Estimation of Extreme Flood Quantiles Using Classical and Machine Learning Methods for the Sites of North-Western Areas of Pakistan

By:

Muhammad Shafeeq ul Rehman Khan
Reg. No. 02-FBAS/PHDST/F16

Department of Mathematics and Statistics
Faculty of Basic and Applied Sciences
International Islamic University, Islamabad
Pakistan
2022

Estimation of Extreme Flood Quantiles Using Classical and Machine Learning Methods for the Sites of North-Western Areas of Pakistan

By:

Muhammad Shafeeq ul Rehman Khan
Reg. No. 02-FBAS/PHDST/F16

Supervised By:

Dr. Ishfaq Ahmad

Co-Supervised By:

Dr. Zamir Hussain

Department of Mathematics and Statistics
Faculty of Basic and Applied Sciences
International Islamic University, Islamabad
Pakistan
2022

Estimation of Extreme Flood Quantiles Using Classical and Machine Learning Methods for the Sites of North-Western Areas of Pakistan

By:

Muhammad Shafeeq ul Rehman Khan
Reg. No. 02-FBAS/PHDST/F16

A DISSERTATION
SUBMITTED IN THE PARTIAL FULFILLMENT OF THE
REQUIRMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN
STATISTICS

Supervised By:

Dr. Ishfaq Ahmad

Co-Supervised By:

Dr. Zamir Hussain

Department of Mathematics and Statistics
Faculty of Basic and Applied Sciences
International Islamic University, Islamabad
Pakistan
2022

Author's Declaration

I, **Muhammad Shafeeq ul Rehman Khan** Reg. No. **02-FBAS/PHDST/F16** hereby state that my Ph.D. thesis titled: **Estimation of Extreme Flood Quintiles Using Classical and Machine Learning Methods for the Sites of North-Western Areas of Pakistan** is my own work and has not been submitted previously by me for taking any degree from this university, **International Islamic University, Sector H-10, Islamabad, Pakistan** or anywhere else in the country/world.

At any time if my statement is found to be incorrect even after my Graduation the university has the right to withdraw my Ph.D. degree.

Name of Student: ***(Muhammad Shafeeq ul Rehman Khan)***
Reg. No. **02-FBAS/PHDST/F16**
Dated: **30/03/2022**

Plagiarism Undertaking

I solemnly declare that research work presented in the thesis titled: **Estimation of Extreme Flood Quintiles Using Classical and Machine Learning Methods for the Sites of North-Western Areas of Pakistan** is solely my research work with no significant contribution from any other person. Small contribution/help wherever taken has been duly acknowledged and that complete thesis has been written by me.

I understand the zero tolerance policy of the HEC and University, **International Islamic University, Sector H-10, Islamabad, Pakistan** towards plagiarism. Therefore, I as an Author of the above titled thesis declare that no portion of my thesis has been plagiarized and any material used as reference is properly referred/cited.

I undertake that if I am found guilty of any formal plagiarism in the above titled thesis even after award of Ph.D. degree, the university reserves the rights to withdraw/revoke my Ph.D. degree and that HEC and the University has the right to publish my name on the HEC/University Website on which names of students are placed who submitted plagiarized thesis.

Student/Author Signature: _____

Name: **(Muhammad Shafeeq ul Rehman Khan)**

Certificate of Approval

This is to certify that the research work presented in this thesis, entitled: **Estimation of Extreme Flood Quintiles Using Classical and Machine Learning Methods for the Sites of North-Western Areas of Pakistan** was conducted by **Mr. Muhammad Shafeeq ul Rehman Khan**, Reg. No. **02-FBAS/PHDST/F16** under the supervision of **Dr. Ishfaq Ahmad** no part of this thesis has been submitted anywhere else for any other degree. This thesis is submitted to the **Department of Mathematics & Statistics, FBAS, IIU, Islamabad** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy in Mathematics, Department of Mathematics & Statistics, Faculty of Basic & Applied Science, International Islamic University, Sector H-10, Islamabad, Pakistan.**

Student Name: Muhammad Shafeeq ul Rehman Khan ***Signature:*** _____

Examination Committee:

a) **External Examiner 1:**
Name/Designation/Office Address **Signature:** _____

Dr. Muhammad Hanif
Associate Professor,
Department of Statistics,
Arid Agriculture University,

b) **External Examiner 2:**
Name/Designation/Office Address) **Signature:** _____

Dr. Iftikhar Hussain Adil
Assistant Professor of Statistics,
Department of Economics,
School of Social Sciences & Humanities (S3H),
National University of Sciences & Technology (NUST);
Sector H-12, Islamabad, Pakistan.

c) **Internal Examiner:**
Name/Designation/Office Address) **Signature:** _____

Dr. Muhammad Akbar
Assistant Professor

Supervisor Name:
Dr. Ishfaq Ahmad **Signature:** _____

Co-Supervisor Name:
Dr. Zamir Hussain **Signature:** _____

Name of HOD:
Prof. Dr. Tariq Javed **Signature:** _____

Name of Dean:
Prof. Dr. Muhammad Irfan Khan **Signature:** _____

DEDICATED

TO

MY

PARENTS

AND

TEACHERS

Acknowledgements

First of all, I pay special thanks to Almighty Allah, who assisted me in all walks of life and also assisted and guided me throughout my PhD work and in the accomplishment of this thesis too. Many Salat-o-Salam on our Holy Prophet Hazrat Muhammad (PBUH) who is a source of knowledge and blessings for the entire creations.

I would like to acknowledge the worth mentioning supervision of **Dr. Ishfaq Ahmad** and Co-supervision of **Dr. Zamir Hussain** who guided me and supported me during my whole research work. Moreover, their supervision enabled me to develop an understanding of the field. Without their sincere efforts I was unable to complete this hard task of my life. Really I am thankful to **Dr. Ishafaq Ahmad and Dr. Zamir Hussain** for their inspiration and encouragement in every field of life especially in education and teaching. **Almighty Allah** may bless them with long life, health, happiness and knowledge. Moreover, I would like to express my sincere thanks to all faculty members of Department of Mathematics and Statistics IIU Islamabad especially to **Dr. Muhammad Akbar Awan.** I would also like to thank all other faculty member of my university for their sincere appreciation, comments and suggestions. I express my thanks to all staff of Mathematics and Statistics Department, IIU, for their various services. I am thankful to my all friends and would like to mention the names of some friends, Mr. Muhammad Amjad, Mr. Atta Muhammad Asif, Mr. Ehtasham ul Haq, Mr. Nasir Ali, and Hamdullah Khan, for their moral support. They always remain ready to assist me when I was needed.

I especially want to acknowledge efforts and prayers of my, parents, brother (Mr. Ateeq ul Rehman Khan), sisters, nephews (M. Dawar Khan, Ali Nawab Khan, Aahil Khan and Hassan Nawab Khan), my beloved wife Shazia Zaheer and my son Orhan Khan for

their love, care and support in my life, which has been directly encouraging me for my study.

(Muhammad Shafeeq ul Rehman Khan)

Preface

This study is an application of statistical modeling for the estimation of extreme flood quantiles. The data series consists of Annual Maximum Peak Flows (AMPF) (maximum value extracted from the daily data series for each year) of thirty-six gauging sites. These sites are located on various streams/rivers of north-western areas of Pakistan. The focus of this study is two-fold; firstly, the application of L-moments based regional flood frequency analysis (RFFA) coupled with machine learning methods to estimate accurate and reliable flood quantiles considering different return periods for gauged as well as ungauged sites. Secondly, assessing suitability of different estimation methods for fitting Pearson Type-3 distribution. The assessment procedure is based on simulation experiments considering variations in the sample size and shape characteristics of the distribution of data series.

For the application of RFFA, as a preprocessing step, necessary assumptions with respect to recorded data series at each site are validated through various graphical and formal statistical tests. These include time-series graphs, run test, rank-sum test and Wald-Wolfowitz test. This graphical and non-parametric analysis show that data series of all thirty-six sites is random, independent, identically distributed and free of significant trends. In the next stage, an established step-wise methodology of RFFA has been used including identification of discordant sites, formation of homogeneous region(s), assessing goodness of fit of a distribution for identified homogenous region(s) and estimation of quantiles for gauged and ungauged sites. According to results of discordancy measure, two sites "Badri" and "Chilah" are observed as discordant sites. For further investigation, their observed data series have been analyzed and found that high outliers are the main reason within the increase of their discordancy values. Considering the existence of such high extremes in the data series is random as

well as important in RFFA. The sites "Badri" and "Chilah" are retained for further analysis. The initial cluster of thirty-six sites is heterogeneous as showed by L-moments based heterogeneity measures. Therefore, it is subdivided into four homogeneous regions considering the most influential site characteristic, i.e., "Latitude" among available, using wards clustering method and Euclidean distance. Each homogeneous region consists of a different number of sites, twelve in Region 1, nine in Region 2, nine in Region 3 and Region 4 has six sites. The geography of the sites located in Region 1, Region 2 and Region 3 is sub-mountainous while the sites located in Region 4 have mountainous land. After the formation of homogeneous regions, $|Z^{Dist}|$ statistic and L-moment ratio diagram are used as goodness of fit measures. As at-least two distributions have passed goodness of fit criteria for most of the regions. Therefore, a simulation-based assessment analysis has been performed to identify most suitable robust distribution for each region. The details are: Generalized Normal (GNO) distribution is robust distribution for Region 1, Generalized Pareto (GPA) for Region 2 as well as for Region 3, and Generalized Logistic (GLO) for Region 4. Regional and at-site flood quantiles for various return periods are estimated using quantile functions of respective robust regional distributions. Backpropagation neural networks (BPNN), radial bases function (RBF) and regression models with robust and OLS estimation methods are used for the estimation of quantiles at ungauged sites within each homogeneous region. Model evaluation criteria's (error comparison of predicted values) show that RBF is suitable method for Region 1 while BPNN is more appropriate for Region 2, Region 3 and Region 4. The predictive ability of the model for T-year flood quantiles at ungauged sites for each region is verified through historical comparison of the highest recorded values of AMPF at each corresponding site for shorter as well as longer return periods.

Accurate fitting of a probability distribution to annual maxima's is a progressive area of extreme values analysis. Different methods of estimation like L-moments, maximum likelihood, method of moments and maximum product of spacing (MPS) are available with choice of various models like PE3, Generalized Extreme Value, etc. Success depends on the size and span of sample, severity of shape characteristics of the distribution of data series and many others. This study provides comparison of three estimation methods namely L-moments, maximum likelihood and MPS using PE3 distribution. Simulations experiments are designed considering variations in the size of sample and values of skewness and kurtosis of the data. The results of simulation experiments and a case study show that MPS estimation method is a reasonable alternative and provides efficient estimates, especially when the data shows large skewness and kurtosis with small to moderate size of sample.

Major details of the chapter are provided below:

Chapter 1 introduces the problem of research being addressed in this thesis. This chapter includes discussion of some basic issues (distributional choice, regional homogeneity and inter-site dependence) related to flood quantile estimation. Recent literature related to the estimation of flood quantile for gauged and ungauged sites is also part of this chapter.

Chapter 2 provides theoretical/methodological details of different measures/methods used in this study. Description of the study area and details of sample data of AMPF of 36 sites with their respective site characteristics are also given in this chapter.

Chapter 3 includes results of preprocessing of the data series at each site.

Chapter 4 provides stepwise analysis of RFFA. This chapter covers comprehensive details of application of L-moments based RFFA on the sites of north western areas of

Pakistan. Notable point is that this is the first application of L-moments based RFFA to the sites of this study area, which we believe a unique contribution of this thesis.

Chapter 5 presents the results of quantiles estimation for ungauged sites. For these, regression models are developed using OLS and robust estimation methods as well as machine learning methods (BPNN and RBF). Introduction of artificial intelligence or machine learning techniques in analyses of extreme values will bring flexibility and improve efficiency of the estimates, if handled properly. Three research papers based on the findings of Chapter 4 and 5 are published in worthy journals. The details of published papers are given below:

Applied Ecology and Environmental Research (2019) 17(3) 6937-6959.

Applied Ecology and Environmental Research (2020) 19(1) 471-489.

Journal of Flood Risk Management (2021) 14(4) 1-21.

Chapter 6 investigates the effects of three methods of estimation namely LM, MLE and MPS considering PE3 distribution. Assessment is based on a two-step approach. The first step uses simulation experiments while the second is based on empirical analyses, by varying size and shape characteristics of the sample. The results of this chapter provide useful guidelines for fitting PE3 distribution, especially in modeling of extreme values. The findings of this chapter are also published in a reputable journal with following publication details:

Water Resources Management (2021) 35(5) 1415–1431.

Table of Contents

Research Profile

Total four research articles are published from this thesis. The details of these articles are as follows.

Khan, M. S. R., Hussain, Z., & Ahmad, I. (2019). A comparison of quadratic regression and artificial neural networks for the estimation of quantiles at ungauged sites in regional frequency analysis. Applied Ecology and Environmental Research, 17(3), 6937-6959. http://www.aloki.hu/indvol17_3.htm

Khan, M. S. R., Hussain, Z., & Ahmad, I. (2020). Regional flood frequency analysis, using l-moments, artificial neural networks and OLS regression, of various sites of Khyber-Pakhtunkhwa, Pakistan. Applied Ecology and Environmental Research, 19(1), 471-489. http://www.aloki.hu/indvol19_1.htm

Khan, M. S. R., Hussain, Z., & Ahmad, I. (2021). Effects of L-Moments, Maximum Likelihood and Maximum Product of Spacing Estimation Methods in Using Pearson Type-3 Distribution for Modeling Extreme Values. Water Resources Management, 35(5) 1415–1431. https://link.springer.com/article/10.1007/s11269-021-02767-w

Khan, M. S. R., Hussain, Z., & Ahmad, I. (2021). Modelling of flood extremes using regional frequency analysis of sites of Khyber Pakhtunkhwa, Pakistan. Journal of Flood Risk Management, 14(4) 1-21. https://doi.org/10.1111/jfr3.12751

List of Abbreviations

AMPF:	Annual maxima's of peak flows
RFFA:	Regional flood frequency analysis
RFA:	Regional frequency analysis
EVA:	Extreme value analysis
Long:	Longitude
Lat:	Latitude
Ele:	Elevation
AARF:	Average annual rainfall
ARMS:	Average rainfall in monsoon
AAT:	Average annual temperature
GEV:	Generalized Extreme Value Distribution
GNO:	Generalized Normal Distribution
GLO:	Generalized Logistic Distribution
PE3:	Pearson Type-3 Distribution
GPA:	Generalized Pareto Distribution
ANN:	Artificial neural network
BPNN:	Backpropagation neural network
RBF:	Radial Basis function
MPS:	Maximum product of spacing
LM:	Method of L-moments
MLE:	Maximum likelihood estimates
LOOCV:	Leave one out cross validation
RMSE:	Root mean square error
MAP:	Mean absolute percentage error

LR:	Linear regression
OLS:	Ordinary least square
QR:	Quadratic regression
$ARMS_{R1}$:	Vector of values of average rainfall in monsoon for Region 1
$ARMS_{R2}$:	Vector of values of average rainfall in monsoon for Region 2
$ARMS_{R3}$:	Vector of values of average rainfall in monsoon for Region 3
$ARMS_{R4}$:	Vector of values of average rainfall in monsoon for Region 4
UB:	Upper error bound
LB:	Lower error bound
KPK:	Khyber-Pakhtunkhwa
GB:	Grubbs and Beck test
CVM:	Cramer Von-Mises test
SEF:	Standard error of fit
L-CV:	L-moment based coefficient of variation

List of Notations

n:	Sample size
l_1:	First sample L-moment
τ:	First sample L-moment ratio (L-CV)
τ_3:	Second sample L-moment ratio (L-Skewness)
τ_4:	Third sample L-moment ratio (L-kurtosis)
τ_5:	Fourth sample L-moment ration based on 5^{th} sample L-moment
D_i:	Discordancy measures
$\lvert Z^{Dist} \rvert$:	Test statistics for best fitted regional distribution
H_1:	Heterogeneity measure based on L-CV
H_2:	Heterogeneity measure based on L-Skewness
H_3:	Heterogeneity measure based on L-Kurtosis
$\hat{l}_{1R1}$:	Vector of first sample L-moments of sites of Region 1
$\hat{l}_{1R2}$:	Vector of first sample L-moments of sites of Region 2
$\hat{l}_{1R3}$:	Vector of first sample L-moments of sites of Region 3
$\hat{l}_{1R4}$:	Vector of first sample L-moments of sites of Region 4

List of Figures

List of Tables

Chapter-1

Introduction:

Frequency analysis of extreme events like floods, rainfall, winds and droughts is necessary for effective planning and management against these natural disasters. It is also useful for the design and development of hydrological structures such as dams, barrages, culverts and bridges, to ensure public safety and efficient utilization of available water resources and so on. The available literature includes a variety of methods for the frequency analysis of extreme events. These methods are mainly divided into at-site and regional frequency analysis (RFA). Both these methods have certain advantages/disadvantages associated to them. However, at-site frequency analysis may not be a preferred choice for the estimation of quantiles due to the availability of a shorter or limited span of observed data series with respect to extreme events at any site. Additionally, the estimates cannot be interpolated/extrapolated effectively for any specific site with no observed record (such sites are commonly referred as ungauged site). Estimates using at-site frequency analysis may suffer from sampling variability especially with the shorter span of observed data while estimation for longer return periods (Cunnane, 1988; Hosking and Wallis, 1993). In this scenario, RFA is an optimum choice, i.e., pooling data of different sites based on similar site characteristics. Major advantages of using RFA include robust estimates of quantiles at gauged sites and estimation or improvement of quantiles at ungauged or partially/poorly gauged sites within the homogeneous region(s). Keeping in view the advantages of RFA, this study is designed to apply a standard methodology of L-moments based RFA available in Hosking and Wallis (1997) to a new and important

study area of Pakistan. The observed variable of analysis at different sites is annual maximum peak flows because L-moments based RFA is most suited and utilized for annual maximum series rather than Peaks-over a threshold or any other monthly/quarterly/seasonal series (Cook, 1985; Hosking and Wallis, 1997; Palutikof et al., 1999 and Ferreira and de Haan, 2015).

Another important consideration in frequency analysis of extreme events is the choice of models or probability distributions for fitting the data series. Many probability distributions with two, three, four and five parameters are available and used for the fitting of extreme values related to floods, rainfall and winds. Moreover, interesting debate is available in literature with respect to modeling of extreme values and generally there is a consensus that distributions with three to five parameters are appropriate to consider as candidates. The use of two-parameter probability distributions resulted in biased estimates of tail quantiles when the shape of frequency distribution is not well estimated by the fitted distribution (Hosking and Wallis 1997). Secondly, with respect to the estimation methods, a wide variety is available including method of moments, maximum likelihood, probability weighted moments, L-moments, trimmed L-moments, L-H moments, maximum product of spacing, etc. There are no universal criteria linked with the use of a single parameter estimation method. However, useful guideline suggest that success depends on the span of data and shape characteristics especially skewness and kurtosis of the data series at different sites. This study has used a set of five three-parameter distributions namely Generalized Extreme Values (GEV), Generalized Pareto (GPA), Generalized Normal (GNO), Generalized Logistic (GLO) and Pearson Type-3 (PE3), with L-moments as estimation method.

Estimation of quantiles at ungauged sites using the estimates of gauged sites is an important part of RFA analysis. For ungauged T-year flood quantiles estimates using

estimates of gauged sites, various methods are in practice including regression analysis with linear as well as non-linear approaches, artificial neural networks (ANN), etc. None of these methods, however, received universal consensus. Adequacy of the method depends upon the quality and availability of the site characteristics and the type of relationship with the observed records at various sites of a homogeneous region. Therefore, keeping in view the available number of sites of analysis, span of data and limited site characteristics, we have used regression models using ordinary least square (OLS) and robust estimation methods. Additionally, two machine learning approaches (back propagation neural network (BPNN) and radial basis function (RBF)) have been used for the estimation of quantiles at ungauged sites within homogeneous region(s). A comparative analysis has been provided to identify most suitable method in given scenarios. A comparison with historic values is also illustrated to highlight the practical utility of the estimated quantiles.

Choice of model and estimation methods is also a debatable issue in at-site frequency analysis especially when the region understudy is showing extreme heterogeneity and not suitable to perform RFA (Soukissian and Tsalis, 2015). The analysis exploring utility of at-site frequency analysis, especially in the presence of small sample size and high skewness in the observed data series, is also part of this dissertation. PE3 distribution is selected as candidate distribution because it is important probability distribution which is used for the for modeling of extreme events. Comparison of three estimation methods (two common/popular methods L-moments and maximum likelihood while one relatively rarely used method maximum product of spacing) has been provided by varying size of sample and shape characteristics using simulation experiments and applied example.

1.1 Background of the study

Pakistan, with integrated river basins, has a long history of floods since 1947 (from the year of its independence). Twenty-four major floods have occurred in the country from 1947 to 2016. Resultantly, the country has suffered direct economic loss of about 38.171 billion USD, approximately 12,502 lives lost, around 197,273 villages destroyed/damaged and over 616,598 Sq.km area affected. Floods are increasing in frequency and intensity in the country by the year 2000 and their trends are alarming from 2010 onwards. The flood that occurred in 2010 was the worst one in the region during the last 80 years or so (Government of Pakistan, 2017). These floods mainly occur due to heavy monsoon rainfall which results into massive water inflows in the main rivers, and main hill torrents nullahs/streams having sharp slopes which significantly enhance the flood intensity and destroy the banks severely.

Indus River is a major river of Pakistan. Indus River System is known as the world's biggest river system due to its very large basin area. This river system has two major sections of tributaries. One section (commonly known as eastern tributaries) consists of rivers Jhelum, Chenab, Ravi, Beas and Sutlej and other section has Kabul, Swat, Panjkora, Bara, Shah Alam and Jundi as north-western tributaries. Eastern tributaries originate from Jamun and Kashmir (A disputed territory between India and Pakistan) and flows from north to south in the Punjab province. Western tributaries of Indus River mostly originate from northern areas of Pakistan except Kabul River (its origin is in Afghanistan, the neighboring country of Pakistan) and flows from northwest to south in the area of Khyber Pakhtunkhwa (KPK). These rivers and streams of the region of KPK has natural flow (less effected from man made changes). Therefore, are suitable to perform L-moments based RFA methods using data series of different gauging sites.

The geography of KPK is mountainous from north side and sub-mountain in south. Therefore, area of KPK is vulnerable to flash flooding due to its steep geography and uncontrolled flow of rivers/streams. This region is badly affected due to flash floods in 1992 and 2010 (Pakistan Meteorological Department, 2012). Therefore, the need of modeling extreme flow behavior of the observed records of various sites in this region is immense. The flood estimates of modeling procedure can be used for effective preventive measures against these natural disasters, generation of flood risk maps, management of stream water and feasibilities/designing of new hydraulic structures. Some pictures of 2010 flood are given in Fig (1.1).

Fig. 1.1: Some pictures of the 2010 flood disaster in KPK Pakistan.

Another important aspect is that Pakistan is a developing country with agriculture sector as a major contributor to its gross domestic product (GDP). Agriculture sector contributes about 24 percent in the GDP of Pakistan (Pakistan Bureau of Statistics, 2018). Therefore, for sustainable economic growth and to ensure food security in the country analyzing the available river water in the area of KPK, using standard methodologies is a primary need of time. The resultant future estimates of floods

quantiles through this study will be useful for agriculture water management and optimum utilization of the available water resources.

1.2 Literature review

1.2.1 Application of L-moments based RFFA

RFA using L-moments is a well-documented popular methodology with application in several case studies around the world. Highlights of few studies of RFFA are provided below:

In Malaysia, Lim and Lye (2003) used RFFA to analyze annual maxima series of 23 gauging sites of Sarawak River basin. For the division of group of 23 sites into homogeneous groups, 6 site characteristics namely basin area, specific discharge, return-period storms with duration of 12 h (T5, T10, T20, T50), mean annual rainfall, longitude and latitude were used. The study area was divided into two homogeneous regions using Euclidean Distance. The results showed that Generalized Extreme Value (GEV) and GLO were best fitted distributions for homogeneous regions.

In India, Kumar and Chatterjee (2005) used annual maxima series of 13 gauging sites of North Brahmaputra region. The considered region was homogeneous in nature as showed by the results of L-moments based heterogeneity measure. GEV distribution was identified as robust distribution for the estimation of quantiles.

Alam et al. (2016) selected 18 sites to perform L-moments based RFFA. Two sites having large values of discordancy measure were dropped from the analysis and the region consists of remaining 16 sites was homogeneous. The Z^{Dist} goodness of fit methods showed that GEV and Gumbel distributions were identified as good fit regional distributions.

In Turkey, Saf, (2009) used K-mean cluster analysis with first five L-moments site statistics to divide the 47 gauging sites into three homogeneous regions. Findings of the

study showed that Pearson type-3 (PE3) is best fitted distribution for Antalya and Lower-West Mediterranean regions and GLO for Upper-West Mediterranean region. Aydoğan et al. (2016) divided the 29 gauging sites of Çoruh Basin of Turkey into four homogeneous regions using geographical suitability method. Results of the study showed that PE3 is best fitted distribution for Region-1 and Region-4, GPA for Region-2 and GEV for Region-3.

In China, Yang et al. (2010) utilized geographical and statistical attributes to divided 19 gauging sites of Pearl River into three homogeneous regions. Outcomes of the study showed that GEV is best fitted distribution for first region, Wakeby distribution for second and GLO distribution for third region.

In Iran, Mosaffaie (2015) divided 15 gauging sites of rivers located in Qazvin province into two homogenous regions. Initially, based on the factor analysis basin area is identified as an important site characteristic that have significant affects on the homogeneity of regions among others like basin slope, perimeter, main channel slope and main channel length. Then basin area is used for sub division of gauging sites into homogeneous region using Ward Clustering with Euclidian Distance. Finding of this study showed that GLO distribution is robust distribution for first region while GPA distribution for second region. Mesbahzadeh et al., (2019) considered a region of 9 sites of Loot River basin to perform RFA., as the results show that entire region is homogeneous and goodness of fit Z^{Dist} statistics show that the Log-Pearson type-3 was the best fitted regional distribution.

In Korea, Lee and Kim (2019) used data of 20 gauging sites of Chungju dam basin to performed RFFA and three distributions GLO, GEV and GNO have passed the criteria of goodness of fit. A simulation analysis was performed and resultantly GNO was identified as robust regional distribution. Similarly, many other applications of RFFA

are available in the literature advocating applicability, significance and effectiveness of this methodology. This methodology has also been used for the quantile estimation of other extreme events like rainfall, droughts and winds. Two important studies providing inter-comparison of various regional flood estimation procedures are by GREHYS (1996a, b). A brief of the development in RFA has been illustrated in Malekinezhad and Zare-Garizi (2014).

1.2.1.1 A brief of application of RFA using L-moments in Pakistan

RFA has also been applied in few published studies related to extreme values of rainfall, floods and winds in Pakistan. For instance, for rainfall (Ahmad et al. 2013; Shahzadi et al. 2013; Ahmad et al. 2016; Ahmad et al. 2017a; Hussain et al. 2017; Khan et al. 2017), for floods (Hussain and Pasha, 2009; Hussain, 2011; Ahmad et al. 2016; Ahmad et al. 2017b; Hussain, 2017; Batool, 2017), for wind (Fawad et al. 2018; Fawad et al. 2019). Since this study has a focus on flood frequency analysis. Therefore, highlights of the published literature regarding flood frequency analysis are provided in the following section:

Hussain and Pasha (2009) identified Generalized Normal (GNO) Distribution as robust probability distribution considering annual maxima's of river discharges for seven sites of three major rivers of Punjab namely Jhelum, Ravi and Chenab. Hussain (2011) used annual maxima's of river discharges of seven sites located on the Indus River to perform RFA. The results showed that Pearson Type-III (PE3) is a robust distribution for the upper half of the Indus River while Generalized Logistic (GLO) is a robust distribution for the lower half. Ahmad et al. (2017b) performed RFA using 10 days average of low flows of nine sites located on different rivers of Pakistan. The study area, consisting of nine sites, was divided into two homogeneous regions. Region 1 having sites Tunsa, Tarbela, Nowshera, and Kalabagh while Region-2 includes

Chashma, Guddu, Mangla and Marala. GNO distribution for Region 1 and Generalized Pareto (GPA) distribution for Region-2 were identified as best fitted distributions. Hussain (2017) performed RFA using annual maxima's of river discharges considering gauging stations of major rivers of Punjab, Pakistan, namely Ravi, Sutlej, Jhelum and Chenab. Two homogeneous regions were defined for the study area. Region-1 contains sites Mangla, Rasul, Marala, Khanki, Qadirabad, Balloki, Sidhnai, Suleimanki and Islam, while Region-2 have only sites Trimmu and Panjnad. The results showed that PE3 is most suitable for Region-1 while GNO distribution is best choice for the sites of Region-2.

These aforementioned details reveal that the sites of major rivers of Punjab and the Indus River have been the focus of studies so far. Hence, there is a need to analyze extreme values of sites located on the rivers and streams of other parts of the country, especially Khyber Pakhtunkhwa (the north-western area of Pakistan). Another important reason of modeling annual maximas of different sites of the study area is that rivers and stream located in this area are the second major sources of river water in Pakistan.

1.2.2 Estimation of quantiles at ungauged sites

Another interesting dimension of analysis in this study is the development of models for the estimation of quantiles at ungauged sites. In the existing literation, a broader division of estimation methods is development of linear models using ordinary least squares, non-linear models using different estimation methods and machine learning techniques including artificial neural network, random forest regression, etc. Brief details of few studies are provided below:

1.2.2.1 Studies using linear regression models

Jingyi and Hall (2004) proposed a multiple linear regression model for ungauged flood quantiles estimation. The selection of explanatory variables was performed through backward elimination method. Resultantly, four site characteristics (catchment area, weighted mean river slope, average annual rainfall, mean annual maximum catchment 1-day rainfall) out of eight were selected for the development of model. Griffis and Stedinger (2007) proposed a generalized least square (GLS) method for the estimation of regression model. The study showed that GLS method gives more efficient and reliable estimates of hydrological regression model parameters as compared to OLS and weighted least square (WLS). Zaman et al., (2012) used two variables design rainfall intensity and catchment area as regressors for the development of regional forecast model for ungauged sites. Smith et al., (2015) selected average annual rainfall and catchment area as explanatory variables for the estimation of mean annual flood with in the homogeneous region. Komi et al., (2016) developed a regression model using three site characteristics drainage area, mean annual rainfall and mean basin slope as regressors for the ungauged flood quantiles estimation. Yang (2016) developed a forecast equation for ungauged sites using drainage area as independent variable. Hailegeorgis and Alfredsen (2017) developed linear regression models using catchment area as explanatory variable.

The above-mentioned studies developed linear models between mean of observed data series at different sites (dependent variable) and their respective site characteristic(s) (independent variable(s)), but only few of them have illustrated complete theoretical and statistical justifications of the developed models.

1.2.2.2 Studies using non-linear models

Scarce literature is available with respect to the development of nonlinear relationship between site characteristics and observed data of gauged sites for ungauged quantiles estimation. Anilan et al., (2018) used the site characteristics drainage area, mainstream slope, mean annual rainfall, stream density, elevation, and rainfall intensity as explanatory variables for development of regional forecast model for ungauged sites. The results show that non-linear model gives reliable estimates as compared to linear models. Cassalho, et al., (2019) developed non-linear regression model by using site characteristics area, mean slope, stream gradient, and flow length as independent variables. The findings of the study showed that nonlinear models performed better than linear models. Khan et al., (2019) proposed a quadratic regression model based on single explanatory variable "average rainfall in monsoon" and used robust estimation method for estimation of model parameters. Durocher et al., (2019) introduced nonparametric regression methods for ungauged flood quantiles estimation. Few other studies have showed that nonlinear relationships may provide more accurate estimates of flood quantiles at ungauged sites as compared to linear relationships (Sivakumar and Singh, 2012; Ouali et al., 2017). In an important study,

Anilan et al. (2016) illustrated details of commonly used site characteristics as independent variables in different studies around the world for development of regression models. These include drainage area, slope of stream, and mean annual rainfall. Adding to this point, we emphasize that identification of the most influential site characteristics, among available, having strong correlation with the dependent variable is an ongoing area of research. Development of an adequate model depends on plenty of features including availability of data with respect to site characteristics and the nature of relationship of site characteristics with observed data series.

1.2.2.3 Studies using machine learning models

Due to recent developments in computational resources, another evolving choice in estimation methods, especially for estimating flood quantiles at ungauged sites, is machine learning models including artificial neural networks (ANN). ANN is a nonparametric approach that works like biological operative of a human brain (Rumelhart et al., 1985). This method provides reliable results over other estimation techniques including regression analyses (Liu et al., 2009; Landi et al., 2010). ANN methods are used to handle various hydrological problems such as river/stream flood forecasting and rainfall modelling (Govindaraju, 2000; Dawson and Wilby, 2001; Abrahart et al., 2004). In a study, Dawson et al., (2006) made a comparison between ANN models and regression model to predict ungauged flood quantiles and the value of index flood. The results show that the accuracy of ANN estimates was better than regression. Shu and Ouarda, (2007) proposed canonical correlation based ANN model for ungauged quantiles and compered its efficiency with single ANN models. The final recommendation is that ANN model based on canonical correlation provides better estimates of quantiles. Aziz et al., (2013) compared the ANN model with fuzzy-based methods and gene expression programming for the prediction of ungauged flood quantiles. In another study, Aziz et al., (2014) performed an RFFA and made a comparison between ANN model and quantile regression for ungauged flood estimates. Anilan et al., (2016) compared the accuracy of ungauged flood quantiles obtained through various regression models with ANN. Ouali et al., (2017) illustrated that the relationship between site statistic and site characteristics is strongly nonlinear. Therefore, ANN methods can give more reliable estimates of ungauged flood quantiles. All these studies cited above may lead to a conclusion that ANN models can provide better estimates of quantiles for ungauged sites as compared to regression models. One

major advantage of ANN is its capacity to identify complex nonlinear relationships and before numerical analysis, there is no need to express such a relationship in mathematical form as the data itself recognizes the model form through the use of artificial intelligence (Hjelmfelt and Wang, 1996).

In hydrological analysis, among different methods of ANN, the radial basis function (RBF) network is a preferred choice because of its accuracy to estimate non-linear and complex functions (Ham and Kostanic, 2001). Allahbakhshian-Farsani (2020) suggested that support vector regression model based on RBF provides more reliable estimates of flood quantiles relative to other machine learning methods. In another study, Haddad and Rahman (2020) showed that RBF method gives more consistent quantile estimates for ungauged sites. A brief of the predictive ability of RBF network in extreme floods is available in Lin and Chen, 2004; Lin et al., 2009 and El-Shafie et al., 2009. Therefore, there is a clear margin of use of machine learning methods for development of relationship for estimation of flood quantiles at ungauged sites in Pakistan. In this study, we have used linear, non-linear and machine learning methods for development of functional relationship for prediction of flood quantiles at ungauged sites. Comparison has been made using various accuracy measures to obtain the most suitable method.

1.2.2.4 At-site frequency analysis in case of heterogeneous region(s): Choice of estimation methods

In RFA application, homogeneity in characteristics of observed record a part from a site-specific scale factor for a group of sites or a region is a critical and essential requirement. In situations, when it is difficult to form homogenous region(s) of the study area, at-site frequency analysis becomes alternate choice. This scenario generates various other challenges including choice of model (probability distribution) and

estimation methods. Success depends on various factors like size and span of the observed record and shape of the distribution of observed data at a site. In this study, we have analyzed the suitability of PE3 distribution for fitting extreme values with three estimation methods namely L-moments, maximum likelihood and maximum product of spacing. PE3 is an important probability distribution for modeling of variety of extreme events. In terms of estimation methods, L-moments and MLE more commonly used for the distribution fitting as compared to MPS. Recently, in some case studies, accuracy of MPS method has been compared with LM, MLE and few others. For example; Soukissian and Tsalis, (2015) illustrated a comparison between different methods of estimation like LM, MLE, MPS and others using GEV distribution for extreme winds quantiles estimation. Their results showed that the estimates based on MPS were better than MLE, LM and others in terms of bias and root mean square error. Asquith et al. (2017) performed a study to assess the uncertainties associated with smaller return period flood quantiles. They reported that accuracy of the estimates based on MPS and LM is comparable to each other. Due to these interesting facts, we have also analyzed the effects of MPS, MLE and LM in the case of fitting PE3 distribution considering a variety of sample sizes and shape characteristics using simulation experiments and a case study. This portion of the study is a novel contribution in the literature of modeling of extreme values. Further details are provided in chapter 6.

1.3 Gaps in the literature

These aforementioned details reveal that there exist few interesting areas of research which needs further investigation with respect to the estimation of flood quantiles, especially in Pakistan

a. L-moments based RFA has never been used for the estimation of flood quantiles for the sites of KPK and no such methods/approaches are available for the reliable

estimation of flood quantiles at ungauged or poorly/partially gauged sites within the study area.

b. Choice of model and estimation method for at-site analysis in case the region under study is heterogeneous

1.4 Objectives of the study

The current study is designed to achieve the following goals.

- To observe and discuss the general trends and tendencies of extremes of floods, i.e. AMPF at various sites of north-western areas of Pakistan using descriptive statistics in terms of L-moments.
- To obtain regional and at-site estimates of flood quantiles using L-moments under RFFA for various return periods.
- To assess the accuracy and reliability of the estimated regional and at-site quantiles using simulation experiments.
- To develop functional relationship between variables for estimation of flood quantiles at ungauged sites using linear, non-linear and machine learning methods, etc.
- To provide useful guidelines for choice of estimation methods in at-site frequency analysis using PE3 distribution, especially when the region understudy is heterogeneous.

1.5 Organization of subsequent chapters

Rest of the thesis organized as follows. chapter 2 describes different methods used for analysis, chapter 3 provides results of data screening for frequency analysis, chapter 4 elaborates detailed application of RFA, chapter 5 illustrates the development of forecast models for the estimation of quantiles at ungauged sites and chapter 6 covers guidelines for choice of estimation methods among LM, MLE and MPS in case of PE3 distribution

for at-site frequency analysis considering variations in size of sample and shape characteristics. Last section provides summary and conclusions of the study and some recommendations for future research.

Chapter 2

Material and Methods

2.1 General

This chapter gives a detailed description of methodology used in this study. As per the sequence of execution, general description of methods/steps include: 1) Application of non-parametric tests for pre-processing of the data series at each site 2) L-moments for descriptive analysis and RFA including formation of homogeneous region(s), distribution fitting and quantiles estimation 3) Assessment analysis of the estimates using simulation experiments 4) Application of linear, non-linear and machine learning methods for the estimation of quantiles at ungauged sites 5) assessment of different estimation methods (L-moments, maximum likelihood (MLE) and maximum product of spacing (MPS)) in at-site frequency analysis using PE3 distribution in presence of different sample sizes and shape parameters/characteristics.

Details of the each adopted method and technique for the analysis have been given in the following sections of this chapter.

2.2 Nonparametric tests

While dealing with data in any applied study, preprocessing of the available data is crucial as it directly impacts the quality of estimates. In a statistical analysis preprocessing generally includes but not limited to cleaning of data, checking and estimation of missing values, detection of outliers, randomness of data series, and independently and identically distributed sample free from significant trends. This study is based on secondary data of annual maximum peak flows at different gauging sites for the application of L-moments based RFA. Therefore, non-parametric tests have

been used to deal with sensitivity of underlying assumptions of parametric methods as samples of extreme values usually have small span and non-normal behavior.

Details of each non-parametric test have been provided below.

2.2.1 Grubbs and Beck Test

Grubbs and Beck (1972) introduced a test for the detection of outliers within the sample data. This test is successfully applied for detection of outliers within the annual maxima and annual minima series of hydrological data (WRC, 1981).

For the application of this test, initially, the sample data sets are transformed by taking the natural logarithm. Then the values of lower and upper bounds from the ranked data sample are calculated using:

$$X_L = EXP\left(\bar{Y} - K_{N,\alpha} S_y\right) \quad (2.1)$$

$$X_U = EXP\left(\bar{Y} + K_{N,\alpha} S_y\right) \quad (2.2)$$

Where $\bar{Y}$ is mean and S_y is the standard deviation of transformed sample data, $K_{N,\alpha}$ is the critical value of the GB and N is the sample size of a random variable.

2.2.2 Runs test

Run test of randomness given in (Bradley, 1968; Hirsch et al., 1992) has been used to check the randomness of observed data set of each site. The test statistics of the Run test for the large sample size is given below:

$$Z = \frac{R - E(R)}{S.E.(R)} \quad (2.3)$$

Here, R is the total number of runs, $E(R)$ is the expected value of R, S.E.(R) is the standard error of R.

2.2.3 Rank Sum test

To validate the assumption of the identical distribution of the data set of each site, Rank-Sum has been used. The details of this test are available in (Hirsch et al., 1992):

For small samples, i.e. if n_1 and n_2 are less than 10, following test statistics is used

$$W = mini(W_1 - W_2) \quad (2.4)$$

where, W_1 is the sum of the rank of the first group and W_2 is the sum of the rank of the second group.

In the case of dealing with large sample, i.e. greater than 10, the test statistic is

$$Z_w = \frac{W-\mu_w}{\sigma_w} \quad (2.5)$$

Here

$$\mu_w = \frac{n_1(n_1+n_2+1)}{2} \text{ and } \sigma_w = \sqrt{\frac{n_1 n_2(n_1+n_2+1)}{12}}$$

2.2.4 Wald-Wolfowitz test

For the validation of an important assumption regarding the data series at each site that it is independent and free from significant trends, Wald-Wolfowitz test (Wald and Wolfowitz, 1943) has been used. For a sample size of less than 10, the test statistic is given as:

$$K = \sum_{i=1}^{n-1} x_i x_{i+1} + x_1 x_n \quad (2.6)$$

The expected mean and variance is

$$\mu_k = \frac{s_1^2-s_2}{n-1}, \text{ and } \sigma_k^2 = \frac{s_2^2-s_4}{n-1} - E(K)^2 + \frac{s_1^4-4s_1^2s_2+4s_1s_3+s_2^2-2s_4}{(n-1)(n-2)},$$

with $s_t = \sum_{i=1}^{n} x_i^t, t = 1,2,3,4$

For a large sample, i.e. greater than 10, the test statistic is computed as;

$$Z_{ww} = \frac{K-\mu_k}{\sigma_k} \quad (2.7)$$

2.3 Method of L-moments

Hosking (1990) introduced a method of estimation known as L-moments (MLM) based on order statistics of the observed data series. MLM gives reliable and robust estimates of parameters of probability distribution especially in case of small sample.

Suppose "X" is a continuous random variable with known probability distribution function. Its first four population L-moments can be derived as flows.

$$\lambda_1 = E(X_{1:1}) \tag{2.8}$$

$$\lambda_2 = \frac{1}{2}E(X_{2:2} - X_{1:2}) \tag{2.9}$$

$$\lambda_3 = \frac{1}{3}E(X_{3:3} - 2X_{2:3} + X_{1:3}) \tag{2.10}$$

$$\lambda_4 = \frac{1}{4}E(X_{4:4} - 3X_{3:4} + 3X_{2:4} - X_{1:4}) \tag{2.11}$$

The general form of the above equations can be written as.

$$\lambda_r = r^{-1}\sum_{j=0}^{r-1}(-1)^j \binom{r-1}{j} E(X_{r-j:r}) \tag{2.12}$$

The expression $E(X_{r:n})$ is defined as

$$E(X_{r:n}) = \frac{n!}{(r-1)!\,(n-r)!}\int_0^1 x(F)F^{r-1}(1-F)^{n-r}dF \tag{2.13}$$

Here F is the cumulative distribution function (CDF) of random variable X.

For the quantities defined in Eq. (2.8 and 2.9), λ_1 is the location parameter while λ_2 is the scale parameter of probability distribution.

The first four sample L-moments analogues to Eq. (2.12) can be obtained as follows.

$$l_1 = b_0 \tag{2.14}$$

$$l_2 = 2b_1 - b_0 \tag{2.15}$$

$$l_3 = 6b_2 - 6b_1 + b_0 \tag{2.16}$$

$$l_4 = 20b_3 - 30b_2 + 12b_1 - b_0 \tag{2.17}$$

Here

$$b_r = n^{-1}\sum_{j=r+1}^{n}\frac{(j-1)(j-2)\ldots(j-r)}{(n-1)(n-2)\ldots(n-r)}x_{j:n} \qquad r = 0,1,2 \ldots \ldots n-1 \tag{2.18}$$

L-moment ratios are defined as.

L-cv, $T = \frac{\lambda_2}{\lambda_1}$ the distribution having positive values than $0 \leq T < 1$

L-skewness, $T_3 = \frac{\lambda_3}{\lambda_2}$ the distribution having positive values then $2T - 1 \leq T_3 < 1$

L-kurtosis, $T_4 = \frac{\lambda_4}{\lambda_2}$ the distribution having positive values then $\frac{1}{4}(5T_3^2 - 1) \leq T_4 < 1$

For higher-order L-moment ratios, the general expression is, $T_r = \frac{\lambda_r}{\lambda_2}$ $r \geq 3$

Sample estimates of location, scale, L-cv, L-skewness and L-kurtosis are represented through l_1, l_2, t, t_3 *and* t_4 respectively. Numerical estimates of L-moment ratios are obtained by replacing the sample estimates of lambdas (λ) in the above expressions.

2.4 Steps and measures in RFA

This study has used L-moments based RFA, proposed by Hosking and Wallis (1997) and applied in various case studies around the world. It is a stepwise approach including calculation of discordancy measure, formation of homogenous region(s), identification of best suited probability distribution and estimation of parameters and quantiles. A summary of few measures of the procedure is given below:

1) To recognize the discordant site(s) *i* in a group of N sites, we define a dissimilarity measure:

$$D_i = \frac{1}{3}N(u_i - \overline{u})^T S^{-1}(u_i - \overline{u}), \qquad i = 1,2,3,\ldots..N \qquad (2.19)$$

Where $S = \sum_{i=1}^{N}(u_i - \overline{u})\,(u_i - \overline{u})^T$

u_i represents a vector of sample L-moments ratios of site *i*, $\overline{u}$ is their mean and *N* is the total number of sites.

2) An important step in RFA is the formation/identification of homogeneous region, i.e. grouping sites with similar site characteristics. Heterogeneity measures based on sample L-moment ratios L-CV, L-Skewness and L-Kurtosis are used to test the regional heterogeneity. If the value of heterogeneity test is less than one the region is considered as homogeneous, if it lies between one and two the region is possibly homogeneous and the region is regarded as definitely heterogeneous if the value of the

test is greater than two. The values of heterogeneity tests based on the sample L-skewness and L-kurtosis rarely exceed from two for a complete heterogeneous region and both tests have less power to differentiate between homogeneous and heterogeneous regions. Consequently, the heterogeneity test based on L-CV considered to be more power full than the tests based on L-skewness and L-kurtosis (Hosking and Wallis, 1997; Satyanarayana and Srinivas, 2008). Therefore, in this study, L-CV based heterogeneity test is used to test the regional heterogeneity. The statistic to compute heterogeneity measure (*H*) is:

$$H_1 = \frac{V-\mu_v}{\sigma_v} \tag{2.20}$$

where $V = \left[\frac{\sum_{i=1}^{N} n_i (t^i - t^R)^2}{\sum_{i=1}^{N} n_i}\right]^{\frac{1}{2}}$ and μ_v and σ_v are respectively the mean and standard deviation of computed inter-site variation obtained through simulations, "t" is the sample L-CV and $t^R = \sum_{i=1}^{N} n_i t^{(i)} / \sum_{i=1}^{N} n_i$.

3) The next step is identification of best-fitted distribution from a set of different three-parameter probability distributions for the defined homogeneous region(s). L-moments ratio diagram and $|Z^{Dist}|$ statistic are used as goodness of fit measures. The formula for $|Z^{Dist}|$ statistic is:

$$|Z^{Dist}| = \frac{\tau_4^{Dist} - t_4^R + \beta_4}{\sigma_4} \tag{2.21}$$

where τ_4^{Dist} is the L-kurtosis of the candidate probability distribution, t_4^R is the regional L-kurtosis, σ_4 is standard deviation and β_4 is the bias of t_4^R obtained through simulations. Further details related to these measures can be found in Hosking and Wallis (1997). A popular set of five three-parameter probability distributions GEV, GPA, GNO, GLO and PE3 have been used as candidates of regional distribution. Their

probability density function (PDF) and cumulative density function (CDF) are provided in Appendix A-1.

4) Estimation of parameters and quantiles of the best-fit distribution is the next obvious step. These regional quantiles are used to estimate at-site quantiles within the homogeneous region using the following relationship:

$$\hat{Q}_i(F) = l_1^{(i)} \hat{q}(F) \tag{2.22}$$

where $\hat{Q}_i(F)$ represents estimated quantile for site i, $l_1^{(i)}$ denotes first sample L-moment of a site i and $\hat{q}(F)$ represents estimated regional quantiles for any return period.

5) It is possible in RFA that two or more probability distributions fulfil the criteria of goodness-of-fit. In such scenario an obvious requirement is to identify a robust probability distribution among successful candidates. For this purpose, a simulations based assessment can be performed to obtain 95% error bounds and root mean square error (RMSE) of the flood quantiles estimates.

RMSE can be calculated using:

$$R_i(F) = \left[M^{-1} \sum_{m=1}^{M} \left\{ \frac{\hat{Q}_i^{[m]}(F) - \hat{Q}_i(F)}{\hat{Q}_i(F)} \right\}^2 \right]^{1/2} \tag{2.23}$$

$\hat{Q}_i^{[m]}(F)$ is the estimated quantile of site-i for non-exceedance probability F at m^{th} repetition. Averaging over the complete region gives:

$$RMSE = N^{-1} \sum_{i=1}^{N} R_i(F) \tag{2.24}$$

Where N represents the total number of gauging stations in a study area. For the growth curve of regional quantiles the quantities $\hat{Q}_i^{[m]}(F)$ and $\hat{Q}_i(F)$ are replaced by $\hat{q}_i^{[m]}(F)$ and $\hat{q}_i(F)$ respectively. The 95% error bounds for $\hat{q}_i(F)$ are.

$$\frac{\hat{q}(F)}{U_{0.025}(F)} \leq q(F) \leq \frac{\hat{q}(F)}{L_{0.025}(F)} \tag{2.25}$$

Where $L_{0.025}(F)$ and $U_{0.025}(F)$ are the values between which approximately 95% of the distribution of simulated values of the ratio of estimated to true values of regional quantile function ($\hat{q}_t(F)/q(F)$) lies.

2.5 Methods of estimation of quantiles at ungauged sites

For the estimation of flood quantiles at ungauged a functional relationship between l_1 (mean of observed AMPF at given sites) and their corresponding available site characteristics has been developed within the homogeneous region. This functional relationship will provide the estimates of at-site mean (for ungauged sites) within the homogeneous region for the estimation of T-years flood quantiles.

Various methods are in practice for the ungauged flood quantiles estimation. These methods include regression with linear/non-linear approaches (Griffis and Stedinger, 2007; Sivakumar and Singh, 2012; Hailegeorgis and Alfredsen, 2017; Ouali et al., 2017), artificial neural networks (Aziz et al., 2014; Anilan et al., 2016), satellite precipitation products (Gado et al., 2017), remotely sensed precipitation information (Faridzad et al., 2018), etc. The given literature show that the relationship between l_1 and the site characteristics is complex. Therefore, none of the adopted method(s) so far received universal acceptability; however, success depends on the availability and suitability of gauged site characteristics. In this study, regression and ANN methods have been used to develop a predictive model of each respective homogeneous region to estimate ungauged flood quantiles. Regression analysis based on linear and polynomial models with OLS and robust estimation methods. These regression models and OLS estimation method are well known in the literature. Their details have been given in Appendix A-2. The description of the ANN and robust regression methods has been given in the following subsections.

2.5.1 Robust regression methods

2.5.1.1 M-estimation method

Huber (1964) introduced the M-estimation method for the estimation of regression model parameters. It is based on maximum likelihood estimation and it gives efficient results as OLS. Fox (2002) suggested that the M-estimation method is the most commonly used method of robust regression. The M-estimation method gives robust estimates of regression parameters when extreme observations or outliers are present in the data sets. In the M-estimation procedure, parametric estimates are obtained by minimizing the residuals function. The objective function for the M-estimate of the regression parameter is given below.

$$\hat{B}_M = min \sum_{i=1}^{n} \rho \left(Y_i - \sum_{j=1}^{k} x'_{ij} B_j\right) \qquad (2.26)$$

To solve this objective function for the estimates of "B" Hampel et al., 2011 proposed a system of normal equations. These normal equations are obtained by taking partial derivatives with respect to unknown parameters and equating them with zero. The final form of M-estimators for regression parameter is given below.

$$\hat{B}_M = (x'wX)^{-1}(x'wY) \qquad (2.27)$$

where w is the matrix having diagonal values of the weight matrix.

$$w = \frac{\varphi\left(\frac{y_i - \sum_{j=0}^{k} x_{ij}\beta}{\hat{\sigma}}\right)}{\left(\frac{y_i - \sum_{j=0}^{k} x_{ij}\beta}{\hat{\sigma}}\right)}, \qquad \text{where } \hat{\sigma} = \frac{median|e_i - median(e_i)|}{0.6745}$$

Because $u_i = \frac{e_i}{\hat{\sigma}}$ than w is equals to the following

$$w = \begin{cases} \left[1 - \left(\frac{u_i}{c}\right)^2\right]^2, & |u_i| \le c \\ 0 & |u_i| > c \end{cases}$$

The procedure of weighted least square (WLS) is adopted to estimate $\hat{B}_M$ using "w" as weight.

2.5.1.2 S-estimation Method

Rousseeuw and Yohai (1984) introduced S-estimation method for the estimation of regression coefficient. S-estimators of regression coefficients are derived from the generalization of two methods Least Median of Squares (LMS) and Least Trimmed Squares (LTS). LMS and LTS also have similar asymptotic properties as the M-estimation method and can tackle half of the extreme (outliers) observations that exist in the data. The residual standard deviation is used by Susanti and Pratiwi (2014) to solve the shortcomings of the median used in the S-estimation process. The S-estimator of regression coefficients is given as follows.

$$\hat{B}_s = min_B\, \hat{\sigma}_s(e_1, e_2, e_1, \ldots\ldots e_n) \tag{2.28}$$

To determine the smallest robust scale, estimator

$$min \sum_{i=1}^{n} \delta\left(\frac{z_i - \sum_{i=1}^{n} y_{ij}B}{\hat{\sigma}_s}\right) \tag{2.29}$$

where

$\hat{\sigma}_s = \frac{median|e_i - median(e_i)|}{0.6747}$, when iteration = 1

$\hat{\sigma}_s = \sqrt{\frac{1}{nk}\sum_{i=1}^{n} \omega_i e_i^2}$, when iteration > 1

$$\sum_{i=1}^{n} y_{ij} \emptyset\left(\frac{z_i - \sum_{i=0}^{k} y_{ij}B}{\hat{\sigma}_s}\right) = 0, \quad j = 0,1,2,3 \ldots\ldots k \tag{2.30}$$

$\emptyset$ is a function drived from takin derivative of δ.

2.5.2 Machine learning methods

As mentioned earlier that there may exist a complex non-linear relationship between l_1 and available site characteristics. Therefore, application of machine learning methods can provide accurate estimates of the dependent variable (l_1) to be used for the estimation of quantiles at ungauged sites. This study has applied different machine learning methods namely back-propagation neural network (BPNN) and radial basis

function (RBF) for estimation of quantiles at ungauged sites. The suitability of these techniques for the prediction of river flows is discussed in details in Maier and Dandy, 2000 and Abrahart et al., 2004. Few details of these methods are as follows:

2.5.2.1 Back-propagation neural network (BPNN)

The ANN model based on back-propagation training is called BPNN. Back-propagation is the process of proper tuning of weights of ANN based on the preceding stage (i.e. iteration) error rate. Lower error rates are achieved by fine-tuning the weights, which increases the model's generalization and hence its reliability. Working structure of BPNN model is given below.

BPNN model comprises an input, hidden and output layers. Neurons layers interact via a network of feed-forward weighted connection. For computations, every input of the neurons multiplied by weight which is known as the connection parameter and combined output with some bias is produced. This value is managed with an activation function. A logistic activation function is used because it provides accurate results for river flow prediction (Shamseldin et al., 2002). A typical logistic activation function is given below.

$$f(x) = \frac{1}{1+e^{-x}} \qquad (2.31)$$

In this study, the relationship between the mean of the AMPF of each site (dependent variable) and site characteristics (independent variables) of regions is estimated using the BPNN model. Working configuration of BPNN is given in **Fig. (2.1)**.

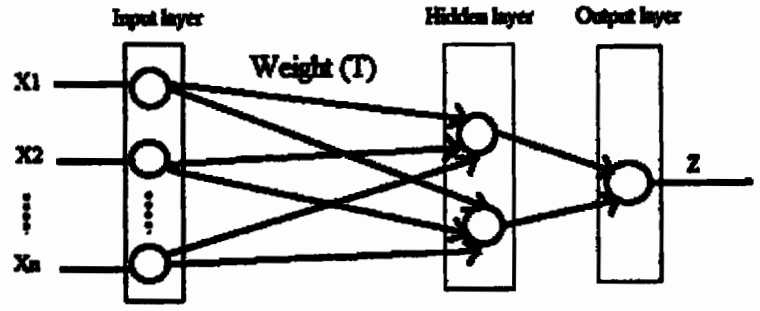


Fig. 2.1: Working configuration of BPNN.

For the application of machine learning methods, the complete data set is usually divided into training, validation and test datasets with the ratio of 60%, 20% and 20%, respectively. This division is useful for large data. In this study, leave one out cross-validation (LOOCV) approach has been used for the training and validation of the sample data set as it is usually considered more useful for smaller data sets. In the LOOCV approach, the data set divides into two parts; if the data set contains n observations, then one observation is used for the validation, i.e. (x_1, y_1) and remaining "n-1" observations $\{(x_2, y_2),(x_3, y_3),,,(x_n, y_n)\}$ are in the training dataset to predict the average value of the dependent variable (which is $\hat{l}_1$ in this case). This process is repeated n times (equals to the total number of observations in the sample) and generate n times mean square error. The estimate of the test means squared error can be obtained from *"n"* test errors as:

$$CV_{(n)} = \frac{1}{n}\sum_{l=1}^{n} MSE_l \qquad (2.32)$$

where MSE denotes mean squared error. For more details of this method see James et al. (2013) and Kuhn and Johnson (2013). The primary objective of training ANN is to reduce the error among the target output and ANN output through adjusting weights

2.5.2.2 Radial basis function (RBF)

RBF is a type of feed-forward neural networks. It is a popular method to solve nonlinear functions because of its easy topological structure and having the capacity to execute

the learning process directly. Furthermore, it has various advantages over the conventional multilayer perceptron like quick convergence, fewer errors and more reliability (Girosi and Poggio, 1990).

The structure of the RBF network is based on three layers; input, hidden and output layers. The input layer divides the input data to the hidden layer without processing the input data. Neurons in a hidden layer of RBF are equal to the historic observation of the predictors. For the estimates of any real-time event, the output of every neuron is the true influence of historic observation. For the input data, every neuron of the hidden layer uses the radial basis function as a non-linear transfer function. The Gaussian function is a commonly used radial basis function. It has two features; center C_j and width H_j. Euclidean distance is used between the center C_j of RBF and input (Y). In the hidden layer, a non-linear transformation is used with RBF as:

$$h_j(Y) = EXP\left(-\left(\|Y - C_j\|^2 / H_j^2\right)\right) \tag{2.33}$$

where h_j is the output of a Jth unit of RBF network, C_j is the center and $H_j$is the width of jth RBF. For the output layer, the following equation is used.

$$Z_k(Y) = \sum_{j=1}^{n} w_{kj} h_j(Y) + B_k \tag{2.34}$$

For any input (Y), Z_k is the kth output unit. Weight connection between jth hidden layer unit and kth output unit is represented by w_{kj}, and B_k represents the bias.

The training of RBF involves the calculation of the weights, spreads and centers. Various mathematical algorithms such as the least square algorithm or genetic algorithm can be used for the selection of centers. After the selection of spread and center of RBF, link weights between output and hidden layer is adjusted using the least square algorithm.

2.6 Method of maximum product of spacing (MPS)

This study is an application of L-moments based RFA but there exists scenarios in which homogeneous regions cannot be identified. An alternate solution then is at site frequency analysis. Quality of quantiles estimates using at-site frequency analysis typically depends on size and span of the sample, distribution characteristics, choice of model and estimation method. This study has also analyzed effectiveness of different estimation methods for fitting Pearson Type 3 distribution in case of at-site frequency analysis through simulations experiments by varying size of sample and shape characteristics. Findings of simulation experiments are validated using real life examples. The estimation methods include L-moments, MLE and maximum products of spacing. Few details of MPS method are provided below.

Cheng and Amin (1983) introduced order statistics-based method of MPS for the estimation of any continuous probability distribution. Suppose $y_{1:n} \leq y_{2:n} \leq \cdots \leq y_{n:n}$ be the ordered sequence of any continuous random variable "Y" having CDF $F(y, \theta)$ and pdf $f(y, \theta)$. The space between two CDF of the consecutive points can be defined as.

$$Z_i(\theta) = F(y_{i:n}, \theta) - F(y_{i-1:n}, \theta) = \int_{y_{i-1:n}}^{y_{i:n}} f(y, \theta) dy, \quad i = 1,2,3 \ldots n+1 \qquad (2.35)$$

Where $y_{0:n} = -\infty$ and $y_{n+1:n} = +\infty$ then the sum of $Z_i(\theta)$ is equals to one.

$$\sum_{i=1}^{n} Z_i(\theta) = 1 \qquad (2.36)$$

and

$$Z_1(\theta) = 1 - F(y_{n:n}, \theta) \qquad (2.37)$$

Method of MPS provides optimum estimates for the value of θ by maximizing the product of probabilities between two adjacent sample points.

The Z_i's are as near as possible to each other. We choose the value of the parameter θ as an estimate that maximizes the log of geometric mean (GM) of $Z_i(\theta)$.

$$K(\theta) = log(\prod_{i=1}^{n+1} Z_i(\theta))^{1/(n+1)} = \frac{1}{n+1}\sum_{i=1}^{n+1} log(Z_i(\theta)) \quad (2.38)$$

The optimum MPS log estimator is given as.

$$K_{opt}(\theta) = log\frac{1}{n+1} \quad (2.39)$$

The relationship given in Eq. (2.23) shows that MPS has advantageous results relative to MLE because log-likelihood can reach to $+\infty$, whereas, MPS estimator is always bounded by $log\frac{1}{n+1}$. It can easily be provided that the maximum and optimum value of GM is only obtained when all Z_i's are equal to $Z_i = i/(n+1)$. For more details, see Ranneby (1984).

Chapter 3

Preprocessing of Data Series

3.1 General

The preprocessing of flood data is very important for the successful application of frequency analysis. Because quality of the results frequency analysis strongly depends on the quality of the available data series. In this study, time series secondary data of AMPF have been used. Time series data usually suffer from missing observations and outliers. Before performing the final analysis filling the missing observations and handling of outliers is necessary to obtain reliable estimates. Moreover, RFA strictly based on the assumptions that the sample data must be random, independent, homogeneous and free from significant trends. Therefore, in this chapter data of AMPF of 36 gauging sites of KPK has been analyzed through non-parametric tests to check the suitability of the data to perform RFA.

Rest of this chapter organized as: section 3.2 describes study area, section 3.3 deals with the availability of the data and missing observations, section 3.4 based on graphical time series analysis, in section 3.5 outlier analysis is performed and in section 3.6 basic assumptions of data series has been tested.

3.2 Study area

The KPK region, having various small rivers and stream, is the second source of river water in Pakistan. KPK has a 101,741 km² area with steep geography and a population of about 35.53 million (as per the population census of 2017 by the Government of Pakistan). The terrain of KPK consists of mountainous in the north, sub mountainous and lands surrounded by hills to the south. Due to the steep geography and mountain

land of KPK, the heavy rainfall usually turns into flash flooding affecting the whole of KPK (Pakistan Meteorological Department, 2012). Southern KPK is the most populated area of the province, and due to its downstream location, it has been affected badly due to heavy floods in 1992 and 2010 (Hashmi et al., 2012). Therefore, preventive measures against these natural disasters are a popular demand of the people of KPK which requires quantification of the frequency associated with these floods. Moreover, KPK has an identified potential of hydroelectricity is about 18698 Mega Watt, as reported in a study on Hydel Potential in Pakistan by National Electric Power Regulatory Authority (NEPRA), Pakistan (NEPRA, 2018) and search for more sites is still on going. None of the published studies so far has used L-moment based RFFA for flood quantiles estimation at various sites of KPK. The current study has performed RFFA using AMPF in cusecs of 36 gauging sites situated in KPK and at-site analysis of some selected sites having variation in sample size and shape characteristics using PE3 distribution. Geographical locations of the 36 gauging sites of KPK have been given in **Fig. (3.1).**

3.3 Availability of data and missing observations

This study has used "annual maximum peaks flow" (AMPF) of 36 gauging sites of important rivers/streams of KPK. The length of recorded data sets on all 36 sites varies between 15 to 55 years. The data sets have been gotten from the flood section of the Irrigation Department of KPK. Few details of the sites along with their respective characteristics namely longitude (Long), latitude (Lat), elevation (Ele) in meters, average annual rainfall (AARF) in millimeter, average rainfall in monsoon season (ARMS) in millimeter and average annual temperature (AAT) in degree Celsius has been given in **Table 3.1.**

Few missing observations has been found in the data series of some gauging sites. If these shortfalls of data series not filled properly can produced inaccurate flood frequency estimates. Missing data filling techniques generally involves deletion of gaps or imputation of single value arithmetic mean or median (Peugh and Enders 2004). Filling out missing observations by using a single value (arithmetic mean or median) never effect the original size of data sample (Ekeu-wei et al., 2018). Therefore, in this study, missing values are estimated and filled with the arithmetic means of AMPF of each respective site where the missing observations found. Similar treatment with missing observation of AMPF was performed by Hussain (2011).

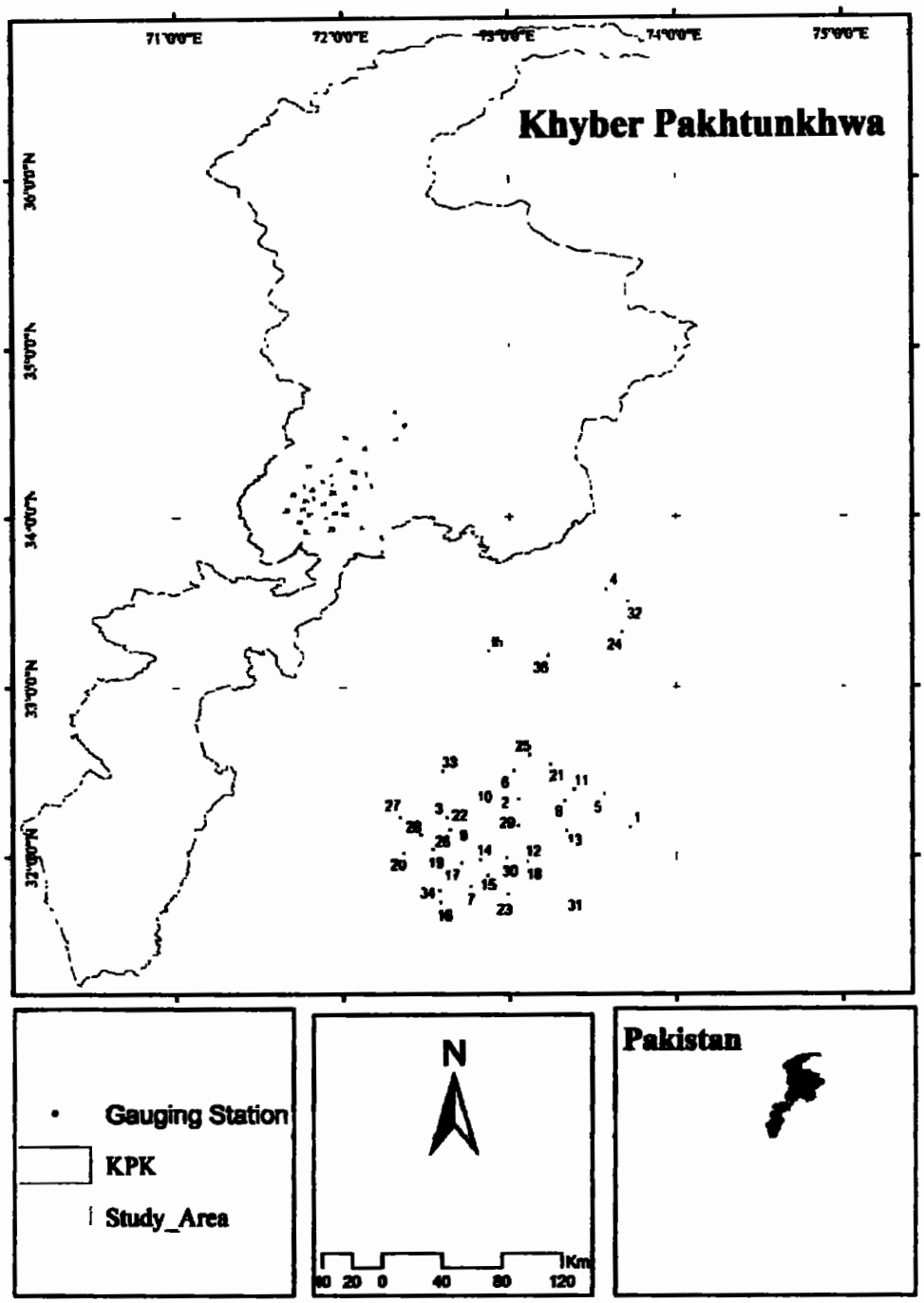


Fig. 3.1: Geographical locations of 36 gauging stations of Khyber Pakhtunkhwa.

Table 3.1: Site characteristics of 36 gauging sites of the study area.

S. No.	Site name	Lat (N)	Long (E)	Ele (m)	AARF (mm)	ARMS (mm)	AAT (c)
1	Budni	34.1307	72.4648	334	639	272	22.7
2	Shahi Bala	34.1858	71.7661	300	460	151	22.7
3	Dallus	34.165	71.5931	310	460	151	22.7
4	Badri	34.9866	72.352	1243	639	272	22.2
5	Naranji	34.2475	72.3432	356	639	272	22.2
6	Kalpani Raisalpur	34.3303	71.9085	345	556	222	22.2
7	Kalpani Deheri	33.9928	71.746	303	559	255	22.2
8	Bagiari	34.2254	72.1543	313	559	227	22.2
9	Katlongi	34.096	71.7416	389	460	151	22.5
10	Chprial	34.1998	71.7584	306	478	212	19.9
11	Jani Khwar	34.2653	72.1963	330	384	105	22.7
12	Shahban	34.0918	72.0388	288	559	227	22.2
13	Muqam	34.1078	72.0505	291	559	227	22.2
14	Chinkar	34.014	71.7538	301	400	119	22.7
15	Wazir Gahri	33.993	71.746	303	400	119	22.7
16	Bara Kohat Road	33.8637	71.5637	413	400	119	22.7
17	Bara Tarnab	34.0165	71.7035	305	400	119	22.7
18	Lund Khwar East	34.0064	71.9777	285	559	255	22.2
19	Kalpani Saidabad	34.0512	71.528	314	559	255	22.2
20	Dagi	34.0865	71.4749	328	384	105	22.7
21	Garandi	34.3571	72.0845	384	532	212	22.4
22	Hakim Gahri	34.1432	71.7053	296	460	151	22.5
23	Khuderzai	34.0116	71.7741	300	532	212	22.4
24	Kabul Nowshera	34.8337	72.4253	985	532	212	22.4
25	Chilah	34.3918	71.9862	375	532	212	22.4
26	Kabul Adezai	34.122	71.6078	305	532	212	22.4
27	Shah Alam	34.1664	71.3689	397	384	105	22.7
28	Panjkora	34.1019	71.4672	328	460	151	22.5
29	Kabul Naguman	34.114	71.7523	292	384	105	22.7
30	Jundi Utmanzai	34.0099	71.8327	294	460	170	22.5
31	Jundi Tangi	33.8965	72.235	266	460	170	22.5
32	Jundi River	34.9422	72.4528	1099	460	151	22.5
33	Swat Khaili	34.3307	71.5706	365	460	151	22.5
34	Swat Ningolai	33.9042	71.5583	379	743	221	19.9
35	Swat khawazakhela	34.7677	71.7924	665	743	221	19.9
36	Swat Munda Head	34.7507	72.0767	923	743	221	19.9

3.4 Time series plots of data sets

Climate change affects the frequency and size of floods in different parts of the world (Maghsood, et al., 2019). Pakistan is also experiencing the impact of climate change, ranking 10th in the long-term climate change index (Kreft, et al., 2015). Flooding has become more common in Pakistan over the last decade, and the country has experienced flooding almost every year since 2010 (Government of Pakistan 2018). As a result, graphical time series analysis has been used to see if any significant trend or pattern exists in the AMPF data sample of each gauging site under investigation. The results of this graphical time series analysis are also used in the RFFA's subsequent measures. Time series plots of AMPF of all 36 gauging stations are presented in **Fig. (3.2)**.

Fig. (3.2) shows that AMPF is randomly distributed along their average line of gauging sites Shahi Bala, Dallus, Badri, Kalpani Deheri, Bagiari, Katlongi, Chaprial, Jani Khwar, Shahban, and Muaqm Shah Alam, with no discernible upward or downward trend. For the period 1994 to 1998, only at site Kalpani Deheri AMPF values highly scattered along the mean line. Some plots of gauging sites in **Fig. (3.2)** indicate that AMPF values have a larger magnitude and deviate significantly from their average line. The details of the time series plots of some sites given in **Fig. (3.2)** are provided in **Table (3.2). Table (3.2)** gives the names of the gauging sites as well as the occurrence time of the value/s that deviate significantly from their average line. The AMPF are scattered along with the average line of each site listed in **Table (3.2)** at random. Furthermore, an upward trend was observed in the AMPF sample data of site Naranji from 1980 to 2010, and at site Kalpani Raisalpur from 2003 to 2010.

Table 3.2: Names of gauging sites, as well as the years in which they deviated significantly from their mean values.

Site Name	Time in year	Site Name	Time in year
Budni	2008	Bara Tarnab	2013
Chinkar	2010	Lund Khwar East	1997
Wazir Ghari	1982, 1998	Dagi	2009
Bara Kohat Road	2010	Garandi	2003
Hakim Ghari	1983	Khuderzai	1984
Kabul Nowshera	2010	Chilah	1979
Panjkora	2010	Kabul Naguman	2010
Swat Khaili	2010	Swat Khawazakhela	2010
Swat Munda Head	2010		

After 2005, the AMPF data sample of sites Kalpani Saidabad, Kabul Adezai, Jundi Utmanzai, Jundi Tangi, and Jundi River shows a rising pattern, as does the data sample of site Swat Ningolai after 2010. The reason for the upward and downward trend in the data set is may be due to the irregular pattern of monsoon season rainfall in the region.

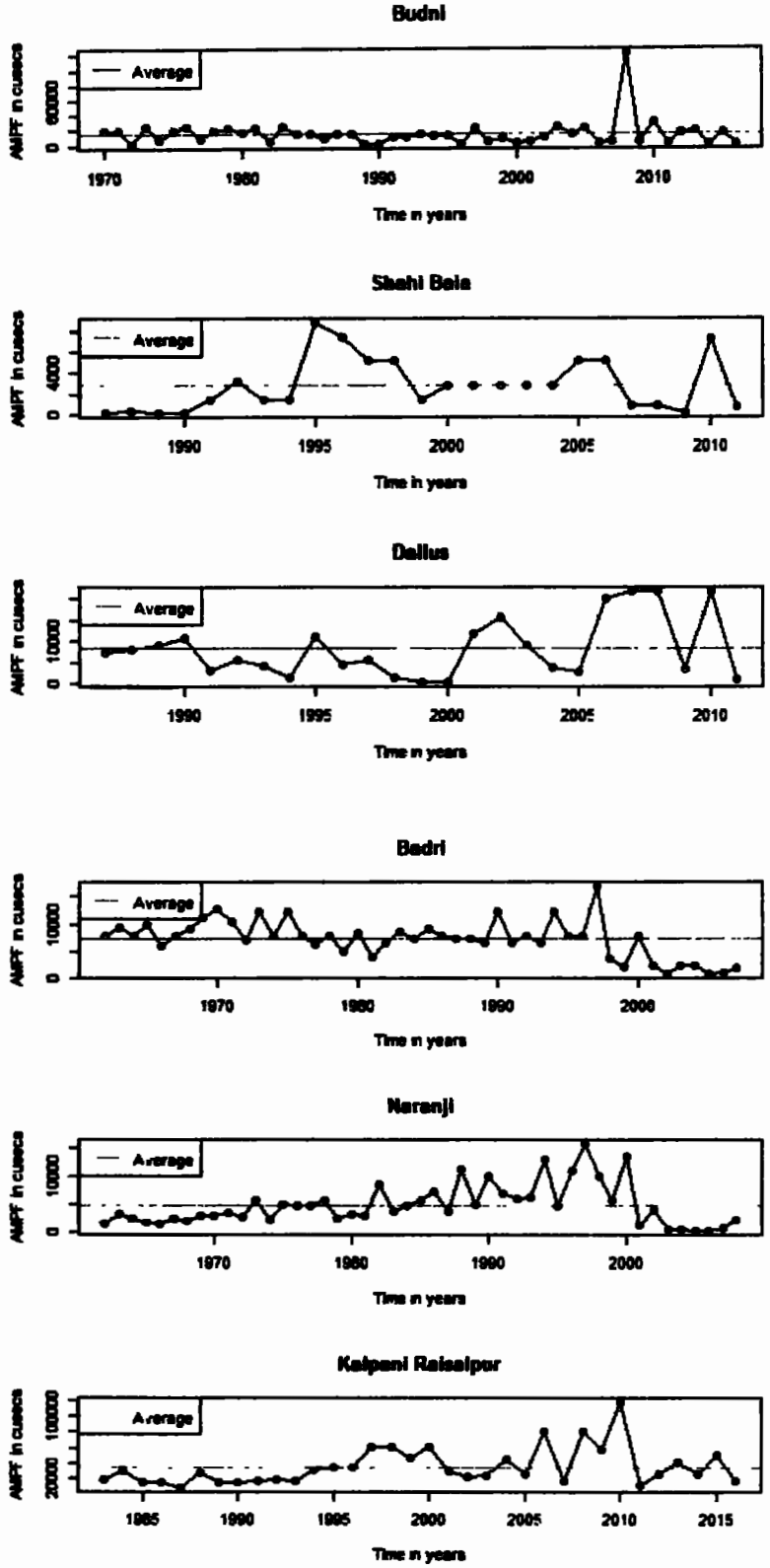
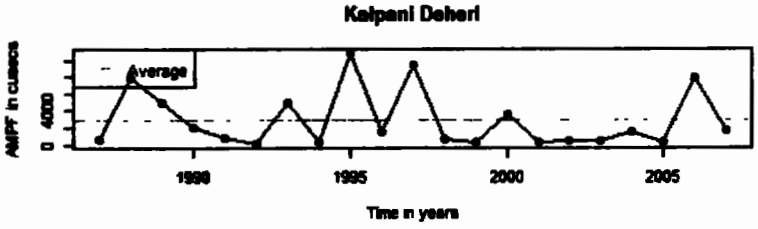
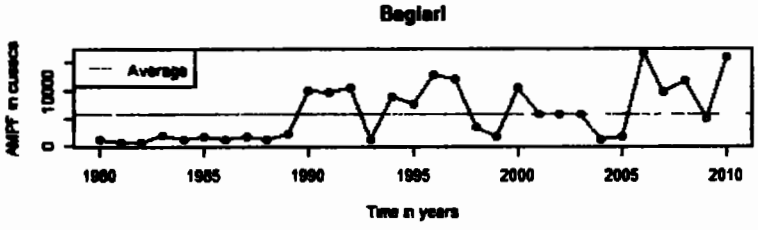
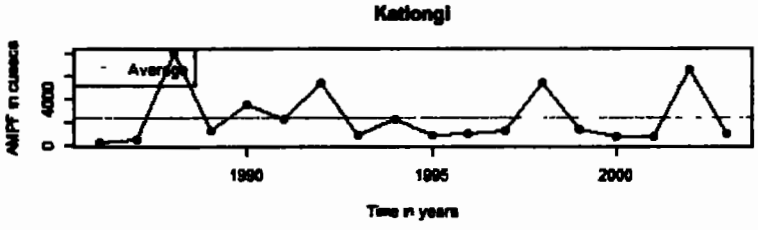
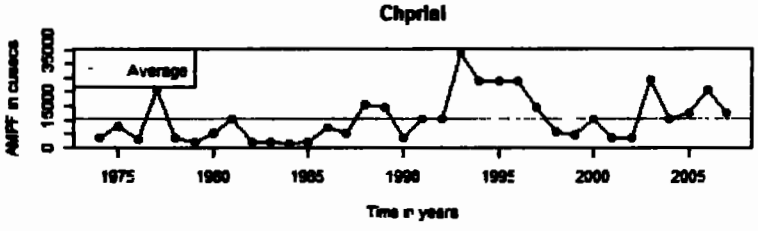


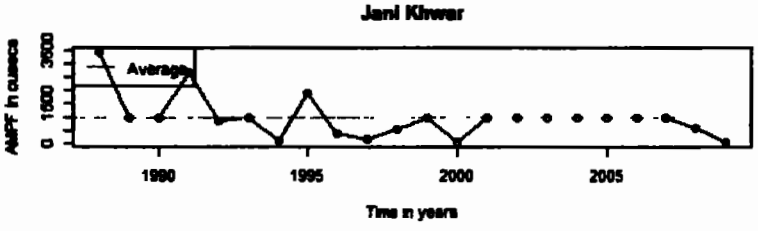
Fig. 3.2: Time series plots of AMPF of 36 gauging sites. *Continued... ...*

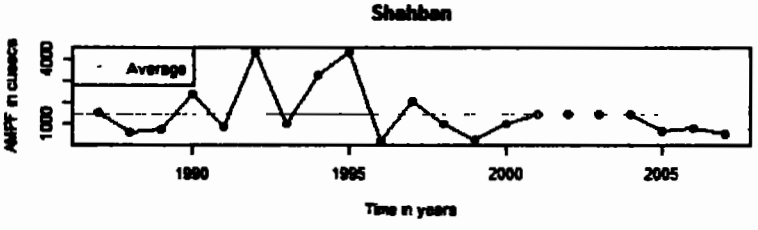

Kalpani Deheri
Average
AMPF in cusecs
0
4000
1990
1995
2000
2005
Time in years


Bagiari
Average
AMPF in cusecs
0
10000
1980
1985
1990
1995
2000
2005
2010
Time in years

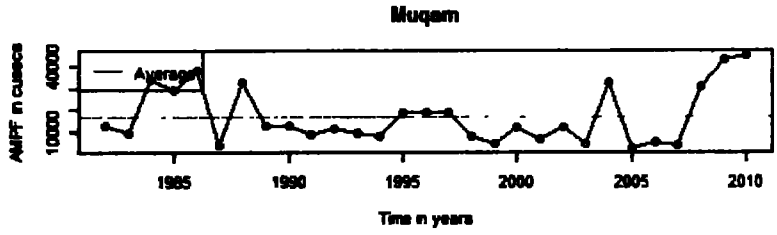

Katlongi
Average
AMPF in cusecs
0
4000
1990
1995
2000
Time in years

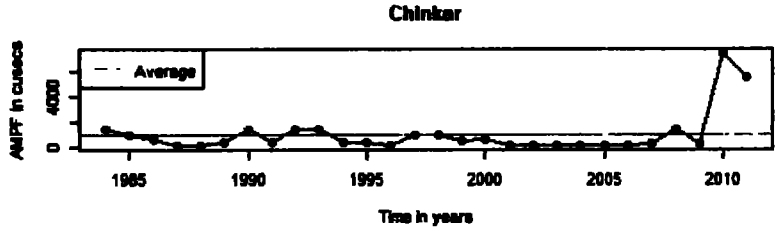

Chpriai
Average
AMPF in cusecs
0
15000
35000
1975
1980
1985
1990
1995
2000
2005
Time in years

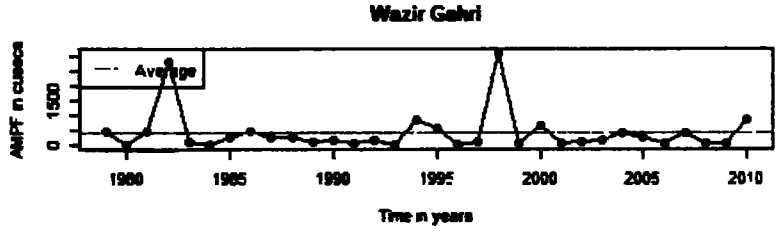

Jani Khwar
Average
AMPF in cusecs
0
1500
3000
1990
1995
2000
2005
Time in years

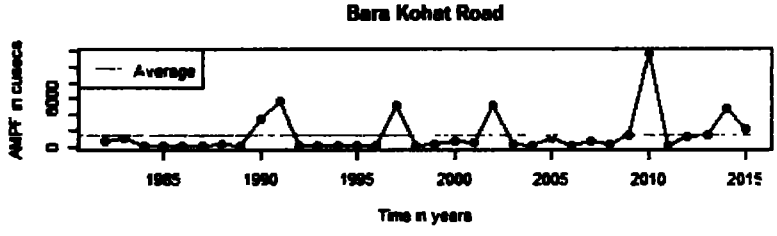

Shahban
Average
AMPF in cusecs
1000
4000
1990
1995
2000
2005
Time in years

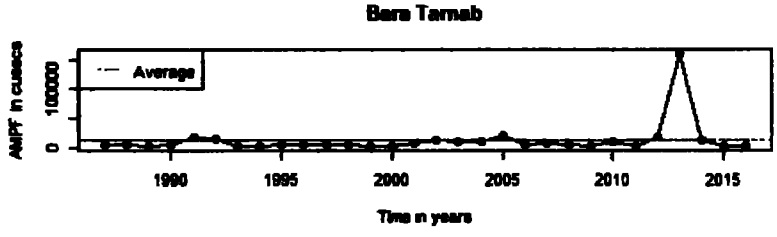
Continued

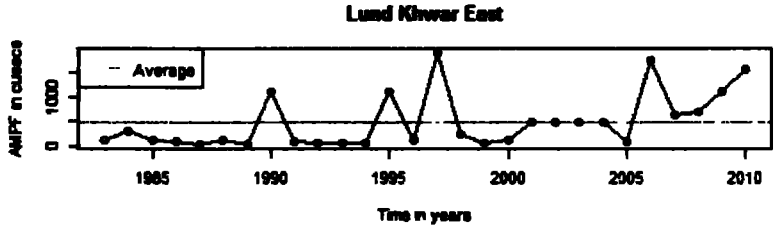

Muqam
AMPF in cusecs
10000
40000
Average
1985
1990
1995
2000
2005
2010
Time in years


Chinkar
AMPF in cusecs
0
4000
Average
1985
1990
1995
2000
2005
2010
Time in years

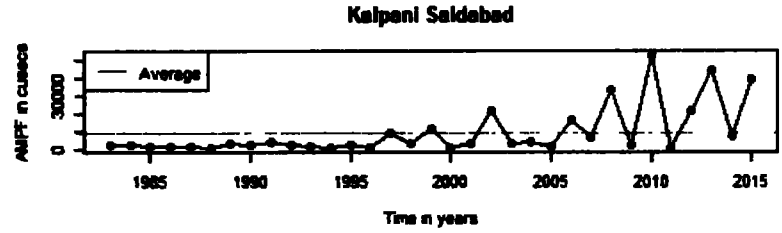

Wazir Gahri
AMPF in cusecs
0
1500
Average
1980
1985
1990
1995
2000
2005
2010
Time in years

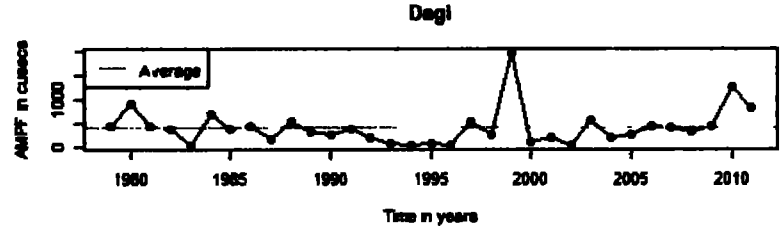

Bara Kohat Road
AMPF in cusecs
0
6000
Average
1985
1990
1995
2000
2005
2010
2015
Time in years

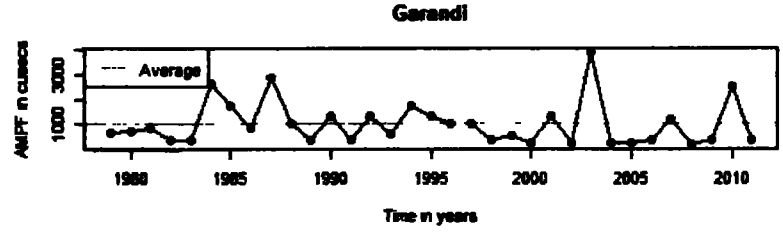

Bara Tarnab
AMPF in cusecs
0
100000
Average
1990
1995
2000
2005
2010
2015
Time in years

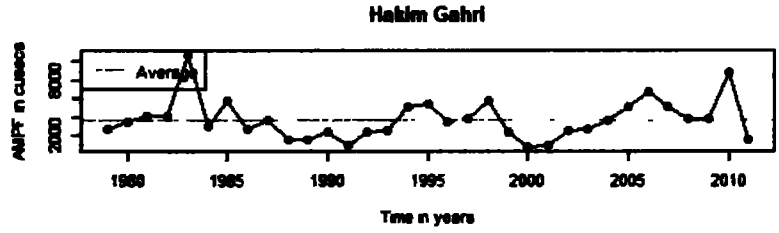

Lund Khwar East
AMPF in cusecs
0
1000
Average
1985
1990
1995
2000
2005
2010
Time in years

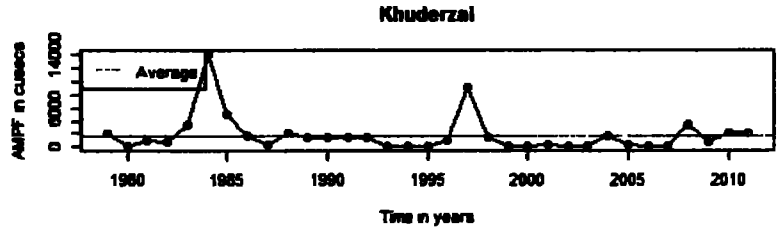
Continued

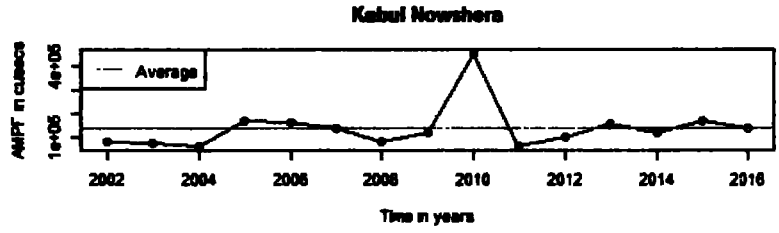

Kalpani Saidabad
Average
AMPF in cusecs
0
30000
1985
1990
1995
2000
2005
2010
2015
Time in years


Dagi
Average
AMPF in cusecs
0
1000
1980
1985
1990
1995
2000
2005
2010
Time in years

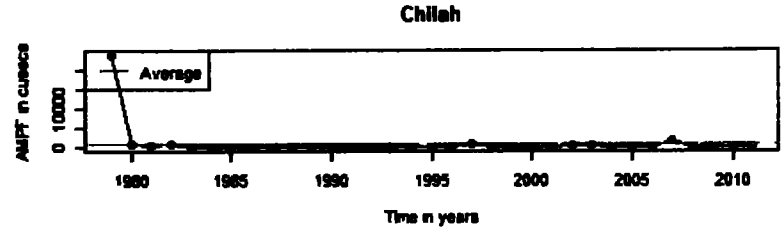

Garandi
Average
AMPF in cusecs
1000
3000
1980
1985
1990
1995
2000
2005
2010
Time in years

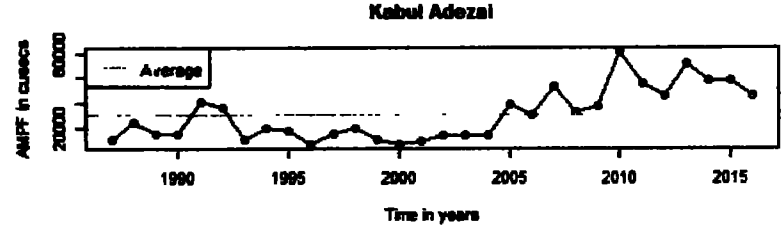

Hakim Gahri
Average
AMPF in cusecs
2000
8000
1980
1985
1990
1995
2000
2005
2010
Time in years

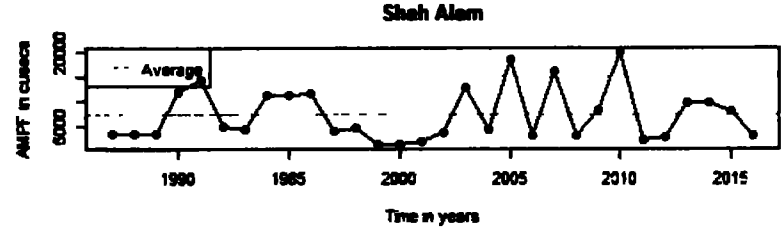

Khuderzai
Average
AMPF in cusecs
0
6000
14000
1980
1985
1990
1995
2000
2005
2010
Time in years

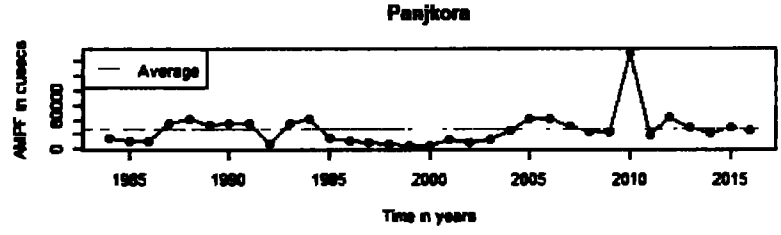

Kabul Nowshera
Average
AMPF in cusecs
1e+05
4e+05
2002
2004
2006
2008
2010
2012
2014
2016
Time in years

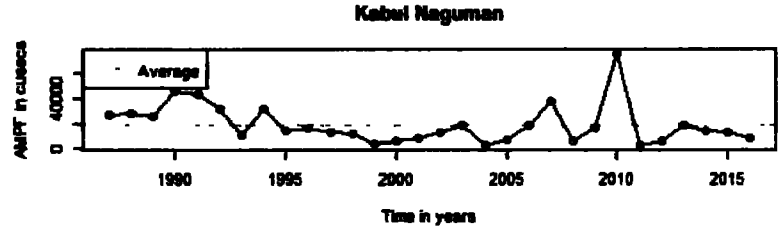
Continued

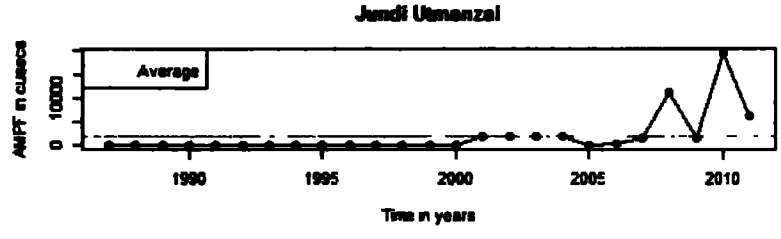

Chilah
Average
AMPF in cusecs
0
10000
1980
1985
1990
1995
2000
2005
2010
Time in years


Kabul Adezai
Average
AMPF in cusecs
20000
60000
1990
1995
2000
2005
2010
2015
Time in years


Shah Alam
Average
AMPF in cusecs
6000
20000
1990
1995
2000
2005
2010
2015
Time in years


Panjkora
Average
AMPF in cusecs
0
60000
1985
1990
1995
2000
2005
2010
2015
Time in years


Kabul Naguman
Average
AMPF in cusecs
0
40000
1990
1995
2000
2005
2010
2015
Time in years


Jandi Utmanzai
Average
AMPF in cusecs
0
10000
1990
1995
2000
2005
2010
Time in years

Continued

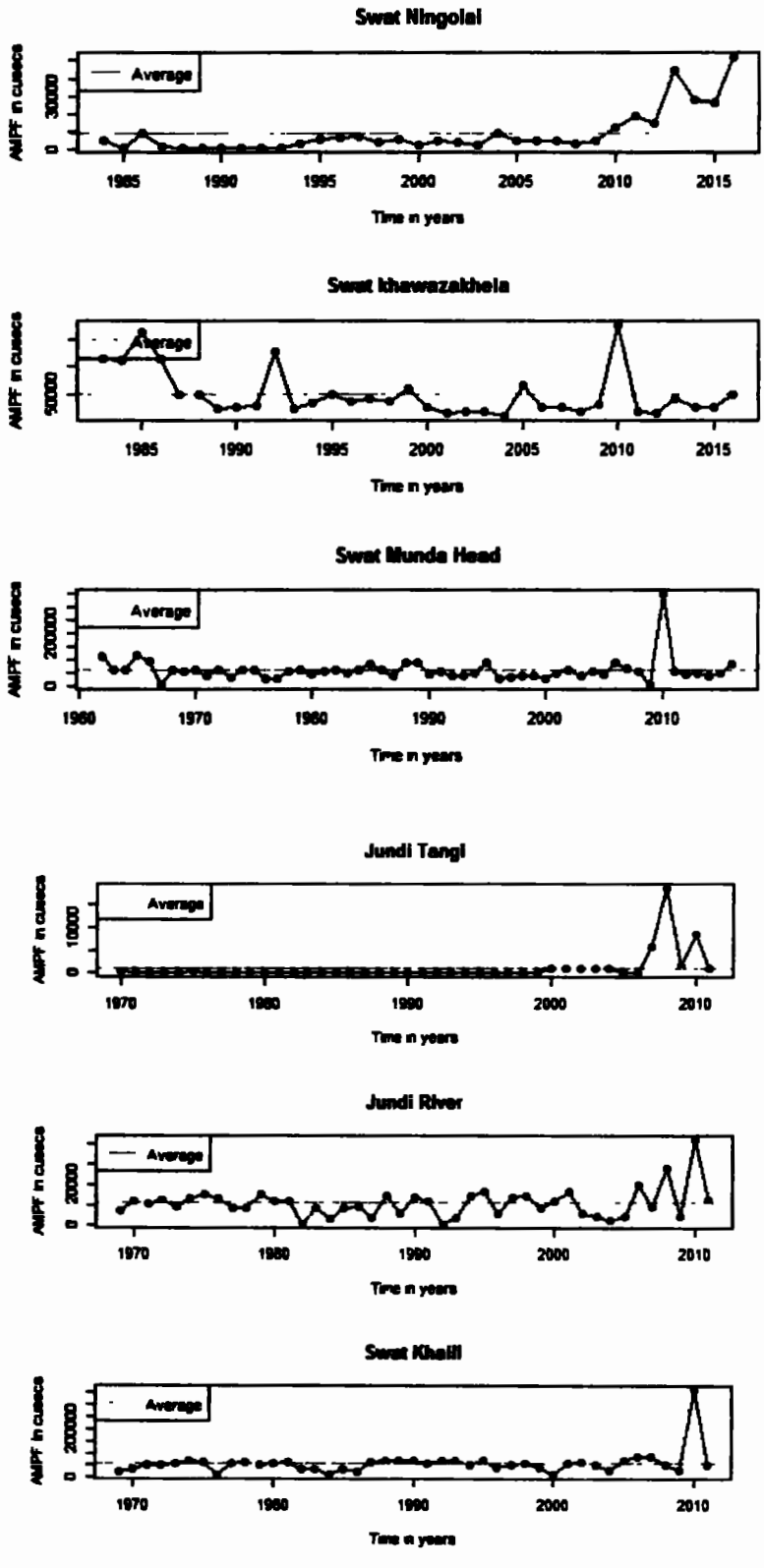


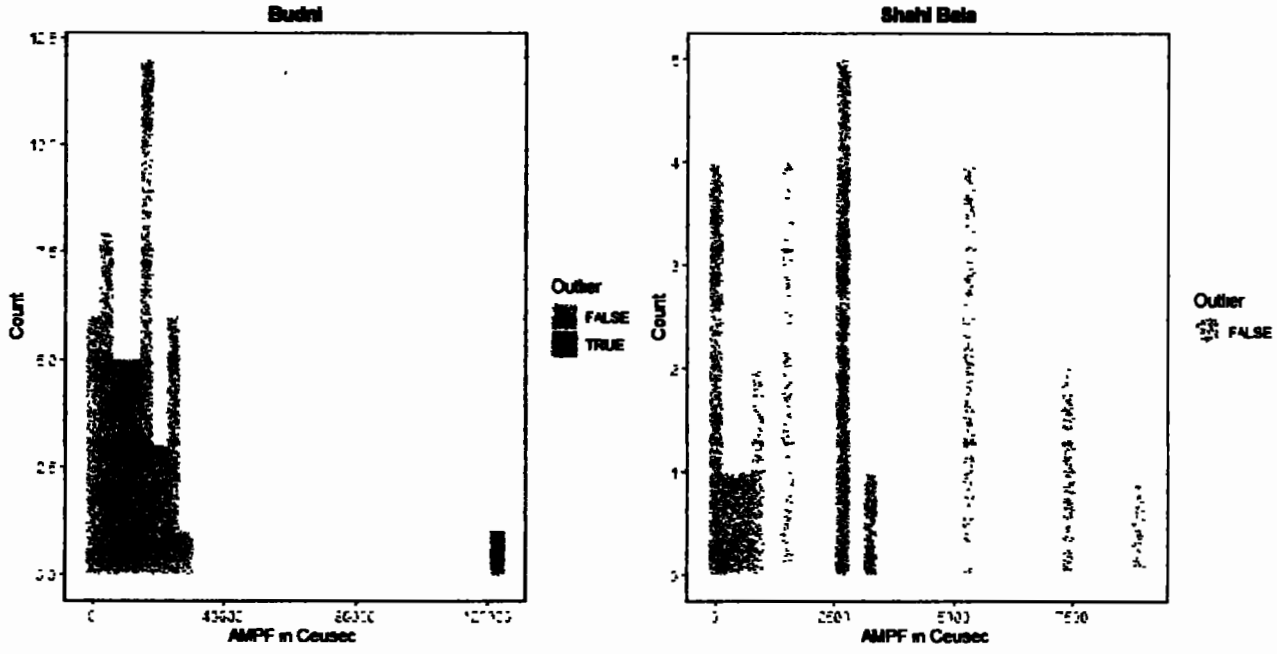
Fig. 3.2: Time series plots of 36 gauging sites.

3.5 Detection of outliers

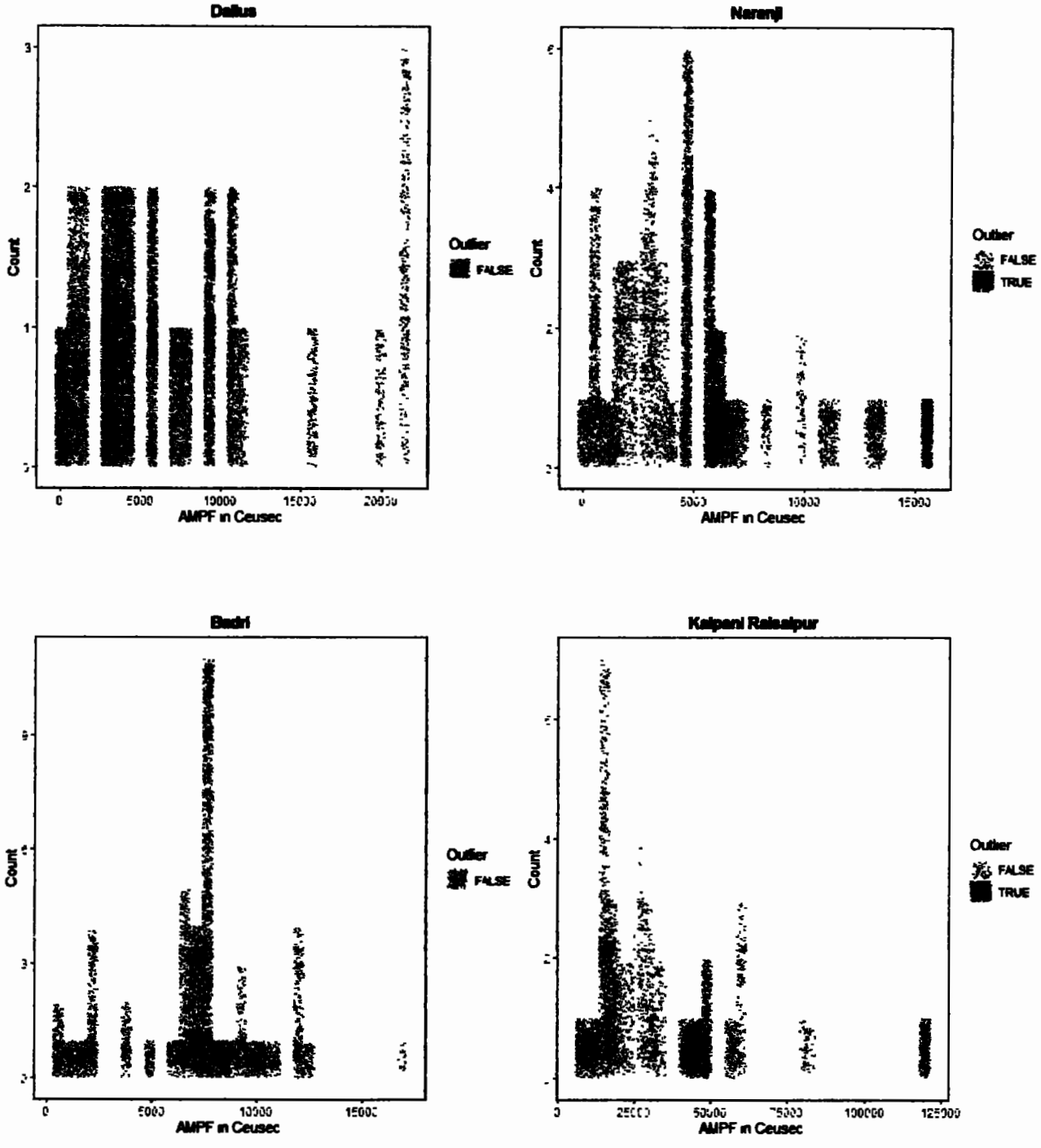
Outlier detection is a frequent subject for hydrological data that has gained much attention in univariate context (Chebana, et al., 2012). The identification of low and high outliers within the data set is evident in extreme hydrological modelling. And when a flood frequency curve is fitted to the annual maxima series, low outliers have a substantial impact on the performance. If there is any low outlier in the data set of the annual maxima series, this must be treated separately before the frequency analysis (Cohn et al., 2013; England et al., 2019). Outliers, in general, may have a negative impact on the selection of an accurate probability model and the parametric estimates associated with it. Therefore, in this study, outlier detection analysis has been performed on sample data of AMPF of 36 gauging sites to avoid the shortcoming of outlier on the results.

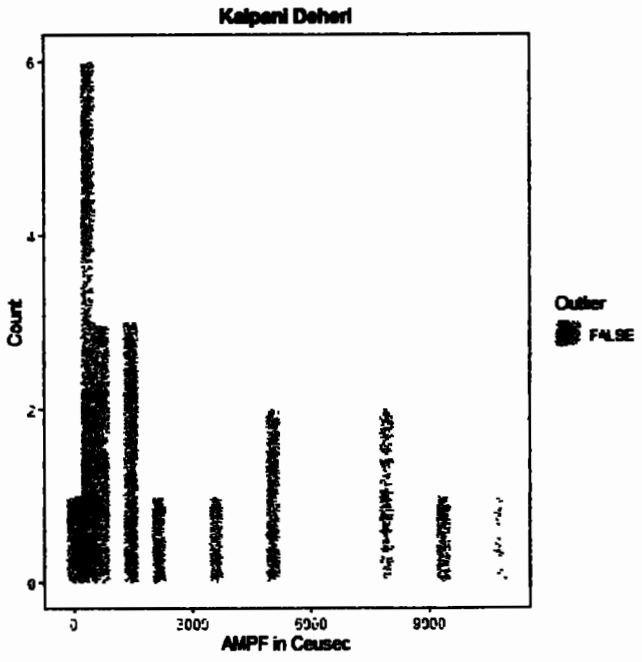
The Grubbs and Beck (GB) test has been used to identify outliers in the AMPF data set of each gauging station. Every value of the data set of each gauging site has been tested using the GB test. The decision has been taken whether the checked finding is an outlier or not based on the GB test results. The GB test results for each gauging station are given in **Fig. (3.3)**. **Fig. (3.3)** shows that although no low outliers have been identified in the data sets of 36 gauging sites, high outliers have been found in the data set of sites Budni, Naranji, Kalpani Raisalpur, Jani Khwar, Wazir Ghari, Chinkar, Bara Khoat Road, Bara Tarnab, Kalpani Saidabad, Dagi, Garandi, Khuderzai, Hakim Ghari, Kabul Nowshera, Chilah, Panjkora, Kabul Naguman, Jundi Tangi, Jundi Utmanzai, Jundi River, Swat Khaili, Swat Khawazakhela, Swat Ningolai and Swat Munda Head. Within the flood data set, there are three main explanations for outliers: 1) an erroneous calculation, 2) a change in the unit of measurement, and 3) a storm or other unusual hydro-meteorological occurrence (Hosking and Wallis 1997; Chebana et al., 2012;

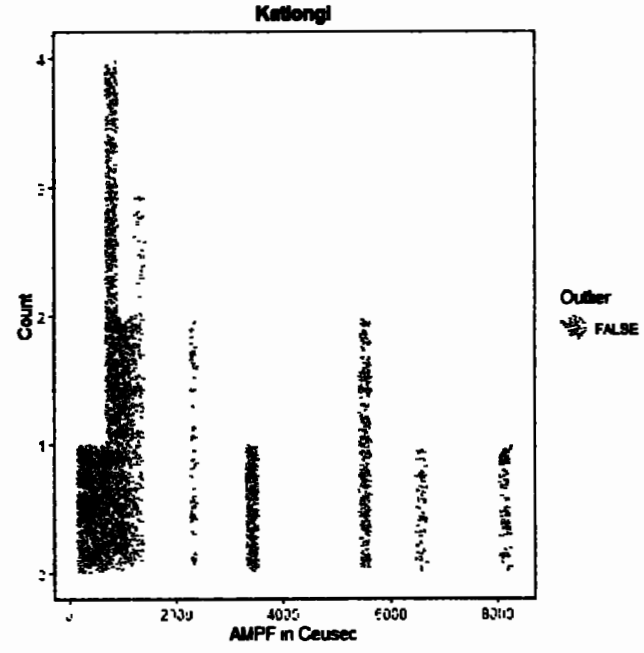
Naghettini, 2017). After diligent visualization, no irregularities found in the data sets of each site where outliers are observed. All of the study area's gauging sites are in an area that is known as more vulnerable to climate change. During the monsoon season, this area experiences erratic rainfall (June to August). Due to climate variability, such events (floods) of low and high magnitude can occur at any time and any place. High outliers provide extremely valuable information for the development of a flood risk map. Furthermore, where estimates of the extreme upper tail quantiles of the frequency distribution of flood are the primary interest, high outlier values must remain within the sample data (Naghettini, 2017; Hussain 2017; Khan et al., 2020). Because of the reasons mentioned above, the values of those high outliers remain within the data sets to perform RFFA.

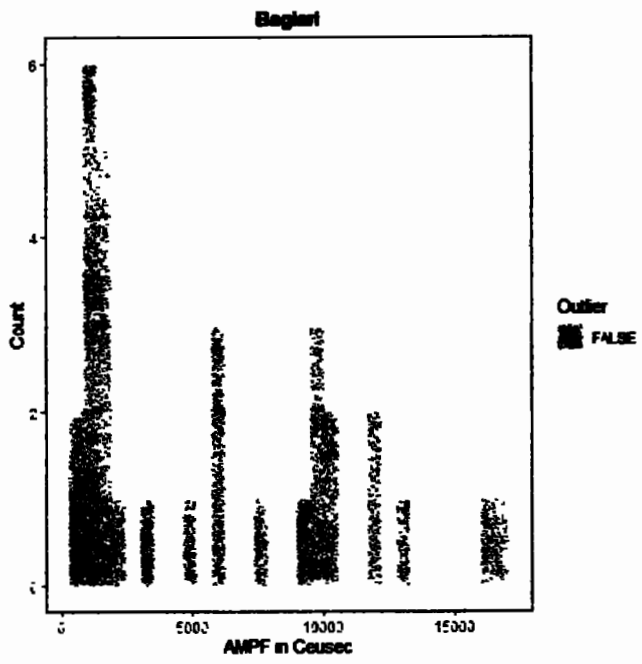


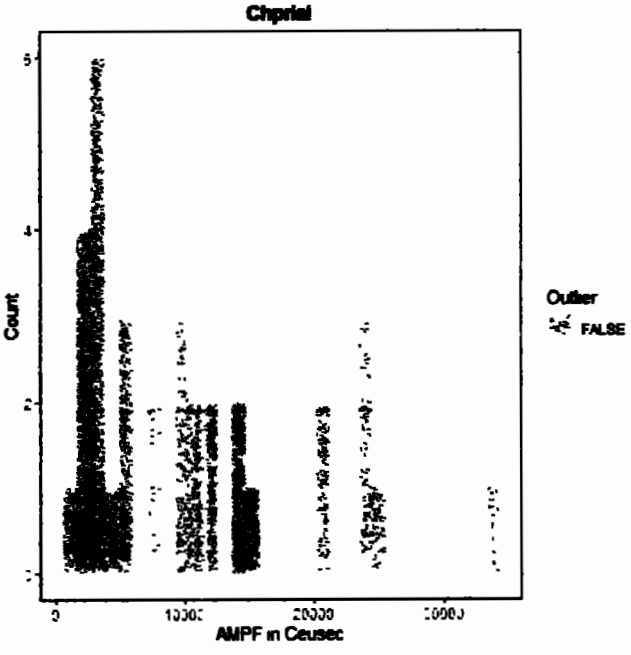
Continued...


Dallus
Count
AMPF in Ceusec
Outlier
FALSE
Naranji
Count
AMPF in Ceusec
Outlier
FALSE
TRUE
Badri
Count
AMPF in Ceusec
Outlier
FALSE
Kalpani Rajaipur
Count
AMPF in Ceusec
Outlier
FALSE
TRUE

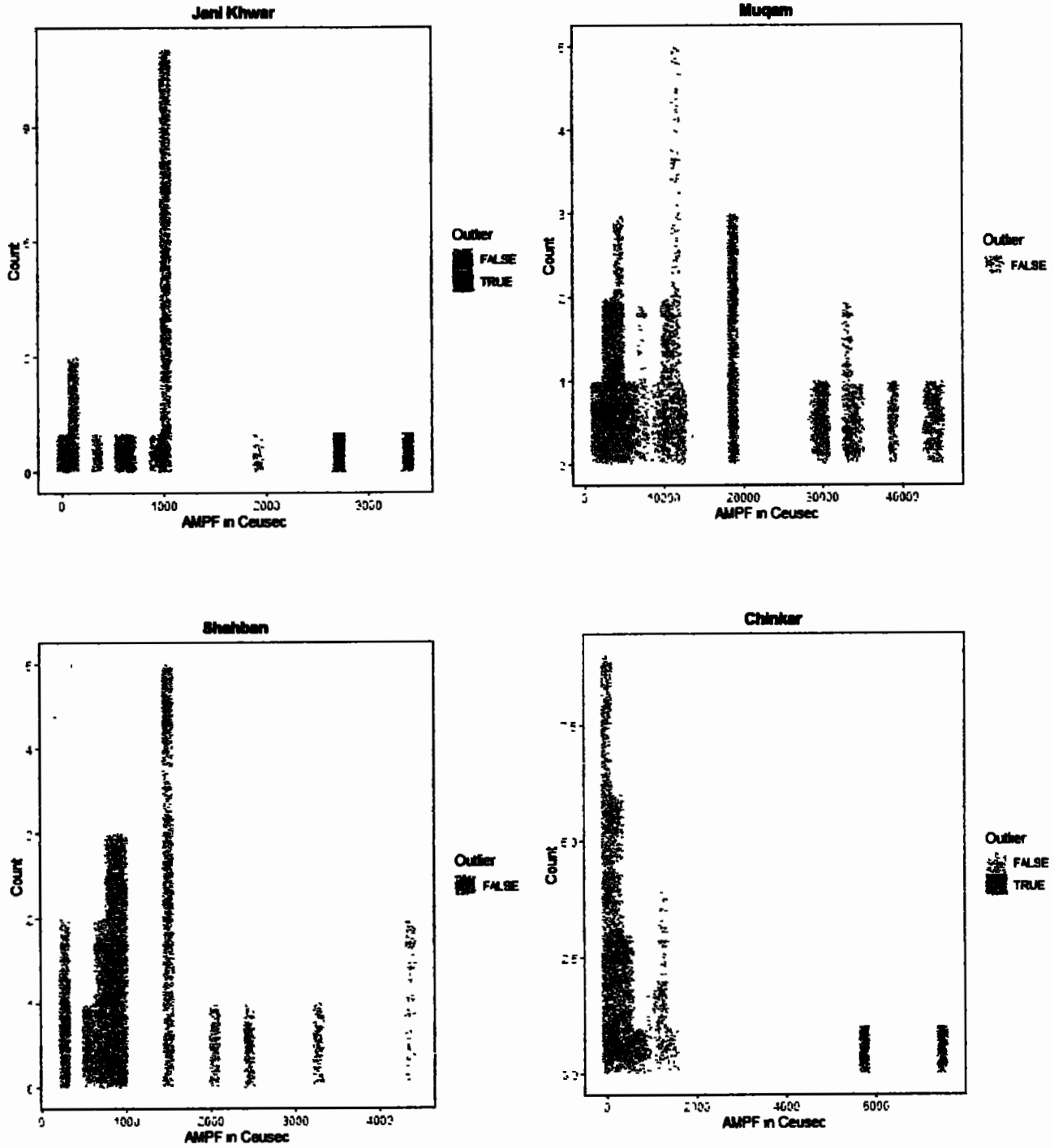

Kalpani Deheri
Count
AMPF in Ceusec
Outlier
FALSE


Katlongi
Count
AMPF in Ceusec
Outlier
FALSE

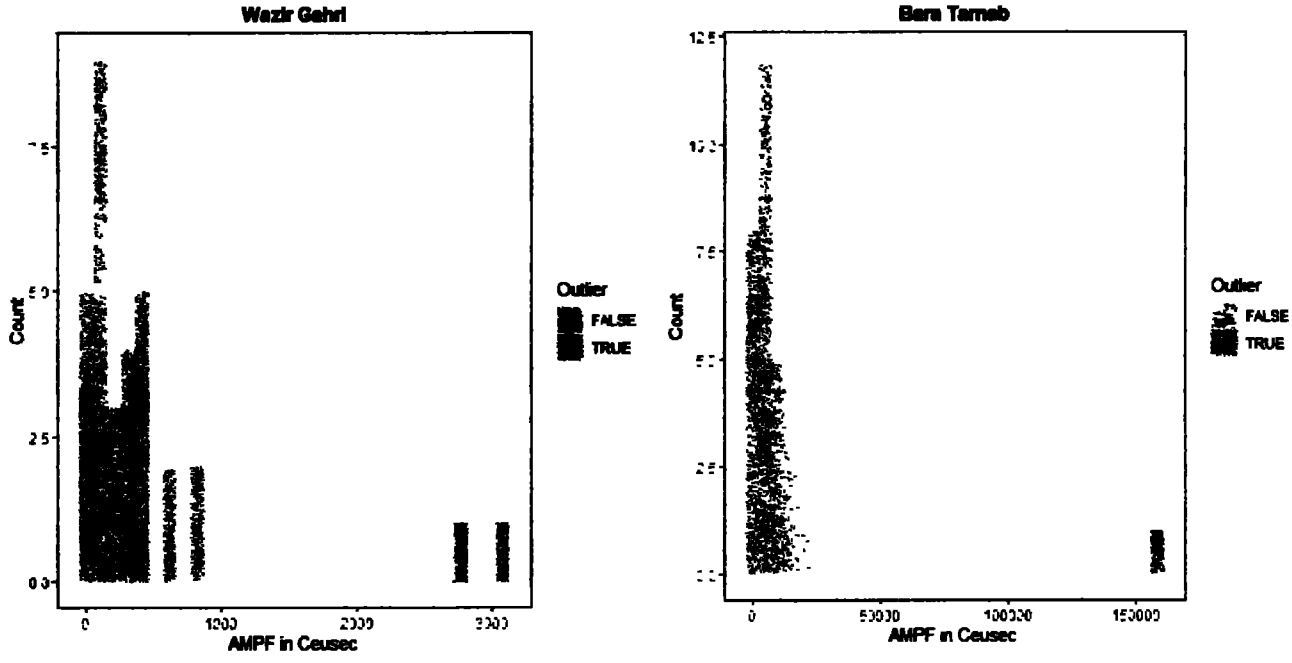

Bagiari
Count
AMPF in Ceusec
Outlier
FALSE

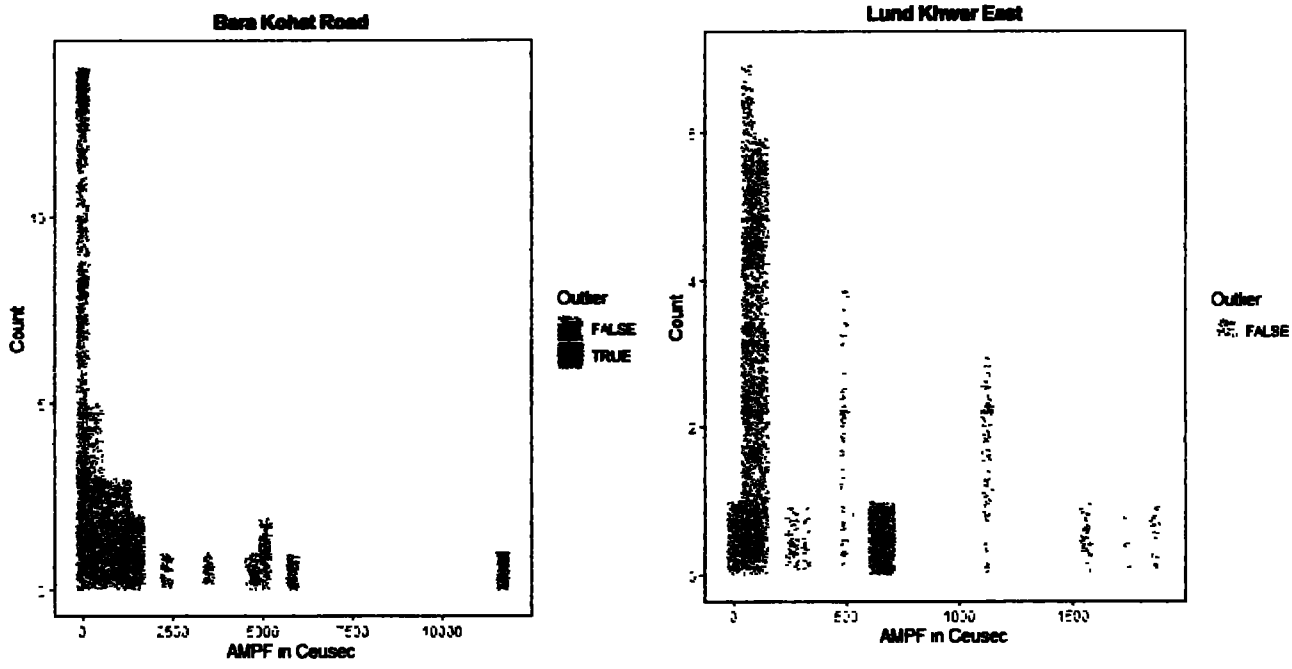

Chpriai
Count
AMPF in Ceusec
Outlier
FALSE

Continued

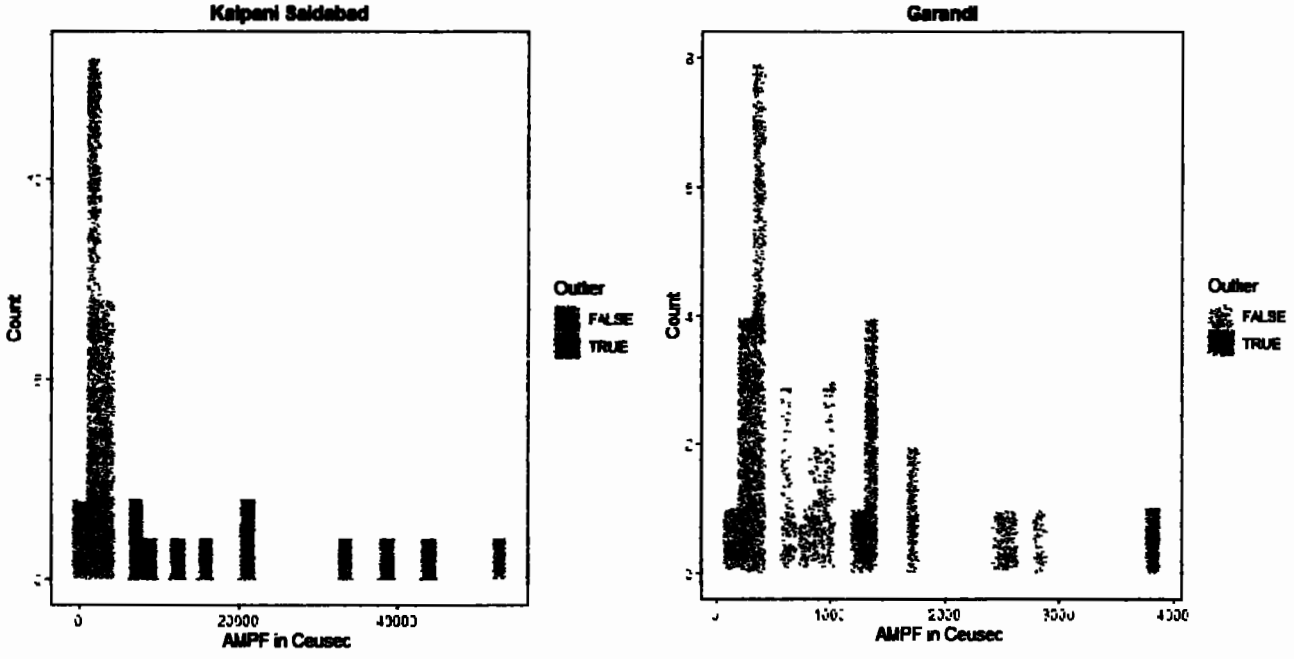

Jeni Khwar
Count
AMPF in Ceusec
Outlier
FALSE
TRUE
Muqam
Count
AMPF in Ceusec
Outlier
FALSE
Shahban
Count
AMPF in Ceusec
Outlier
FALSE
Chinkar
Count
AMPF in Ceusec
Outlier
FALSE
TRUE

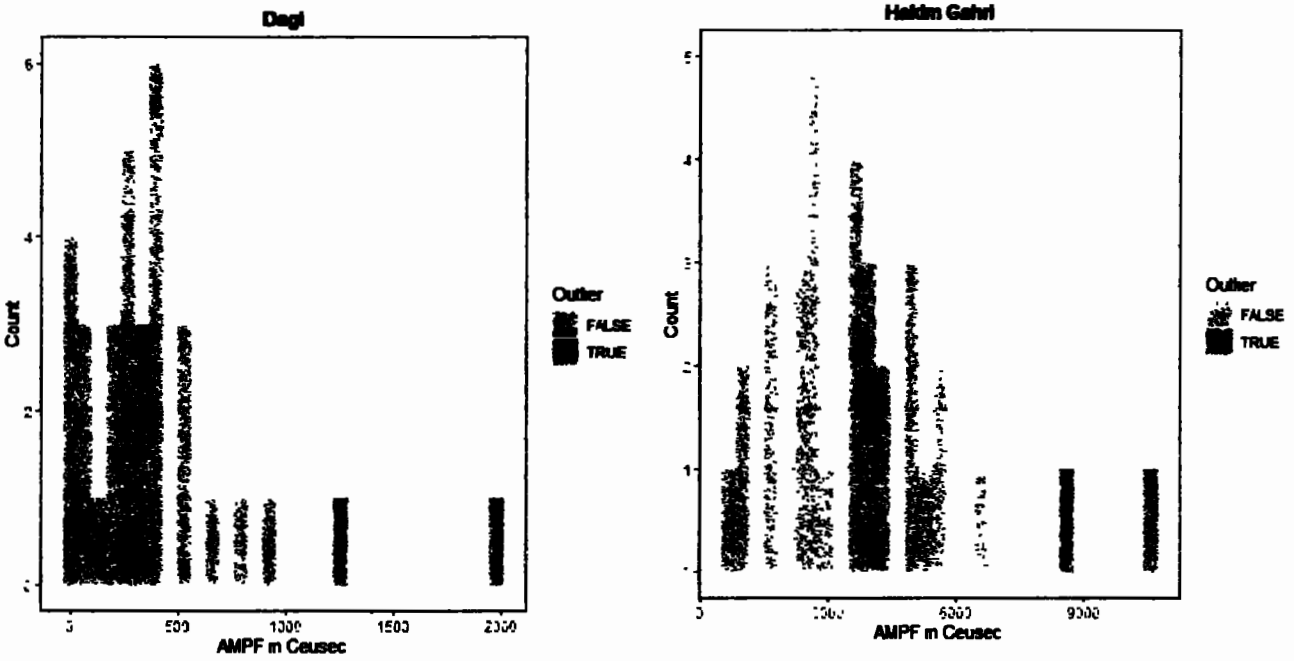
Continued


Wazir Gahri
Count
AMPF in Ceusec
Outlier
FALSE
TRUE
Bara Tarnab
Count
AMPF in Ceusec
Outlier
FALSE
TRUE

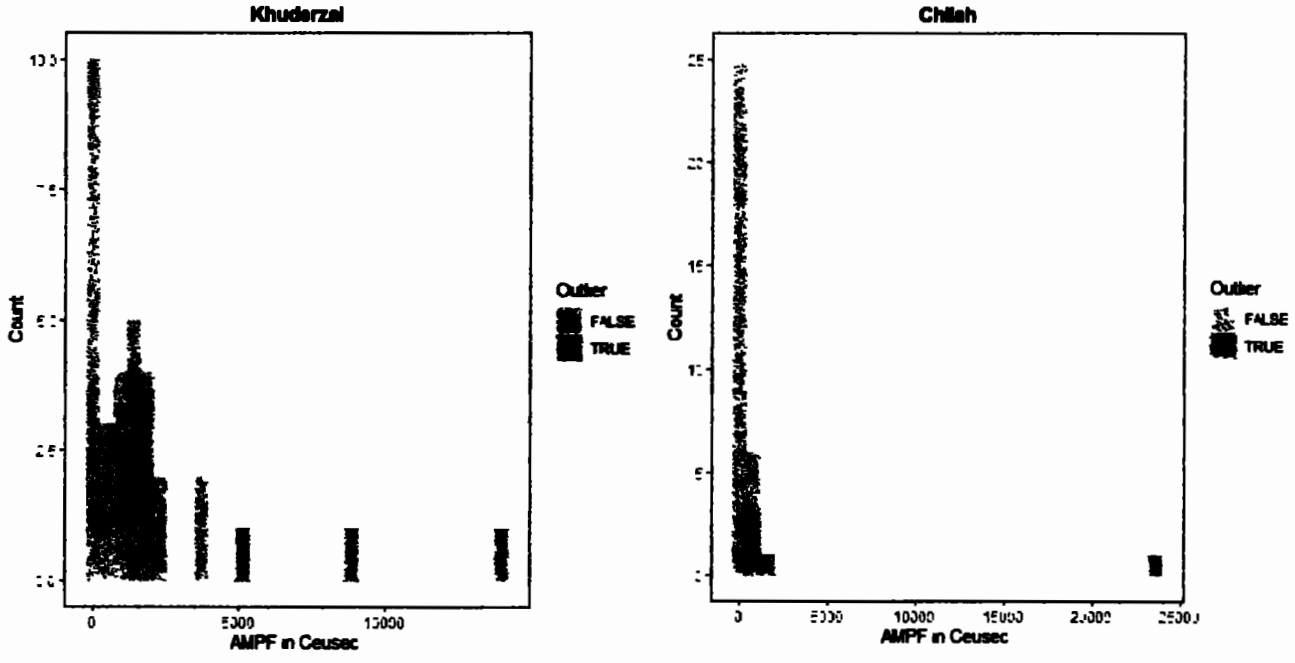

Bara Kohat Road
Count
AMPF in Ceusec
Outlier
FALSE
TRUE
Lund Khwar East
Count
AMPF in Ceusec
Outlier
FALSE

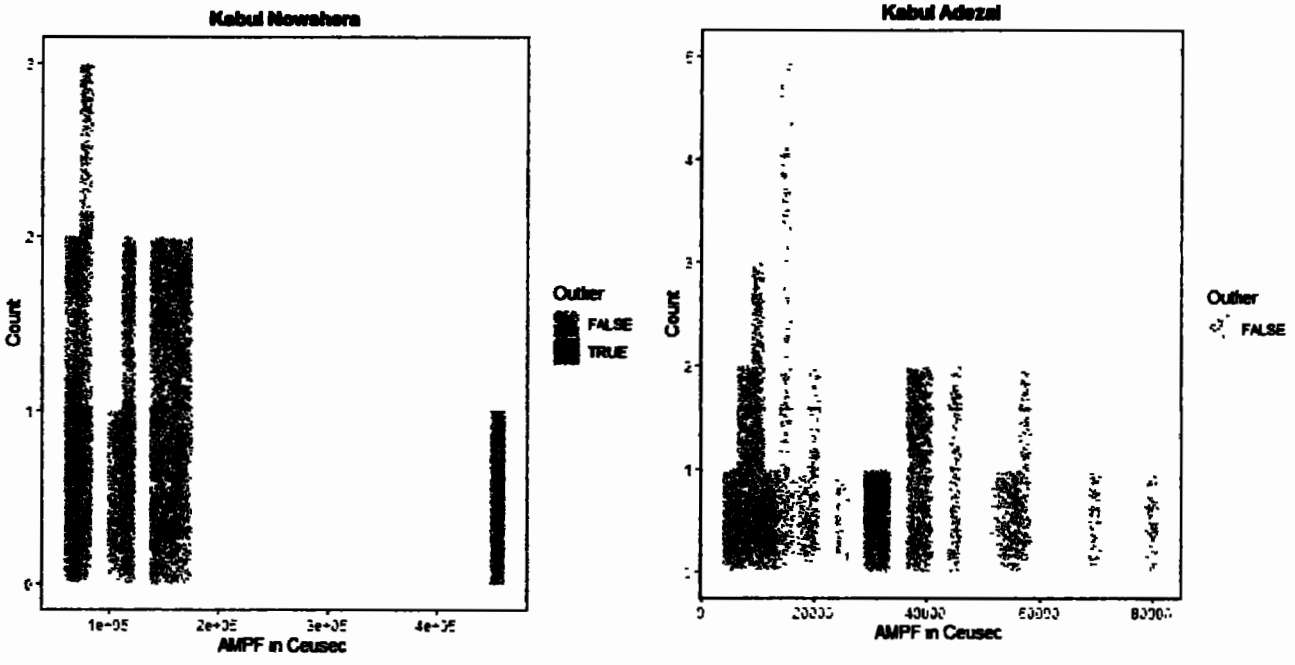
Continued


Kalpani Saidabad
Count
AMPF in Ceusec
Outlier
FALSE
TRUE
Garandi
Count
AMPF in Ceusec
Outlier
FALSE
TRUE

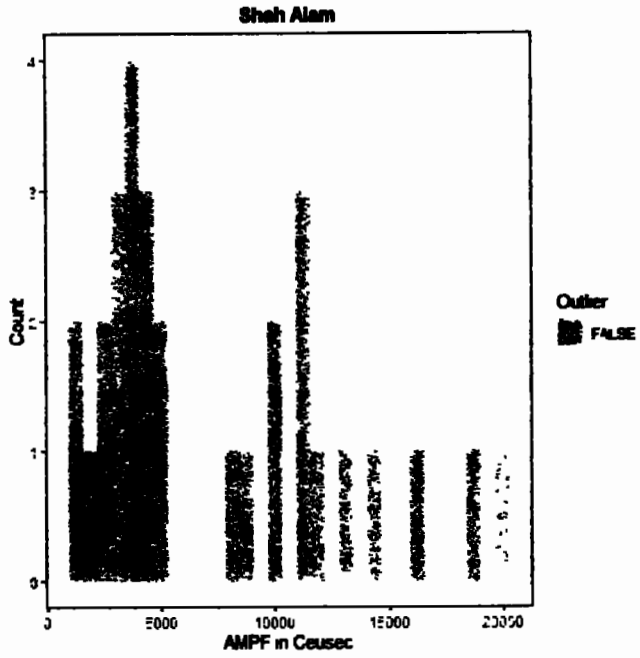

Dagi
Count
AMPF in Ceusec
Outlier
FALSE
TRUE
Haldm Gahri
Count
AMPF in Ceusec
Outlier
FALSE
TRUE

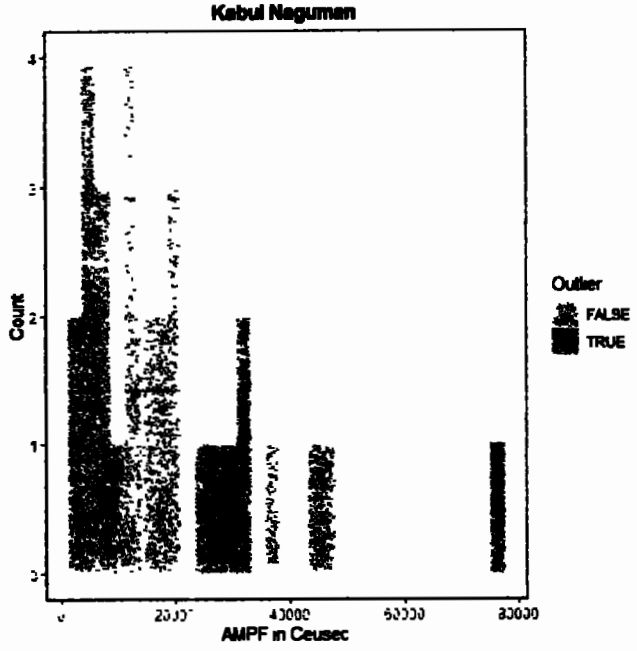
Continued

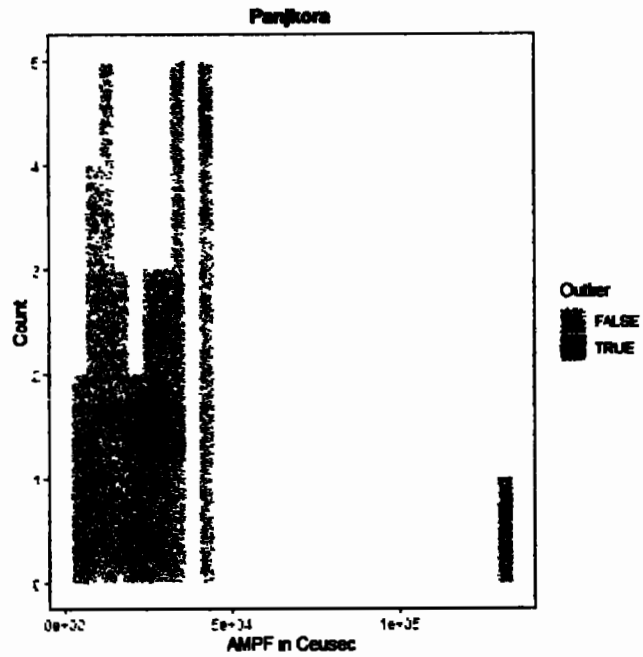

Khuderzai
Count
AMPF in Ceusec
Outlier
FALSE
TRUE
Chilah
Count
AMPF in Ceusec
Outlier
FALSE
TRUE

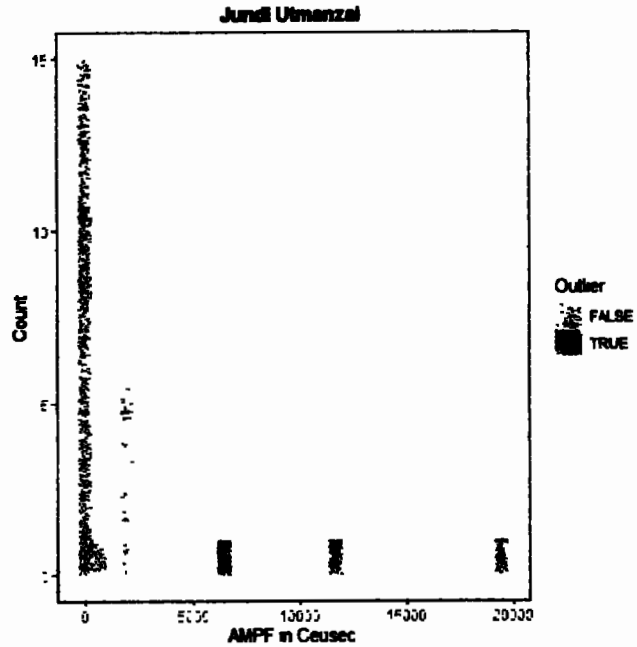

Kabul Nowshera
Count
AMPF in Ceusec
Outlier
FALSE
TRUE
Kabul Adezai
Count
AMPF in Ceusec
Outlier
FALSE

Continued

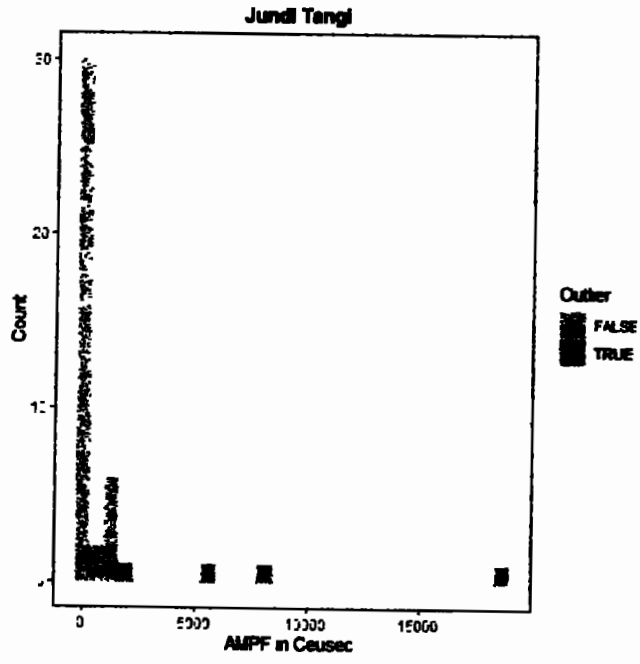

Shah Alam
Count
AMPF in Ceusec
Outlier
FALSE

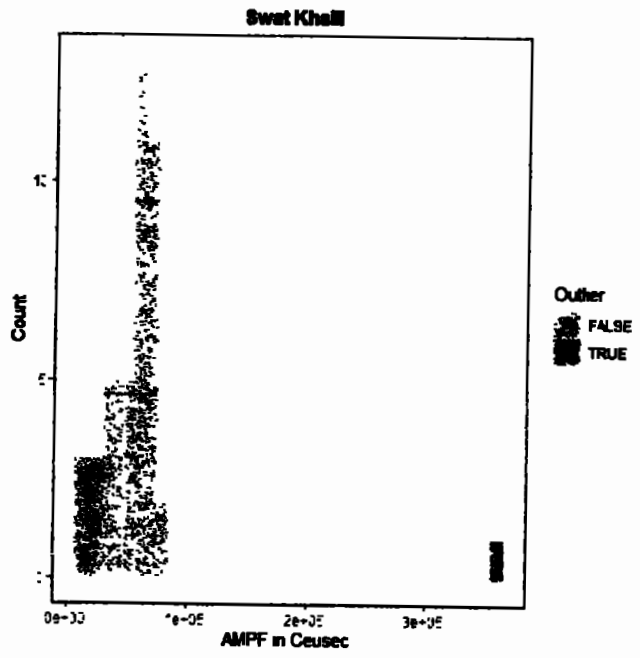

Kabul Nagunan
Count
AMPF in Ceusec
Outlier
FALSE
TRUE

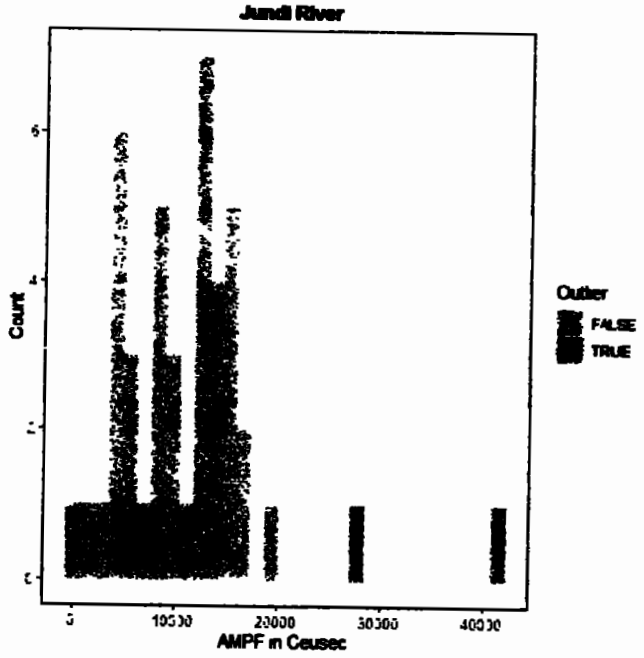

Panjkora
Count
AMPF in Ceusec
Outlier
FALSE
TRUE

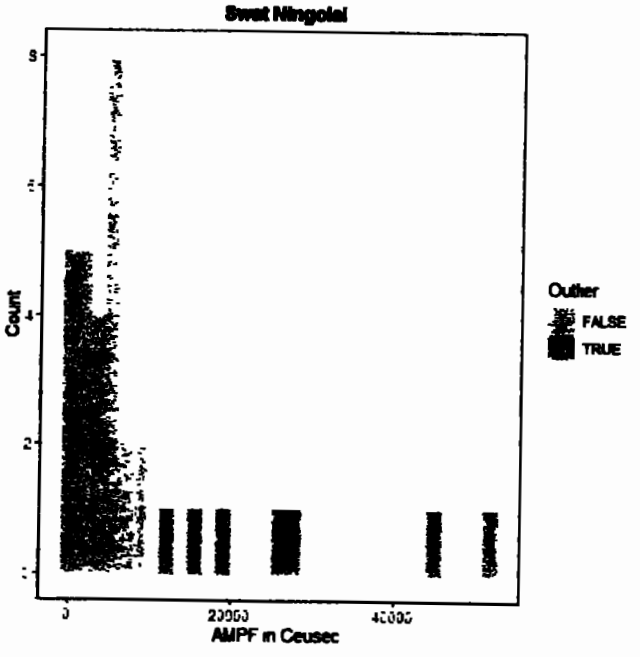

Jundi Utmanzai
Count
AMPF in Ceusec
Outlier
FALSE
TRUE

Continued


Jundi Tangi
Count
AMPF in Ceusec
Outlier
FALSE
TRUE


Swat Khalil
Count
AMPF in Ceusec
Outlier
FALSE
TRUE


Jundi River
Count
AMPF in Ceusec
Outlier
FALSE
TRUE


Swat Ningolai
Count
AMPF in Ceusec
Outlier
FALSE
TRUE

Continued

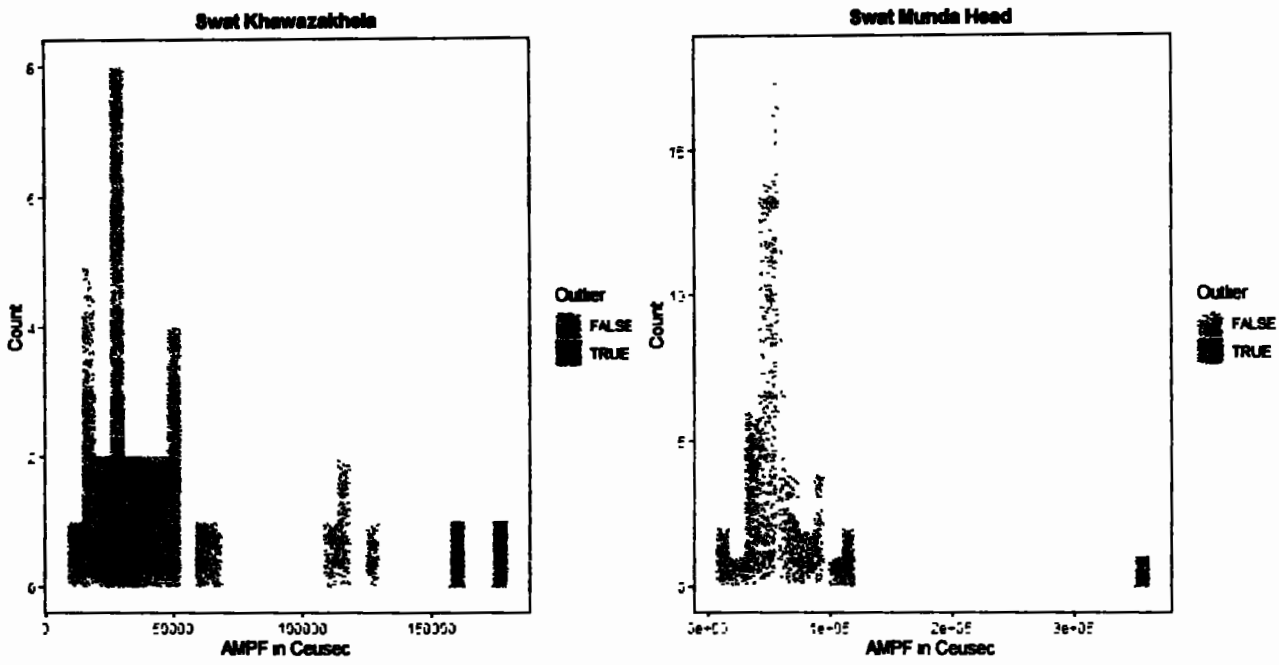


Fig 3.3: Results of Grubbs and Beck test for all 36 gauging sites.

3.6 Assumptions of randomness, homogeneity, independence and stationarity

RFFA based on few assumptions related to data series. These assumptions have been tested for the data series of each gauging site and results have been given in **Table (3.3).** The detailed discussion related to each assumption has been given in the following points.

1) Randomness of data series of each site is tested using Run test (Bradley, 1968). The results of run test given in **Table (3.3)** show that the data series of each site is random at 5% level of significance. This shows that sample data of each gauging site is drawn at random from the population, with each sample point has the same probability of being chosen.

2) Rank-sum test is used to test the homogeneity (Hirsch et al., 1992) of the data series at various sites. The results of **Table (3.3)** show that data series of each site fulfill the criteria of Rank-sum test at 5% level of significance. This mean that all of the sample points of each respective site are drawn from the same population.

3) Wald-Wolfowitz test (Wald and Wolfowitz, 1943; Rai et al., 2013) has been used to test independence and stationarity of data series of each site. The data series of each respective gauging site pass Wald-Wolfowitz test of independence and stationarity at 5% level of significance as shown in **Table (3.3)**. Therefore, we conclude that no single data point in the sample of each respective gauging site would have an impact on the non-occurrence or occurrence of any other element in the sample. Moreover, data set of each gauging site is stationarity and free from significant trends.

The results of **Table (3.3)** show that AMPFs at 36 sites are random, independent, free from significant trends and identically distributed. As a result, the data series of 36 gauging site is ideal for performing RFFA.

Table 3.3: Calculated values of test statistics and corresponding p-values of Run Test, Rank Sum Test and Wald-Wolfowitz Test.

S. No.	Site name	Rank-Sum		Run Test		Wald-Wolfowitz	
		Test Statistic	P-Value	Test Statistic	P-Value	Test Statistic	P-Value
1	Budni	-0.1444	0.8852	0.883	0.3772	-1.7913	0.0732
2	Shahi Bala	-0.7675	0.4427	-0.408	0.6833	1.9036	0.0570
3	Dallus	-1.059	0.2892	-0.69	0.4902	1.2807	0.2003
4	Badri	-0.408	0.6833	-0.644	0.5194	1.899	0.0576
5	Naranji	-0.446	0.6556	-1.656	0.0977	0.9353	0.3496
6	Kalpani Raisalpur	-0.3483	0.7276	1.467	0.1424	1.096	0.2729
7	Kalpani Deheri	0.739	0.4599	0.459	0.6459	-0.955	0.3392
8	Bagiari	0.293	0.7695	1.34	0.1802	1.6385	0.1013
9	Katlongi	0.4859	0.627	0.265	0.7910	-1.047	0.2951
10	Chprial	-1.089	0.2762	1.601	0.1094	1.512	0.1305
11	Jani Khwar	0.525	0.5996	-1.253	0.2100	-0.782	0.4337
12	Shahban	0.0547	0.9564	0.81	0.4179	-0.1173	0.9066
13	Muqam	-1.5407	0.1234	0.676	0.4990	1.4452	0.1484
14	Chinkar	1.126	0.2602	-1.201	0.2298	0.309	0.7573
15	Wazir Gahri	-0.132	0.895	0.574	0.5656	-0.678	0.4975
16	Bara Kohat Road	-1.171	0.2416	-0.542	0.5878	-0.083	0.9339
17	Bara Tarnab	-1.659	0.0971	-1.858	0.0631	1.251	0.2109
18	Lund Khwar East	-1.44	0.1499	-1.155	0.2479	0.109	0.9131
19	Kalpani Saidabad	-1.008	0.3135	-1.797	0.0723	0.765	0.4443
20	Dagi	-0.3484	0.7275	-0.522	0.6017	0.1914	0.8482
21	Garandi	0.7188	0.4723	1.135	0.2564	-1.042	0.2973
22	Hakim Gahri	-1.913	0.0557	-1.027	0.3044	1.34	0.1802
23	Khuderzai	0.955	0.3396	-1.274	0.2026	0.268	0.7887
24	Kabul Nowshera	-0.868	0.3854	-1.112	0.2658	-0.788	0.4307
25	Chilah	-0.594	0.5525	-1.070	0.2846	1.304	0.1921
26	Kabul Adezai	-1.716	0.0862	-1.318	0.1875	1.591	0.1116
27	Shah Alam	0.3716	0.7102	-0.664	0.5067	-0.23	0.8181
28	Panjkora	-1.797	0.0723	-0.991	0.3217	-0.621	0.5346
29	Kabul Naguman	-1.4864	0.1372	1.037	0.2997	1.204	0.2286
30	Jundi Utmanzai	0.578	0.5633	-0.481	0.6305	1.851	0.0641
31	Jundi Tangi	-1.49	0.1362	1.436	0.151	1.182	0.2372
32	Jundi River	-0.79	0.2495	0.291	0.7711	-0.563	0.5730
33	Swat Khaili	-1.64	0.1010	-0.937	0.3486	-1.08	0.2801
34	Swat Ningolai	1.864	0.0623	-0.371	0.7106	1.758	0.0787
35	Swat khawazakhela	1.724	0.0847	-1.741	0.0815	1.814	0.0695
36	Swat Munda Head	1.423	0.1547	-0.141	0.8875	-1.033	0.3015

Summary

The data of AMPF must fulfil the basic assumptions for valid and realistic estimates of flood quantiles. The current adopted technique for frequency analysis is also focused on these fundamental assumptions about data sets. As a result, in this chapter, preliminary data screening has been carried out to prepare data sets of 36 gauging sites under study. The key findings of this chapter given below.

I. In the data of many gauging sites, missing observations have been observed. The average AMPF value of their respective site has been used to filling the missing values. In a study (Hussain, 2017), similar handling has been done with missing values within the data of flood.

II. Graphical analysis (time series plot) has been performed to observe the trend within the data sets of AMPF of 36 gauging sites. AMPFs of some sites show an upward trend after 2010. Among the data sets of many sites, there have been occurred some high magnitude values. The Wald-Wolfowitz test has been used to assess the significance of trend within each site's data set, and no noticeable trend has been discovered in the AMPF across all sites.

III. The Grubbs and Beck test has been used to identify outliers, and as a result, high outliers have been identified but no low outliers found within the data of 36 sites. These high outliers have not been discarded from the sample data and RFFA analysis has been performed with these high outlier values.

IV. Few critical assumptions associated to the observed data series at various sites for RFFA has been tested and validated through various statistical tests. The results have revealed that the observed data at each site is random, independent, homogeneous, stationarity and free of regular trends.

Chapter 4

Regional Frequency Analysis of Sites of Khyber Pakhtunkhwa, Pakistan

4.1 Introduction

Reliable estimation of extreme hydrological events is needed for designing and building of hydrological structures on stream channels. These structures are important to provide protection against floods and to regulate the supply of available water. Various approaches including at-site and regional are in practice for flood frequency analysis in different parts of the world. However, regional approach has advantageous results in terms of accuracy and efficiency relative to at-site analysis. Therefore, in this chapter we perform application of L-moments based RFFA of 36 gauging sites of KPK Pakistan. Results, discussion and finding of this chapter given below.

4.2 Results and discussion

4.2.1 Discordancy measure

Descriptive Statistics in term of L-moments and values of D_i for each site by using Eq. (2.24) are given in **Table (4.1)**. The results show that two sites, "BADRI" and "CHILAH", are discordant, i.e. their D_i values are greater than 3. Therefore, possible options may be; either to drop these two sites at this stage or investigate the reasons for their large D_i values. These sites may be retained if there are abrupt variations or outliers in the data series at these sites (Hussain, 2011). For data visualization, time series plots of these two sites are illustrated in **Fig. (3.1)**. For site Badri, the distribution of the data around the average value is approximately symmetrical. However, a downward trend

exists in the values of the last seven years or so. This distribution of high and low values of AMPF is obvious in the shape of the distribution of the data series being negatively skewed (-0.0211 as shown in **Table (4.1)**). The time series plot of site Chilah is showing a flood of a very high magnitude in 1979. Grubbs and Beck test is also applied to detect outliers in the data series at these two sites and the results are presented in **Fig. (3.2)**. For site Chilah, six observations can be considered as high outliers within the data series. These high outliers are a major reason for the increase in its discordancy value. Such events of low and high magnitude can occur at any site due to climate variability and are random. Therefore, these two sites are retained in the group for further analysis.

Table 4.1: Descriptive statistics of each site in terms of L-moments and values of discordancy measure.

S. No.	Sites	n	l_1	T	t_3	t_4	t_5	D_i
1	Budni	47	14810.39	0.4678	0.3093	0.2978	0.3299	0.48
2	Shahi Bala	25	2792.4	0.5145	0.242	0.0675	0.0431	1.05
3	Dallus	25	8196.84	0.4744	0.2517	0.0657	-0.0298	0.54
4	Badri	46	7229	0.2701	-0.0211	0.195	0.0686	4.18*
5	Naranji	47	4836.136	0.4104	0.2587	0.167	0.0579	0.17
6	Kalpani Raisalpur	34	34773.34	0.3682	0.3357	0.1593	0.0755	0.97
7	Kalpani Deheri	21	2856.61	0.6275	0.4218	0.0894	-0.0475	1.03
8	Bagiari	31	5767.035	0.4877	0.218	-0.0367	0.0129	1.19
9	Katlongi	18	2396.555	0.5172	0.4206	0.1441	-0.0617	0.59
10	Chprial	34	10479.75	0.442	0.2632	0.0652	0.0265	0.39
11	Jani Khwar	22	984.918	0.3928	0.2438	0.4011	0.3221	1.90
12	Shahban	21	1515.763	0.4052	0.3454	0.2173	0.0561	0.32
13	Muqam	29	16669.1	0.4302	0.2711	0.0577	-0.0417	0.45
14	Chinkar	28	922.269	0.7141	0.6098	0.4117	0.339	0.63
15	Wazir Gahri	32	426.466	0.6457	0.5863	0.4113	0.3193	0.33
16	Bara Kohat Road	34	1453	0.7242	0.5958	0.3248	0.1661	0.75
17	Bara Tarnab	30	11884.18	0.6479	0.7283	0.6424	0.594	1.51
18	Lund Khwar East	28	484.082	0.5902	0.4183	0.1165	0.0081	0.66
19	Kalpani Saidabad	33	9408.818	0.6698	0.578	0.2948	0.0894	0.57
20	Dagi	33	390.818	0.4852	0.3351	0.2931	0.2535	0.30
21	Garandi	33	1004.636	0.4494	0.3741	0.1716	0.1078	0.41
22	Hakim Gahri	33	3713.903	0.3123	0.2035	0.1995	0.1137	0.51
23	Khuderzai	33	1758.374	0.6765	0.5319	0.3615	0.3326	0.57
24	Kabul Nowshera	15	138870.7	0.3059	0.4014	0.1818	0.1376	2.23
25	Chilah	33	1029.687	0.8349	0.8908	0.8475	0.8089	3.20*
26	Kabul Adezai	30	30027.69	0.3877	0.228	0.0258	0.0126	0.60
27	Shah Alam	30	7343.067	0.3997	0.2649	0.048	-0.0109	0.61
28	Panjkora	33	26271.79	0.3897	0.2744	0.2225	0.2408	0.18
29	Kabul Naguman	30	19227.27	0.4279	0.3195	0.2095	0.1169	0.08
30	Jundi Utmanzai	25	2052.571	0.8037	0.7232	0.5031	0.3593	1.27
31	Jundi Tangi	42	1104.653	0.8156	0.8131	0.6704	0.5421	1.84
32	Jundi River	43	11060.14	0.3295	0.1764	0.2337	0.1906	0.83
33	Swat Khaili	43	59534.23	0.2852	0.3021	0.2516	0.2072	1.82
34	Swat Ningolai	33	8933.677	0.6162	0.5349	0.3514	0.2009	0.19
35	Swat khawazakhela	34	50834.78	0.4153	0.4363	0.2169	0.0572	1.55
36	Swat Munda Head	55	62730.52	0.2687	0.3371	0.3179	0.1755	2.07

4.2.2 Formation of homogeneous regions

Formation/identification of homogeneous region(s) is an important and critical step in RFA. There exist a variety of objective and subjective techniques in the literature to delineate a study area into homogeneous regions if required. In this regard, elementary linkage analysis is used to define the homogeneous regions on the bases of cross correlation matrix calculated using the observed data of gauging sites. Each site is assigned to a region on the bases of highest index of correlation with other site. Every site in a region is highly correlated with the other sites in the region. Another correlation based technique that is used to delineate homogeneous regions is spatial correlation analysis. In this approach, correlation is considered as the function of distance between sites and direction. The two-dimensional correlogram is used to plot the correlation function. Smooth appearance of correlogram show the homogeneity of the region while irregularities indicate the heterogeneity of the region. Similar subjective approach for the defining of homogeneous regions is principal component analysis. Spatial configurations of prominent principal components are examined to delineate coherent homogeneous regions.

To minimize the subjectivity in the formation of homogeneous regions cluster analysis approach is gained popularity in RFA. Hosking and Wallis (1997) suggested cluster analysis based on site characteristics for the formation of homogeneous regions. Rao and Srinivas (2008) also provided useful details of hierarchical cluster analysis for the identification of homogeneous regions in RFA. Few other studies using hierarchical cluster analysis for identification of homogenous regions are Arellano-Lara and Escalante-Sandoval (2014) and Rasheed et al., 2019. This study has used hierarchical clustering based on site characteristics with few subjective adjustments to partition the

group of thirty-six sites into four homogeneous regions. Complete details are provided in the following section.

For initial estimate of degree of homogeneity in the group of 36 sites, heterogeneity measures based on L-CV, L-skewness and L-kurtosis are estimated as 8.58, 5.82 and 3.82, respectively, showing that the region is definitely heterogeneous and requires subdivision.

There are six available site characteristics which can be used for partitioning this heterogeneous group into homogeneous regions. It is obvious that each site characteristic has a different degree of relationship with observed data series. Therefore, to identify the most influential or significant site characteristic, at first step, the Pearson Correlation Coefficient is calculated between the average value of the AMPF at different sites (l_1) and the available site characteristics. This correlation matrix is illustrated in **Table (4.2)**, which shows that "latitude" has strongest positive significant correlation with l_1. Therefore, it is used to perform cluster analysis with Ward's linkage method and Euclidean distance measure. The dendrogram of cluster analysis is provided in **Fig. (4.1)**, which shows subdivision into seven clusters at first step. Heterogeneity measures based on L-CV is calculated to check the degree of homogeneity in each subdivided group. The details are: from left to right, first group with 8 sites (H is -0.48), second group with 3 sites (H is 0.95), third group with 4 sites (H is 4.91), forth group with 9 sites (H is 0.51), fifth group with 7 sites (H is 0.11), sixth group with 2 sites (H is 0.33) and seventh group with 3 sites (H is 1.22).

Table 4.2: Estimates of correlations between l_1 and site characteristics.

	l_1	Latitude	longitude	Elevation	AARF	ARMS	AAT
l_1	1	0.5469 (0.0006)	0.1922 (0.2614)	0.4881 (0.0025)	0.2731 (0.1071)	0.1361 (0.4287)	-0.2490 (0.1431)
Latitude		1	0.5298 (0.0009)	0.8930 (0.0001)	0.3778 (0.0231)	0.2449 (0.1500)	-0.2696 (0.1118)
longitude			1	0.4901 (0.0024)	0.3447 (0.0395)	0.4594 (0.0048)	0.0187 (0.9138)
Elevation				1	0.3539 (0.0342)	0.2017 (0.2381)	-0.2490 (0.1431)
AARF					1	0.8286 (0.0001)	-0.6881 (0.0001)
ARMS						1	-0.3901 (0.0187)

Note: Here values without parenthesis are the estimates of correlation coefficients and values in parenthesis are the corresponding p-values for testing the significance of correlation coefficient.

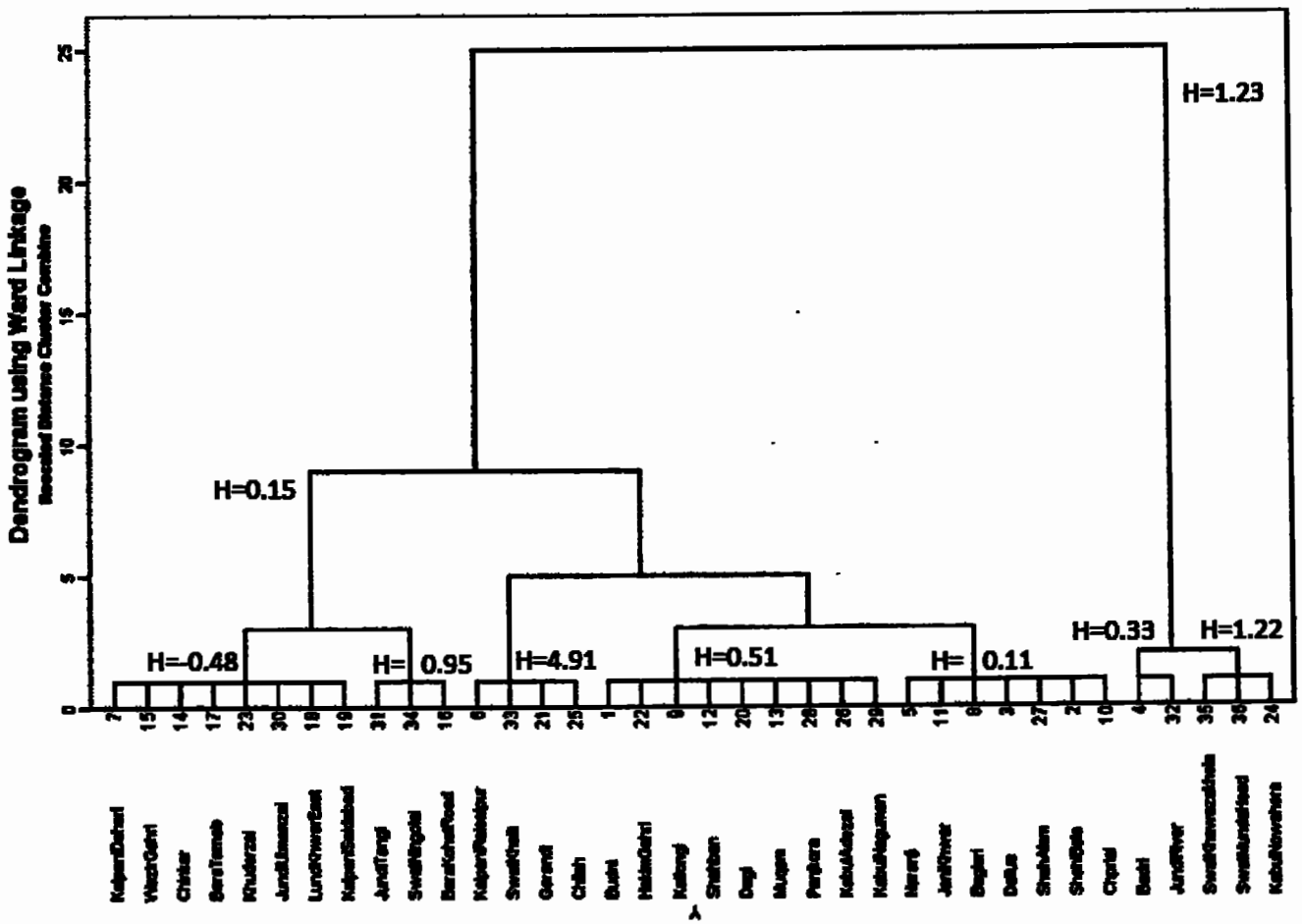


Fig. 4.1: Dendrogram which showing the division of 36 gauging sites in sub groups.

Keeping in view an inclusion of a reasonable number of sites in a group to perform RFA; this division of seven groups/clusters is subjectively adjusted to form fewer clusters with a larger number of sites and values of heterogeneity measure (H) less than 1. Neighboring clusters are combined to form fewer clusters as guided in the dendrogram for next step (like combining first group with second, fourth and fifth groups remains separate regions, and sixth group combined with seventh). The value of heterogeneity test is very high for third group and its combination with other groups also effects their homogeneity. The sites of third group are relocated to other homogeneous groups as suggested by (Hosking and Wallis, 1997; Satyanarayana and Srinivas, 2008). This relocation of sites is based on the variation (L-CV) of the observed data of the site which is being transferred to other region and the sites already in that region. This is done that the sites have similar variation to each other remain in the same group and eventually the homogeneity of the region is not effected (Khan et al., 2021). Relocation of sites of third group is given as: sites having similar values of L-CV; like shifting site "Chillah" from third group to the combination of first and second group, sites "Garandi" and "Kalpani Raisalpur" from third group to fifth group, and site "Swat Khaili" from third group to the combination of sixth and seventh group. Details of delineation of study region into homogenous groups are illustrated in **Table (4.3)** showing that the subdivided regions are homogeneous and ready to perform further steps in RFA.

Table 4.3: Details of delineation of study area into homogeneous regions.

Region identification	Combinations	Number of sites	Site names	Heterogeneity Measures
Region 1	First group + second group + Site Chillah	12	Kalpani Deheri, Wazir Ghari, Chinkar, Bara Tarnab, Khuderzai, Jundi Utmanzai, Lund Khwar East, Kalpani Saidabad, Jundi Tangi, Swat Ningolai, Bara Kohat Road, Chillah	H = 0.26
Region 2	Forth group	9	Budni, Hakim Ghari, Katlongi, Shahban, Dagi, Muqam, Panjkora, Adezai, Naguman	H = 0.54
Region 3	Fifth group + Site Garandi + Site Kalpani Raisalpur	9	Naranji, Bagiari, Dallus, Shah Alam, Shahi Bala, Chprial, Garandi, Kalpani Raisalpur, Jani Khwar	H = 0.14
Region 4	Sixth group + Seventh group + Site Swat Khaili	6	Badri, Jundi River, Swat Khawazakhela, Swat Munda Head, Kabul Nowshera, Swat Khaili	H = 0.91

4.2.3 Fitting of regional probability distribution

For the regional distribution five three parameter distributions (GLO, GEV, GPA, GNO and PE3) have been used. The main reason for the inclusion of only these five distributions is because this is a unique set of distributions that have location scale and shape parameters among the class of three parameter distributions.

L-moment ratio diagrams of the four regions are illustrated in **Fig. (4.2)**. A probability distribution is assumed to be fit if the regional average of L-skewness and L-kurtosis lies closest to its theoretical lines so as the tendency of the individual points. Based on these principles, details of good fit distribution(s) for each region are: Region 1 has GNO and GPA distributions; Region 2 has GNO, PE3 and GPA distributions; Region 3 has GPA distribution; and Region 4 has GLO distribution.

The calculated values of $|Z^{Dist}|$ statistic, for the four regions, are illustrated in **Table (4.4)**. Details of the distributions passing the goodness of fit criteria are: Region 1 has GLO, GEV, GNO and GPA distributions; Region 2 has GLO, GEV, GNO, GPA and PE3 distributions; Region 3 has PE3 and GPA distributions; while Region 4 has GLO distribution.

The two goodness-of-fit methods are in fair agreement to each other with respect to the identification of successful regional distributions. However, the results of $|Z^{Dist}|$ statistic, being a quantitative method based on simulations, are taken for further analysis.

Table 4.4: Values of $|Z^{Dist}|$ statistic for candidate distributions. * Indicates the calculated values exceeding critical value, i.e. 1.64.

S. No.	Region identification	GLO	GEV	GNO	PE3	GPA
1	Region 1	0.06	0.11	1.14	2.76*	0.76
2	Region 2	1.38	0.62	0.01	1.06	1.47
3	Region 3	3.47*	2.49*	1.85*	0.7	0.11
4	Region 4	1.55	2.41*	2.8*	3.52*	4.51*

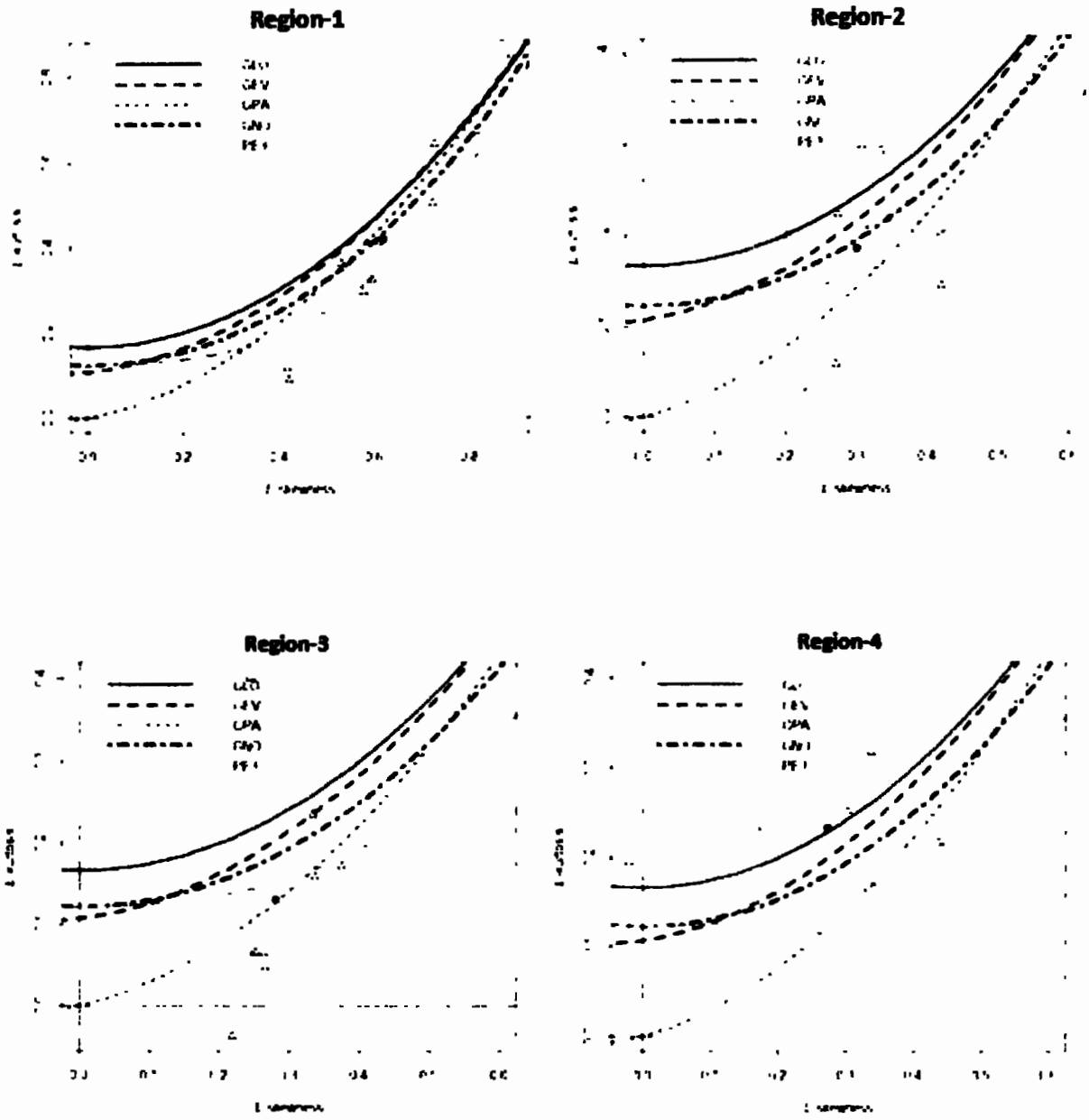


Fig. 4.2: L-moment Ratio diagrams of four homogeneous regions. Red dot (●) shows the regional average of L-skewness and L-kurtosis.

4.2.4 Identification of a robust regional distribution

The goodness-of-fit methods have identified two or more probability distributions as successful candidates for three of the four regions. Therefore, an assessment analyses

using simulations is required to identify the robust probability distribution for each region. The details of the development of these simulations experiments are stated in Hosking and Wallis (1997). A brief of setting up of a base/artificial region for simulations is described as follows:

The first step is the development of an artificial region similar to the actual/study region concerning the number of sites, observations at each site and regional average estimates of L-moment ratios. In addition, L-moment ratios for each site should be chosen in such a way that the value of heterogeneity measure H remains close to the value calculated using actual/observed data series. To check for the inter-site dependence between sites, a correlation matrix is calculated. The average values of inter-site correlation for Region 1, Region 2, Region 3 and Region 4 are -0.014, 0.122, 0.259 and -0.055, respectively. This indicates weak inter-site dependence between sites of all the regions. This may be because these sites are located on different streams/rivers. For the development of artificial regions, the details of the linear variations in the values of L-CV with incremental effect for each site, the chosen values of L-skewness for each site and the estimated values of heterogeneity measure for each region are summarized in **Table (4.5)**.

The developed artificial regions are showing a comparable degree of homogeneity relative to their actual counterparts. Therefore, they can be used for simulations to calculate the accuracy measures for the identification of the robust regional distribution. For instance, using the base region of Region 1, 5000 realizations are performed, and every time the successful distribution is fitted through a method of L-moments. This process continues for GPA, GLO, GEV and GNO distributions. The relative root mean square error (RMSE) of regional quantiles is calculated from these simulations and the results are shown in **Table (4.6)**. These results indicate that, in general, the estimates

of quantiles for GNO distribution have minimum RMSE. Moreover, regional growth curves with 95% error bounds for GLO, GEV, GNO and GPA distribution are given in **Fig (4.3)**. The graph shows that, in general, the growth curve of GNO distribution has the shortest 95% error bounds, especially for longer return periods. Secondly, the growth curve of GNO distribution remains within the limits of 95 percent error bounds, while growth curves of GLO, GPA and GEV distributions are below the lower limits of 95% error bounds for longer return periods. Therefore, GNO distribution is the most stable and robust distribution for Region 1.

Similar, the methodology has been adopted for the identification of robust distributions for Region 2, Region 3 and Region 4. For Region 4, accuracy measures are calculated for GLO distribution as being the only good-fit distribution. The estimates of regional quantiles using successful candidate's distributions and their RMSE for Region 2, Region 3 and Region 4 are given in **Table (4.7), (4.8) and (4.9)**, respectively. Regional growth curves for Region 2, Region 3 and Region 4 with their respective 95% error bounds are given in **Fig. (4.4), (4.5) and (4.6)**, respectively. These results show that the most stable and robust distribution for Region 2 is GPA (especially for longer return periods), Region 3 is also GPA and Region 4 is GLO.

After the identification of robust distribution for each region, at-site flood quantiles using the estimated regional quantiles (based on most stable regional distribution) of each respective region, their RMSE and 95% error bounds are given in **Table (4.10)** for Region 1, **Table (4.11)** for Region 2, **Table (4.12)** for Region 3 and **Table (4.13)** for Region 4. The at-site flood quantiles are given in **Table (4.10)** for Region 1, **Table (4.11)** for Region 2, **Table (4.12)** for Region 3 and **Table (4.13)** for region 4 are estimated by multiplying average values of each site within the each homogeneous

region with their regional quantiles. The accuracy measures RMSE and 95% error bounds are estimated through simulation process.

The results of **Table (4.10)** show that in terms of magnitude of flood the site Wazir Gahri is the smaller and the site Bara Tarnab is the larger site within Region 1. The results of site Wazir Gahri (smaller site of Region 1) show that in next 15 years the flood with magnitude of 1320 Cusecs occurs at most one time has 0.93 probability of non-excedence. For next 30, 50, 100, 150 and 200 years the flood with magnitudes of 2119, 2951, 4377, 5290 and 6279 Cusecs occur at most onetime have probabilities of non-excedences 0.96, 0.98, 0.99, 0.993 and 0.995 respectively. At-site Bara Tarnab (larger site of Region 1) in next 15 years the flood with magnitude of 36789 Cusec occurs at most one time has 0.93 probability of non-excedence. For next 30, 50, 100, 150 and 200 years the flood with magnitudes of 50049, 82229, 121980, 147403 and 174966 Cusec occur at most onetime have probabilities of non-excedences 0.96, 0.98, 0.99, 0.993 and 0.995 respectively. Estimated at-site flood quantiles of each site within the Region 1 are greater than the average value of their corresponding site. For each site of Region 1, estimated at-site flood quantiles for small to large return periods lies within the 95% error bounds limits. Similar, finding are observed from the results of Region 2, Region 3 and Region 4 which are given in **Table (4.11)**, **Table (4.12)** and **Table (4.13).**

These estimates are useful for the scientists, hydrologists and government officials dealing with designing and developing proposed and existing hydrological structures as well as water resources management and flood protection planning of the region. The accuracy measures of these at-site quantiles would be helpful for future studies to compare the quality of the estimates using alternative methods of modelling.

Table 4.5 Information of base regions used for the assessment analyses.

S. No.	Region name	Number of sites	Linear variation in the values of L-CV	Increment at each step	L-skewness	Estimated value of H
1	Region 1	12	0.5903 at site 1 to 0.8433 at site 12	0.0230	0.6194	0.22
2	Region 2	9	0.2806 at site 1 to 0.5628 at site 9	0.0227	0.2938	0.58
3	Region 3	9	0.3680 at site 1 to 0.4936 at site 9	0.0157	0.2807	0.19
4	Region 4	6	0.2686 at site 1 to 0.4286 at site 6	0.0320	0.2720	0.94

Table 4.6: Estimated quantiles and their RMSE for Region 1.

Return Periods	Distributions							
	GPA		GEV		GLO		GNO	
	$\hat{q}$	RMSE	$\hat{q}$	RMSE	$\hat{q}$	RMSE	$\hat{q}$	RMSE
15	2.8425	0.2852	2.6404	0.2795	2.6002	0.2572	3.0956	0.2491
30	4.5067	0.4121	4.2174	0.3829	4.1399	0.3715	4.9687	0.4830
50	6.314	0.678	5.9914	0.6411	5.8832	0.6362	6.9191	0.8052
100	9.6199	1.4218	9.3656	1.4111	9.2277	1.3837	10.264	1.3711
150	11.8761	2.0486	11.745	2.0802	11.6044	2.0248	12.4032	1.9487
200	14.4476	2.8429	14.518	2.9456	14.3906	2.8528	14.7225	2.5008

Table 4.7: Estimated quantiles and their RMSE for Region 2.

Return periods	Distributions									
	GPA		GEV		GLO		GNO		PE3	
	$\hat{q}$	RMSE	$\hat{q}$	RMSE	$\hat{q}$	RMSE	$\hat{q}$	RMSE	$\hat{q}$	RMSE
15	2.4281	0.3299	2.3217	0.3034	2.2572	0.2894	2.3587	0.3145	2.4033	0.3205
30	2.9241	0.4608	2.9269	0.4539	2.8859	0.4457	2.9399	0.4617	2.9306	0.4602
50	3.2911	0.5665	3.4508	0.5923	3.4635	0.5971	3.4224	0.5891	3.3395	0.5748
100	3.7444	0.7101	4.2133	0.8074	4.3595	0.8458	4.0933	0.7734	3.8701	0.7313
150	3.9666	0.7866	4.6441	0.9363	4.8945	1.0017	4.4576	0.8768	4.1418	0.8147
200	4.1697	0.8607	5.0769	1.0711	5.4523	1.1699	4.8138	0.9801	4.3975	0.8952

Table 4.8: Estimated quantiles and their RMSE for Region 3.

Return Periods	Distributions			
	GPA		PE3	
	$\hat{q}$	RMSE	$\hat{q}$	RMSE
15	2.4838	0.1596	2.4552	0.3274
30	2.9888	0.2427	2.9976	0.4707
50	3.3596	0.3189	3.4176	0.5882
100	3.8139	0.4322	3.9621	0.7486
150	4.0351	0.4962	4.2407	0.8342
200	4.2363	0.5598	4.5029	0.9167

Table 4.9: Estimated quantiles and their RMSE for Region 4.

Return Periods	GLO Distribution	
	$\hat{q}$	RMSE
15	1.9063	0.1457
30	2.3225	0.2093
50	2.6945	0.2692
100	3.256	0.407
150	3.5836	0.5291
200	3.9201	0.6907

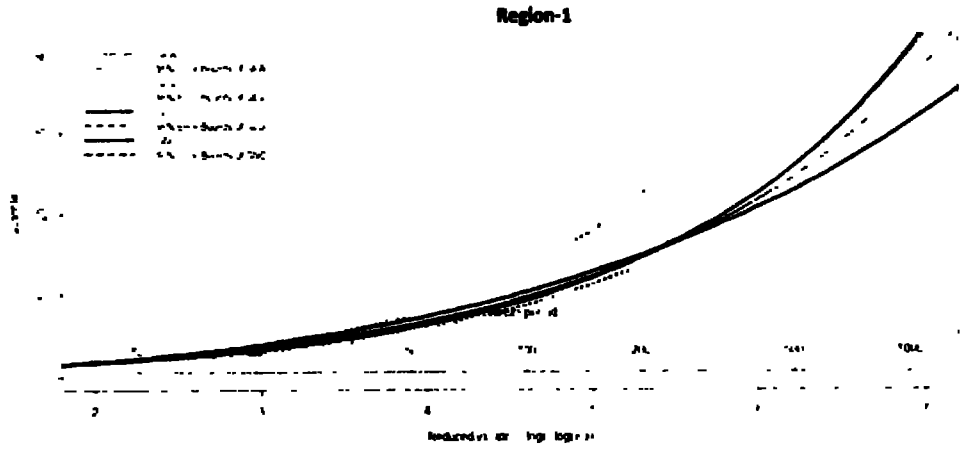


Fig 4.3: Regional growth curves of successful distributions of Region 1 with their 95% error bounds.

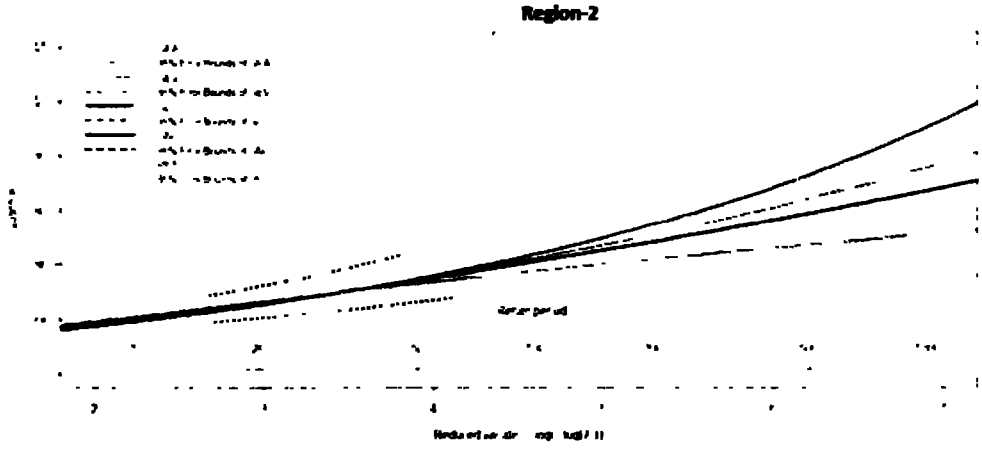


Fig. 4.4: Regional growth curves of successful distributions of Region 2 with their 95% error bounds.

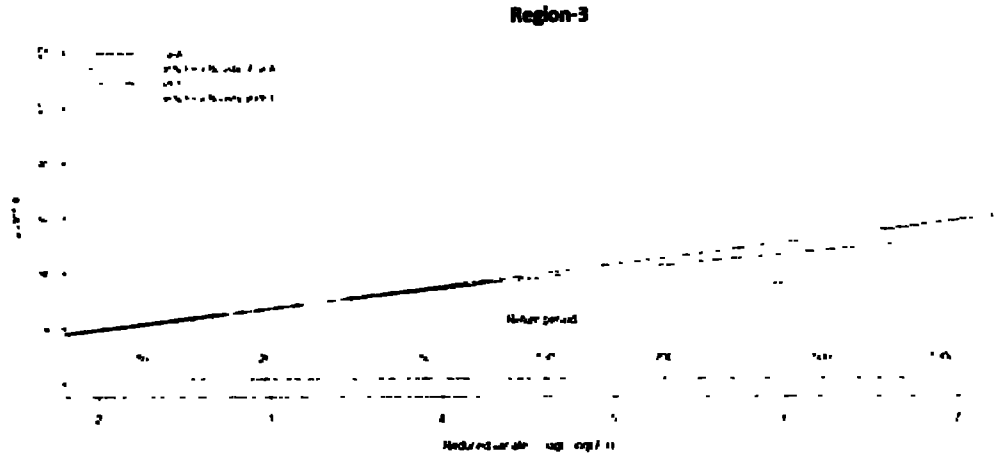


Fig. 4.5: Regional growth curves of successful distributions of Region 2 with their 95% error bounds.

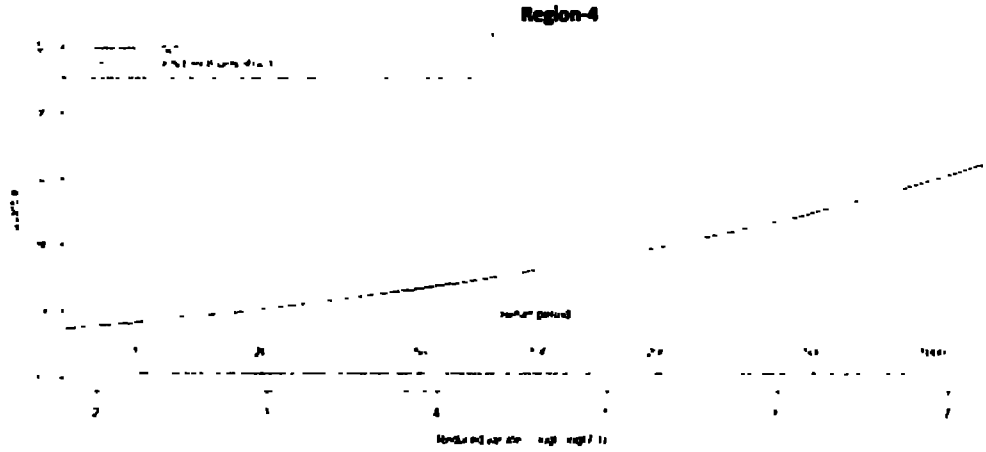


Fig. 4.6: Regional growth curve of GLO distribution (for Region 4) with their 95% error bounds.

Table 4.10: Estimated at site flood quantiles with RMSE and 95% error bounds of Region 1 using GNO distribution.

Site Names	Measures	15	30	50	100	150	200
Kalpani Deheri	$\hat{Q}$	8843	14194	19765	29320	35431	42057
	RMSE	5054	8009	11093	16419	19851	23595
	LB	4555	7514	10623	16130	19541	23413
	UB	21432	36082	52322	80882	99912	121932
Wazir Gahri	$\hat{Q}$	1320	2119	2951	4377	5290	6279
	RMSE	598	952	1324	1972	2391	2851
	LB	731	1201	1690	2563	3106	3708
	UB	2740	4634	6696	10454	12934	15740
Chinkar	$\hat{Q}$	2855	4583	6381	9466	11439	13578
	RMSE	1418	2240	3100	4588	5549	6599
	LB	1545	2538	3605	5458	6668	7941
	UB	6256	10599	15386	23944	29719	35923
Bara Tarnab	$\hat{Q}$	36789	59049	82229	121980	147403	174966
	RMSE	20402	32011	44087	64884	78265	92846
	LB	19772	32455	45937	69531	85156	102456
	UB	78095	132085	190298	296608	367379	446644
Khuderzai	$\hat{Q}$	5443	8737	12166	18048	21810	25888
	RMSE	2785	4450	6204	9262	11246	13419
	LB	2997	4909	6928	10476	12733	15186
	UB	11177	18860	27392	42449	52861	64141
Jundi Utmanzai	$\hat{Q}$	6354	10199	14202	21068	25459	30219
	RMSE	3548	5662	7892	11783	14309	17078
	LB	3303	5421	7711	11590	14097	16910
	UB	14288	24269	34984	54354	67139	81082
Lund Khwar East	$\hat{Q}$	1499	2405	3349	4969	6004	7127
	RMSE	802	1268	1755	2596	3138	3729
	LB	791	1307	1848	2810	3420	4093
	UB	3221	5469	7911	12312	15311	18596
Kalpani Saidabad	$\hat{Q}$	29127	46750	65101	96573	116700	138522
	RMSE	13407	21348	29713	44293	53761	64138
	LB	16147	26324	37265	56639	69029	82375
	UB	58837	99778	144389	224167	278063	337619
Jundi Tangi	$\hat{Q}$	3420	5489	7643	11338	13701	16263
	RMSE	1431	2291	3202	4799	5841	6986
	LB	1970	3227	4553	6867	8359	10008
	UB	6610	11135	16126	25213	31317	38182
Swat Ningolai	$\hat{Q}$	27656	44389	61814	91696	110807	131527
	RMSE	12525	19993	27863	41587	50499	60267
	LB	15301	25167	35794	53845	65473	78420
	UB	57299	96827	140640	220431	272911	331770
Bara Kohat Road	$\hat{Q}$	4498	7220	10054	14914	18022	21392
	RMSE	2021	3239	4523	6762	8216	9810
	LB	2492	4067	5763	8724	10599	12610
	UB	9071	15390	22070	34291	42344	51175
Chillah	$\hat{Q}$	3188	5116	7125	10569	12772	15160
	RMSE	1733	2815	3968	5992	7313	8765
	LB	1772	2930	4153	6262	7640	9163
	UB	6644	11319	16414	25608	31667	38282

Table 4.11: Estimated at site flood quantiles with RMSE and 95% error bounds of Region 2 using GPA distribution.

Site Names	Measures	15	30	50	100	150	200
Budni	$\hat{Q}$	35962	43307	48743	55456	58748	61756
	RMSE	6119	8014	9539	11599	12693	13748
	LB	27752	32607	35958	39979	41953	43619
	UB	48476	59634	67852	78461	83634	88541
Hakim Gahri	$\hat{Q}$	9018	10860	12223	13906	14732	15486
	RMSE	1731	2226	2618	3141	3416	3680
	LB	6808	8019	8905	9967	10464	10908
	UB	12866	15746	17982	20756	22131	23338
Katlongi	$\hat{Q}$	5819	7008	7887	8974	9506	9993
	RMSE	1357	1719	2004	2378	2573	2759
	LB	4155	4923	5453	6089	6390	6656
	UB	9227	11328	12909	14864	15855	16807
Shahban	$\hat{Q}$	3681	4432	4989	5676	6013	6320
	RMSE	818	1044	1224	1461	1585	1704
	LB	2649	3131	3478	3873	4062	4232
	UB	5597	6851	7798	9013	9633	10224
Dagi	$\hat{Q}$	949	1143	1286	1463	1550	1630
	RMSE	179	231	271	325	354	381
	LB	719	847	939	1047	1099	1144
	UB	1347	1646	1880	2180	2332	2469
Muqam	$\hat{Q}$	40476	48743	54860	62416	66121	69507
	RMSE	8210	10535	12371	14807	16084	17305
	LB	29941	35335	39186	43659	45850	47750
	UB	58853	72499	82443	95240	101839	108228
Panjkora	$\hat{Q}$	63793	76822	86464	98373	104212	109548
	RMSE	12192	15732	18548	22311	24293	26196
	LB	48279	56881	62909	70180	73538	76612
	UB	90884	112080	127535	147341	157502	167023
Kabul Adezai	$\hat{Q}$	72913	87805	98825	112436	119110	125209
	RMSE	14371	18532	21830	26217	28519	30725
	LB	54510	64403	71210	79293	83368	86887
	UB	104078	127630	145766	168103	179856	190529
Kabul Naguman	$\hat{Q}$	46687	56223	63280	71995	76269	80174
	RMSE	9291	11924	14006	16772	18223	19613
	LB	34845	41187	45639	50819	53363	55700
	UB	67966	83367	95118	110033	117305	123867

Table 4.12: Estimated at site flood quantiles with RMSE and 95% error bounds of Region 3 based using GPA distribution.

Site Names	Measures	15	30	50	100	150	200
Naranji	$\hat{Q}$	13529	16281	18300	20775	21979	23075
	RMSE	1673	2096	2461	2992	3290	3587
	LB	11224	13384	14859	16619	17424	18130
	UB	16829	20477	23296	26895	28671	30327
Bagiari	$\hat{Q}$	14324	17237	19375	21995	23271	24431
	RMSE	2239	2752	3172	3758	4078	4393
	LB	11283	13541	15128	16945	17776	18524
	UB	19120	23239	26236	30070	32033	33853
Dallus	$\hat{Q}$	20359	24499	27539	31262	33075	34725
	RMSE	3450	4220	4843	5699	6163	6617
	LB	15846	18960	21216	23805	25014	26097
	UB	27773	33532	37960	43708	46513	49108
Shah Alam	$\hat{Q}$	18239	21948	24670	28006	29630	31108
	RMSE	2840	3491	4025	4771	5179	5580
	LB	14500	17333	19381	21771	22838	23759
	UB	24465	29642	33530	38491	41040	43354
Shabi Bala	$\hat{Q}$	6936	8346	9382	10650	11268	11830
	RMSE	1173	1434	1644	1934	2090	2244
	LB	5399	6460	7229	8128	8551	8906
	UB	9476	11472	12996	14955	15937	16836
Chprial	$\hat{Q}$	26030	31323	35209	39969	42287	44396
	RMSE	3873	4795	5560	6636	7227	7808
	LB	20816	24940	27797	31181	32722	34170
	UB	34321	41661	47195	54306	57968	61385
Garandi	$\hat{Q}$	2495	3003	3375	3832	4054	4256
	RMSE	376	462	534	634	689	743
	LB	1991	2381	2658	2990	3144	3277
	UB	3274	3971	4503	5177	5527	5843
Kalpani Raisalpur	$\hat{Q}$	86371	103934	116827	132623	140315	147312
	RMSE	13115	16231	18807	22415	24391	26335
	LB	68524	81818	91166	102302	107272	111613
	UB	113623	138560	157133	180703	192981	203724
Jani Khwar	$\hat{Q}$	2446	2944	3309	3756	3974	4172
	RMSE	444	543	622	730	787	843
	LB	1860	2235	2502	2817	2964	3090
	UB	3414	4136	4697	5360	5715	6039

Table 4.13: Estimated at site flood quantiles with RMSE and 95% error bounds of Region 4 using GLO distribution.

Site Names	Measures	15	30	50	100	150	200
Badri	$\hat{Q}$	13781	16790	19479	23538	25907	28339
	RMSE	1972	2683	3377	4517	5229	5993
	LB	11105	13217	15094	17725	19257	20743
	UB	17760	22356	26732	33386	37411	41676
Jundi River	$\hat{Q}$	21085	25688	29802	36012	39636	43357
	RMSE	3060	4152	5216	6963	8053	9223
	LB	16898	20077	22902	26968	29303	31697
	UB	27091	34079	40701	51146	57514	64252
Swat Khawazakhela	$\hat{Q}$	96910	118066	136979	165520	182176	199279
	RMSE	15439	20795	25982	34435	39685	45302
	LB	76772	91248	103912	122470	132943	143843
	UB	129020	162222	193014	241063	270290	300767
Swat Munda Head	$\hat{Q}$	119588	145694	169033	204253	224807	245912
	RMSE	16260	22417	28461	38432	44681	51403
	LB	97305	115724	131439	154804	167828	181087
	UB	151246	190871	227053	283550	317760	353271
Kabul Nowshera	$\hat{Q}$	264739	322533	374199	452169	497670	544391
	RMSE	57311	73814	89509	114688	130150	146578
	LB	191030	229720	262901	311784	339714	367850
	UB	387385	486291	579803	723338	808678	900271
Swat Khaili	$\hat{Q}$	113494	138271	160420	193846	213353	233382
	RMSE	16628	22590	28399	37913	43842	50198
	LB	90580	107638	122819	144595	156694	169249
	UB	146931	185429	220863	276271	310299	346232

Summary

This chapter of the study is based on the application of RFFA for estimating flood quantiles considering AMPF of 36 sites located on important streams/rivers of KPK, Pakistan. Systematic, detailed and comprehensive application of a standard procedure to a new study area is an important contribution of this study. Some major findings are summarized below:

i. L-moments based descriptive statistics show that there exist variations in the data series at 36 sites. The L-kurtosis values, however, are relatively lower than

the values of L-skewness. This shows that the variation in the data series at different sites is following a specific pattern or there is a frequent flooding at different sites of the study area. A possible reason for these fluctuations may be the erratic cycles of monsoon rainfall, because floods in Pakistan rely mostly on heavy monsoon rainfall. Hussain (2017) reported similar results for major river sites in Punjab, Pakistan.

ii. Due to the existence of heterogeneity in the study region, it is divided into four homogeneous regions. Wards clustering method using Euclidean distance based on the most significant site characteristic is used for the sub division of study area into homogenous regions.

iii. Five commonly used probability distributions have been used as candidates for regional distribution. The goodness of fit criterion of $|Z^{Dist}|$ statistic and L-moment ratio diagram show that two or more distributions have passed goodness-of-fit criteria for three of the four regions. Therefore, an assessment analyses using simulations is performed to identify the robust distribution. The results of different accuracy measures (95 percent error bounds and RMSE of estimated quantiles) show that GNO distribution for Region 1, GPA distribution for Region 2 and Region 3, and GLO distribution for Region 4 are robust distributions. Identification of dissimilar regional distributions indicates differences in the trends, tendencies and shape of the distribution of the data series in different areas.

iv. Using regional quantiles of each region at-site quantiles are estimated. These estimated at sites quantiles are larger than average values of each site within each region which shows the rising trend of flood in future.

Chapter 5

Flood Quantiles Estimation at Ungauged Sites

5.1 Introduction

RFFA basically involves two principal steps, (1) identification of regions having similar site characteristics and (2) development of forecast equations for the estimation of flood quantiles at gauged and ungauged sites. In RFFA, several methods are in practice including regression techniques, rational methods, ANNs, etc. for the development of forecast equations to estimate flood quantiles at ungauged sites. In this chapter, regression models with robust estimation methods and machine learning methods (BPNN and RBF) have been used for the development of models for ungauged flood quantiles estimation. The details are given as follows.

5.2 Results and discussion

Major advantages of using RFA include robust estimates of quantiles at gauged sites and estimation or improvement of quantiles at ungauged or partially/poorly gauged sites within the homogeneous region(s). For this purpose, following steps are followed: a) formation of homogenous region(s) of gauging sites b) identification of suitable regional distribution for each homogeneous region c) estimation of dimensionless regional quantiles of regional distribution, d) development of forecasting model for each homogeneous region using site statistics l_1 (average data value of each site within the homogeneous region) as dependent variable and characteristics of all sites located within the homogeneous region as independent variable(s). The developed regional forecast model is used for the estimation of sites statistic by putting site characteristics of that particular ungauged site in the regional forecast model. The estimated site

statistic for ungauged site is multiplied with regional quantiles to estimate the flood quantiles for that particular ungauged site (Hosking and Wallis, 1997; Hussain, 2017; Khan et al., 2019; Khan et al., 2020). We have discussed first three steps in Chapter 4 and the details of step d are given in the following sections.

5.2.1 Regression based models

Regression models are developed for the prediction of ungauged flood quantiles. For such purpose, from the available site characteristics, selection of the most influential one that has a significant impact on floods within the region is very important. Such site characteristic(s) is/are used as a predictor variable(s) within the regression model. Anilan, et al. (2016) gives the details of some commonly used site characteristics as independent variables in regression models around the world for estimating ungauged sites flood quantiles. These site characteristics are drainage area, the slope of stream, and mean annual rainfall. Availability and identification of the most influential site characteristics that can be used for the prediction of flood at an ungauged site within the region is an ongoing area of research. The development of an adequate regression model depends on the site characteristics that show a significant relationship with the recorded data sets of gauged sites. Therefore, first, we i

dentify the most influential site characteristic that may be used as a predictor variable within the regression model of each homogeneous region for the prediction of ungauged flood quantiles.

Floods in Pakistan are mostly occurred in the monsoon season due to heavy monsoon rainfall within the region (Government of Pakistan, 2017). The frequency in the percentage of recorded AMPF of gauging sites of Region 1, Region 2, Region 3 and Region 4 during a year (summer (monsoon), autumn, winter and spring) have been calculated to identify in which season most of the AMPF values occurred. The

percentages of the frequencies of AMPF are illustrated in **Table (5.1), (5.2), (5.3)** and **(5.4)**. The values of **Table (5.1), (5.2), (5.3)** and **(5.4)** show the highest percentage of AMPF occurred in the monsoon season for all regions. Therefore, ARMS is the most suitable site characteristic that will be used as the independent variable for the development of regression models for Region1, Region 2, Region 3 and Region 4.

Table 5.1: Percentage (%) of frequency of AMPF (Annual Maximum Peak Flows) in four seasons at each site of Region 1.

S. No.	Site name	n_i	Monsoon (%)	Autumn (%)	Winter (%)	Spring (%)
1	Kalpani Deheri	21	85.7	4.7	0.0	9.5
2	Wazir Ghari	32	46.8	12.5	12.5	21.8
3	Chinkar	28	64.2	7.1	10.7	10.7
4	Bara Tarnab	30	56.6	3.3	3.3	30.0
5	Khuderzai	33	51.5	12.1	9.1	24.2
6	Jundi Utmanzai	25	80.0	0.0	0.0	4.0
7	Lund Khwar East	28	64.2	17.8	0.0	3.5
8	Kalpani Saidabad	33	84.8	12.1	3.0	0.0
9	Jundi Tangi	42	69.04	9.5	0.0	9.5
10	Swat Ningolai	33	81.0	10.0	0.0	3.0
11	Bara Kohat Road	34	83.0	5.0	0.0	12.0
12	Chillah	32	60	6	13	21

Note: The time period for Monsoon is from June to August, autumn is from September to November, winter is from December to February and spring is from March to May.

Table 5.2: Percentage (%) of frequency of AMPF (Annual Maximum Peak Flows) in four seasons at each site of Region 2.

S. No.	Site names	*nl*	Monsoon (%)	Autumn (%)	Winter (%)	Spring (%)
1	Budni	47	73	11	7	9
2	Hakim Gahri	33	52	19	10	19
3	Katlongi	18	94	6	0	0
4	Shahban	21	59	0	6	35
5	Dagi	33	64	4	7	25
6	Muqam	29	88	3	3	6
7	Panjkora	33	52	12	0	36
8	Kabul Adezai	30	93	0	0	7
9	Kabul Naguman	30	83	7	0	10

Table 5.3: Percentage (%) of frequency of AMPF (Annual Maximum Peak Flows) in four seasons at each site of Region 3.

S. No.	Site names	*nl*	Monsoon (%)	Autumn (%)	Winter (%)	Spring (%)
1	Naranji	52	84	12	2	2
2	Bagiari	31	80	4	10	6
3	Dallus	25	60	8	8	24
4	Shah Alam	30	80	0	0	20
5	Shahi Bala	25	74	12	0	14
6	Chprial	34	59	0	3	38
7	Garandi	33	67	9	9	15
8	Kalpani Raisalpur	34	82	3	6	9
9	Jani Khwar	22	51	16	0	33

Table 5.4: Percentage (%) of frequency of AMPF (Annual Maximum Peak Flows) in four seasons at each site of Region 4.

S. No.	Site names	*nl*	Monsoon (%)	Autumn (%)	Winter (%)	Spring (%)
1	Badri	46	88	7	5	0
2	Jundi River	43	93	4	0	3
3	Swat Khawazakhela	34	94	3	0	3
4	Swat Munda Head	55	76	6	2	16
5	Kabul Nowshera	15	87	0	0	13
6	Swat Khaili	43	79	5	0	16

For the identification of the functional relationship between dependent and independent variables for all four regions scatter plots are given in **Fig. (5.1)**. For Region 1, a scatter plot between l_1 (at site mean of AMPF) and ARMS shows that a U-shaped relationship exists between dependent and independent variables. Moreover, for Region 2, Region 3 and Region 4 no such pattern is observed between l_1 and ARMS.

Based on the aforementioned details, for Region 1, it is appropriate to use a quadratic form of ARMS as an explanatory variable for the regression model. For Region 2, Region 3 and Region 4 the linear form of ARMS has been used within the regression models.

Plots of Region 1 and Region 2 show that Few observations in the data series do not follow the usual pattern of the data and create high scatter within the data. In classical regression modelling, such observations (outliers) create problems of estimation and in such a situation it is very difficult to fulfil the critical assumptions of classical regression (normality and constant variance of the error term). Therefore, for Region 1 and Region 2 robust estimation methods are the obvious choice for the model estimation. Due to the existence of high scatter between the independent and dependent variable, the M-estimation method (Huber, 1973) for Region 1 and the S-estimation method for Region 2 has been used.

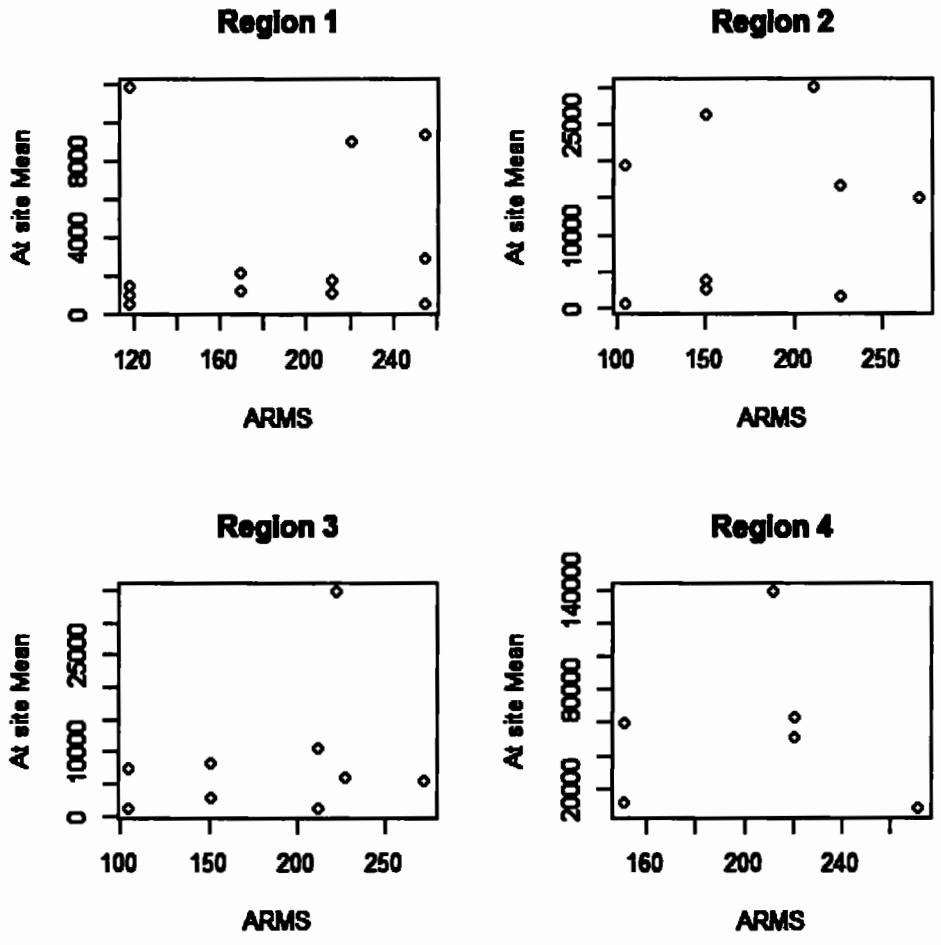


Fig. 5.1: Scatter plots for four homogeneous regions between At-site mean and ARMS.

Thus the fitted regression models for Region 1 and Region 2 based on the aforementioned details are.

Model for Region 1

$$\hat{l}_{1R1} = 9220.21 - 104.20\, ARMS_{R1} + 0.3107\, (ARMS_{R1})^2 \quad (5.1)$$

Model for Region 2

$$\hat{l}_{1R2} = 40.5682\, ARMS_{R2} \quad (5.2)$$

Values of the coefficients, standard errors of estimates, t-calculated (to validate the statistical significance of the provided coefficients) and their corresponding p-values for the model of Region 1 and Region 2 are given in **Table (5.5)** and **(5.6)**.

Table 5.5: Estimated values of the fitted regression model for Region 1, coefficients and their corresponding standard errors (S.E.), t-values and P-values.

S. No.	Independent variables	Coefficients	S.E.	t-value	P-value
1	$ARMS_{R1}$	−104.20	32.01	-3.2554	0.001
2	$(ARMS_{R1})^2$	0.3107	0.0719	4.3186	0.000

Note: ARMS (Average rainfall in monsoon)

Table 5.6: Estimated values of the fitted regression model for Region 2, coefficients and their corresponding standard errors (S.E.), t-values and P-value.

S. No.	Independent variables	Coefficients	S.E.	t-value	P-value
1	$ARMS_{R2}$	40.5682	14.9308	2.7172	0.006

The results show that the coefficients of the fitted regression models in Eq. (5.1) and Eq. (5.2) are highly significant at 5% level of significance. The quadratic term of the model (Eq. (5.1)) have a positive impact on flood flows when rainfall during the monsoon season is increased from its average value. The value of adjusted R_w^2 for model given in Eq. (5.1) is 0.89 and 0.47 for the model given in Eq. (5.2).

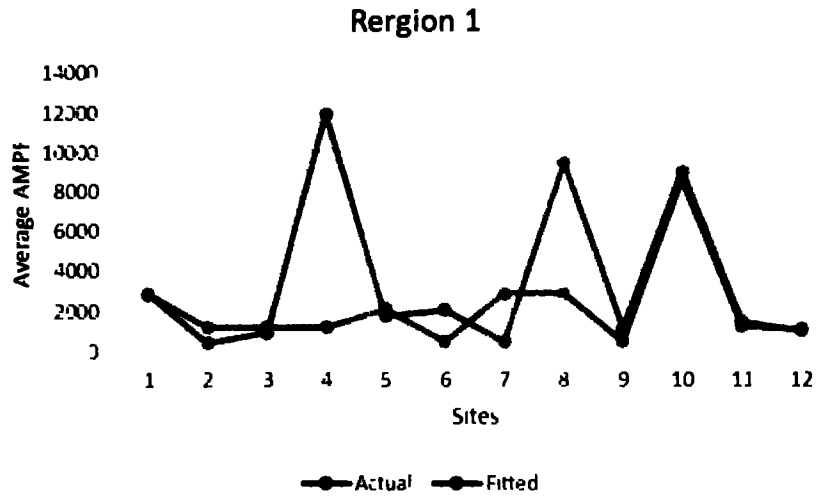


Fig. 5.2: Comparison of fitted and original values of l_1 estimated through QR and M-estimation method for Region 1.

Fig. 5.3: Comparison of fitted and original values of l_1 estimated through linear regression and the S-estimation method for Region 2.

Comparison of predicted values through Eq. (5.1), (5.2) and their corresponding original values of dependent variable have been given in Fig. (5.2), (5.3).

For Region 3 and Region 4, log-transformed linear models with the OLS estimation method are developed. Both regression models for Region 3 and Region 2 are given below:

Model for Region 3

$$ln(\hat{l}_{1R3}) = 1.6702\ ln(ARMS_{R3}) \quad (5.3)$$

Model for Region 4

$$ln(\hat{l}_{1R4}) = 1.9748\ ln(ARMS_{R4}) \quad (5.4)$$

The intercept term is not included in both models of Region 1 and Region 2 as being statistically insignificant (at 5% level of significance), high standard error and practically insignificant, i.e., there is supposed to be no flood in the region with the value of ARMS as zero (the floods in Pakistan are usually dependent on the monsoon rainfall (Hussain and Pasha, 2009). For the estimated model in Eq. (5.3) the value of R^2 (coefficient of determination) is 0.9931 and adjusted-R^2 is 0.9921. For the fitted model given in Eq. (5.4) the value of R^2 is 0.9862 and adjusted- R^2 is 0.9801. This show that the linear regression lines are given in Eq (5.3), (5.4) fits the data of Region 3 and Region 4 sites adequately. The estimated regression coefficient, standard error of the estimate, t-calculated and its corresponding p-value of models Eq. (5.3), (5.4) are given in **Table (5.7), (5.8)**. Results of **Table (5.7), (5.8)** show that the estimated regression coefficients of Eq. (5.3), (5.4) models are statistically significant with low standard errors.

Table 5.7: Results of the fitted regression model in Eq. (5.3).

Independent variable	Coefficient	S.E.	t-value	P-value
$ln(ARMS_{R3})$	1.6702	0.0525	31.81	0.0000

Table 5.8: Results of the fitted regression model in Eq. (5.4).

Independent variable	Coefficients	S.E.	t-value	P-value
$\ln(ARMS_{R4})$	1.9748	0.0978	20.1756	0.0000

These details show that the fitted models in Eq. (5.3) and (5.4) are adequate, still, assumptions related to the error term (normality, zero mean and homoscedasticity) are requisite (for details, see Gujarati, 2003). To check these assumptions, Jarque-Bera test with the null hypothesis that "the residuals follow normal distribution" has been applied for residuals of fitted models in Eq. (5.3) and (5.4). The calculated value of the Jarque-Bera test statistic for the Eq. (5.3) model's residuals is 0.2725 with its corresponding p-value as 0.8725 and for Eq. (5.4) value of the Jarque-Bera test statistic is 0.7480, and P-value is 0.6879. As the p-value exceeds 5% level of significance for Eq. (5.3), (5.4) models residuals; therefore, we are unable to reject the null hypothesis for Eq. (5.3), (5.4) models, hence, we can conclude that the error terms of Eq. (5.3), (5.4) models follow the normal distribution. To check for the homoscedasticity of the error term, White's Test for heteroscedasticity has been applied under the null hypothesis that the "variances for the errors are equal". For a model of Eq. (5.3), the corresponding test statistic for White's test is W = 0.0260 with the corresponding p-value as 0.88 and for Eq. (5.4) the value for White's test is W = 1.3378, and its P-value is 0.2519. This shows that we are unable to reject that the residuals of Eq. (5.3) and Eq. (5.4) are homoscedastic. All these details show that the estimated regression models in Eq. (5.3), (5.4) are an adequate fit. Therefore, the fitted models in Eq. (5.3), (5.4) can be used to predict l_1 for each ungauged site within the homogeneous Region 3 and Region 4 respectively.

For Region 1, Region 2, Region 3 and Region 4 estimated values of l_1 through different methods of regression have been given in **Table (5.9)**. Comparison of predicted values through Eq. (5.3), (5.4) and their corresponding original values of dependent variable have been given in **Fig. (5.4), (5.5).**

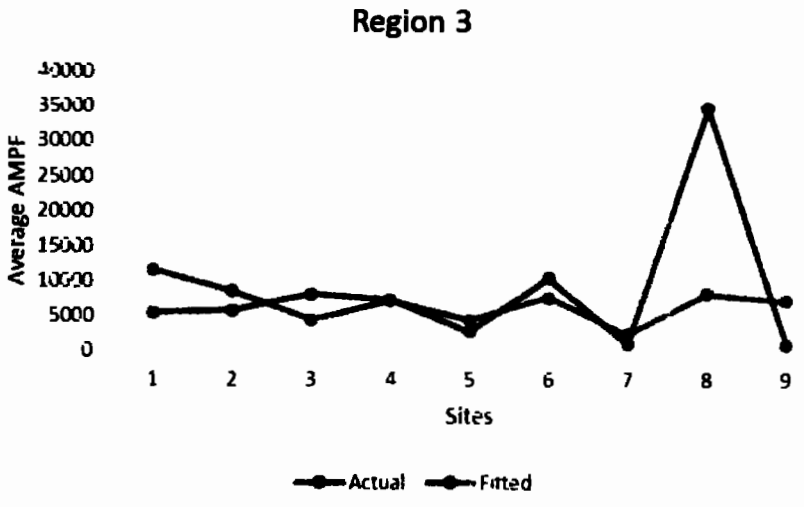


Fig. 5.4: Comparison of fitted and original values of l_1 estimated through linear regression and the OLS method for Region 3.

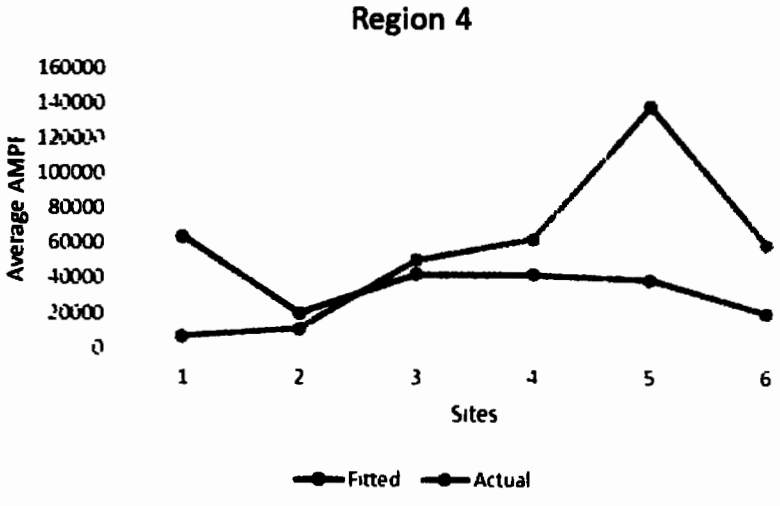


Fig. 5.5: Comparison of fitted and original Values of l_1 estimated through linear regression and the OLS method for Region 4.

Table 5.9: Fitted and original values of l_1 through differnet methods of regssion for Region 1, Region 2, Region 3 and Region 4.

Region 1			Region 2			Region 3			Region 4		
Site Name	l_1	l_1Fitted	Site Name	l_1	l_1Fitted	Site Name	l_1	l_1Fitted	Site Name	l_1	l_1Fitted
Kalpani Deheri	2857	2855	Budni	14810	11034	Naranji	5447	11648	Badri	7229	64259
Wazir Ghari	426	1221	Hakim Gahri	3714	6125	Bagiari	5767	8611	Jundi River	11060	20099
Chinkar	922	1221	Katlongi	2397	6125	Dallus	8197	4359	Swat Khawazakhela	50835	42643
Bara Tarnab	11884	1221	Shahban	1516	9208	Shah Alam	7343	7204	Swat Munda Head	62731	42643
Khuderzai	1758	2111	Dagi	391	4259	Shahi Bala	2792	4359	Kabul Nowshera	138871	39281
Jundi Utmanzai	2053	486	Muqam	16669	9208	Chprial	10480	7682	Swat Khaili	59534	20099
Lund Khwar East	484	2855	Panjkora	26272	6125	Garandi	1005	2376			
Kalpani Saidabad	9409	2855	Kabul Adezai	30028	8600	Kalpani Raisalpur	34773	8297			
Jundi Tangi	1105	486	Kabul Naguman	19227	4259	Jani Khwar	985	7204			
Swat Ningolai	8934	8473									
Bara Kohat Road	1453	1221									
Chillah	1030	1094									

5.2.2 Back-propagation neural network (BPNN) model

The primary objective of training ANN is to reduce the error among the target output and ANN output through adjusting weights. The "caret" package of R-language has been used for the training of BPNN. To select the best BPNN model, different combinations of hidden layers and neurons have been observed against the MSE of observed and fitted mean values ($\hat{l}_1$) of the Region 1, Region 2, Region 3 and Region 4. For each homogeneous region (Region 1, Region 2, Region 3 and Region 4) the model with minimum MSE relative to other models has been selected for the prediction of at-site mean values (l_1). BPNN algorithm with two hidden layers, five neurons in the first layer and three in the second layer, have been used. Few of the published studies have also developed such BPNN model with only two input variables and three hidden layers for the estimation of floods, for instance, Aziz et al. (2014). To avoid over fitting of the model to ensure the quality of the developed BPNN model, testing MSE and training MSE have been compared. Training of BPNN has been terminated for an observed increase in the test MSE or even a decrease in the training MSE.

Six input and one output variables have been used for the prediction of the dependent variable (average value of the observed AMPF at various sites (l_1)). Site characteristics, such as, "Lat", "Long", "Ele", "AARF", "ARMS" and "AAT" have been used as input variables of the BPNN model. The variable l_1 has been used as the dependent variable of the model. The functional relationship of the dependent and independent variables is given as

$$f(l_1) = g(Lat, Long, Ele, AARF, ARMS, AAT) \quad (5.5)$$

Predicted values of $l_1$by using BPNN methods for Region 1, Region 2, Region 3 and Region 4 are provided in **Table (5.10)** and in **Fig. (5.6) (5.7) (5.8)** and **(5.9)**. The results of **Table (5.10)** show that in Region 1, the sites having a larger magnitude of AMPF

(Bara Tarnab, Kalpani Saidabad and Swat Ningolai) BPNN model closely predicted the values of l_1. Moreover, in Region 1 BPNN model under and overestimate the values of l_1 for the sites having a smaller magnitude of AMPF, larger variation and skewness within the data series. Similar findings observed in the results of Region 2 and Region 3. The results of Region 4 show that the BPNN model closely estimated the values of l_1 for all sites in the region. The reason is that the magnitude of AMPF of all sites included in Region 4 is larger and variation within the data sets is small as compared to Region 1, Region 2 and Region 3.

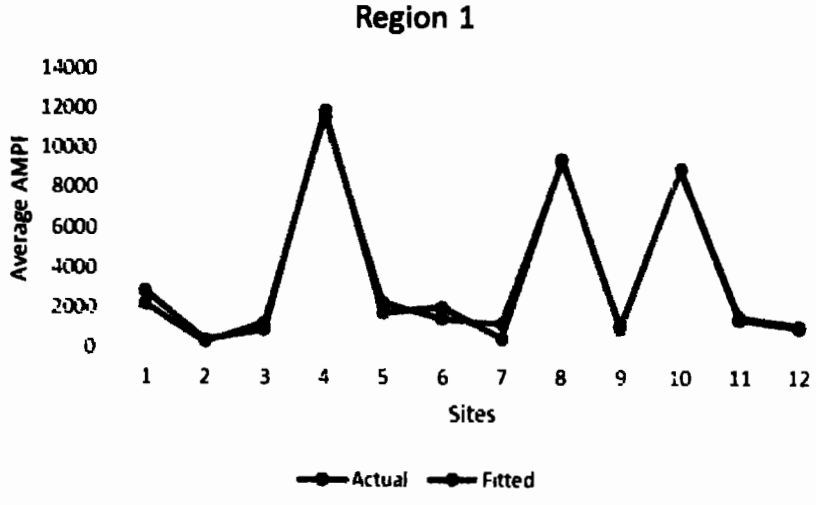


Fig. 5.6: Comparison between fitted values estimated through BPNN and observed values of l_1 for Region 1

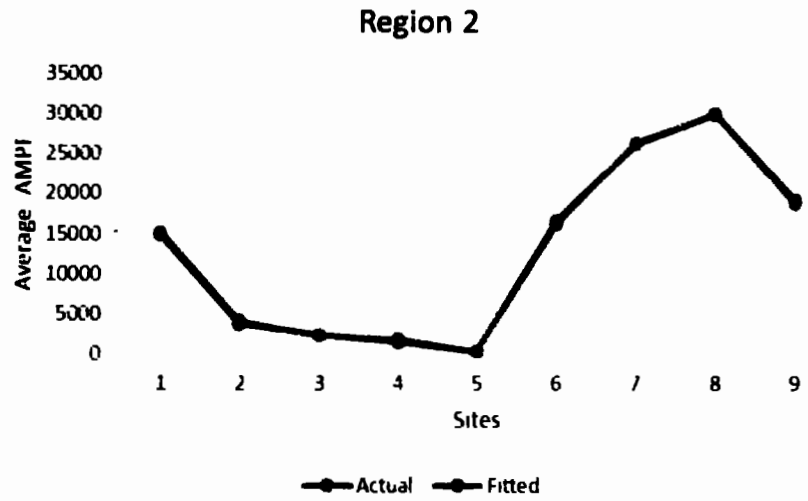


Fig. 5.7: Comparison between fitted values estimated through BPNN and observed values of l_i for Region 2.

Fig. 5.8: Comparison between fitted values estimated through BPNN and observed values of l_i for Region 3.

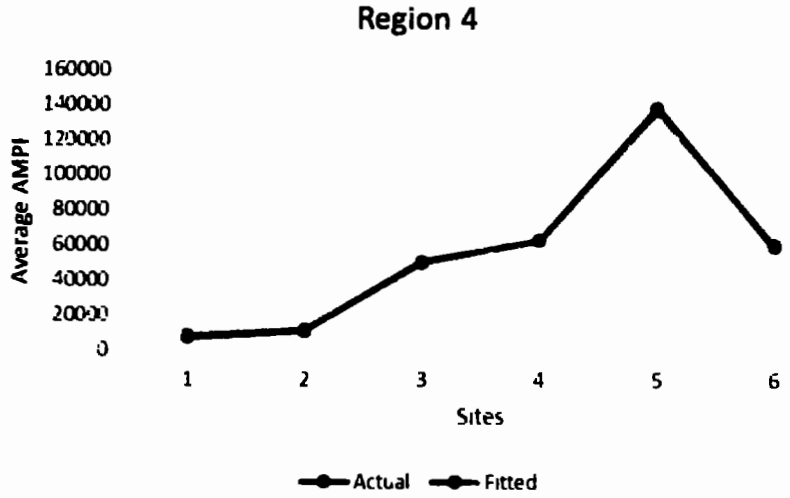


Fig. 5.9: Comparison between fitted values estimated through BPNN and observed values of l_1 for Region 4.

Table 5.10: Fitted and original values of l_1 through BPNN for Region 1, Region 2, Region 3 and Region 4.

Region 1			Region 2			Region 3			Region 4		
Site Name	l_1	l_1Fitted	Site Name	l_1	l_1Fitted	Site Name	l_1	l_1Fitted	Site Name	l_1	l_1Fitted
Kalpani Deheri	2857	2187	Budni	14810	15202	Naranji	5447	4721	Badri	7229	8433
Wazir Ghari	426	327	Hakim Gahri	3714	4094	Bagiari	5767	6249	Jundi River	11060	12009
Chinkar	922	1259	Katlongi	2397	2324	Dallus	8197	8337	Swat Khawazakhela	50835	50299
Bara Tarnab	11884	11568	Shahban	1516	1861	Shah Alam	7343	7372	Swat Munda Head	62731	62986
Khuderzai	1758	2296	Dagi	391	296	Shahi Bala	2792	2566	Kabul Nowshera	138871	137049
Jundi Utmanzai	2053	1456	Muqam	16669	16265	Chprial	10480	10492	Swat Khaili	59534	60766
Lund Khwar East	484	1224	Panjkora	26272	26470	Garandi	1005	2336			
Kalpani Saidabad	9409	9356	Kabul Adezai	30028	29898	Kalpani Raisalpur	34773	34140			
Jundi Tangi	1105	917	Kabul Naguman	19227	18798	Jani Khwar	985	524			
Swat Ningolai	8934	8989									
Bara Kohat Road	1453	1602									
Chillah	1030	1145									

5.2.3 Radial base function (RBF)

RBF method gives accurate estimates in flood prediction modeling. This method has been used in various studies for short term stream flow forecasting, for example, Kagoda et al., 2010; Uysal, 2016 and Sahoo et al., 2019. Therefore, in this study, RBF method is used for ungauged flood quantile estimation in Region 1, Region 2, Region 3 and Region 4. The details of the adopted procedure are:

For the application of the RBF, the dependent and independent variables are rescaled and their standardized form has been used for the training of the model for the four regions. A random partition of 70% and 30% have been used for the training and testing of the model. There are six units (independent variables) in the input layer. The hidden layer has the same number of units as in the input layer and the output layer containing only one unit. In this analysis Gaussian function has been used as a link function between hidden and input layers. Sum of squares error and relative error are used for the model evaluation criteria's. Model summary of the training and testing phases for each region are provided in **Table (5.11)**. A graphical comparison of the fitted values against observed values of the dependent variable is illustrated in **Fig. (5.10), (5.11), (5.12), (5.13)**. Furthermore, predicted values of dependent variable l_1 of each region given in **Table (5.12)**. The results of **Table (5.12)** and **Fig. (5.10), (5.11), (5.12), (5.13)** show that the RBF model more accurately predicts the values of l_1 for Region 1 and Region 4. Results of Region 2 shows large variation in the estimates for the sites "Dagi" and "Kabul Naguman". The results of Region 3 show that predicted values of l_1 significantly varies from their original values for the first four sites "Naranji", "Bagiari", "Dallus" and "Shah Alam" and after that predicted values accurately follow the pattern of original values. Therefore, RBF estimates can be coupled with regional

quantiles of the respective region for the estimation of T-years quantiles at the ungauged site.

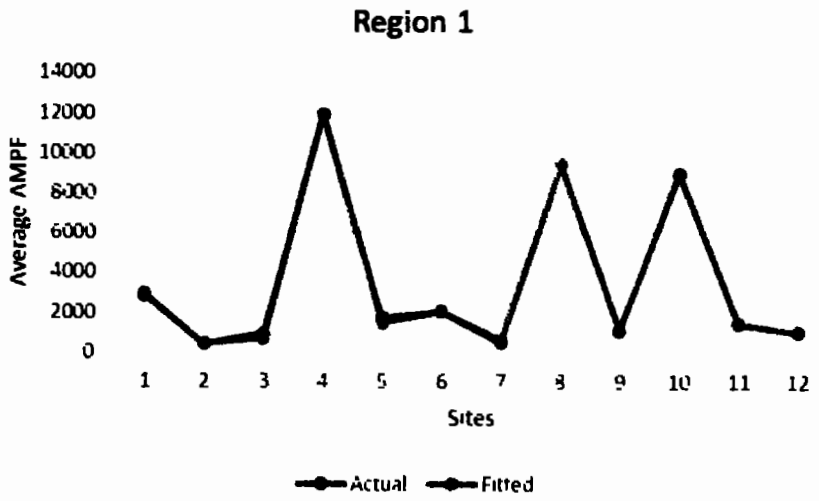


Fig. 5.10: Comparison between fitted values estimated through RBF and observed values of l_1 for Region 1.

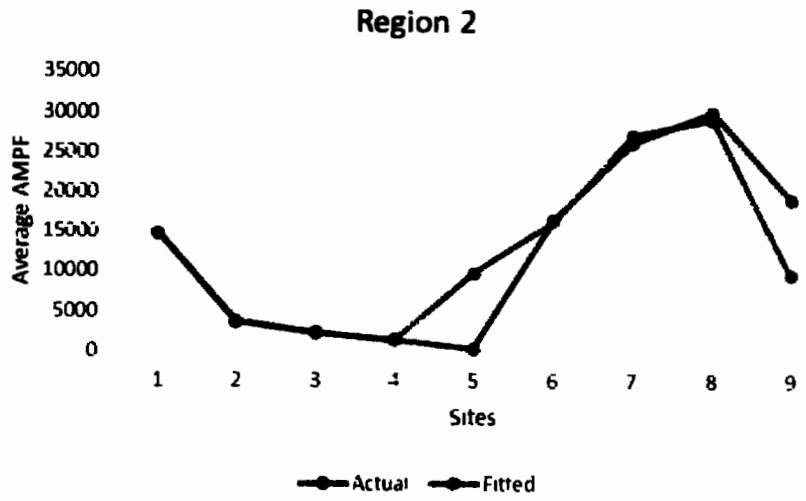


Fig. 5.11: Comparison between fitted values estimated through RBF and observed values of l_1 for Region 2.

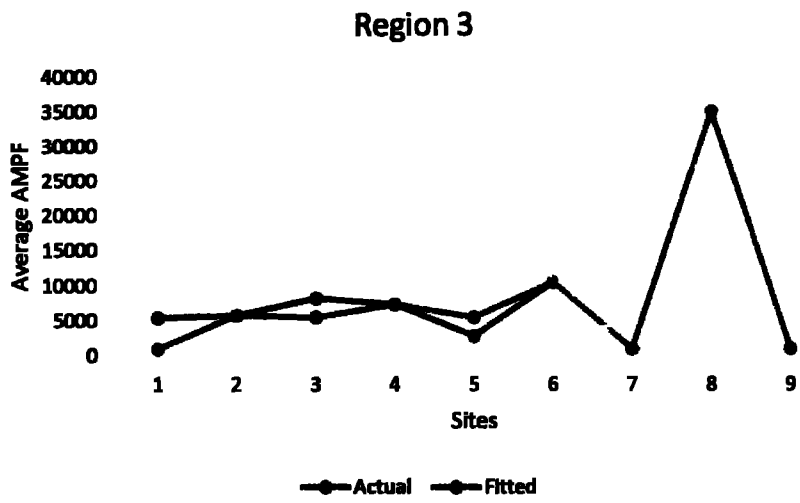


Fig. 5.12: Comparison between fitted values estimated through RBF and observed values of l_j for Region 3.

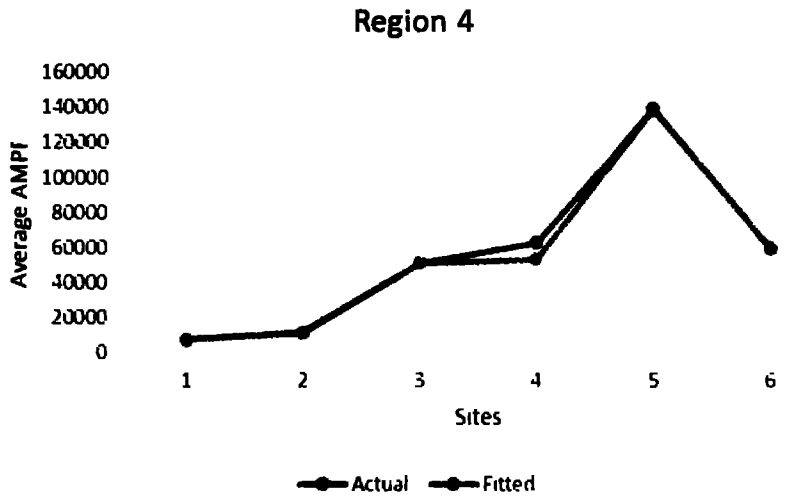


Fig. 5.13: Comparison between fitted values estimated through RBF and observed values of l_j for Region 4.

Table 5.11: Model summaries of RBF during the training and testing phase of each region.

Model Summary		Region 1	Region 2	Region 3	Region 4
Training	Sum of Squares Error	0.033	1.045	0.052	0.012
	Relative Error	0.008	0.298	0.017	0.006
Testing	Sum of Squares Error	0.035	0.681	0.079	0.015
	Relative Error	0.477	0.573	1.762	1.541

Table 5.12: Observed and fitted values of l_1 estimated through RBF of four homogeneous regions.

Region 1			Region 2			Region 3			Region 4		
Site Name	l_1	l_1 Fitted	Site Name	l_1	l_1 Fitted	Site Name	l_1	l_1 Fitted	Site Name	l_1	l_1 Fitted
Kalpani Deheri	2857	2968	Budni	14810	14810	Naranji	5447	985	Badri	7229	7229
Wazir Ghari	426	426	Hakim Gahri	3714	3817	Bagiari	5767	5767	Jundi River	11060	11260
Chinkar	922	698	Katlongi	2397	2374	Dallus	8197	5495	Swat Khawazakhela	50835	50835
Bara Tarnab	11884	11884	Shahban	1516	1519	Shah Alam	7343	7343	Swat Munda Head	62731	53009
Khuderzai	1758	1473	Dagi	391	9861	Shahi Bala	2792	5495	Kabul Nowshera	138871	138671
Jundi Utmanzai	2053	2060	Muqam	16669	16432	Chprial	10480	10480	Swat Khaili	59534	59434
Lund Khwar East	484	655	Panjkora	26272	27180	Garandi	1005	985			
Kalpani Saidabad	9409	9405	Kabul Adezai	30028	29136	Kalpani Raisalpur	34773	34773			
Jundi Tangi	1105	1105	Kabul Naguman	19227	9657	Jani Khwar	985	985			
Swat Ningolai	8934	8934									
Bara Kohat Road	1453	1453									
Chillah	1030	1030									

5.3 Assessment analysis of RBF, BPNN and regression methods

The three estimation methods (RBF, BPNN and regression) have been used for T-year flood quantiles estimation at ungauged sites within Region 1, Region 2, Region 3 and Region 4. The developed models based on the three methods (BPNN, RBF and regression) are theoretically justified for each region, but still, an assessment analysis is required to check the relative accuracy of estimated flood quantiles. To do so, R^2, RMSE and MAPE have been calculated using the information of **Table (5.9), (5.10), (5.12)** and the results are given in **Table (5.13)**. Their corresponding formulas are:

$$R^2 = 1 - \frac{\Sigma(l_1 - \hat{l}_1)^2}{\Sigma(l_1 - \bar{l}_1)^2} \quad (5.6)$$

$$RMSE = \sqrt{\frac{\Sigma(l_1 - \hat{l}_1)^2}{n}} \quad (5.7)$$

$$MAPE = \left(\frac{1}{n}\sum\frac{|l_1 - \hat{l}_1|}{l_1}\right) \times 100 \quad (5.8)$$

Table (5.13) shows that the values of two evaluation metrics, R^2 and MAPE, for Region 1 confirms that RBF estimates are more reliable than BPNN and regression. In comparison to RBF and regression analysis, the values of three assessment parameters in **Table (5.13)** indicate that the BPNN approach provides accurate and reliable estimates for Region 2, Region 3, and Region 4. Therefore, RBF can be the preferred choice to predict the flood quantiles for ungauged and poorly gauged sites of Region 1 and BPNN for Region 2, Region 3 and Region 4.

Table 5.13: Accuracy measures of three estimation methods for Region 1, Region 2, Region 3 and Region 4.

Estimation Method	Region 1			Region 2		
	R^2	RMSE	MAPE	R^2	RMSE	MAPE
Regression	0.0841	4681.15	87.34	0.2396	16693.67	223.69
BPNN	0.9892	3.17	28.17	0.9991	61.33	7.64
RBF	0.9990	64.66	6.67	0.8189	79.33	275.94
	Region 3			Region 4		
Regression	0.0519	5016.33	126.51	0.3467	41329.42	176.12
BPNN	0.9962	17.33	23.67	0.9993	523.37	5.01
RBF	0.9597	1493.67	23.74	0.9916	4009.81	2.93

5.4 Practical validation of estimated quantiles

The estimates of quantiles for various return periods through RFFA for gauged and ungauged sites are statistically sound but their practical validation is still required. A comparison of quantiles estimates obtained using regression model, BPNN and RBF of 15, 50 and 100-year return periods have been performed with highest values of observed/historic AMPF (first and second as per their order of magnitude along with their year of occurrence) for all sites results have been illustrated in **Table (5.14), (5.15), (5.16), (5.17).**

The results of Region 1 given in **Table (5.14)** show that the highest values of observed AMPFs have closely been estimated through RBF for the sites Chinkar, Bara Tarnab, Khuderzai and Jundi Utmanzai at 100-year return period quantile. A similar comparison is seen for the sites Wazir Ghari, Kalpani Saidabad and at Swat Ningolai on 50-year quantile and at the sites Lund Khwar East on 15-year return period quantiles. The RBF

estimates for sites Kalpani Deheri and Chillah at 15-year return period and site Jundi Tangi at 50-year quantile estimate are comparable to the second-highest value of observed AMPF. Therefore, based on the above discussion, for Region 1, the RBF model gives more practical results as compared to BPNN and QR model.

Results of Region 2 given in **Table (5.15)** show that the RBF model closely predicts the highest values of AMPF of sites Hakim Ghari and Katlongi at a 50-year quantile. The second highest value is estimated through RBF at 15-year quantile for the sites Budni and Kabul Adezai. BPNN model gives a close estimate to the highest values of sites Shahban at 15-year quantile, for site Muqam at 50-year quantile and site Kabul Naguman at 100-year quantile. The second highest value of site Dagi is also estimated at 100-year quantile through BPNN. Therefore, for Region 2 BPNN and RBF models give more reliable estimates as compared to regression model for the ungauged sites.

For Region 3, results given in **Table (5.16)** illustrate that the BPNN model gives comparable estimates of quantiles with the highest values of AMPF for all sites except two sites Garandi and Jani Khwar. Moreover, the comparison of the estimated quantiles using RBF for Region 3 show that the estimates are accurate and close to the highest values of AMPF for all the sites. For Region 4, the results of RBF and BPNN provided in **Table (5.17)** show similar behavior for the estimation of quantiles for all sites. RBF estimates for the 100-year return period are closer to the highest values of AMPF as compared to BPNN. Therefore, the RBF model is the better choice for the estimation of the ungauged quantile within Region 3 and Region 4.

Results of **Table (5.16) (5.17)** show that the predicted quantiles through OLS regression for smaller return period (15 and 50 years) are comparable with the highest values of observed AMPF for sites Naranji, Bagiari, Shah Alam, Shahi Bala, Chprial and Garandi sites of Region 3. Similar results has been observed for the sites Badri and Jundi River

of Region 4. A notable point is that the OLS regression analysis provides reasonably close estimates of flood quantiles within the span of the observed data. The estimated quantiles using OLS regression for longer return periods (100 years) or outside the available span of the data, show large deviations from the highest values of observed AMPF for all the sites. This is a major disadvantage of using OLS regression for estimating flood quantiles on longer return periods or beyond the span of the observed data series.

Table 5.14: Practical validation of estimated flood quantiles through RBF, BPNN and regression methods for Region 1.

S. No.	Site name	Highest values of observed AMPF (year)		Estimated quantiles through QR			Estimated quantiles through BPNN			Estimated quantiles through RBF		
		1st Highest	2nd Highest	15	50	100	15	50	100	15	50	100
1	Kalpani Deheri	10845 (1995)	9380 (1997)	8838	19753	29303	6771	15134	22451	9187	20534	30461
2	Wazir Ghari	3080 (1998)	2784 (1982)	3778	8445	12528	1012	2261	3355	1320	2951	4377
3	Chinkar	7500 (2010)	5625 (2011)	3778	8445	12528	3898	8712	12923	2160	4827	7161
4	Bara Tarnab	159100(2013)	20300 (2005)	3778	8445	12528	35811	80042	118736	36789	82228	121979
5	Khuderzai	13952 (1984)	9000 (1997)	6534	14604	21663	7107	15886	23566	4559	10191	15117
6	Jundi Utmanzai	19433 (2010)	11360 (2008)	1506	3365	4992	4509	10078	14949	6377	14253	21144
7	Lund Khwar East	1891 (1997)	1735 (2006)	8838	19753	29303	3789	8469	12563	2027	4532	6722
8	Kalpani Saidabad	52912 (2010)	44321 (2013)	8838	19753	29303	28962	64735	96030	29114	65074	96532
9	Jundi Tangi	18679 (2008)	8382 (2010)	1506	3365	4992	2839	6346	9413	3420	7643	11338
10	Swat Ningolai	52098 (2016)	44870 (2013)	26230	58627	86969	27827	62197	92264	27655	61813	91695
11	Bara Kohat Road	11698 (2010)	5688 (1991)	3778	8445	12528	4960	11086	16445	4498	10053	14914
12	Chillah	23620 (1979)	1800 (2007)	3387	7571	11231	3545	7923	11753	3187	7125	10569

Table 1.15: Practical validation of estimated flood quantiles through RBF, BPNN and regression methods for Region 2.

S. No.	Site name	Highest values of observed AMPF (year)		Estimated quantiles through LR			Estimated quantiles through BPNN			Estimated quantiles through RBF		
		1st Highest	2nd Highest	15	50	100	15	50	100	15	50	100
1	Budni	123000 (2008)	30000 (2010)	26791	36313	41314	36912	50031	56922	35961	48742	55456
2	Hakim Gahri	10650 (1983)	8662 (2010)	14873	20159	22935	9940	13473	15329	9268	12563	14293
3	Katlongi	8060 (1988)	6510 (2002)	14873	20159	22935	5643	7649	8702	5765	7814	8891
4	Shahban	4309 (1995)	3290 (1994)	22358	30305	34479	4519	6125	6968	3687	4998	5686
5	Dagi	1980 (1999)	1237 (2010)	10342	14018	15948	718	973	1107	23943	32453	36923
6	Muqam	45000 (2010)	43215 (2009)	22358	30305	34479	39494	53531	60905	39899	54080	61529
7	Panjkora	130936 (2010)	41774 (2005)	14873	20159	22935	64271	87114	99112	65995	89452	101772
8	Kabul Adezai	80700 (2010)	70700 (2013)	20881	28302	32201	72595	98396	111949	70744	95888	109096
9	Kabul Naguman	75575 (2010)	46425 (1990)	10342	14018	15948	45643	61865	70386	23449	31783	36161

Table 5.16: Practical validation of estimated flood quantiles through RBF, BPNN and regression methods for Region 3.

S. No.	Site name	Highest values of observed AMPF (year)		Estimated quantiles through LR			Estimated quantiles through BPNN			Estimated quantiles through RBF		
		1st Highest	2nd Highest	15	50	100	15	50	100	15	50	100
1	Naranji	30000 (2010)	15704 (1997)	28931	39133	44424	11726	15860	18005	2446	3309	3756
2	Bagiari	16688 (2006)	16023 (2010)	21388	28930	32841	15522	20995	23834	14324	19375	21995
3	Dallus	21700 (2010)	19984 (2006)	10827	14644	16625	20708	28010	31798	13648	18460	20956
4	Shah Alam	20000 (2010)	18513 (2005)	17893	24203	27475	18310	24766	28115	18239	24670	28006
5	Shahi Bala	8911 (1995)	7427 (1996)	10827	14644	16625	6374	8622	9788	13648	18460	20956
6	Chprial	33836 (1993)	24639 (2003)	19081	25808	29298	26061	35250	40016	26030	35208	39969
7	Garandi	3869 (2003)	2600 (1984)	5902	7982	9062	5802	7848	8909	2446	3309	3756
8	Kalpani Raisalpur	118606 (2010)	80615 (2008)	20608	27875	31644	84796	114696	130206	86370	116825	132622
9	Jani Khwar	3424 (1988)	2655 (1991)	17893	24203	27475	1302	1760	1998	2447	3309	3757

Table 5.17: Practical validation of estimated flood quantiles through RBF, BPNN and regression methods for Region 4.

S.#	Site name	Highest values of observed AMPF (year)		Estimated quantiles through LR			Estimated quantiles through BPNN			Estimated quantiles through RBF		
		1st Highest	2nd Highest	15	50	100	15	50	100	15	50	100
1	Badri	16968 (1997)	12735 (1970)	122497	173145	209227	16077	22724	27459	13781	19479	23538
2	Jundi River	42000 (2010)	27800 (2008)	38315	54157	65442	22894	32359	39103	21465	30340	36663
3	Swat Khawazakhela	175546 (2010)	160958 (1985)	81290	114901	138845	95885	135531	163774	96906	136974	165518
4	Swat Munda Head	350000 (2010)	119500 (1965)	81290	114901	138845	120071	169717	205084	101051	142833	172597
5	Kabul Nowshera	450000 (2010)	169600 (2005)	74882	105844	127900	261257	369279	446232	264348	373648	451512
6	Swat Khaili	360000 (2010)	85000 (2006)	38315	54157	65442	115837	163733	197853	113299	160146	193518

Summary

In this chapter, for ungauged flood quantiles estimation within four homogeneous regions, machine learning (RBF and BPNN) and regression models with robust and OLS estimation methods have been used to develop a functional relationship between l_1 values and their respective site characteristics. The key findings of this chapter are given below.

I. For Region 1 Region 2, Region 3 and Region 4 results show that the estimates obtained using machine learning (RBF and BPNN) methods are better as compared to regression analysis. RBF is the preferred method for ungauged flood estimation in Region 1. Moreover, BPNN is more suitable for flood quantiles forecasting at ungauged and poorly/partially gauged sites in Region 2, Region 3 and Region 4.

II. A comparison has been demonstrated using the first and second highest values of the observed AMPF at all the gauging sites located in Region 1, Region 2, Region 3, and Region 4 to determine the functional validity of the given estimates for various return periods, using RBF, BPNN, and regression methods. Most of the sites in Region 1 have RBF estimates that are closer to the highest observed values of AMPF (though only for gauged sites). Therefore, RBF is the reliable method for ungauged flood estimation in Region 1. The estimated quantiles from BPNN at the several sites of Region 2, Region 3, and Region 4, for small to large return period, are closer to the largest and second high value of AMPF. Therefore, BPNN would be a preferred approach as compared to regression models and RBF for flood quantiles estimation at ungauged sites (especially for longer return periods) within the homogeneous Region 2, 3 and 4.

The results of this chapter will not only assist the officials dealing with flood risks management but will also be useful for the management of agricultural water and the

design capacity of existing and proposed hydrologic structures in the study region. For instance, a proposed project of the provincial government of KPK is the site Bara Dam (an ungauged site of the study area located in Region 1).The RBF model of Region 1 can be used successfully for flood quantiles estimation at proposed Bara Dam site by including the values of site characteristics in the model. These findings may also be useful in improving the consistency of quantiles at the study area's poorly gauged sites.

Chapter 6

Choice of Estimation Methods in at-Site Frequency Analysis Using Pearson Type-3 distribution

6.1 Introduction

Univariate modeling using probability distribution(s) also known as at-site frequency analysis is a popular area of research. Fitting a probability distribution to a series of values generated from a random process can provide accurate and reliable estimates of quantiles if modeled properly. Probabilistic modeling could be a challenge, especially when dealing with extreme values because of their non-normal or skewed behavior and the availability of a limited span of data series at a site. There exists a variety of probability distributions for modeling a random variable with different methods of estimation. Success depends on the type, trends and tendencies, and shape of the variable under study along with the available size of the sample including many others. Details of dealing with extreme values in the area of statistical hydrology, meteorology and wind are available in Beirlant et al., (2004) and Naghettini (2017). A review of different methods of estimation of parameters in extreme values analyses (EVA) and few recommendations for best practices have been discussed in Palutikof et al., 1999 and Arns et al., 2013. In EVA, the choice of a method for parameters estimation is not the only source of variation for considering a distribution. Few stochastic uncertainties that belong to a specific method/model for the estimation of quantiles of T-year return period are a selection of threshold in a peak over threshold method, selection of block size (daily, monthly, quarterly, yearly) for maxima and minima analysis (Coles et al., 2001).

In EVA, the dealing variable usually consists of minima's or maxima's (hourly, weekly, monthly, quarterly or yearly) or peaks over a threshold (POT). Each series of values have certain advantages or disadvantages associated with it. The series of annual maxima's with a record length of at least twenty years usually provide accurate estimates of T-year quantiles as compared to POT (Cook 1985; Palutikof et al., 1999; Ferreira and de Haan 2015). The variable in the focus of this study will be annual maxima's as being a common variable in frequency analysis of extreme events. Major limitations attached with annual maxima's include the availability of a limited number of observations (span of the data series) and skewed shape of the distribution.

Pearson Type III (PE3) distribution is an important distribution in EVA as illustrated in several case studies around the world including IACWD, 1982; Chang and Moore 1983; Hussain et al., 2017; Asquith et al., 2017; Li et al., 2017; Lei et al., 2018. Its inclusion in a set of five candidate distributions in the famous methodology of regional frequency analysis proposed by Hosking and Wallis (1997) is also advocating its significance and adequacy to models extreme values.

For fitting PE3 distribution, different methods of estimation have been proposed. For example, Song and Ding (1988) suggested using the probability weighted moments method while Hosking and Wallis (1997) provided L-moments (LM) estimators of its parameters. A comparison of maximum likelihood estimation (MLE) and the method of moments is illustrated in Bobee and Ashkar (1991). Koutrouvelis and Canavos (1999) proposed mixed moment estimators for PE3 distribution using exponentially transformed data. Jan and Shabri (2017) provided a comparison of L-moments and TL-moments estimators of PE3 distribution for river flow predictions in Johor, Malaysia. The results of these mentioned studies reveal the following important facts that there does not exist a universal method of estimation of parameters of PE3 distribution and

the choice of the estimation method is a critical factor, which depends heavily on the size and shape characteristics of the sample. Therefore, this study is designed to compare the performance of three estimation methods (MLE, LM and Maximum product of spacing (MPS)) for fitting PE3 distribution using simulations and empirical analyses by varying size and shape characteristics of the sample. A concise discussion on the use of these three methods is provided below:

In EVA, the efficiency of estimates using MLE (especially quantiles of T-year return period) is linked to the size of the sample (Katz et al. 2002). To overcome this issue, LM derived by Hosking (1990) is a preferred choice. A large number of studies have used LM in regional frequency analysis while modeling annual maxima's, for instance (Lee and Kim, 2019; Vivekanandan, 2015; Drissia et al., 2019; Hussain, 2017; Khan et al., 2019; Rutkowska, 2018).

Another, relatively less common choice, especially with a small sample size is MPS. Few studies like Wong and Li, 2006; Singh et al., 2014; Kumar Singh et al., 2016; Murage et al., 2019 and El-Sherpieny et al., 2020 showed that the method of MPS is a better choice of estimation than traditional methods. Asquith et al. (2017) reported that MPS and LM methods provide nearly identical parameter estimates of PE3 distribution. The study of Soukissian and Tsalis, (2015) illustrated that estimate using the MPS method are better in terms of bias, mean square error and variance of the estimated parameters for modeling extreme wind speed. But none of the studies so far has provided a comparison of LM, MLE and MPS methods for estimation of PE3 distribution.

This study is designed to check the performance of LM, MLE and MPS estimation methods for fitting PE3 distribution. A two-step approach is adopted. In the first step, simulation experiments have been performed by introducing linear variation in the

shape parameter of the distribution (while selecting standard values of the location and scale parameters) for different sample sizes. These variations will help in the assessment of the estimation methods concerning different combinations of size and shape characteristics of the sample. The second step deals with empirical analysis, i.e. estimating flood quantiles considering annuals maxima's of peak flows (AMPF) of four sites of Khyber Pakhtunkhwa (KPK), Pakistan. These sites have been selected keeping in view the sample size, scale and shape of the distribution associated with the observed data series. The findings of this study will provide useful information for the choice of PE3 distribution with an adequate estimation method in EVA.

6.2 Maximum product of spacing estimates of PE3 distribution

PE3 distribution also known as generalized gamma distribution is an important probability distribution in EVA. For a random variable Y having PE3 distribution, its density function is

$$f(Y) = \frac{1}{a\Gamma(b)}\left(\frac{y-\varepsilon}{a}\right)^{b-1} e^{-\left(\frac{y-\varepsilon}{a}\right)} \qquad (6.1)$$

Where a, b and ε are scale, shape and location parameters respectively. If $y \geq \varepsilon$ and $a > 0$, the shape of PE3 distribution is positively skewed. If $y \leq \varepsilon$ and $a < 0$, its shape is negatively skewed.

The key standardizations of a random variable "Y" having PE3 distribution with (a, b, ε) are given as:

$$Z = \frac{Y-\varepsilon}{a}, \qquad K = \frac{Y-\varepsilon-ab}{ab^{1/2}} \qquad (6.2)$$

Here the random variable Z has Gamma distribution with one parameter which is equal to the PE3 (1, b, 0), and "K" is the frequency factor with mean "zero", variance "one" and skewness "$2b^{1/2}$".

The relationship of Eq. (2.35) can be used to obtain the estimates of parameters of PE3 distribution using the MPS method. By using the pdf given in Eq. (6.1), the following equation is obtained.

$$K_{opt}(a,b,\varepsilon) = \frac{1}{n+1}\sum_{i=1}^{n+1} log\left[\int_{y_{i-1}}^{y_i}\left(\frac{1}{a\Gamma(b)}\left(\frac{y-\varepsilon}{a}\right)^{b-1}e^{-\left(\frac{y-\varepsilon}{a}\right)}\right)dy\right] \qquad (6.3)$$

Estimates of parameters a, b and ε can be obtained by maximizing the MPS log estimator is given in Eq. (6.3). A Closed-form solution of Eq. (6.3) is not available. Therefore, non-linear optimization is used to obtain the numerical solutions of MPS estimators.

6.3. Simulation experiments

The analyses include two steps for evaluating the adequacy of three estimation methods for fitting PE3 distribution. The first step includes simulation experiments while the second step is based on empirical analysis using AMPF of four sites of KPK. In the first step, 1000 repetitions for different sample sizes, i.e. small, moderate and large (20, 40, 75 and 100) from PE3 distribution have been generated in each case of the estimation of parameters. The standardized form of PE3 distribution has been used as recommended by Wang 1990; NIST/SEMATECH 2012; Jan and Shabri 2017. For standard form location and scale parameters are set equals to zero and one respectively, the values of shape parameters varies arbitrarily. If the value of shape parameter increases than the skewness of PE3 distribution increases and its density curve become flatter with low kurtosis. Root mean square error (RMSE) and bias have been used as accurate measures of the estimates. A brief description of different steps in simulation experiments is as follows:

1) Random samples of size 20, 40, 75 and 100 have been generated, 1000 times, by setting the values of parameters of PE3 distribution as $\varepsilon = 0$, a=1 and b. A

linear variation in the shape parameter "b" has been introduced with a unit difference for the range of 1.5 to 6.5. This variation will help in assessing the performance of the three estimation methods at different levels of skewness.

2) For each case of 1000 samples, parameters of PE3 distribution have been estimated through LM, MLE and MPS. In few cases, MLE and MPS fail to produce the estimates of parameters at some repetitions during simulations due to the convergence issue of the optimization algorithm.

3) Vectors of estimated parameters through simulations have been obtained. These 1000 estimates have been used to calculate RMSE and bias associated with each parameter using the following expressions:

$$bias = E(\hat{\theta}) - \theta \tag{6.2}$$

$$RMSE = \sqrt{Var(\hat{\theta}) + \left(bias(\hat{\theta})\right)^2} \tag{6.3}$$

Where θ is the actual/assumed value of the parameter and $E(\hat{\theta})$ is the expected value of the estimated values of the parameter calculated through simulations. The results of these simulation experiments for sample size 20 and 40 are illustrated in **Table (6.1)** while sample size 75 and 100 are presented in **Table (6.2)**.

For a relatively small sample size, i.e. n = 20, the LM estimation method provides estimates with low bias and RMSE relative to MLE and MPS for location and scale parameters. However, for the shape parameter, the estimates provided by the LM method have more bias and are less efficient. The method of MPS provides more accurate and efficient estimates as the value of the shape parameter increases. Almost similar trends are obvious for a moderate sample size of n=40. In few cases, while dealing with extreme values, especially on an annual scale, a sample size of 40 may be considered as a large sample size.

For a sample size of n=75 and n=100, the LM method provides estimates with less bias for location, scale and shape parameters in comparison to MLE and MPS. However, the estimates of shape parameter are less efficient. The method of MPS is a preferred choice in terms of RMSE, especially when the data exhibits a large value of shape parameter. Another notable point is that the method of MLE provides low values of RMSE in comparison to the LM method for the estimation of shape parameter when the sample size is quite large, i.e. 75 and 100.

In general, the LM method provides estimates with a low bias for location, scale and shape parameters for small, moderate and large sample sizes. Therefore, it can be concluded that LM estimates are relatively stable in terms of bias for estimating the parameters of PE3 distribution. The method of MPS is a preferred choice for estimating the scale parameter of PE3 distribution for all the sample sizes and values of the shape parameters. The efficiency of the MPS method, for estimating the shape parameter, increases with the increase in the value of the shape parameter. The method of MLE provides comparable values of bias and RMSE for relatively large sample sizes, i.e. 75 and 100 and low values of the shape parameter, i.e. for b=1.5.

Table 6.1: Values of bias and RMSE of the parameters estimated through LM, MLE and MPS for sample size 20 and 40.

"b"		n=20								
		LM			MLE			MPS		
		Location (ε)	Scale (a)	Shape (b)	Location (ε)	Scale (a)	Shape (b)	Location (ε)	Scale (a)	Shape (b)
1.5	RMSE	0.2267	0.2453	0.6401	0.2385	0.2759	0.5892	0.2348	0.2835	0.5676
	Bias	-0.0044	0.0053	-0.0951	-0.0319	0.0448	0.2573	0.0589	0.1351	0.0791
2.5	RMSE	0.2151	0.3591	0.8334	0.2136	0.3764	0.4987	0.2457	0.3863	0.5086
	Bias	0.0142	0.0253	-0.0442	0.0449	0.0865	0.0778	0.1090	0.1888	0.0916
3.5	RMSE	0.2033	0.4118	0.5938	0.2494	0.4725	0.5055	0.2699	0.5290	0.3036
	Bias	0.0127	-0.0363	-0.2787	0.0884	0.0634	-0.2925	0.1358	0.1918	-0.2008
4.5	RMSE	0.2065	0.5657	1.4380	0.2280	0.7245	1.3598	0.2238	0.6921	1.3129
	Bias	-0.0045	0.0286	0.1723	0.1184	0.2938	-0.0228	0.0431	0.1046	-0.1640
5.5	RMSE	0.2398	0.9972	2.2015	0.2795	1.1368	1.7119	0.2612	1.1228	1.6181
	Bias	0.0175	0.1727	0.4095	0.1158	0.3938	0.0228	0.0994	0.4063	0.3233
6.5	RMSE	0.2256	1.3475	4.2180	0.2708	1.5568	3.3284	0.2299	1.0058	2.5620
	Bias	0.0208	0.2618	0.9481	0.1144	0.5899	0.4289	0.0712	0.2534	0.0310
n=40										
1.5	RMSE	0.1493	0.1723	0.4968	0.1535	0.1904	0.4018	0.1550	0.1829	0.3600
	Bias	-0.0096	0.0127	0.0115	-0.0176	0.0498	0.1877	0.0302	0.0830	0.0720
2.5	RMSE	0.1623	0.2659	0.5269	0.1671	0.3098	0.3506	0.1918	0.3116	0.2827
	Bias	0.0001	0.0208	0.0244	0.0489	0.0838	0.0405	0.0708	0.1452	0.1191
3.5	RMSE	0.1443	0.2993	0.6944	0.1849	0.3818	0.5432	0.1874	0.3649	0.3550
	Bias	-0.0076	0.0052	0.0263	0.1067	0.1415	-0.1539	0.0444	0.0432	-0.1453
4.5	RMSE	0.1616	0.4659	1.0067	0.2190	0.6303	0.9439	0.1973	0.5771	0.9708
	Bias	0.0140	0.0748	0.1513	0.1364	0.2931	-0.1122	0.0838	0.2640	0.2248
5.5	RMSE	0.1635	0.5347	1.2858	0.2195	0.7056	1.1741	0.2080	0.6684	1.1048
	Bias	0.0093	0.0755	0.2138	0.1029	0.2904	-0.0550	0.0859	0.2516	0.0054
6.5	RMSE	0.1648	0.7520	1.7845	0.2089	1.0445	1.5147	0.1965	0.9100	1.4068
	Bias	0.0042	0.0579	0.1274	0.0788	0.3015	-0.1037	0.0629	0.1939	-0.2521

Table 6.2: Values of bias and RMSE of the parameters estimated through LM, MLE and MPS for sample size 75 and 100.

"b"		n=75								
		LM			MLE			MPS		
		Location (ε)	Scale (a)	Shape (b)	Location (ε)	Scale (a)	Shape (b)	Location (ε)	Scale (a)	Shape (b)
1.5	RMSE	0.1139	0.1291	0.3259	0.1158	0.1325	0.2074	0.1164	0.1296	0.2042
	Bias	-0.0053	0.0067	-0.0018	-0.0059	0.0253	0.1089	0.0191	0.0507	0.0540
2.5	RMSE	0.1111	0.1755	0.3644	0.1367	0.2384	0.2488	0.1330	0.2083	0.1903
	Bias	0.0181	0.0359	0.0251	0.0753	0.1045	0.0146	0.0645	0.1208	0.0884
3.5	RMSE	0.1170	0.2341	0.5089	0.1716	0.3224	0.3743	0.2258	0.4333	0.3202
	Bias	-0.0031	0.0045	0.0136	0.1324	0.1516	-0.2380	0.0982	0.1404	-0.1370
4.5	RMSE	0.1250	0.3137	0.6595	0.1998	0.4653	0.5511	0.1892	0.4633	0.5289
	Bias	0.0026	0.0181	0.0520	0.1494	0.2512	-0.2922	0.1131	0.2386	-0.0667
5.5	RMSE	0.1138	0.3540	0.9044	0.1939	0.5447	0.7813	0.1594	0.4657	0.7790
	Bias	-0.0086	0.0004	0.1259	0.0959	0.2261	-0.1668	0.0632	0.1449	-0.1486
6.5	RMSE	0.1212	0.4422	1.0621	0.1706	0.6020	0.9172	0.1559	0.5477	0.8432
	Bias	0.0023	0.0131	0.0084	0.0878	0.2597	-0.1948	0.0567	0.1483	-0.2618
		n=100								
1.5	RMSE	0.0920	0.1008	0.2729	0.0920	0.1024	0.1666	0.0932	0.1006	0.1613
	Bias	0.0014	0.0067	0.0142	0.0015	0.0184	0.0732	0.0209	0.0384	0.0286
2.5	RMSE	0.1052	0.1648	0.3081	0.1297	0.2154	0.2189	0.1263	0.1888	0.1580
	Bias	0.0069	0.0186	0.0170	0.0655	0.0867	0.0045	0.0425	0.0829	0.0680
3.5	RMSE	0.1010	0.1952	0.4160	0.1545	0.2764	0.3180	0.1823	0.3018	0.2457
	Bias	0.0084	0.0204	0.0156	0.1315	0.1595	-0.1989	0.1393	0.2324	-0.0170
4.5	RMSE	0.1135	0.2831	0.5569	0.1781	0.4107	0.5016	0.1782	0.4240	0.4417
	Bias	0.0014	0.0273	0.1128	0.1344	0.2414	-0.2030	0.1149	0.2410	-0.0657
5.5	RMSE	0.0980	0.3139	0.7489	0.1852	0.5304	0.6599	0.1710	0.4980	0.6548
	Bias	-0.0013	0.0053	0.0270	0.1318	0.3077	-0.2466	0.0858	0.1899	-0.2300
6.5	RMSE	0.1050	0.4071	0.9570	0.1783	0.6251	0.8275	0.1592	0.5582	0.7654
	Bias	0.0029	0.0451	0.1807	0.0980	0.3142	-0.0788	0.0703	0.2156	-0.1191

6.4 Empirical analysis

In the second step, the performance of three estimation methods (LM, MLE and MPS) for fitting PE3 distribution has also been tested using a real-life data set. AMPF in cubic feet per second of four sites in KPK, Pakistan have been used. The secondary data is

provided by the hydrology section of the Irrigation Department of KPK. These sites are selected keeping in view the variations in the sample size, trends and tendencies of scale and shape characteristics (skewness and kurtosis) of the observed data series. None of the published studies so far (to the best of authors' knowledge) performed an at-site frequency analysis of AMPF using the PE3 distribution of these four sites. Geographical coordinates and record length of observed data series at each site is provided in **Table (6.3)**. Few details of the fitting procedure are:

Time series plots of AMPF at each site used in this chapter are illustrated in **Fig. (3.2)**. These graphs show that there exists random variation in the observed data series at each site. Moreover, the occasional occurrence of a large magnitude of a flood is also obvious. To observe the general trends and tendencies of AMPF at each site, few descriptive measures are calculated and presented in **Table (6.4)**. The information reveals that for the four sites, the sample size varies from 21 to 34. Data exhibits variation as shown by standard deviation as a scale statistic. The shape of the data series at four sites is positively skewed with the values of skewness ranging from 1.45 to 3.19 and leptokurtic behavior as kurtosis values are showing more spread ranging from 1.48 to 10.67. These descriptive statistics show that the trends and tendencies of AMPF at the four sites are different from each other, especially in terms of the shape of the distribution. Therefore, the data is suitable to evaluate the effectiveness of different estimation methods for fitting the PE3 distribution.

The estimated parameters along with their RMSE and bias are given in **Table (6.5)**. The two accuracy measures, i.e. RMSE and bias have been calculated using simulation experiments. For this purpose, estimated parameters of each site have been used to generate 1000 random samples from PE3 with a sample size equal to its observed counterpart. Then for each site and generated sample, PE3 distribution is fitted using

LM, MLE and MPS methods. These simulated values (of each parameter) have been used to calculate the RMSE and bias. The results of **Table (6.5)**, for each site, are discussed below:

i. For the site Wazir Ghari having relatively moderate sample size and high skewness and kurtosis values, the estimates of location parameter have nearly comparable RMSE for LM and MPS methods but the LM method is showing less bias. However, for the estimates of scale and shape parameters, the MPS method has obvious fewer values of RMSE while the LM method is showing less bias.

ii. For the site Jundi Utmanzai, having a relatively small sample size but high skewness and kurtosis values, estimates of the MPS method for all the three parameters are showing less RMSE values. The estimates using the LM method are showing less bias; but, interestingly, the estimates of the MPS method are showing comparable bias for the estimation of the shape parameter. Therefore, the performance of the MPS method in the case of small sample size with high skewness and kurtosis is comparable with the LM method while estimating scale and shape parameters in terms of bias.

iii. For the site Bara Kohat Road having a relatively large sample size with moderate skewness and high kurtosis, the LM method provides estimates for location and scale parameters with low bias; however, the bias is comparable for LM and MPS methods for the estimation of the shape parameter. Here MPS method provides estimates with low RMSE for all the three parameters of PE3 distribution.

iv. For the site Shahban having a relatively small sample size and the distribution of observed data series is showing low skewness and kurtosis values, the LM method provides estimates with low bias and RMSE for location and scale parameters.

However, for the estimates of the shape parameter, the MPS method provides the lowest RMSE value in comparison to LM and MLE.

The above discussion reveals that the method of LM generally provides estimates with a low bias for all the parameters of PE3 distribution but the MPS method provides efficient estimates, i.e. having low values of RMSE especially for the scale and shape parameters. The efficiency of the LM method increases for relatively small sample size and small to moderate values of L-skewness and L-kurtosis, for instance like the site Shahban. However, as the severity of skewness and kurtosis increases, the MPS method is more suitable to estimate the scale and shape parameters of PE3 distribution. The results of this application also show that the MLE method is not a reasonable choice of estimation in EVA for fitting PE3 distribution with small to moderate size sample.

For a general numerical assessment of goodness-of-fit of the considered methods, two measures have been used namely standard error of fit (SEF) and Cramer Von-Mises (CVM) test.

The SEF is defined in Kite, 1988 as:

$$SEF = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \vartheta_i)^2}{n - n_p}} \qquad (6.4)$$

Where y_i are the observed values of the data series at each site, ϑ_i are the fitted values generated through fitted distribution, "n" is the sample size and n_p is the number of estimated parameters.

CVM test is another goodness of fit test. In this test, F_0 (a cumulative distribution function (CDF)) is compared with a given F_n (empirical CDF). Suppose $X_1, X_2 \ldots . X_n \sim F$, then a null hypothesis $H_0: F = F_0$ is tested against the two-tailed alternative hypothesis. The test statistics of the CVM test uses squared difference between F_n and F_0.

$$T = n\int\left(F_n(x) - F_0(x)\right)^2 dF_0(x) \qquad (6.5)$$

After simplification of Eq. (6.5), the following expression is used for the computations of T.

$$T = \frac{1}{12n} + \sum_{j=1}^{n}\left(F_0\left(X_{(j)}\right) - \frac{2j-1}{2n}\right)^2 \qquad (6.6)$$

For a small value of T, the null hypothesis holds otherwise rejected. For more details of the CVM test see Csörgő and Faraway (1996).

The values of these goodness-of-fit measures are given in **Table (6.6)**. Comparable performances have been observed for LM and MPS methods. For the sites with relatively large sample size, skewness and kurtosis values, i.e. Wazir Ghari and Bara Kohat Road, the method of MPS is a preferred choice; however, the CVM method favors of LM method for the site Wazir Ghari. For the site Jundi Utmanzai having a relatively small sample size and high skewness and kurtosis values, the MPS method is the preferred choice. For the site Shahban having the smallest sample size with low skewness and kurtosis values, the LM method has the least SEF value and highest corresponding p-value of CVM. These results again show that method of MPS can be a preferred choice of estimation of PE3 distribution for sites having moderate to large sample size and high values of skewness and kurtosis. However, the LM method provides better results for fitting PE3 distribution for a small sample size with moderate skewness and kurtosis associated with the distribution of observed data series.

Flood quantiles for return periods of 20, 50 and 100 years have been estimated using the quantile function of PE3 distribution. These estimated quantiles along with their RMSE and bias are given in **Table (6.7)**. Similar trends are obvious from the values of accuracy measures of the estimated quantiles. In general, the estimates using the MPS method have low RMSE values while with LM method have low bias. The performance

of the LM method is superior for the site having a small sample size and low values of skewness and kurtosis (for the site Shahban).

Another way of measuring the performance of estimation methods in the fitting of a model is the calculation of 95% confidence intervals of the estimated quantiles. R-package "lmomco" provided by Asquith (2020) has been used for the development of 95% confidence intervals for the estimated quantiles. These intervals along with growth curves of the estimated flood quantiles are illustrated in **Fig. (6.1)**. This figure shows that the MPS estimation method provides the shortest 95% confidence interval, especially a stable upper bound in the extreme upper tail of the distribution.

Table 6.3: Geographical coordinates and record length of four sites.

S. No.	Site Name	Latitude (North)	Longitude (East)	Record length in years
1	Wazir Ghari	33.9845	71.7749	1979-2010
2	Jundi Utmanzai	34.0094	71.8328	1987-2011
3	Bara Kohat Road	33.8638	71.5635	1982-2015
4	Shahban	34.0919	72.0391	1987-2007

Table 6.4: Descriptive statistics of AMRD of four sites. Here n is the number of observations, Min and Max are the minimum and maximum values in the data series, Skewness and Kurtosis are moments measures of skewness of kurtosis.

S. No.	Site Name	N	Min	Max	Mean	S. D.	C.V.	Skewness	Kurtosis
1	Wazir Ghari	32	5	3080	427	697.73	163.61	3.16	10.03
2	Jundi Utmanzai	25	35	19433	2053	4407.79	214.74	3.19	10.67
3	Bara Kohat Road	34	24	11698	1453	2456	169.03	2.71	8.58
4	Shahban	21	218	4309	1516	1171.86	77.31	1.45	1.48

Table 6.5: Estimates of parameters of PE3 distribution along with their RMSE and Bias.

Estimation Method		Wazir Ghari			Jundi Utmanzai			Bara Kohat Road			Shahban		
	Parameter	Estimate	RMSE	Bias	Estimate	RMSE	Bias	Estimate	RMSE	Bias	Estimate	RMSE	Bias
LM	Location (ε)	426.47	125.53	-0.35	2052.61	1072	-47.89	1453.00	486.94	1.23	1515.76	253.41	3.10
	Scale (a)	697.43	278.93	11.26	5135.75	3062.8	-36.94	2697.11	1128.06	60.30	1238.83	381.65	29.75
	Shape (b)	3.7573	0.8305	0.011	5.2349	1.3625	0.0500	3.8403	0.8270	0.0204	2.0732	0.7478	-0.0087
MLE	Location (ε)	423.85	136.64	-6.99	2075.69	1259.1	124.23	1453.03	494.27	45.78	1554.06	313.49	28.20
	Scale (a)	773.01	330.7	45.69	5615.70	4079	899.5	2924.01	1184.4	284.6	1405.40	496.5	173.2
	Shape (b)	3.6911	0.8107	0.284	5.5037	1.9993	0.6402	4.0923	0.9343	0.2982	2.1038	0.4968	0.2919
MPS	Location (ε)	517.10	126.73	51.46	2315.12	718.52	-12.35	1481.48	426.01	17.26	1573.92	299.97	96.71
	Scale (a)	678.62	196.89	101.7	4133.00	1529.3	-44.26	2457.26	791.08	44.97	1237.17	408.87	177.92
	Shape (b)	2.6355	0.2954	0.121	3.6242	0.5394	-0.0319	3.3717	0.4810	0.0355	1.6276	0.4529	0.0548

Table 6.6: Values of goodness-of-fit measures for PE3 distribution. Here bold values indicate best fit method.

Estimation Methods	Wazir Ghari			Jundi Utmanzai		
	SEF	CVM	P-value	SEF	CVM	P-value
LM	1009.51	**0.1811**	**0.3082**	6783.48	0.2960	0.1384
MLE	1060.71	0.6939	0.0126	7124.32	0.8168	0.0061
MPS	**996.59**	0.2212	0.2303	**6208.14**	**0.2376**	**0.2053**
	Bara Kohat Road			**Shahban**		
LM	3703.27	0.1009	0.5828	**1793.43**	**0.0689**	**0.7638**
MLE	3859.99	0.2815	0.1525	1917.02	0.1029	0.5748
MPS	**3549.94**	**0.0696**	**0.7570**	1795.14	0.0765	0.7177

Table 6.7: Predicted flood quantiles for various return periods (in years) along with their RMSE and Bias.

Method	Site name	Wazir Ghari			Jundi Utmanzai			Bara Kohat Road			Shahban		
	Return period	20	50	100	20	50	100	20	50	100	20	50	100
LM	Quantile	1785	2696	3427	10946	18972	25714	6681	10250	13119	3993	5150	6028
	RMSE	607.92	1067.47	1329.67	5061.44	10055.32	16705.13	2242.83	4036.10	5450.48	946.56	1438.33	2024.56
	Bias	-29.28	11.14	-17.06	1214.00	-366.20	-1014.00	-164.60	-541.90	-55.46	-29.99	-11.15	-369.40
MLE	Quantile	1934	2934	3735	11522	20516	28136	7035	11044	14293	4367	5690	6695
	RMSE	661.05	1043.29	1647.86	6036.78	12816.27	17888.04	2677.70	4520.81	6561.35	1192.52	1713.46	2327
	Bias	-65.43	-49.62	-352.90	719.20	-844.20	-2752.50	-243.55	-785.70	-1062.00	-447.00	-425.84	-769.2
MPS	Quantile	1883	2607	3167	10418	15709	19934	6351	9369	11761	4005	5025	5785
	RMSE	494.41	790.46	993.25	4669.49	7590.64	7840.88	2068.11	3089.41	4018.14	1006.12	1452.33	1681.82
	Bias	-216.3	-277.4	-401.8	-543.9	168.90	-465.80	-370.60	-577.50	-354.80	-577.80	-606.30	-744

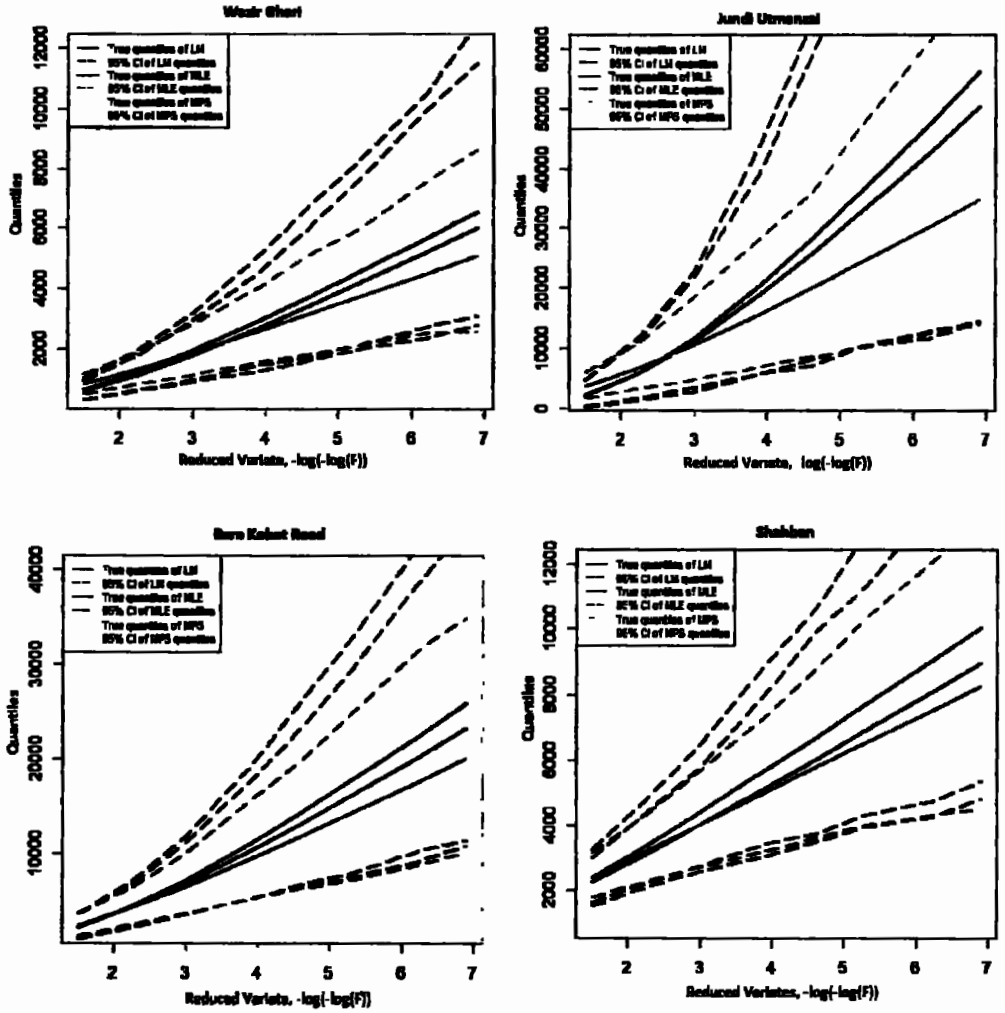


Fig. 6.1: Growth curves of predicted flood quantiles with 95 % confidences intervals.

Summary

The aims of this chapter is to assess the performance of three estimation methods MLE, LM and MPS for fitting PE3 distribution by varying size and shape characteristics of the sample. Two of the three methods (MLE and LM) are quite popular while scarce literature is available with respect to the application of MPS method. The assessment is based on simulations and empirical analyses. The fitting of PE3 distribution has been tested through two goodness-of-fit measures SEF and CVM. RMSE, bias and 95%

confidence intervals of the estimated quantiles have been calculated for assessment analyses of estimation methods. Major findings of the study are:

i. The results of simulation experiments reveal that the estimates using LM method have low bias. The method of MPS is a preferred choice for estimating scale parameter of PE3 distribution for all the considered sample sizes and values of the shape parameters. The efficiency of MPS method, for estimating the shape parameter, increases with increase in the value of shape parameter. The method of MLE provides comparable values of bias and RMSE for relatively large sample sizes, i.e. 75 and 100 and low values of shape parameter, i.e. for b=1.5.

ii. Similar tendencies of assessment measures have been observed through empirical analyses. These include that LM method is a preferred choice in case of small sample size and low or moderate skewness and kurtosis values of the observed data series. Alternatively, MPS method provides efficient estimates for moderate to large sample size and high values of skewness and kurtosis. The stability and efficiency of MPS method is obvious for the estimation of shape parameter of PE3 distribution. Moreover, MLE method is not a preferred method of estimation for fitting PE3 distribution for relatively small to moderate sample size.

Therefore, the study concluded that for fitting PE3 distribution, estimates using LM method have low bias in case of small sample and when data exhibits small to moderate skewness and kurtosis. MPS is a reasonable alternative and provides efficient estimates, especially when the data shows large skewness and kurtosis. MLE is useful in case of very large sample size with low values of shape characteristics of data. The results of this study provide useful guidelines for fitting PE3 distribution to extreme values. These results can be improved in future by considering different probability distributions from

the family of extreme value distributions, sample sizes and variations in terms of location, scale and shape parameters of the distributions.

Chapter 7

Summary and Conclusions

For a developing state with agro centered economy and facing problems of water shortage along with high variations in the stream flows, the importance of precise flood estimates becomes vast, especially for the small streams and rivers which originates with in the country. The results of this study contribute in terms of unique area of study for the application of L-moments based RFA, emphases on the justification of basic assumptions associated to RFA, application of machine learning methods and regression analysis to estimate floods and so on. Moreover, the scenarios in which homogeneous regions cannot be identified an alternate solution then is at site frequency analysis. Quality of quantiles estimates using at-site frequency analysis typically depends on size and span of the sample, distribution characteristics, choice of model and estimation method. This study has also analyzed effectiveness of different estimation methods for fitting Pearson Type 3 distribution in case of at-site frequency analysis through simulations experiments by varying size of sample and shape characteristics. Findings of simulation experiments are validated using real life examples.

The key finding of this study are given below.

- Flood frequency analysis based on the necessary assumption related to data series. Data set of 36 gauging sites has been tested through nonparametric tests and found that data of all sites random, independent, homogeneous and free from significant trends. Therefore, data set is suitable for RFFA and results are reliable for policy

making related to management and efficient utilization of flood water in the study area.

- The estimates of L-moment ratios of all 36 sites showed that there exist deviations in the recorded data series at various sites. However, the L-kurtosis values are comparatively small then the L-skewness values. One possible reason for these fluctuations is the erratic cycles of monsoon rainfall because floods in Pakistan usually rely on the extreme spells of monsoon rainfall. Hussain (2017) found similar results for the sites of river basins in Punjab, Pakistan. The overall shape of the distribution of flood within the study area is flat tope curve with high skewness (heavy upper tail). Therefore, in future, the threats of bigger floods in the study area are very high.
- Larger heterogeneity exists within the group of 36 sites. Therefore, four homogeneous regions have been defined using hierarchical wards clustering method and Euclidean distance to achieve regional homogeneity among the sites. The most relevant site characteristic has been used for the division of gauging sites in to homogeneous regions.
- After identifying the best fitted distribution(s) for Region 1, Region 2, Region 3 and Region 4 a simulation based assessment analysis has been performed to identify robust regional distribution for each region. The results of different accuracy measures and 95% error bounds show that GNO distribution for Region 1, GPA distribution for Region 2 and Region 3, and GLO distribution for Region 4 has been identified robust distributions. These identified divergent regional distributions for each region are indicating dissimilarities in trends, tendencies and shape associated with data series in different areas. Hence delineation of the study area into smaller homogeneous region appears suitable.

- At-sites flood quantiles of each site within the study area are greater than their average flood values and show rising trend. This shows alarming situation related to future flood events. Therefore, serious planning is needed to overcome the damages of future flood disaster.
- Linear/non-linear regression, BPNN and RBF methods have been used to predict site statistic (the average of AMPF) for ungauged sites. This site statistic is used for the estimation of flood quantiles at ungauged sites with in homogeneous region(s). Error evaluation and historical comparison of estimated quantiles with the highest recorded values of AMPF show that RBF model gives efficient estimates for Region 1 and BPNN for Region 2, Region 3 and Region 4.
- In at-site frequency analysis accuracy of estimates strongly depends on the model choice, sample size, shape characteristics of sample data. In this regard, PE3 is selected and its parameters have been estimated using method of LM, MPS and MLE. The accuracy of the estimates has been tested through simulation experiments. The simulation results show that MPS method provides more accurate estimates for the shape parameter of PE3 distribution as compared to LM and MLE. Moreover, the findings show that LM method is a preferred choice in the case when observed sample is small and data series has low or moderate values of shape characteristics (Skewness and Kurtosis). MPS method provides efficient estimates for moderate to large sample size and high values of skewness and kurtosis.

The flood estimates of the study are beneficial for the authorities concerning flood risk management, water resources management, irrigation, planning and development of existing and potential hydraulic structures in the study area. The flood event observed in past and forecasted in this study show that frequency and magnitude of flood increases in future in the study area. Therefore, it is primary need of the time to built

dams and flood protection barriers for the efficient utilization of flood water and to protect the infrastructure and human's life in the study area.

7.1 Future work Recommendations

For future studies, the focus would be to adopt different modelling approaches of analyzing extreme events (like Bayesian approach) by varying estimation methods (like maximum product spacing's). We may also consider various novel techniques of formation of homogeneous regions in RFA. Secondly, the inclusion of the data of more rivers sites that are located in the Potohar region and Koh-e-Suleman mountain range to expand the current study for large data sets. All the rivers/streams located in these areas are known as the part of Indus river basin. Moreover, inclusion of few other site characteristics for the development of models to estimate quantiles at ungauged sites can improve the quality of estimates. Another important area is to perform RFA using variables other than AMPF like 3 days, 5 days or 7 days maxima's to add more data for the application of RFA. Supposedly, it will improve the quality and usefulness of the estimates.

References

Abrahart, R., Kneale, P. E., & See, L. M. (eds.) (2004). Neural networks for hydrological modeling. CRC Press.

Ahmad, I., Abbas, A., Fawad, M., & Saghir, A. (2017a). Regional Frequency Analysis of Annual Total Rainfall in Pakistan Using L-Moments. NUST Journal of Engineering Sciences, 10(1), 19-29.

Ahmad, I., Fawad, M., Akbar, M., Abbas, A., & Zafar, H. (2016). Regional Frequency Analysis of Annual Peak Flows in Pakistan Using Linear Combination of Order Statistics. Polish Journal of Environmental Studies, 25(6).

Ahmad, I., Shah, S. F., Mahmood, I., & Ahmad, Z. (2013). Modeling of monsoon rainfall in Pakistan based on Kappa distribution. Sci. Int. (Lahore), 25(2), 333-336.

Ahmad, I., Yasin, M., Fawad, M., & Saghir, A. (2017b). Regional frequency analysis of low flows using L-Moments for Indus basin, in Pakistan. Pakistan Journal of Science, 69(1), 75.

Alam, J., Muzzammil, M., & Khan, M. K. (2016). Regional flood frequency analysis: comparison of L-moment and conventional approaches for an Indian catchment. ISH Journal of Hydraulic Engineering, 22(3), 247-253.

Allahbakhshian-Farsani, P., Vafakhah, M., Khosravi-Farsani, H., & Hertig, E. (2020). Regional flood frequency analysis through some machine learning models in semi-arid regions. Water Resources Management, 34(9), 2887-2909.

Anilan, T., Satilmis, U., Kankal, M., & Yuksek, O. (2016). Application of Artificial Neural Networks and regression analysis to L-moments based regional frequency analysis in the Eastern Black Sea Basin, Turkey. KSCE Journal of Civil Engineering, 20(5), 2082-2092.

Anilan, T., Uzlu, E., Kankal, M., & Yuksek, O. (2018). The estimation of flood quantiles in ungauged sites using teaching-learning based optimization and artificial bee colony algorithms. Scientia Iranica, 25(2), 632-645.

Arellano-Lara, F., & Escalante-Sandoval, C. A. (2014). Multivariate delineation of rainfall homogeneous regions for estimating quantiles of maximum daily rainfall: A case study of northwestern Mexico. Atmósfera, 27(1), 47-60.

Arns, A., Wahl, T., Haigh, I. D., Jensen, J., & Pattiaratchi, C. (2013). Estimating extreme water level probabilities: A comparison of the direct methods and recommendations for best practise. Coastal Engineering, 81, 51-66.

Asquith, W. H., Kiang, J. E., & Cohn, T. A. (2017). Application of at-site peak-streamflow frequency analyses for very low annual exceedance probabilities (No. 2017-5038). US Geological Survey.

Asquith, W.H., (2020). L-Moments, Censored L-Moments, Trimmed L-Moments, L-Comoments, and Many Distributions. R Package Version 2.3.6, Texas Tech University, Lubbock, Texas. https://cran.r-project.org/web/packages/lmomco/lmomco.pdf

Aydoğan, D., Kankal, M., & Önsoy, H. (2016). Regional flood frequency analysis for Çoruh Basin of Turkey with L-moments approach. Journal of Flood Risk Management, 9(1), 69-86.

Aziz, K., Rahman, A., Fang, G., & Shrestha, S. (2014). Application of artificial neural networks in regional flood frequency analysis: a case study for Australia. Stochastic environmental research and risk assessment, 28(3), 541-554.

Aziz, K., Rahman, A., Shamseldin, A., & Shoaib, M. (2013, December). Regional flood estimation in Australia: Application of gene expression programming and artificial neural network techniques. In Proceedings of the 20th International Congress on Modelling and Simulation, Adelaide, Australia (pp. 1-6).

Batool, Z. (2017). Flood Frequency Analysis of Stream Flow in Pakistan Using L-Moments and TL-Moments. International Journal of Advance Research, Ideas and Innovations in Technology, 3(4), 136-142.

Beirlant, J., Goegebeur, Y., Teugels, J., & Segers, J. (2004). Statistics of extremes. Wiley series in probability and statistics.

Bobee, B., & Ashkar, F. (1991). The Gamma family and derived distributions applied in hydrology (no. GB656. 2. M34. B63 1991).

Bradley, J. V. (1968). Distribution-free statistical tests. No. 04; QA278. 8, B7.

Cassalho, F., Beskow, S., de Mello, C. R., & de Moura, M. M. (2019). Regional flood frequency analysis using L-moments for geographically defined regions: An assessment in Brazil. Journal of Flood Risk Management, 12(2), e12453.

Chang, S. K., & Moore, S. M. (1983). Flood frequency analysis for mall watersheds in southern Illinois. Water Resources Research, 19(2), 277-282.

Chebana, F., Dabo-Niang, S., & Ouarda, T. B. (2012). Exploratory functional flood frequency analysis and outlier detection. Water Resources Research, 48(4).

Cheng, R. C. H., & Amin, N. A. K. (1983). Estimating parameters in continuous univariate distributions with a shifted origin. Journal of the Royal Statistical Society: Series B (Methodological), 45(3), 394-403.

Cohn, T. A., England, J. F., Berenbrock, C. E., Mason, R. R., Stedinger, J. R., & Lamontagne, J. R. (2013). A generalized Grubbs-Beck test statistic for detecting multiple potentially influential low outliers in flood series. Water Resources Research, 49(8), 5047-5058.

Coles, S., Bawa, J., Trenner, L., & Dorazio, P. (2001). An introduction to statistical modeling of extreme values (Vol. 208, p. 208). London: Springer.

Cook, N. J. (1985). The designer's guide to wind loading of building structure's part 1: background. Damage survey, wind data and structural classification building research establishment, Garston and Butterworths London.

Csörgő, S., & Faraway, J. J. (1996). The exact and asymptotic distributions of Cramér-von Mises statistics. Journal of the Royal Statistical Society: Series B (Methodological), 58(1), 221-234.

Cunnane, C. (1988). Methods and merits of regional flood frequency analysis. Journal of Hydrology, 100(1-3), 269-290.

Dawson, C. W., & Wilby, R. L. (2001). Hydrological modelling using artificial neural networks. Progress in physical Geography, 25(1), 80-108.

Dawson, C. W., Abrahart, R. J., Shamseldin, A. Y., & Wilby, R. L. (2006). Flood estimation at ungauged sites using artificial neural networks. Journal of hydrology, 319(1-4), 391-409.

Debele, S. E., Strupczewski, W. G., & Bogdanowicz, E. (2017). A comparison of three approaches to non-stationary flood frequency analysis. Acta Geophysica, 65(4), 863-883.

Development Advocate Pakistan (2017): Water Security in Pakistan: Issues and Challenges 3(4).http://www.pk.undp.org/content/pakistan/en/home/library

Drissia, T. K., Jothiprakash, V., & Anitha, A. B. (2019). Flood Frequency Analysis Using L Moments: a Comparison between At-Site and Regional Approach. Water resources management, 33(3), 1013-1037.

Durocher, M., Burn, D. H., Mostofi Zadeh, S., & Ashkar, F. (2019). Estimating flood quantiles at ungauged sites using nonparametric regression methods with spatial components. Hydrological Sciences Journal, 64(9), 1056-1070.

Ekeu-wei, I. T., Blackburn, G. A., & Pedruco, P. (2018). Infilling missing data in hydrology: Solutions using satellite radar altimetry and multiple imputation for data-sparse regions. Water, 10(10), 1483.

El-Shafie, A., Abdin, A. E., Noureldin, A., & Taha, M. R. (2009). Enhancing inflow forecasting model at Aswan high dam utilizing radial basis neural network and upstream monitoring stations measurements. Water resources management, 23(11), 2289-2315.

El-Sherpieny, E. S. A., Almetwally, E. M., & Muhammed, H. Z. (2020). Progressive Type-II hybrid censored schemes based on maximum product spacing with application to Power Lomax distribution. Physica A: Statistical Mechanics and its Applications, 124251.

England Jr, J. F., Cohn, T. A., Faber, B. A., Stedinger, J. R., Thomas Jr, W. O., Veilleux, A. G., & Mason Jr, R. R. (2019). Guidelines for determining flood flow frequency, Bulletin 17C (No. 4-B5). US Geological Survey.

Fawad, M., Ahmad, I., Nadeem, F. A., Yan, T., & Abbas, A. (2018). Estimation of wind speed using regional frequency analysis based on linear-moments. International Journal of Climatology, 38(12), 4431-4444.

Fawad, M., Yan, T., Chen, L., Huang, K., & Singh, V. P. (2019). Multiparameter probability distributions for at-site frequency analysis of annual maximum wind speed with L-moments for parameter estimation. Energy, 181, 724-737.

Ferreira, A., & De Haan, L. (2015). On the block maxima method in extreme value theory: PWM estimators. Annals of Statistics, 43(1), 276-298.

Fox, J. (2002). Robust regression: An R and S-PLUS companion to applied regression.

Girosi, F., & Poggio, T. (1990). Networks and the best approximation property. Biological cybernetics, 63(3), 169-176.

Government of Pakistan (2017): Annual flood report 2017. Ministry of Water and Power, Office of the Chief Engineer Advisor and Chairman, Federal Flood Commission, Islamabad. [Available at :http://www.ffc.gov.pk/download/AFR/Annual%20Flood%20Report%202016.pdf.]

Government of Pakistan (2018). Annual flood report 2017: Ministry of Water Resources.https://mowr.gov.pk/wp-content/uploads/2018/06/Annual-Flood-Report-of-FFC-2017.pdf

Govindaraju, R. S. (2000). Artificial neural networks in hydrology. II: hydrologic applications. Journal of Hydrologic Engineering, 5(2), 124-137.

GREH, G. D. R. E. H. S. (1996b). Inter-comparison of regional flood frequency procedures for Canadian rivers. Journal of hydrology (Amsterdam), 186(1-4), 85-103.

GREHY, G. D. R. E. S. (1996a). Presentation and review of some methods for regional flood frequency analysis. Journal of hydrology (Amsterdam), 186(1-4), 63-84.

Griffis, V. W., & Stedinger, J. R. (2007). The use of GLS regression in regional hydrologic analyses. Journal of Hydrology, 344(1-2), 82-95.

Grubbs, F. E., & Beck, G. (1972). Extension of sample sizes and percentage points for significance tests of outlying observations. Technometrics, 14(4), 847-854.

Gujarati, D. N. (2003). Basic Econometrics. – McGraw-Hill, New York.

Haddad, K., & Rahman, A. (2020). Regional flood frequency analysis: evaluation of regions in cluster space using support vector regression. Natural Hazards, 102(1), 489-517.

Hailegeorgis, T. T., & Alfredsen, K. (2017). Regional flood frequency analysis and prediction in ungauged basins including estimation of major uncertainties for mid-Norway. Journal of Hydrology: Regional Studies, 9, 104-126.

Ham, F., & Kostanic, I. (2001). Fundamental neurocomputing concepts. Principles of Neuro computing for Science and Engineering. McGraw-Hill.

Hamed, K., & Rao, A. R. (1999). Flood frequency analysis. CRC Press, Boca Raton.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., & Stahel, W. A. (2011). Robust statistics: the approach based on influence functions (Vol. 196). John Wiley & Sons.

Hashmi, H. N., Siddiqui, Q. T. M., Ghumman, A. R., & Kamal, M. A. (2012). A critical analysis of 2010 floods in Pakistan. African Journal of Agricultural Research, 7(7), 1054-1067.

Hassan, M. U., Hayat, O., & Noreen, Z. (2019). Selecting the best probability distribution for at-site flood frequency analysis; a study of Torne River. SN Applied Sciences, 1(12), 1-10.

Hirsch, R. M., Helsel, D. R., Cohn, T. A., & Gilroy, E. J. (1992). Statistical analysis of hydrologic data. In: Maidment, D. R. (ed.) Handbook of Hydrology, Chapter 17. McGraw-Hill, New York.

Hjelmfelt, A. T., & Wang, M. (1996). Predicting runoff using artificial neural networks. In Proceedings of the International Conference on Hydrology and Water Resources, New Delhi, India, December 1993, Springer, Dordrecht, 233-244.

Hosking, J. R. (1990). L-moments: Analysis and estimation of distributions using linear combinations of order statistics. Journal of the Royal Statistical Society: Series B (Methodological), 52(1), 105-124.

Hosking, J. R. M., & Wallis, J. R. (1993). Some statistics useful in regional frequency analysis. Water resources research, 29(2), 271-281.

Hosking, J. R. M., & Wallis, J. R. (1997). Regional frequency analysis: an approach based on L-moments. Cambridge University Press.

Hounkpè, J., Diekkrüger, B., Badou, D. F., & Afouda, A. A. (2015). Non-stationary flood frequency analysis in the Ouémé River Basin, Benin Republic. Hydrology, 2(4), 210-229.

Huber, P. J. (1964). Robust Estimation of a Location Parameter. The Annals of Mathematical Statistics, 35 (1), 73-101.

Hussain, Z. (2011). Application of the regional flood frequency analysis to the upper and lower basins of the Indus River, Pakistan. Water resources management, 25(11), 2797-2822.

Hussain, Z. (2017). Estimation of flood quantiles at gauged and ungauged sites of the four major rivers of Punjab, Pakistan. Natural hazards, 86(1), 107-123.

Hussain, Z., & Pasha, G. R. (2009). Regional flood frequency analysis of the seven sites of Punjab, Pakistan, using L-moments. Water resources management, 23(10), 1917-1933.

Hussain, Z., Shahzad, M. N., & Abbas, K. (2017). Application of regional rainfall frequency analysis on seven sites of Sindh, Pakistan. KSCE Journal of Civil Engineering, 21(5), 1812-1819.

Interagency Advisory Committee on Water Data (IACWD) (1982). Guidelines for determining flood flow frequency: Bulletin 17b of the hydrology subcommittee, office of water data coordination, U.S. geological survey, Reston, Va., 183.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning with Applications in R. New York: Springer.

Jan, N. A. M., & Shabri, A. (2017). Estimating distribution parameters of annual maximum stream flows in Johor, Malaysia using TL-moments approach. Theoretical and applied climatology, 127(1-2), 213-227.

Jingyi, Z., & Hall, M. J. (2004). Regional flood frequency analysis for the Gan-Ming River basin in China. Journal of hydrology, 296(1-4), 98-117.

Kagoda, P. A., Ndiritu, J., Ntuli, C., & Mwaka, B. (2010). Application of radial basis function neural networks to short-term streamflow forecasting. Physics and Chemistry of the Earth, Parts A/B/C, 35, 571-581.

Katz, R. W., Parlange, M. B., & Naveau, P. (2002). Statistics of extremes in hydrology. Advances in water resources, 25(8-12), 1287-1304.

Khan, M. S. R., Hussain, Z., & Ahmad, I. (2019). A comparison of quadratic regression and artificial neural networks for the estimation of quantiles at ungauged sites in regional frequency analysis. Applied Ecology and Environmental Research, 17(3), 6937-6959.

Khan, M. S. R., Hussain, Z., & Ahmad, I. (2020). Regional flood frequency analysis, using l-moments, artificial neural networks and ols regression, of various sites of Khyber-Pakhtunkhwa, Pakistan. Applied Ecology and Environmental Research, 19(1), 471-489.

Khan, M. S. U. R., Hussain, Z., Ahmad, I., & Noor, F. (2021). Modeling of flood extremes using regional frequency analysis of sites of Khyber Pakhtunkhwa, Pakistan. *Journal of Flood Risk Management*, *14*(4), e12751.

Khan, S. A., Hussain, I., Hussain, T., Faisal, M., Muhammad, Y. S., & Mohamd Shoukry, A. (2017). Regional Frequency Analysis of Extremes Precipitation Using L-Moments and Partial L-Moments. Advances in Meteorology.

Kite, G. W. (1988). Frequency and risk analyses in hydrology. Water Resources Publications, Littleton.

Komi, K., Amisigo, B. A., Diekkrüger, B., & Hountondji, F. C. (2016). Regional flood frequency analysis in the Volta River basin, West Africa. Hydrology, 3(1), 5.

Koop, G. (2008). Introduction to Econometrics. John Wiley & Sons, Ltd, West Sussex.

Koutrouvelis, I. A., & Canavos, G. C. (1999). Estimation in the Pearson type 3 distribution. Water resources research, 35(9), 2693-2704.

Kreft, S., Eckstein, D., Dorsch, L., & Fischer, L. (2015). Global climate risk index 2016: who suffers most from extreme weather events? Weather-related loss events in 2014 and 1995 to 2014. Germanwatch Nord-Süd Initiative eV.

Kuhn, M., & Johnson, K. (2013). Applied predictive modeling. Springer, New York.

Kumar Singh, R., Kumar Singh, S., & Singh, U. (2016). Maximum product spacings method for the estimation of parameters of generalized inverted exponential distribution under Progressive Type II Censoring. Journal of Statistics and Management Systems, 19(2), 219-245.

Kumar, R., & Chatterjee, C. (2005). Regional flood frequency analysis using L-moments for North Brahmaputra region of India. Journal of Hydrologic Engineering, 10(1), 1-7.

Landi, A., Piaggi, P., Laurino, M., & Menicucci, D. (2010). Artificial neural networks for nonlinear regression and classification. In Intelligent Systems Design and Applications (ISDA), 10th International Conference IEEE, 115-120.

Lee, D. H., & Kim, N. W. (2019). Regional Flood Frequency Analysis for a Poorly Gauged Basin Using the Simulated Flood Data and L-Moment Method. Water, 11(8), 1717.

Lei, G. J., Yin, J. X., Wang, W. C., & Wang, H. (2018). The Analysis and Improvement of the Fuzzy Weighted Optimum Curve-Fitting Method of Pearson–Type III Distribution. Water Resources Management, 32(14), 4511-4526.

Li, W., Zhou, J., Sun, H., Feng, K., Zhang, H., & Tayyab, M. (2017). Impact of distribution type in Bayes probability flood forecasting. Water Resources Management, 31(3), 961-977.

Lim, Y.E.O.H., & Lye, L.M. (2003). Regional flood estimation for ungauged basins in Sarawak, Malaysia. Hydrological Sciences Journal, 48(1), 79-94.

Lin, G. F., & Chen, L. H. (2004). A non-linear rainfall-runoff model using radial basis function network. Journal of Hydrology, 289(1-4), 1-8.

Lin, G. F., Wu, M. C., Chen, G. R., & Tsai, F. Y. (2009). An RBF-based model with an information processor for forecasting hourly reservoir inflow during typhoons. Hydrological Processes: An International Journal, 23(25), 3598-3609.

Liu, D., Yuan, Y., & Liao, S. (2009). Artificial neural network vs. nonlinear regression for gold content estimation in pyrometallurgy. Expert Systems with Applications, 36(7), 10397-10400.

Maghsood, F. F., Moradi, H., Massah Bavani, A. R., Panahi, M., Berndtsson, R., & Hashemi, H. (2019). Climate change impact on flood frequency and source area in northern Iran under CMIP5 scenarios. Water, 11(2), 273.

Maier, H. R., & Dandy, G. C. (2000). Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications. Environmental modelling & software, 15(1), 101-124.

Malekinezhad, H., & Zare-Garizi, A. (2014). Regional frequency analysis of daily rainfall extremes using L-moments approach. Atmósfera, 27(4), 411-427.

Mesbahzadeh, T., Soleimani Sardoo, F., & Kouhestani, S. (2019). Flood frequency analysis for the Iranian interior deserts using the method of L-moments: A case study in the Loot River Basin. Natural Resource Modeling, 32(2), e12208.

Mosaffaie, J. (2015). Comparison of two methods of regional flood frequency analysis by using L-moments. Water Resources, 42(3), 313-321.

Murage, P., Mung'atu, J., & Odero, E. (2019). Optimal Threshold Determination for the Maximum Product of Spacing Methodology with Ties for Extreme Events. Open Journal of Modelling and Simulation, 7(03), 149.

Naghettini, M. (Ed.). (2017). Fundamentals of statistical hydrology. Cham: Springer International Publishing.

National Electric Power Regulatory Authority (NEPRA) (2018): Hydel Potential in Pakistan.http://www.nepra.org.pk/Policies/Hydel%20Potential%20in%20Pakistan.pdf

NIST/SEMATECH (2012) e-handbook of statistical methods., NIST/SEMATECH 2012. http://www.itl.nist. gov/div898/handbook/. Accessed in Dec 2019

Okoli, K., Breinl, K., Mazzoleni, M., & Di Baldassarre, G. (2019). Design Flood Estimation: Exploring the Potentials and Limitations of Two Alternative Approaches. Water, 11(4), 729.

Ouali, D., Chebana, F., & Ouarda, T. B. (2017). Fully nonlinear statistical and machine-learning approaches for hydrological frequency estimation at ungauged sites. Journal of Advances in Modeling Earth Systems, 9(2), 1292-1306.

Pakistan Meteorological Department (2012). The implementation of diagnostic study for 2010 flood and extreme moon soon rains 2011 in Pakistan under sustainable development through peace building, governance and economic recovery in KP and support landslide IDPs in Hunza Nagar and Gilgit district when UNDP surves as implementing partner. [Available at http://www.pmd.gov.pk/reports/flood_diagnostic_2010_2011.pdf]

Pall, P., Aina, T., Stone, D. A., Stott, P. A., Nozawa, T., Hilberts, A. G., ... & Allen, M. R. (2011). Anthropogenic greenhouse gas contribution to flood risk in England and Wales in autumn 2000. Nature, 470(7334), 382-385.

Palutikof, J. P., Brabson, B. B., Lister, D. H., & Adcock, S. T. (1999). A review of methods to calculate extreme wind speeds. Meteorological Applications: A journal of forecasting, practical applications, training techniques and modelling, 6(2), 119-132.

Peugh, J. L., & Enders, C. K. (2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. Review of educational research, 74(4), 525-556.

Ranneby, B. (1984). The maximum spacing method. An estimation method related to the maximum likelihood method. Scandinavian Journal of Statistics, 93-112.

Rao, A. R., & Hamed, K. H. (2000). The logistic distribution. Flood Frequency Analysis. CRC Press. Boca Raton, Florida, USA, 291-321.

Rao, A. R., & Srinivas, V. V. (2008). Regionalization of watersheds: an approach based on cluster analysis (Vol. 58). Springer Science & Business Media.

Rao, C. R. & Toutenburg, H. (1999). Linear Models: Least Squares and Alternatives, 2 edn, Springer, New York.

Rasheed, A., Egodawatta, P., Goonetilleke, A., & McGree, J. (2019). A novel approach for delineation of homogeneous rainfall regions for water sensitive urban design—a case study in Southeast Queensland. Water, 11(3), 570.

Rousseeuw, P., & Yohai, V. (1984). Robust regression by means of S-estimators. In Robust and nonlinear time series analysis (pp. 256-272). Springer, New York, NY.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1985). Learning internal representations by error propagation (No. ICS-8506). California Univ San Diego La Jolla Inst for Cognitive Science.

Rutkowska, A., Żelazny, M., Kohnová, S., Łyp, M., & Banasik, K. (2018). Regional L-moment-based flood frequency analysis in the Upper Vistula River basin, Poland. In Geo informatics and Atmospheric Science (pp. 243-263). Birkhäuser, Cham.

Saf, B. (2009). Regional flood frequency analysis using L-moments for the West Mediterranean region of Turkey. Water Resources Management, 23(3), 531-551.

Sahoo, A., Samantaray, S., & Ghose, D. K. (2019). Stream Flow Forecasting in Mahanadi River Basin using Artificial Neural Networks. Procedia Computer Science, 157, 168-174.

Satyanarayana, P., & Srinivas, V. V. (2008). Regional frequency analysis of precipitation using large-scale atmospheric variables. *Journal of Geophysical Research: Atmospheres, 113*(D24).

Shahzadi, A., Akhter, A. S., & Saf, B. (2013). Regional frequency analysis of annual maximum rainfall in monsoon region of Pakistan using L-moments. Pakistan Journal of Statistics and Operation Research, 9(1), 111-136.

Shamseldin, A. Y., Nasr, A. E., & O'connor, K. M. (2002). Comparison of different forms of the Multi-layer Feed-Forward Neural Network method used for river flow forecasting. hydrology and earth system sciences, 6(4), 671-684.

Shu, C., & Ouarda, T. B. (2007). Flood frequency analysis at ungauged sites using artificial neural networks in canonical correlation analysis physiographic space. Water Resources Research, 43(7).

Singh, U., Singh, S. K., & Singh, R. K. (2014). A comparative study of traditional estimation methods and maximum product spacings method in generalized inverted exponential distribution. Journal of Statistics Applications & Probability, 3(2), 153.

Sivakumar, B., & Singh, V. P. (2012). Hydrologic system complexity and nonlinear dynamic concepts for a catchment classification framework. Hydrology and Earth System Sciences, 16(11), 4119-4131.

Smith, A., Sampson, C., & Bates, P. (2015). Regional flood frequency analysis at the global scale. Water Resources Research, 51(1), 539-553.

Song, D., & Ding, J. (1988). The application of probability weighted moments in estimating the parameters of the Pearson type three distribution. Journal of hydrology, 101(1-4), 47-61.

Soukissian, T. H., & Tsalis, C. (2015). The effect of the generalized extreme value distribution parameter estimation methods in extreme wind speed prediction. Natural Hazards, 78(3), 1777-1809.

Susanti, Y., & Pratiwi, H. (2014). M-estimation, S-estimation, and MM- estimation in robust regression. International Journal of Pure and Applied Mathematics, 91(3), 349-360.

Todhunter, P. E. (2012). Uncertainty of the assumptions required for estimating the regulatory flood: Red River of the north. Journal of Hydrologic Engineering, 17(9), 1011-1020.

Uysal, G. (2016). Streamflow forecasting using different neural network models with satellite data for a snow dominated region in Turkey. Procedia Engineering, 154, 1185-1192.

Vivekanandan, N. (2015). Flood frequency analysis using method of moments and L-moments of probability distributions. Cogent engineering, 2(1), 1018704.

Wald, A., & Wolfowitz, J. (1943). An exact test for randomness in the non-parametric case based on serial correlation. The Annals of Mathematical Statistics, 14(4), 378-388.

Wang, Q. J. (1990). Estimation of the GEV distribution from censored samples by method of partial probability weighted moments. Journal of Hydrology, 120(1-4), 103-114.

Water and Power Development Authority (WAPDA) (2018).http://www.wapda.gov.pk/index.php/projects/water-sector/future/bara-dam-project

Wong, T. S. T., & Li, W. K. (2006). A note on the estimation of extreme value distributions using maximum product of spacings. In Time Series and Related Topics (pp. 272-283). Institute of Mathematical Statistics.

WRC (1981) Guidelines for determining flood flow frequency, Bulletin 17B (1981). United States Water Resources Council-Hydrology Committee, Washington

Yang, L. (2016). Regional flood frequency analysis for Newfoundland and Labrador using the L-Moments index-flood method (Doctoral dissertation, Memorial University of Newfoundland).

Yang, T., Xu, C. Y., Shao, Q. X., & Chen, X. (2010). Regional flood frequency and spatial patterns analysis in the Pearl River Delta region using L-moments approach. Stochastic Environmental Research and Risk Assessment, 24(2), 165-182.

Zaman, M. A., Rahman, A., & Haddad, K. (2012). Regional flood frequency analysis in arid regions: A case study for Australia. Journal of Hydrology, 475, 74-83.

Zhang, T., Wang, Y., Wang, B., Tan, S., & Feng, P. (2018). Nonstationary flood frequency analysis using univariate and bivariate time-varying models based on GAMLSS. Water, 10(7), 819.

Appendix

A-1

Generalized Logistic Distribution

Parameters: Location (ε), Scale (α), Shape (k)

Range of x is: $-\infty < x \leq \varepsilon + \alpha/k$ if $k > 0$; $-\infty < x < \infty$ if $k = 0$; $\varepsilon + \alpha/k < x \leq \infty$ if $k < 0$.

Probability density function

$$f(x) = \frac{e^{-(1-k)y}}{\alpha(1+e^{-y})^2} \qquad y = \begin{cases} -k^{-1}\log\left(1 - k\left(\frac{x-\varepsilon}{\alpha}\right)\right), & if\ k \neq 0 \\ \left(\frac{x-\varepsilon}{\alpha}\right), & if\ k = 0 \end{cases}$$

CDF

$$F(x) = (1 + e^{-y})^{-1}$$

Generalized Extreme Values Distribution

Parameters: Location (ε), Scale (α), Shape (k)

Range of x is: $-\infty < x \leq \varepsilon + \alpha/k$ if $k > 0$; $-\infty < x < \infty$ if $k = 0$; $\varepsilon + \alpha/k < x \leq \infty$ if $k < 0$.

$$f(x) = \frac{e^{-(1-k)y - e^{-y}}}{\alpha} \qquad y = \begin{cases} -k^{-1}\log\left(1 - k\left(\frac{x-\varepsilon}{\alpha}\right)\right), & if\ k \neq 0 \\ \left(\frac{x-\varepsilon}{\alpha}\right), & if\ k = 0 \end{cases}$$

CDF

$$F(x) = e^{-e^{-y}}$$

Generalized Pareto Distribution

Parameters: Location (ε), Scale (α), Shape (k)

Range of x is: $\varepsilon \leq x \leq \varepsilon + \alpha/k$ if $k > 0$; $\varepsilon \leq x < \infty$ if $k \leq 0$.

$$f(x) = \frac{e^{-(1-k)y}}{\alpha} \qquad y = \begin{cases} -k^{-1}\log\left(1 - k\left(\frac{x-\varepsilon}{\alpha}\right)\right), & if\ k \neq 0 \\ \left(\frac{x-\varepsilon}{\alpha}\right), & if\ k = 0 \end{cases}$$

CDF

$F(x) = 1 - e^{-y}$

Generalized Normal Distribution

Parameters: Location (ε), Scale (α), Shape (k)

Range of x is: $-\infty \le x \le \varepsilon + \alpha/k$ if $k > 0$; $-\infty < x \le \infty$ if $k = 0$; $\varepsilon + \alpha/k \le x \le \infty$ if $k < 0$.

$$f(x) = \frac{\emptyset(y)}{\alpha - k(x-\varepsilon)} \qquad y = \begin{cases} -k^{-1}\log\left(1 - k\left(\frac{x-\varepsilon}{\alpha}\right)\right), & if\ k \neq 0 \\ \left(\frac{x-\varepsilon}{\alpha}\right), & if\ k = 0 \end{cases}$$

$\emptyset$ is the standard normal pdf.

CDF

$F(x) = \Phi(y)$, where Φ is the standard normal CDF.

Pearson Type-3 Distribution

Parameters: Location (ε), Scale (α), Shape (k)

Let $a = \frac{4}{k^2}$, $\beta = \frac{1}{2}\alpha|k|$, and $\mu = \varepsilon - 2\,\alpha/k$

If $k > 0$ than rang of x is $\mu \le x < \infty$ and

$$f(x) = \frac{(x-\mu)^{a-1}e^{-(x-\mu)/\beta}}{\beta^a \Gamma(a)} \qquad F(x) = \frac{G\left(a, \frac{(x-\mu)}{\beta}\right)}{\Gamma(a)}$$

If $k = 0$ than distribution is normal, the range of x is $-\infty < x < \infty$ and

$$f(x) = \emptyset\left(\frac{x-\varepsilon}{\alpha}\right) \qquad F(x) = \Phi\left(\frac{x-\varepsilon}{\alpha}\right)$$

If $k < 0$ than rang of x is $-\infty < x \le \mu$ and

$$f(x) = \frac{(\mu-x)^{a-1}e^{-(\mu-x)/\beta}}{\beta^a \Gamma(a)} \qquad F(x) = 1 - \frac{G\left(a, \frac{(\mu-x)}{\beta}\right)}{\Gamma(a)}$$

A-2

Linear regression

In hydrological regression modelling quantity of interest Y_i of any i site can be written as the linear function of their site characteristics represented through X_i. The equation form of the model is given below.

$$\boldsymbol{Y = XB + \varepsilon} \quad \text{(A2.1)}$$

where $\boldsymbol{X}$ is a matrix of gauging site characteristics of order $(n \times k)$, $\boldsymbol{B}$ is the vector of regression parameters and $\boldsymbol{\varepsilon}$ is the vector of the random error term. The order of $\boldsymbol{B}$ and $\boldsymbol{\varepsilon}$ is $(n \times 1)$.

Polynomial regression

In regression modelling, when dependent and independent variables are not linearly related to each other than non-linear functions (in terms of variables) are used to develop and estimates the model. A polynomial regression model is used when the relationship between independent and dependent variable is curvilinear. A general form of polynomial regression model is given below.

$$Y = \alpha + \sum_{i=1}^{n} \beta_i X^i + \varepsilon \quad \text{(A2.2)}$$

Most of the time, in flood modelling the relationship between flood values and their corresponding site characteristics are nonlinear. In this situation, the curvilinear regression model is used for the estimation of flood values (Khan et al., 2019). Therefore, in this study the first-time quadratic regression is introduced for the estimation of flood quantiles at ungauged sites.

OLS estimation method

For the estimation of regression model parameters, OLS methods of estimation commonly used. OLS estimators for the regression parameters are known as the "best

linear unbiased estimators" (BLUE). OLS estimators for the regression parameters of Eq. (2.26) are given below.

$$\widehat{\boldsymbol{B}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} \tag{A2.3}$$

and

$$Var(\widehat{\boldsymbol{B}}) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1} \tag{A2.4}$$

OLS estimators for $\boldsymbol{B}$ are minimum variance unbiased estimators and does not dependent on $\hat{\sigma}^2$ stated and proven in Gauss-Markov-Aitken theorem see also (Rao and Toutenburg, 1999; Koop, 2005).