808	2.102	0.042
Upper_GelanaYirgachefe	678.24	69.478	1166	2879	1923.234	2.097	0.102
Hammassa_Wajjifo	460.24	44.620	1202	2837	1635.439	2.272	0.097
Hammassa_Humbo	206.44	34.399	1336	2837	1740.311	1.916	0.167
Baddessa	289.76	25.438	1494	3000	2132.246	1.498	0.088
Gidabo_Measso	311.6	33.978	1295	2795	2023.702	1.630	0.109
Kolla	260	33.433	1571	3027	2073.960	2.246	0.129
Gidabo_Apposto	432.92	105.600	1711	3065	2171.178	1.432	0.244
Bilate_Tena	3717.4	75.738	1451	3292	2015.064	2.465	0.085
Tikurwuha_Hawasabridge	273.92	52.266	1708	2828	2110.572	2.400	0.191
Tikurwuha_Dato	369	32.578	1559	2629	2072.140	1.492	0.088
Bilate_Alabakulitto	1831.28	71.590	1699	3292	2210.690	2.463	0.116
Ferfuro	292.2	36.158	1872	2811	2195.761	1.822	0.124
Gedemso	512.36	57.781	1896	4088	2336.611	2.044	0.113
Horakello	1646.68	94.707	1491	4088	2291.049	1.627	0.120
Ketar_Fete	1883.88	52.603	2275	4162	2847.019	1.724	0.074
Chiufa	488.72	32.829	1645	3920	2413.669	1.399	0.092
Ketar_Abura	2800	68.900	1641	4162	2649.572	1.802	0.094
Meki	1099.32	83.000	1636	3638	2208.394	1.892	0.076
Rinzaf	293.28	29.219	1942	3442	2483.711	1.450	0.100
Gombora	108.12	26.935	2187	2794	2501.686	1.856	0.249
Harie	181.16	31.398	1181	3442	2545.260	1.702	0.173
Kulfo	353.04	21.138	1265	3464	2305.322	1.621	0.099
Sagan	5190.64	51.742	867	3121	1393.016	2.244	0.062
Guder	73.72	22.589	2222	2844	2515.407	1.932	0.306
Shafe	164.68	30.439	1170	3017	2075.631	1.961	0.185
Baso	149.96	26.355	1161	3061	2254.703	1.742	0.176
Bisare	617.88	30.742	1241	2645	1647.795	1.707	0.050
Woji	1092.64	31.143	1641	3440	2100.986	1.884	0.029

ANNEX-B: Catchment characteristics (% of Soil texture)

St.name	% F&M	% C	%F	% M&C	% M
Weito	23.67	13.98	7.38	29.98	24.98
Upper_GelanaYirgachefe	78.66	0.00	15.15	6.20	0.00
Hammassa_Wajjifo	30.61	53.75	0.00	15.64	0.00
Hammassa_Humbo	29.27	37.65	0.00	33.08	0.00
Baddessa	100.00	0.00	0.00	0.00	0.00
Gidabo_Measso	100.00	0.00	0.00	0.00	0.00
Kolla	100.00	0.00	0.00	0.00	0.00
Gidabo_Apposto	100.00	0.00	0.00	0.00	0.00
Bilate_Tena	55.88	2.81	3.82	37.49	0.00
Tikurwuha_Hawasabridge	97.48	2.52	0.00	0.00	0.00
Tikurwuha_Dato	89.62	0.66	0.00	9.72	0.00
Bilate_Alabakulitto	62.60	3.93	8.28	25.19	0.00
Ferfuro	69.05	0.00	18.44	12.51	0.00
Gedemso	100.00	0.00	0.00	0.00	0.00
Horakello	80.81	13.47	0.00	5.72	0.00
Ketar_Fete	100.00	0.00	0.00	0.00	0.00
Chiufa	100.00	0.00	0.00	0.00	0.00
Ketar_Abura	98.44	0.00	0.00	1.56	0.00
Meki	16.85	57.51	25.64	0.00	0.00
Rinzaf	45.49	41.49	9.43	3.59	0.00
Gombora	66.62	0.00	0.00	33.38	0.00
Harie	98.70	0.00	0.00	1.29	0.01
Kulfo	53.14	0.03	0.00	0.00	46.83
Sagan	53.83	3.50	1.16	31.62	9.89
Guder	62.05	0.00	0.00	37.95	0.00
Shafe	90.19	0.00	1.37	8.43	0.00
Baso	98.08	0.00	0.00	1.92	0.00
Bisare	48.24	6.80	0.00	44.95	0.00
Woji	56.69	13.64	26.22	3.45	0.00

ANNEX-C: catchment characteristics (Meteorology, Location and Landuse)

St.name	Avg_annual. PPn	Latitude	Longitude	%mod. cult	%int. cult	%Grass land
Weito	1366.795	5.716	37.433	20.25	0.17	0.00
Upper_GelanaYirgachefe	1257.113	6.150	38.183	9.90	67.91	0.00
Hammassa_Wajjifo	1307.569	6.567	37.816	30.69	33.17	0.00
Hammassa_Humbo	1391.222	6.667	37.750	39.84	59.11	0.00
Baddessa	1410.814	6.383	38.300	4.87	95.13	0.00
Gidabo_Measso	1413.032	6.433	38.433	3.12	90.01	0.00
Kolla	1108.947	6.634	38.399	5.84	94.11	0.00
Gidabo_Apposto	1062.918	6.750	38.383	0.00	100.00	0.00
Bilate_Tena	1119.461	6.933	38.133	1.25	90.20	1.01
Tikurwuha_Hawasabridge	978.907	7.083	38.483	0.00	76.71	0.08
Tikurwuha_Dato	970.859	7.100	38.500	0.00	80.63	1.49
Bilate_Alabakulitto	1217.380	7.283	38.066	0.00	92.81	0.00
Ferfuro	1268.594	7.733	38.117	0.00	99.79	0.00
Gedemso	995.913	7.467	38.866	0.00	75.03	0.00
Horakello	1022.113	7.667	38.700	19.30	49.43	0.00
Ketar_Fete	786.481	7.783	39.050	0.00	80.07	3.92
Chiufa	835.991	7.983	39.067	7.18	76.90	0.00
Ketar_Abura	795.121	8.067	39.050	6.43	76.22	2.64
Meki	1137.944	8.185	38.796	52.06	26.02	0.00
Rinzaf	1333.659	8.117	38.367	0.00	67.43	0.00
Gombora	1196.847	7.584	37.916	0.00	100.00	0.00
Harie	793.879	6.067	37.600	87.92	2.21	0.00
Kulfo	1292.031	6.033	37.533	70.61	0.00	0.00
Sagan	893.583	5.244	37.560	13.29	2.55	10.83
Guder	1203.529	7.550	37.866	0.00	100.00	0.00
Shafe	582.258	6.251	37.791	3.88	9.24	0.00
Baso	691.998	6.123	37.640	50.03	11.45	0.00
Bisare	1015.672	6.680	38.040	0.00	73.47	0.00
Woji	1103.942	8.022	38.726	4.81	75.32	0.00

ANNEX-D: Catchment characteristics (% of Landuse)

St.name	%rep. veg	%shrubland	%exp. surf	%forest	%urban area	%afro. alph	%water.b	%woodland	% marshland
Weito	1.64	63.11	0.00	8.21	0.00	0.00	0.00	6.62	0.00
Upper_GelanaYirgachefe	1.88	15.47	0.00	0.30	0.00	0.00	0.00	0.00	4.54
Hamasa_Wajifo	0.00	35.67	0.00	0.47	0.00	0.00	0.00	0.00	0.00
Hamasa_Humbo	0.00	0.00	0.00	1.05	0.00	0.00	0.00	0.00	0.00
Baddessa	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Gidabo_Measso	0.00	5.29	0.00	1.57	0.00	0.00	0.00	0.00	0.00
Kolla	0.00	0.00	0.00	0.05	0.00	0.00	0.00	0.00	0.00
Gidabo_Apposto	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Bilate_Tena	0.00	2.61	2.22	0.00	0.05	0.00	0.17	0.00	2.50
Tikurwuha_Hawasabridge	0.00	0.00	0.00	0.00	9.89	0.00	13.33	0.00	0.00
Tikurwuha_Dato	0.00	0.00	0.00	0.00	2.13	0.00	0.00	0.00	15.74
Bilate_Alabakulto	0.00	0.28	1.92	0.01	0.09	0.00	0.35	0.00	4.54
Ferfuro	0.00	0.21	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Gedemso	0.00	0.00	0.00	19.5	0.00	5.48	0.00	0.00	0.00
Horakello	0.00	0.00	2.17	15.7	0.00	3.04	9.46	0.83	0.00
Ketar_Fete	0.00	0.00	0.00	0.56	0.00	15.44	0.00	0.00	0.00
Chiufa	0.00	0.00	4.19	0.00	0.97	10.24	0.00	0.00	0.52
Ketar_Abura	0.00	0.12	0.86	0.38	0.17	12.20	0.00	0.00	0.97
Meki	0.00	12.41	0.00	9.51	0.00	0.00	0.00	0.00	0.00
Rinzaf	0.00	1.82	0.00	30.7	0.00	0.00	0.00	0.00	0.00
Gombora	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Harie	0.00	7.36	0.00	2.50	0.00	0.00	0.00	0.00	0.00
Kulfo	0.00	29.39	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Sagan	2.32	68.28	0.00	0.58	0.00	0.00	0.00	2.16	0.00
Guder	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Shafe	0.00	86.87	0.00	0.00	0.00	0.00	0.01	0.00	0.00
Baso	0.00	38.30	0.00	0.22	0.00	0.00	0.00	0.00	0.00
Bisare	0.00	26.43	0.00	0.00	0.00	0.00	0.10	0.00	0.00
Woji	0.00	4.33	0.00	8.33	0.00	0.00	1.22	0.00	5.99

ANNEX-E Agglomeration schedule (Hierarchical)

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	21	25	.033	0	0	22
2	5	6	.076	0	0	7
3	16	18	.256	0	0	13
4	9	12	.438	0	0	20
5	22	27	.663	0	0	10
6	3	4	1.009	0	0	11
7	5	7	1.397	2	0	12
8	13	29	1.910	0	0	14
9	14	15	2.677	0	0	18
10	22	26	3.548	5	0	19
11	3	28	4.480	6	0	22
12	5	8	5.434	7	0	15
13	16	17	6.391	3	0	18
14	2	13	7.500	0	8	20
15	5	11	8.677	12	0	21
16	19	20	9.886	0	0	25
17	1	24	11.097	0	0	26
18	14	16	12.499	9	13	27
19	22	23	14.052	10	0	26
20	2	9	15.855	14	4	23
21	5	10	18.038	15	0	24
22	3	21	20.367	11	1	23
23	2	3	23.084	20	22	24
24	2	5	26.428	23	21	25
25	2	19	30.465	24	16	27
26	1	22	36.109	17	19	28
27	2	14	43.381	25	18	28
28	1	2	52.643	26	27	0

ANNEX-F: Analysis of Variance (K-mean)

	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
Area (km^2)	7.311	3	.243	25	30.120	.000
Length of the main channel (L_m)	2.020	3	.878	25	2.302	.102
Minimum elevation	6.855	3	.297	25	23.046	.000
Maximum elevation	6.296	3	.364	25	17.273	.000
Mean elevation	4.783	3	.546	25	8.760	.000
Gravelius index (K_g)	.734	3	1.032	25	.712	.554
Stream density (S_d)	4.193	3	.617	25	6.796	.002
Average annual PPn	.280	3	1.086	25	.258	.855
Latitude	7.653	3	.202	25	37.966	.000
Longitude	5.770	3	.428	25	13.494	.000
%Fine and Medium	1.424	3	.949	25	1.501	.239
%Course	.836	3	1.020	25	.820	.495
%Fine	1.003	3	1.000	25	1.004	.408
%Medium and Course	1.371	3	.955	25	1.435	.256
%Medium	2.076	3	.871	25	2.384	.093
%Moderately cultivated	2.460	3	.825	25	2.983	.050
%Intensively cultivated	4.864	3	.536	25	9.068	.000
%Grassland	3.684	3	.678	25	5.434	.005
%Riparian vegetation	6.283	3	.366	25	17.168	.000
%Shurbland	5.532	3	.456	25	12.128	.000
%Exposed surface	2.905	3	.771	25	3.766	.023
%Forest	2.657	3	.801	25	3.316	.036
%Urban area	1.412	3	.951	25	1.486	.243
%Alpine vegetation	2.880	3	.774	25	3.719	.024
%Water body	.640	3	1.043	25	.614	.613
%Woodland	7.185	3	.258	25	27.870	.000
%Marshland	.539	3	1.055	25	.511	.679

ANNEX-G Iteration History (K-mean)

	Change in Cluster Centers			
Iteration	1	2	3	4
1	3.577	5.458	5.062	5.303
2	0.000	.724	.719	1.008
3	0.000	0.000	.403	.480
4	0.000	.887	.639	0.000
5	0.000	0.000	0.000	0.000