Received: 20 September 2020 | Revised: 10 May 2021 | Accepted: 20 July 2021

DOI: 10.1111/jfr3.12751

ORIGINAL ARTICLE

CIWEM Chartered Institution of Water and Environmental Management | Journal of Flood Risk Management | WILEY

Modeling of flood extremes using regional frequency analysis of sites of Khyber Pakhtunkhwa, Pakistan

Muhammad Shafeeq ul Rehman Khan[1] | **Zamir Hussain**[2] | **Ishfaq Ahmad**[1] | **Farzana Noor**[1]

[1]Department of Mathematics and Statistics, International Islamic University H-10, Islamabad, Pakistan

[2]Research Centre for Modelling and Simulation (RCMS), National University of Sciences and Technology (NUST) H-12, Islamabad, Pakistan

Correspondence
Muhammad Shafeeq ul Rehman Khan, Department of Mathematics and Statistics, International Islamic University, H-10 Islamabad, Pakistan.
Email: shafiqnaizi@gmail.com

Abstract

The study provides results of regional frequency analysis (RFA) using annual maximum peak flows (AMPF) of 36 sites located on various streams and rivers of Khyber-Pakhtunkhwa, Pakistan. Assumptions of randomness, independent and identical distribution regarding AMPF at each site have been validated using different statistical tests. The region of 36 sites is heterogeneous as confirmed by L-moments based heterogeneity measure. Therefore, it is subdivided into four homogeneous regions considering the most influential site characteristic among available using wards clustering method and Euclidean distance. To identify good-fit-regional distribution(s), from a set of popular three-parameter distributions, L-moment ratio diagram and $|Z\text{-Dist}|$ statistic are used as goodness-of-fit criteria. To obtain the most suitable distribution having robust properties, a simulation-based assessment analysis is performed for each homogeneous region considering root mean square error and 95% error bounds of regional quantiles as accuracy measures. Due to non-linearity (in the functional relationship between the mean of AMPF at various sites and their corresponding site characteristics) and the existence of multicollinearity between the site characteristics, radial basis function (RBF) network has been used for the estimation of quantiles at ungauged sites. The results show that the adopted methodology is useful for the estimation of quantiles at gauged and ungauged sites within the defined homogeneous regions.

KEYWORDS

L-moments, quantile estimation, radial basis function network, regional frequency analysis, ungauged site, Ward's clustering method

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. The analysis includes both at-site as well as regional approaches with certain advantages/

disadvantages associated with them. At-site frequency analysis may not be a preferred choice in case of a shorter or limited span of observed data series at any site. Additionally, the estimates cannot be interpolated or extrapolated effectively for a neighboring site with no observed record (commonly known as an 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 & Wallis, 1993). In this scenario, regional frequency analysis (RFA) is an optimum choice in which we pool 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). RFA using L-moments is a popular method and has been used in several case studies around the world. For example; in Korea, Lee and Kim (2019); in Canada, Requena et al. (2017); in Norway, Hailegeorgis and Alfredsen (2017); in India, Alam et al. (2016); in China, Yang et al. (2010); in Iran, Mesbahzadeh et al. (2019); in Turkey, Aydoğan et al. (2016); and many more. Two important studies providing inter-comparison of various regional flood estimation procedures are by GREHYS (1996a, 1996b). A brief of the development in RFA has been illustrated in Malekinezhad and Zare-Garizi (2014).

RFA has also been applied in a few of the published studies in Pakistan. These include: for rainfall (Ahmad et al., 2013; Ahmad et al., 2016a; Ahmad et al., 2017a; Hussain et al., 2017; Khan et al., 2017; Shahzadi et al., 2013), for floods (Ahmad et al., 2016b; Ahmad et al., 2017b; Batool, 2017; Hussain, 2011; Hussain, 2017; Hussain & Pasha, 2009), for wind (Fawad et al., 2018; Fawad et al., 2019). Highlights of some published literature concerning flood frequency analysis in Pakistan are provided in the following section:

The study of Hussain and Pasha (2009), perhaps, was the first application of L-moments based RFA in Pakistan. In their study, the focus area was sites of four major rivers of Punjab. The study concluded that Generalized Normal (GNO) distribution is a robust model for AMPF of the region. In another study by Hussain (2011) considering AMPF of sites of Indus River, the results are in favor of Pearson Type-3 (PE3) distribution for sites of the upper half of Indus River while Generalized Logistic (GLO) distribution for sites belongs to the lower half. The study of Ahmad et al. (2017) used 10 days average of low flows considering nine sites of different rivers of Pakistan. Two homogeneous regions were identified. Region-1 consisting of stations Tunsa, Tarbela, Nowshera, and Kalabagh while Region-2 includes Chashma, Guddu, Mangla, and Marala. The best-suited distribution for Region-1 is GNO while for Region-2 is Generalized Pareto (GPA). In another application of RFA using AMPF, Hussain et al. (2017) considered various stations of major rivers of Punjab, Pakistan, namely Ravi, Sutlej, Jhelum, and Chenab. For the two homogeneous regions, PE3 is the most suitable distribution for Region-1 while GNO is the best-fit distribution for Region-2.

The details of these studies reveal that the focused study areas were the sites of Punjab province (also called the land of five rivers) and the Indus River. However, a complete set of sites of various small rivers and streams of Khyber Pakhtunkhwa (the north-western area of the country) have not been considered concerning the application of RFA. Another interesting fact is that most of the rivers and streams of this area originate in Pakistan with natural flows and are very less affected by man-made changes like the construction of barrages and dams; hence, the sites of this area are most suitable to perform RFA. Therefore, this study is designed to apply a standard methodology available in Hosking and Wallis (1997) to a new study area of Pakistan. Moreover, RFA is a preferred methodology to model AMPF relative to Peaks Over a Threshold especially for a limited span of available data (Cook, 1985; Ferreira & De Haan, 2015; Palutikof et al., 1999).

The development of models to estimate floods quantiles at ungauged sites of the study area is another important area of research in RFA. A variety of techniques have been provided in the literature. For example, in few cases, the relationship between the dependent variable (usually mean of observed data series at different sites) and one or more independent variables (site characteristics) is non-linear and complex (Ouali et al., 2017; Sivakumar & Singh, 2012) and is inexpressible in a mathematical/statistical form. A useful consideration for the development of such models is machine learning methods like decision trees, random forests, and artificial neural networks (ANN) (Anilan et al., 2016; Aziz et al., 2014). Among different methods of ANN, the radial basis function (RBF) network is quite popular due to its accuracy to estimate non-linear and complex functions (Ham & Kostanic, 2001). Allahbakhshian-Farsani et al. (2020) also suggested that the sport vector regression model based on RBF provides more reliable estimates of flood quantiles relative to other machine learning methods. In another study by Haddad and Rahman (2020), the RBF method provides more consistent quantile estimates for ungauged sites. A brief of the predictive ability of the RBF network in extreme floods is available in Lin & Chen, 2004; Lin et al., 2009 and El-Shafie et al., 2009. Keeping these in view, this study has also used the RBF network to estimate ungauged sites flood quantiles.

The rest of the paper is organized as follows. Section 2 describes the study area and data utilized for the analysis, Section 3 provides the stepwise methodology of regional frequency analysis, Section 4 explains findings of the study, Section 5 illustrates the development of the RBF model for the estimation of quantiles at ungauged sites, and Section 6 covers the summary and major conclusions of the study.

2 | STUDY AREA AND DATA DESCRIPTION

Pakistan is a devolving country with an agro-based economy and having a long history of floods. Twenty-four major flood events occurred in the country from 1947 to 2016 and the frequency and intensity of these floods is becoming more and more vulnerable for the last few years or so (Government of Pakistan, 2016). On the other hand, due to the lack of reservoirs, a huge quantity of freshwater flowed down to the Arabian Sea. Resultantly, the country faces severe water shortage and we are rapidly becoming a water deficit country (Development Advocate Pakistan, 2016).

KPK Province has an area of 101,741 km^2 and a population of about 35.53 million (as per the population census of 2017 by the Government of Pakistan). Due to its steep geography and mountain land, the heavy rainfall usually turns into flash floods and usually affect the whole of KPK (Pakistan Meteorological Department, 2012). The southern part of the province, due to its downstream location, is the most flood-affected area Hashmi et al. (2012). Therefore, there is a need to predict the magnitude and frequency of these floods to generate flood risk maps, management of stream water and feasibilities/designing of new hydraulic structures for the rivers and streams.

This study has used the AMPF of 36 sites of important rivers/streams of KPK with the observed information spanning from 15 years to 55 years. Secondary data is provided by the flood section of the Irrigation Department of KPK. Few details of the sites along with their respective site characteristics namely longitude (Long), latitude (Lat), elevation (Ele) in meters, average annual rainfall (AARF) in millimeters, average rainfall in monsoon season (ARMS) in millimeters and average annual temperature (AAT) in degree Celsius are given in Table 1. A map showing the locations of 36 sites is given in Figure 1.

3 | METHODOLOGY

3.1 | Regional frequency analysis

Summarized information of different measures of regional frequency analysis is given below:

1. For the identification of discordant site(s) in a region, a measure denoted by D_i is:

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

where $S = \sum_{i=1}^{N} (u_i - \overline{u})\,(u_i - \overline{u})^T$, and $\overline{u} = \frac{\sum_{i=1}^{N} u_i}{N.}$.

u_i contains the estimates of sample L-moments ratios of site i, $\overline{u}$ is unweighted group average and N is the number of sites in the region.

2. A fundamental requirement in RFA is the formation/identification of homogeneous region(s) of the study area, that is, grouping the sites with homogeneous site characteristics. The statistic to check heterogeneity in a group of sites is:

$$H = \frac{V - \mu_v}{\sigma_v} \qquad (2)$$

where $V = \left[\frac{\sum_{i=1}^{N} n_i (t^i - t^R)^2}{\sum_{i=1}^{N} n_i} \right]^{\frac{1}{2}}$. μ_v is mean and σ_v is the standard deviation of computed inter-site variation obtained through simulations, t is the sample L-cv and, $t^R = \frac{\sum_{i=1}^{N} n_i t^{(i)}}{\sum_{i=1}^{N} n_i}$. Desirably, the value of H should be less than 1 for a homogenous region.

3. Two important goodness of fit measures are the L-moments ratio diagram and $|Z\text{-Dist}|$ statistic. L-moment ratio diagram is a graph of L-skewness Vs L-kurtosis, while $|Z\text{-Dist}|$ measure is calculated as:

$$|Z - \text{Dist}| = \frac{\tau_4^{\text{Dist}} - t_4^R + \beta_4}{\sigma_4} \qquad (3)$$

where τ_4^{Dist} is L-kurtosis of the potential frequency distribution, t_4^R is regional L-kurtosis, β_4 is the bias of t_4^R and σ_4 is the *SD*.

4. For the estimation of at-site quantiles using regional quantiles, the following relationship is available:

$$\widehat{Q}_i(F) = l_1^{(i)} \widehat{q}(F) \qquad (4)$$

where $\widehat{Q}_i(F)$ are at-site quantiles for site i, $l_1^{(i)}$ is average of AMPF at site i, and $\widehat{q}(F)$ are regional quantiles for different return periods.

5. To choose a robust distribution from different good fit distributions, a simulations based assessment procedure is available. This procedure leads to 95% error bounds and root mean square error (RMSE) of the estimated quantiles.

The formula of RMSE is:

TABLE 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	34.1650	71.5931	310	460	151	22.7
4	Badri	34.9866	72.3520	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	71.7460	303	559	255	22.2
8	Bagiari	34.2254	72.1543	313	559	227	22.2
9	Katlongi	34.0960	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	34.0140	71.7538	301	400	119	22.7
15	Wazir Gahri	33.9930	71.7460	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	71.5280	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	34.1220	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	34.1140	71.7523	292	384	105	22.7
30	Jundi Utmanzai	34.0099	71.8327	294	460	170	22.5
31	Jundi Tangi	33.8965	72.2350	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

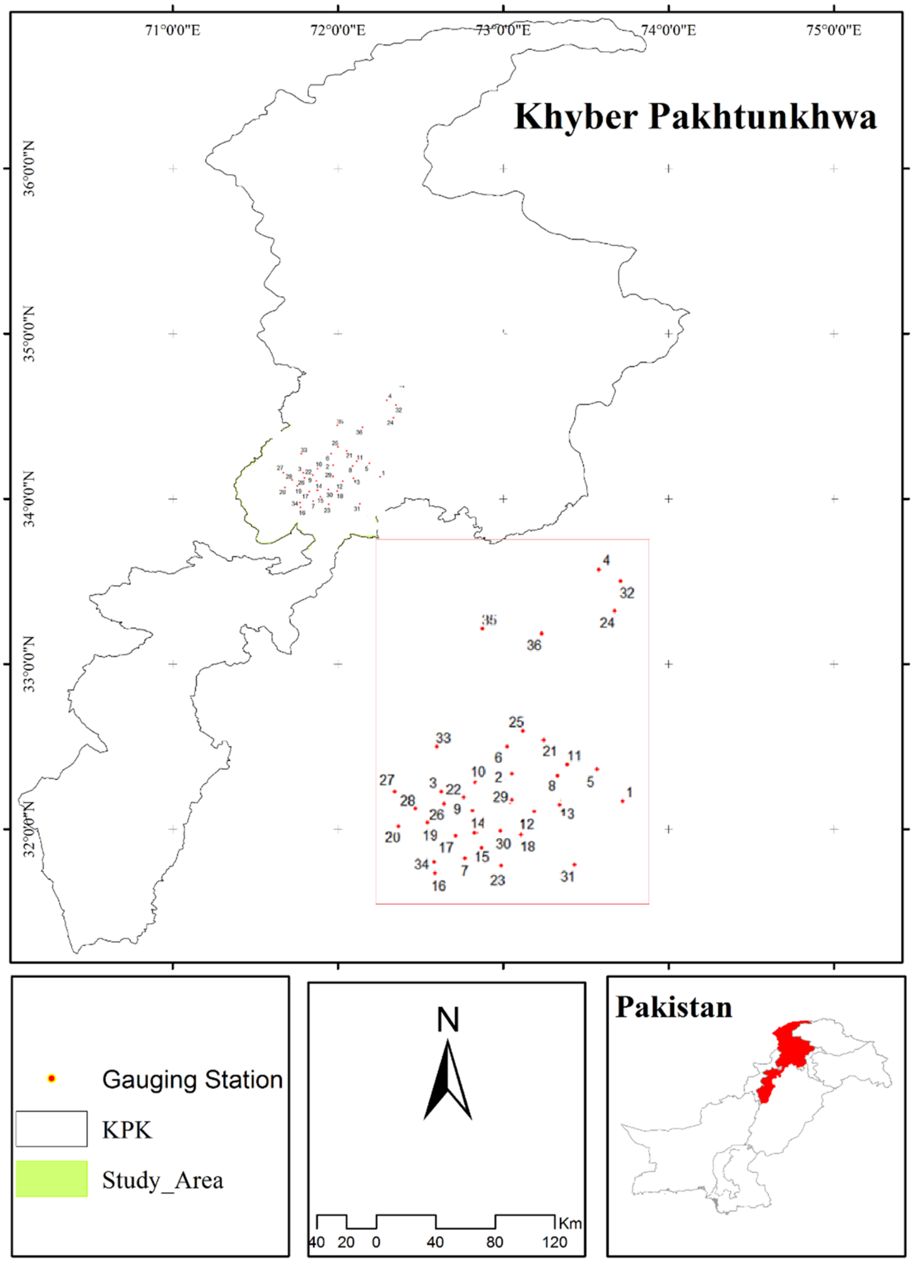
$$R_i(F) = \left[M^{-1} \sum_{m=1}^{M} \left\{ \frac{\widehat{Q}_i^{[m]}(F) - \widehat{Q}_i(F)}{\widehat{Q}_i(F)} \right\}^2 \right]^{½} \quad (5)$$

At *m*th repetition, $\widehat{Q}_i^{[m]}(F)$ is estimated quantile for site-i with a non-exceedance probability *F*. Averaging over the entire region provides:

$$\mathrm{RMSE} = N^{-1} \sum_{i=1}^{N} R_i(F) \quad (6)$$

For the regional growth curves, $\widehat{q}_i^{[m]}(F)$ and $\widehat{q}_i(F)$ can be used instead of $\widehat{Q}_i^{[m]}(F)$ and $\widehat{Q}_i(F)$, respectively. 95% confidence intervals for the growth factors are.

FIGURE 1 Geographical locations of 36 gauging sites of Khyber Pakhtunkhwa



$$\frac{\widehat{q}(F)}{U_{0.025}(F)} \le q(F) \le \frac{\widehat{q}(F)}{L_{0.025}(F)} \tag{7}$$

where $L_{0.025}(F)$ and $U_{0.025}(F)$ are the lower and upper bounds for $\frac{q_i(F)}{q(F)}$.

3.2 | Radial basis function

RBF is a type of feed-forward neural networks having various advantages over the conventional multilayer perceptron like quick convergence, fewer errors, and more reliability (Girosi & Poggio, 1990).

The structure of the RBF network is based on three layers; input, hidden, and output layers. The input layer provides information to the hidden layer without processing the input data. Neurons in the 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 center C_j of RBF and input (Y). In the hidden layer, a non-linear transformation is used with RBF as:

$$h_j(Y) = \mathrm{EXP}\left(-\left(\|Y-C_j\|^2 / H_j^2\right)\right) \tag{8}$$

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.

$$Z_k(Y) = \sum_{j=1}^{n} w_{kj} h_j(Y) + B_k \tag{9}$$

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 a 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 a least square algorithm.

4 | RESULTS AND DISCUSSION

4.1 | Screening of data for RFA

This section provides details of pre-processing or validation of certain assumptions of data values at each site using different measures. For instance, the Run test for randomness (Bradley, 1968), the Wald-Wolfowitz test for independence and stationarity (Rai et al., 2013; Wald & Wolfowitz, 1943), Rank-sum test for homogeneity (Hirsch et al., 1992). Table 2 provides the results of these tests with estimates of test statistics and p-values. The results show that AMPF at 36 sites have passed the preprocessing step as all the p-values are greater than the chosen level of significance, that is, 5%.

4.2 | Discordancy measure

Summary measures of L-moments and estimates of D_i using Equation (1) are provided in Table 3. These results show that two sites, "Badri" and "Chilah," are discordant, that is, 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 random outliers in the data series (Hussain, 2011). For data visualization, time series plots of these two sites are illustrated in Figure 2. For site Badri, the distribution of data around the mean is approximately symmetrical. However, a downward trend exists in the values of the last 7 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 (as shown in Table 3, i.e., −0.0211). The time series plot of site Chilah is showing a flood of a very high magnitude in the year 1979. Grubbs and Beck test (Grubbs & Beck, 1972) is also applied to detect outliers in the data series at these two sites and the results are presented in Figure 3. For the 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.

4.3 | 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. 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. Few other studies using hierarchical cluster analysis for the formation of homogenous regions are Arellano-Lara and Escalante-Sandoval (2014) and Rasheed et al., 2019. This study has used hierarchical cluster analysis based on site characteristics with few subjective adjustments to partition the group of 36 sites into four homogeneous regions. Complete details of this division are provided in the following section:

For the initial estimate of the 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 heterogeneous and requires subdivision.

Six available site characteristics can be used to delineate this heterogeneous group into homogeneous regions. However, each site characteristic has a different degree of relationship with observed data series. Therefore, to identify the most influential or significant site characteristic, at the first step, the Pearson Correlation Coefficient is calculated between the average value of AMPF at different sites (l_1) and the site characteristics. This correlation matrix is illustrated in Table 4, which shows that "latitude" has the 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

TABLE 2 Calculated values of test statistics and corresponding *p*-values of run test, rank sum test, and Wald-Wolfowitz test

		Rank-sum		Run test		Wald-Wolfowitz	
S. no.	**Site name**	**Test statistic**	***p*-value**	**Test statistic**	***p*-value**	**Test statistic**	***p*-value**
1	Budni	−0.1444	0.8852	0.8830	0.3772	−1.7913	0.0732
2	Shahi Bala	−0.7675	0.4427	−0.4080	0.6833	1.9036	0.0570
3	Dallus	−1.0590	0.2892	−0.6900	0.4902	1.2807	0.2003
4	Badri	−0.4080	0.6833	−0.6440	0.5194	1.8990	0.0576
5	Naranji	−0.4460	0.6556	−1.6560	0.0977	0.9353	0.3496
6	Kalpani Raisalpur	−0.3483	0.7276	1.4670	0.1424	1.0960	0.2729
7	Kalpani Deheri	0.7390	0.4599	0.4590	0.6459	−0.9550	0.3392
8	Bagiari	0.2930	0.7695	1.3400	0.1802	1.6385	0.1013
9	Katlongi	0.4859	0.6270	0.2650	0.7910	−1.0470	0.2951
10	Chprial	−1.0890	0.2762	1.6010	0.1094	1.5120	0.1305
11	Jani Khwar	0.5250	0.5996	−1.2530	0.2100	−0.7820	0.4337
12	Shahban	0.0547	0.9564	0.8100	0.4179	−0.1173	0.9066
13	Muqam	−1.5407	0.1234	0.6760	0.4990	1.4452	0.1484
14	Chinkar	1.1260	0.2602	−1.2010	0.2298	0.3090	0.7573
15	Wazir Gahri	−0.1320	0.8950	0.5740	0.5656	−0.6780	0.4975
16	Bara Kohat Road	−1.1710	0.2416	−0.5420	0.5878	−0.0830	0.9339
17	Bara Tarnab	−1.6590	0.0971	−1.8580	0.0631	1.2510	0.2109
18	Lund Khwar East	−1.4400	0.1499	−1.1550	0.2479	0.1090	0.9131
19	Kalpani Saidabad	−1.0080	0.3135	−1.7970	0.0723	0.7650	0.4443
20	Dagi	−0.3484	0.7275	−0.5220	0.6017	0.1914	0.8482
21	Garandi	0.7188	0.4723	1.1350	0.2564	−1.0420	0.2973
22	Hakim Gahri	−1.9130	0.0557	−1.0270	0.3044	1.3400	0.1802
23	Khuderzai	0.9550	0.3396	−1.2740	0.2026	0.2680	0.7887
24	Kabul Nowshera	−0.8680	0.3854	−1.1120	0.2658	−0.7880	0.4307
25	Chilah	−0.5940	0.5525	−1.0700	0.2846	1.3040	0.1921
26	Kabul Adezai	−1.7160	0.0862	−1.3180	0.1875	1.5910	0.1116
27	Shah Alam	0.3716	0.7102	−0.6640	0.5067	−0.2300	0.8181
28	Panjkora	−1.7970	0.0723	−0.9910	0.3217	−0.6210	0.5346
29	Kabul Naguman	−1.4864	0.1372	1.0370	0.2997	1.2040	0.2286
30	Jundi Utmanzai	0.5780	0.5633	−0.4810	0.6305	1.8510	0.0641
31	Jundi Tangi	−1.4900	0.1362	1.4360	0.1510	1.1820	0.2372
32	Jundi River	−0.7900	0.2495	0.2910	0.7711	−0.5630	0.5730
33	Swat Khaili	−1.6400	0.1010	−0.9370	0.3486	−1.0800	0.2801
34	Swat Ningolai	1.8640	0.0623	−0.3710	0.7106	1.7580	0.0787
35	Swat khawazakhela	1.7240	0.0847	−1.7410	0.0815	1.8140	0.0695
36	Swat Munda Head	1.4230	0.1547	−0.1410	0.8875	−1.0330	0.3015

Figure 4; which is indicating a subdivision into seven clusters at the first step. Heterogeneity measure based on L-CV is calculated to check the degree of homogeneity in each subdivided group. The details are: from left to right, the first group with eight sites (H is −0.48), the second group with three sites (H is 0.95), the third group with four sites (H is 4.91), the fourth group with nine sites (H is 0.51), the fifth group with seven sites (H is 0.11), the sixth group with two sites (H is 0.33), and seventh group with three sites (H is 1.22).

TABLE 3 Descriptive statistics and values of discordancy measure (D_i) for 36 gauging sites. Critical value of D_i for 36 sites is 3. Here n is the number of observations at each site, l_1 is first sample L-moment, t is sample L-CV, t_3 is sample L-skewness, t_4 is sample L-kurtosis, t_5 is the 5th sample L-moment ratio, and D_i is discordancy measure

Sites	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	0.2420	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	0.1950	0.0686	4.18[a]
5	Naranji	47	4836.136	0.4104	0.2587	0.1670	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	0.2180	−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	0.4420	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	0.3390	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	0.5940	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	0.5780	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[a]
26	Kabul Adezai	30	30027.69	0.3877	0.2280	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

[a] Indicates values of D_i greater than 3.

Keeping in view the inclusion of a reasonable number of sites in a group to perform RFA; the proposed division of seven groups/clusters is subjectively adjusted to form fewer clusters with a large number of sites and values of heterogeneity measures less than 1. Neighboring clusters are combined to form fewer clusters for the next step (like combining the first group with the second and the sixth group with the seventh). Relocation of sites of

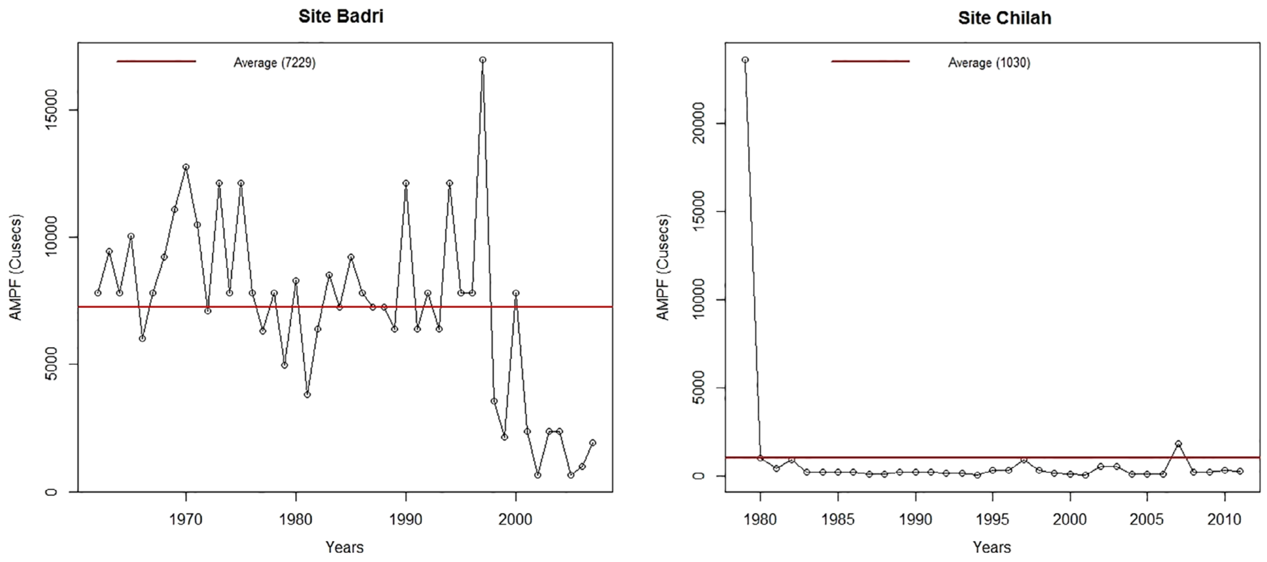


FIGURE 2 Time series plots of discordant sites

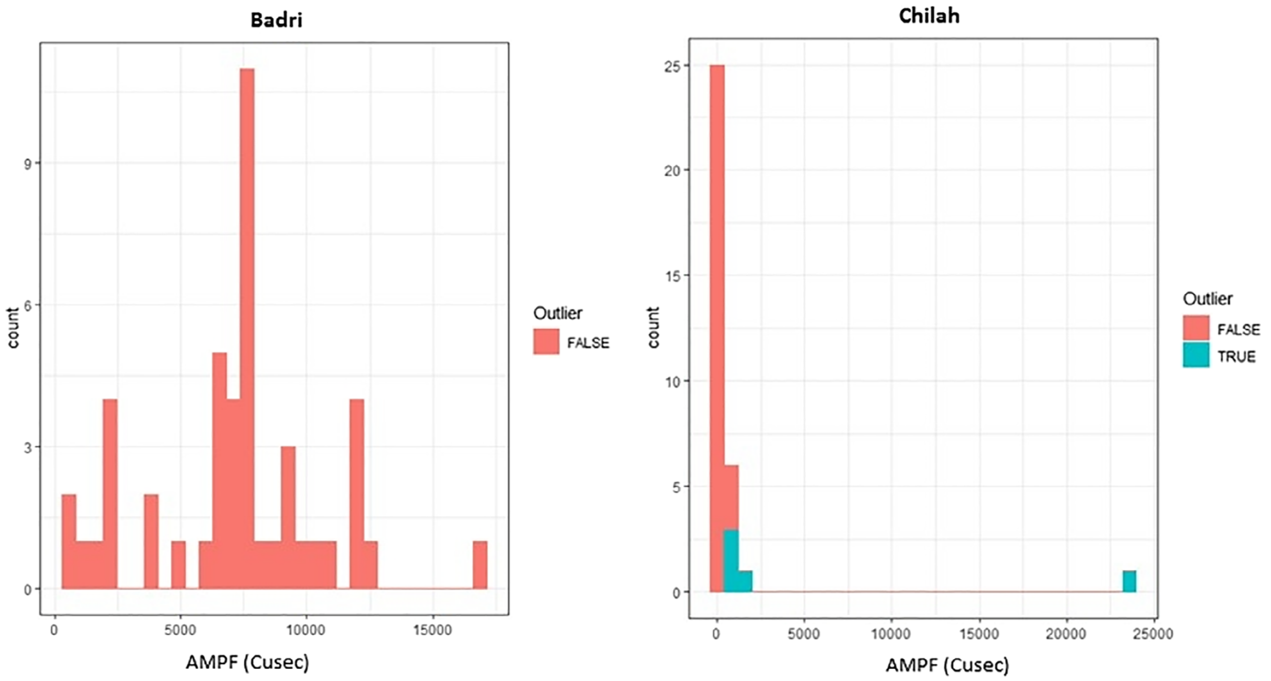


FIGURE 3 Results of Grubbs and Beck test of sites Badri and Chilah

the third group (due to its heterogeneity) to other groups with sites having similar values of L-CV; like shifting site "Chillah" to the combination of the first and second group, sites "Garandi" and "Kalpani Raisalpur" to the fifth group, and site "Swat Khaili" to the combination of the sixth and seventh group. These details of delineation of the study region into four homogenous regions/groups are illustrated in Table 5. The estimates are showing that the four regions are homogeneous and adequate to proceed further in RFA.

4.4 | Fitting of the regional probability distribution

L-moment ratio diagrams of the four regions are illustrated in Figure 5. By visualizing L-moment ratio diagrams and the tendency of the plotted points, good fit distribution(s) for each region are: GNO and GPA for Region 1, GNO, PE3, and GPA for Region 2, GPA for Region 3, and GLO for Region 4.

	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)

TABLE 4 Estimates of coefficient of correlation between l_1 and site characteristics. 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

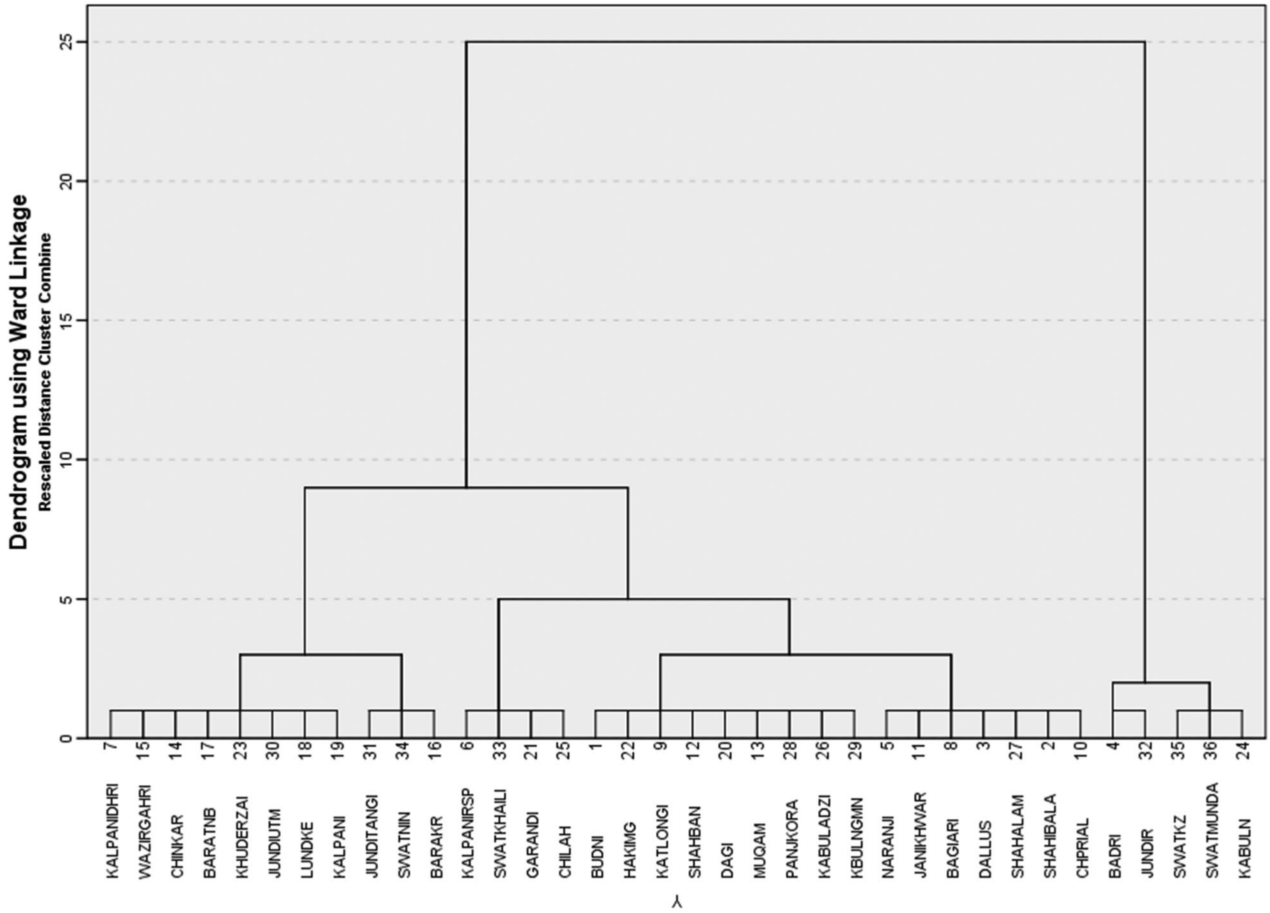


FIGURE 4 Dendrogram showing subdivision of heterogeneous cluster of 36 sites into homogeneous groups

The calculated values of |Z-Dist| statistic, for the four regions, are illustrated in Table 6. Details of the distributions passing this goodness-of-fit criterion are GLO, GEV, GNO, and GPA for Region 1; GLO, GEV, GNO, GPA, and PE3 for Region 2; PE3 and GPA for Region 3 while GLO for Region 4.

The above results show that the two goodness-of-fit methods are in fair agreement with each other. However, the results of |Z-Dist| statistic, being a quantitative method based on simulations, are taken for further analysis.

4.5 | Identification of a robust regional distribution

The |Z-Dist| statistic has identified two or more probability distributions as successful candidates for three of the

TABLE 5 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	$H_1 = 0.26$ $H_2 = 0.79$ $H_3 = 1.21$
Region 2	Forth group	9	Budni, Hakim Ghari, Katlongi, Shahban, Dagi, Muqam, Panjkora, Adezai, Naguman	$H_1 = 0.54$ $H_2 = -1.19$ $H_3 = -0.55$
Region 3	Fifth group + Site Garandi + Site Kalpani Raisalpur	9	Naranji, Bagiari, Dallus, Shah Alam, Shahi Bala, Chprial, Garandi, Kalpani Raisalpur, Jani Khwar	$H_1 = 0.14$ $H_2 = -0.68$ $H_3 = 1.00$
Region 4	Sixth group + Seventh group	6	Badri, Jundi River, Swat Khawazakhela, Swat Munda Head, Kabul Nowshera, Swat Khaili	$H_1 = 0.91$ $H_2 = 2.06$ $H_3 = 1.12$

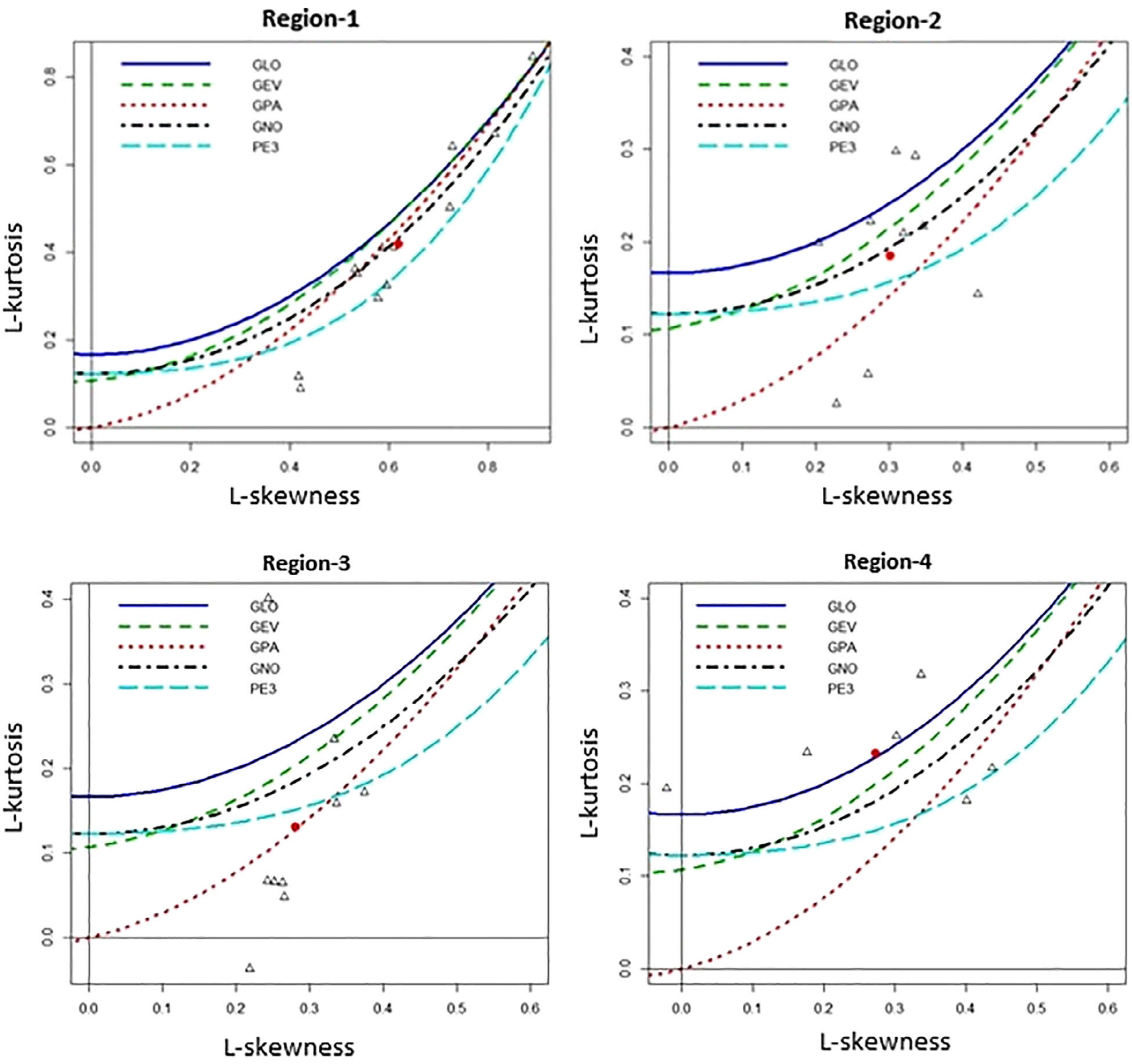


FIGURE 5 L-moment ration diagrams

four regions. Therefore, an assessment analysis using simulations is required to find a robust probability distribution for each region. Complete details of the setting up of simulation experiments are available in Hosking and Wallis (1997). A brief of the base/artificial region for this study is provided below:

An initial requirement is the development of an artificial region mimicking the actual/study region in terms of the number of sites, observations at each site and estimates of regional average L-moment ratios. Moreover, L-moment ratios for each site should be defined like that the heterogeneity of the artificial and the actual region remains comparable. To observe the inter-site dependence, 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. These values indicate weak inter-site dependence for all the regions. This may be because these sites are located on different streams/rivers. Details of choice of linear variations in L-CV with incremental effect for each site, chosen values of L-skewness and estimated values of the heterogeneity measure for each region are summarized in Table 7.

Details of Table 7 are showing a comparable degree of homogeneity between artificial and actual regions. Therefore, these artificial regions are good to find accuracy measures for the identification of a robust distribution for each region. For Region 1, using the artificial/base region, 5000 realizations are performed, considering each successful distribution using the estimation method of L-moments and the process continues for GPA, GLO, GEV, and GNO distributions. Relative root means square error (RMSE) of regional quantiles is calculated from these simulations and the results are shown in Table 8. 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 distributions are given in Figure 6. This graph shows that the growth curve of GNO distribution has the shortest 95% error bounds, especially for longer return periods.

TABLE 6 Values of $|Z\text{-Dist}|$ statistic for candidate distributions

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

[a] Indicates the calculated values exceeding critical value, that is, 1.64.

TABLE 7 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 8 Estimated quantiles and their RMSE for Region 1

	Distributions							
	GPA		GEV		GLO		GNO	
Return periods	$\hat{q}$	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	14.5188	2.9456	14.3906	2.8528	14.7225	2.5008

Importantly, the growth curve of GNO distribution remains within the limits of 95% error bounds, while the growth curves of other distributions are below the lower limits of error bounds for longer return periods. Therefore, it can be concluded that GNO distribution is the most stable and robust distribution for Region 1.

Following a similar scheme, a robust distribution has been identified for Region 2 and Region 3. For Region 4, accuracy measures are calculated for GLO distribution as being the only good-fit distribution. The estimates of quantiles using candidate distributions and their RMSE for Region 2, Region 3, and Region 4 are given in Table 9, Table 10, and Table 11, respectively. Growth curves for Region 2, Region 3, and Region 4 with their respective 95% error bounds are given in Figure 7, Figure 8, and Figure 9, respectively. These results are favoring GPA distribution as robust distribution for Region 2 and Region 3 while GLO distribution for Region 4.

Using regional quantiles of identified robust distributions, at-site quantiles (using Equation (4)), their RMSE and 95% error bounds are given in Table 12 (for Region 1), Table 13 (for Region 2), Table 14 (for Region 3) and Table 15 (for Region 4). These estimates are useful for the scientists, hydrologists, and government officials dealing with designing and developing hydrological structures as well as water resources management and flood protection planning of the region. Accuracy measures of these at-site quantiles would be helpful for future studies to compare the quality of the estimates using alternative methods of modeling extreme values.

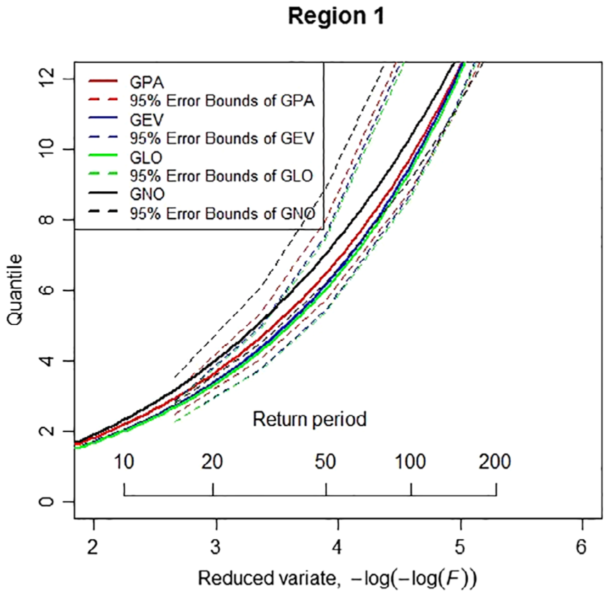


FIGURE 6 Regional growth curves of successful distributions of Region 1 with their 95% error bounds

5 | ESTIMATION OF QUANTILES AT UNGAUGED SITES

RBF network is used to develop a model considering l_1 (as a dependent variable) and site characteristics (as independent variables) for the prediction of quantiles at ungauged sites. This method has been used in various

TABLE 10 Estimated quantiles and their RMSE for Region 3

	Distributions			
	GPA		PE3	
Return periods	$\hat{q}$	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 9 Estimated quantiles and their RMSE for Region 2

	Distributions									
	GPA		GEV		GLO		GNO		PE3	
Return periods	$\hat{q}$	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 11 Estimated quantiles and their RMSE for Region 4

	GLO distribution	
Return periods	$\hat{q}$	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

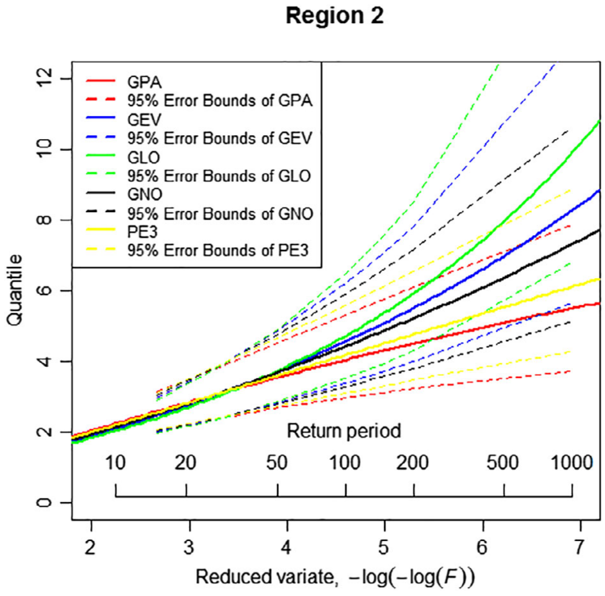


FIGURE 7 Regional growth curves of successful distributions of Region 2 with their 95% error bounds

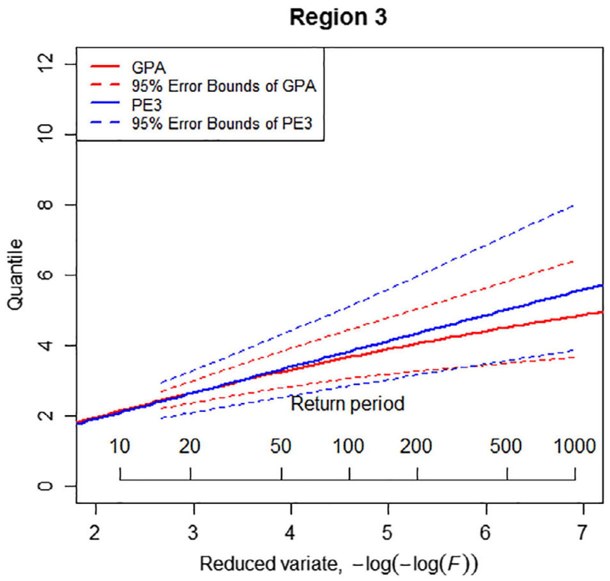


FIGURE 8 Regional growth curves of successful distributions of Region 3 with their 95% error bounds

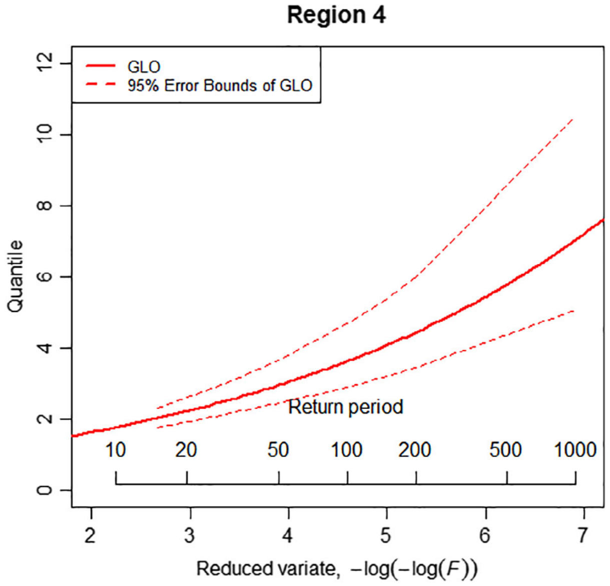


FIGURE 9 Regional growth curve of GLO distribution for Region 4 with its 95% error bounds

studies for short term streamflow forecasting, for example, Kagoda et al., 2010; Uysal, 2016; and Sahoo et al., 2019. Few details of the procedure are:

For the application of the RBF network concerning each homogenous region, all variables are rescaled, i.e. their standardized forms are used for training of the model. A random partition of 70% and 30% is used for training and testing of the model. The input layer consists of six units (independent variables). The hidden layer has the same number of units as the input layer with the Gaussian link function while the output layer has only one unit. A Sum of squares of error and the relative error is used for the performance evaluation criteria of the model. Model summary of the training and testing phases for each region are provided in Table 16. A graphical comparison of the fitted values against observed values of the dependent variable is illustrated in Figure 10. The results of Table 16 and Figure 10 show that the application of RBF network provides adequate results and can be used for the estimation of l_1 for any ungauged site in a particular homogeneous region. These estimates can then be linked with regional quantiles of the respective regions for the estimation of quantiles at the ungauged site for any return period.

TABLE 12 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	14,194	19,765	29,320	35,431	42,057
	RMSE	5054	8009	11,093	16,419	19,851	23,595
	LB	4555	7514	10,623	16,130	19,541	23,413
	UB	21,432	36,082	52,322	80,882	99,912	121,932
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	10,454	12,934	15,740
Chinkar	$\hat{Q}$	2855	4583	6381	9466	11,439	13,578
	RMSE	1418	2240	3100	4588	5549	6599
	LB	1545	2538	3605	5458	6668	7941
	UB	6256	10,599	15,386	23,944	29,719	35,923
Bara Tarnab	$\hat{Q}$	36,789	59,049	82,229	121,980	147,403	174,966
	RMSE	20,402	32,011	44,087	64,884	78,265	92,846
	LB	19,772	32,455	45,937	69,531	85,156	102,456
	UB	78,095	132,085	190,298	296,608	367,379	446,644
Khuderzai	$\hat{Q}$	5443	8737	12,166	18,048	21,810	25,888
	RMSE	2785	4450	6204	9262	11,246	13,419
	LB	2997	4909	6928	10,476	12,733	15,186
	UB	11,177	18,860	27,392	42,449	52,861	64,141
Jundi Utmanzai	$\hat{Q}$	6354	10,199	14,202	21,068	25,459	30,219
	RMSE	3548	5662	7892	11,783	14,309	17,078
	LB	3303	5421	7711	11,590	14,097	16,910
	UB	14,288	24,269	34,984	54,354	67,139	81,082
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	12,312	15,311	18,596
Kalpani Saidabad	$\hat{Q}$	29,127	46,750	65,101	96,573	116,700	138,522
	RMSE	13,407	21,348	29,713	44,293	53,761	64,138
	LB	16,147	26,324	37,265	56,639	69,029	82,375
	UB	58,837	99,778	144,389	224,167	278,063	337,619
Jundi Tangi	$\hat{Q}$	3420	5489	7643	11,338	13,701	16,263
	RMSE	1431	2291	3202	4799	5841	6986
	LB	1970	3227	4553	6867	8359	10,008
	UB	6610	11,135	16,126	25,213	31,317	38,182
Swat Ningolai	$\hat{Q}$	27,656	44,389	61,814	91,696	110,807	131,527
	RMSE	12,525	19,993	27,863	41,587	50,499	60,267
	LB	15,301	25,167	35,794	53,845	65,473	78,420
	UB	57,299	96,827	140,640	220,431	272,911	331,770
Bara Kohat Road	$\hat{Q}$	4498	7220	10,054	14,914	18,022	21,392
	RMSE	2021	3239	4523	6762	8216	9810
	LB	2492	4067	5763	8724	10,599	12,610

(Continues)

TABLE 12 (Continued)

Site names	Measures	15	30	50	100	150	200
	UB	9071	15,390	22,070	34,291	42,344	
Chillah	$\hat{Q}$	3188	5116	7125	10,569	12,772	15,160
	RMSE	1733	2815	3968	5992	7313	8765
	LB	1772	2930	4153	6262	7640	9163
	UB	6644	11,319	16,414	25,608	31,667	38,282

TABLE 13 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	$\hat{Q}$	35,962	43,307	48,743	55,456	58,748	61,756
	RMSE	6119	8014	9539	11,599	12,693	13,748
	LB	27,752	32,607	35,958	39,979	41,953	43,619
	UB	48,476	59,634	67,852	78,461	83,634	88,541
Hakim Gahri	$\hat{Q}$	9018	10,860	12,223	13,906	14,732	15,486
	RMSE	1731	2226	2618	3141	3416	3680
	LB	6808	8019	8905	9967	10,464	10,908
	UB	12,866	15,746	17,982	20,756	22,131	23,338
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	11,328	12,909	14,864	15,855	16,807
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	10,224
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	$\hat{Q}$	40,476	48,743	54,860	62,416	66,121	69,507
	RMSE	8210	10,535	12,371	14,807	16,084	17,305
	LB	29,941	35,335	39,186	43,659	45,850	47,750
	UB	58,853	72,499	82,443	95,240	101,839	108,228
Panjkora	$\hat{Q}$	63,793	76,822	86,464	98,373	104,212	109,548
	RMSE	12,192	15,732	18,548	22,311	24,293	26,196
	LB	48,279	56,881	62,909	70,180	73,538	76,612
	UB	90,884	112,080	127,535	147,341	157,502	167,023
Kabul Adezai	$\hat{Q}$	72,913	87,805	98,825	112,436	119,110	125,209
	RMSE	14,371	18,532	21,830	26,217	28,519	30,725
	LB	54,510	64,403	71,210	79,293	83,368	86,887
	UB	104,078	127,630	145,766	168,103	179,856	190,529

TABLE 13 (Continued)

Site names	Measures	15	30	50	100	150	200
Kabul Naguman	$\widehat{Q}$	46,687	56,223	63,280	71,995	76,269	80,174
	RMSE	9291	11,924	14,006	16,772	18,223	19,613
	LB	34,845	41,187	45,639	50,819	53,363	55,700
	UB	67,966	83,367	95,118	110,033	117,305	123,867

TABLE 14 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	$\widehat{Q}$	13,529	16,281	18,300	20,775	21,979	23,075
	RMSE	1673	2096	2461	2992	3290	3587
	LB	11,224	13,384	14,859	16,619	17,424	18,130
	UB	16,829	20,477	23,296	26,895	28,671	30,327
Bagiari	$\widehat{Q}$	14,324	17,237	19,375	21,995	23,271	24,431
	RMSE	2239	2752	3172	3758	4078	4393
	LB	11,283	13,541	15,128	16,945	17,776	18,524
	UB	19,120	23,239	26,236	30,070	32,033	33,853
Dallus	$\widehat{Q}$	20,359	24,499	27,539	31,262	33,075	34,725
	RMSE	3450	4220	4843	5699	6163	6617
	LB	15,846	18,960	21,216	23,805	25,014	26,097
	UB	27,773	33,532	37,960	43,708	46,513	49,108
Shah Alam	$\widehat{Q}$	18,239	21,948	24,670	28,006	29,630	31,108
	RMSE	2840	3491	4025	4771	5179	5580
	LB	14,500	17,333	19,381	21,771	22,838	23,759
	UB	24,465	29,642	33,530	38,491	41,040	43,354
Shabi Bala	$\widehat{Q}$	6936	8346	9382	10,650	11,268	11,830
	RMSE	1173	1434	1644	1934	2090	2244
	LB	5399	6460	7229	8128	8551	8906
	UB	9476	11,472	12,996	14,955	15,937	16,836
Chprial	$\widehat{Q}$	26,030	31,323	35,209	39,969	42,287	44,396
	RMSE	3873	4795	5560	6636	7227	7808
	LB	20,816	24,940	27,797	31,181	32,722	34,170
	UB	34,321	41,661	47,195	54,306	57,968	61,385
Garandi	$\widehat{Q}$	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	$\widehat{Q}$	86,371	103,934	116,827	132,623	140,315	147,312
	RMSE	13,115	16,231	18,807	22,415	24,391	26,335
	LB	68,524	81,818	91,166	102,302	107,272	111,613
	UB	113,623	138,560	157,133	180,703	192,981	203,724
Jani Khwar	$\widehat{Q}$	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 15 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	$\widehat{Q}$	13,781	16,790	19,479	23,538	25,907	28,339
	RMSE	1972	2683	3377	4517	5229	5993
	LB	11,105	13,217	15,094	17,725	19,257	20,743
	UB	17,760	22,356	26,732	33,386	37,411	41,676
Jundi River	$\widehat{Q}$	21,085	25,688	29,802	36,012	39,636	43,357
	RMSE	3060	4152	5216	6963	8053	9223
	LB	16,898	20,077	22,902	26,968	29,303	31,697
	UB	27,091	34,079	40,701	51,146	57,514	64,252
Swat Khawazakhela	$\widehat{Q}$	96,910	118,066	136,979	165,520	182,176	199,279
	RMSE	15,439	20,795	25,982	34,435	39,685	45,302
	LB	76,772	91,248	103,912	122,470	132,943	143,843
	UB	129,020	162,222	193,014	241,063	270,290	300,767
Swat Munda Head	$\widehat{Q}$	119,588	145,694	169,033	204,253	224,807	245,912
	RMSE	16,260	22,417	28,461	38,432	44,681	51,403
	LB	97,305	115,724	131,439	154,804	167,828	181,087
	UB	151,246	190,871	227,053	283,550	317,760	353,271
Kabul Nowshera	$\widehat{Q}$	264,739	322,533	374,199	452,169	497,670	544,391
	RMSE	57,311	73,814	89,509	114,688	130,150	146,578
	LB	191,030	229,720	262,901	311,784	339,714	367,850
	UB	387,385	486,291	579,803	723,338	808,678	900,271
Swat Khaili	$\widehat{Q}$	113,494	138,271	160,420	193,846	213,353	233,382
	RMSE	16,628	22,590	28,399	37,913	43,842	50,198
	LB	90,580	107,638	122,819	144,595	156,694	169,249
	UB	146,931	185,429	220,863	276,271	310,299	346,232

TABLE 16 Model summaries of RBF during 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

6 | SUMMARY AND CONCLUSIONS

This study is an application of RFA for estimating flood quantiles considering the AMPF of 36 sites located on important streams/rivers of KPK, Pakistan. A systematic, detailed, and comprehensive application of a standard procedure to a new study area and demonstration of radial basis function network for estimation of quantiles at ungauged sites are few important contributions of this study. Some major findings are summarized below:

1. In preprocessing steps of the analysis, results of different statistical measures indicate that data series at each site is random and independently identically distributed.
2. Summary measures of L-moments ratios show that there exist variations in the data series at 36 sites with smaller L-kurtosis values than L-skewness. This is an indication of frequent flooding in the area possibly due to monsoon rainfall. Hussain et al. (2017) reported similar tendencies for the sites of rivers of Punjab, Pakistan.

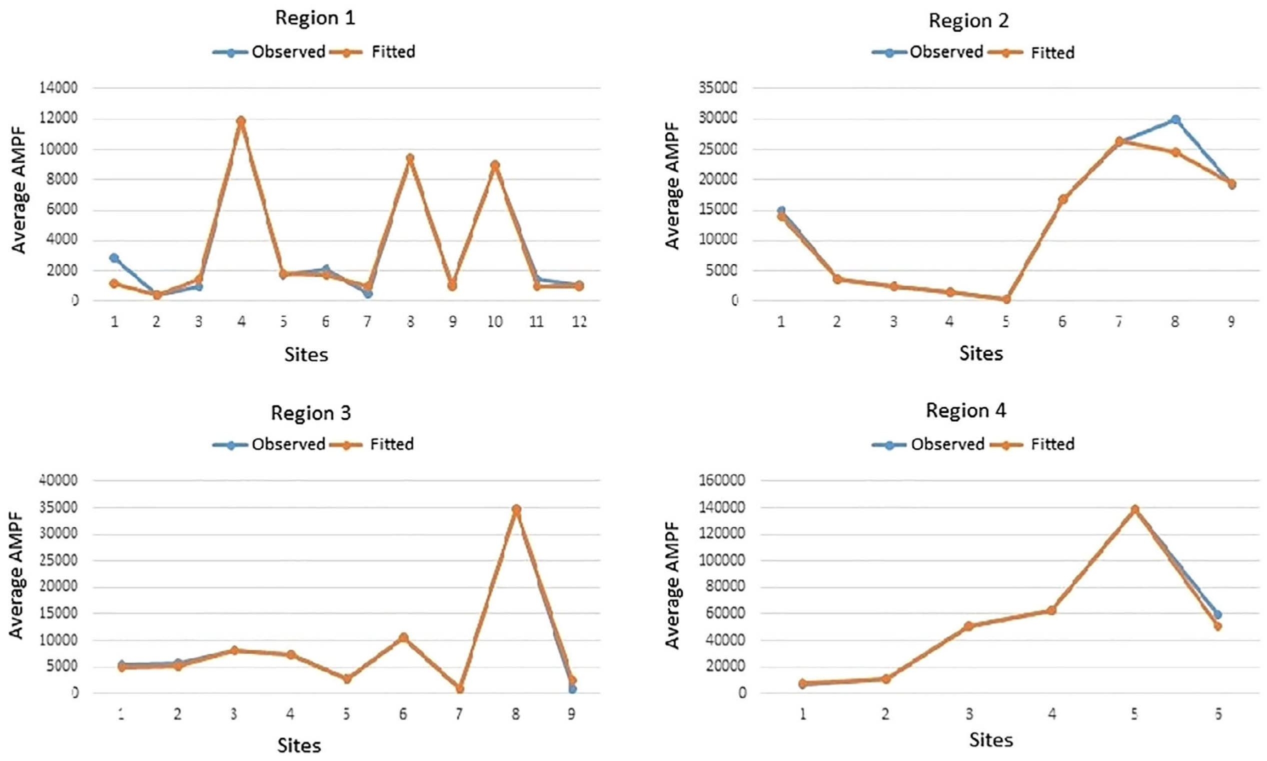


FIGURE 10 A comparison of observed Vs fitted values of the dependent variable for four regions

3. To gain maximum benefits out of RFA, the heterogeneous study region is divided into four homogeneous regions. Wards clustering method with Euclidean distance using the most significant site characteristic, that is, latitude is used for this subdivision.
4. Five commonly used probability distributions are considered as candidates for regional distribution. The goodness of fit methods of $|Z\text{-Dist}|$ statistic and L-moment ratio diagram shows 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 a robust regional distribution. The results of different accuracy measures show that GNO distribution for Region 1, GPA distribution for Region 2 and Region 3, and GLO distribution for Region 4 have robust properties. 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.
5. For the estimation of l_1 for ungauged sites, the RBF network is used. This method is preferred due to inherent correlations between site characteristics, the non-linear nature of variables and estimation problems of a classical linear regression model. The results indicate that the proposed method provides an adequate fit. Therefore, can be used for the estimation of quantiles at ungauged or poorly/partially gauged sites within the respective regions.
6. A major limitation of the study includes the availability of a limited record of values for the demonstration of the RBF network. However, the results can be improved in future considering more data or variables or testing considering other activation functions in the processing. Another important recommendation for future work is the use of a variable(s) other than AMPF at different sites to apply RFA.

The flood estimates of the study are beneficial for the authorities concerning flood risk management, water resources management, irrigation, and planning and development of existing and potential hydraulic structures in the study area. For future studies, the focus would be to adopt different modeling approaches of analyzing extreme events (like Bayesian Information criteria) by varying estimation methods (like maximum product spacings). Second, the 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.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

ORCID

Muhammad Shafeeq ul Rehman Khan https://orcid.org/0000-0003-2367-7893

REFERENCES

Ahmad, I., Abbass, A., Fawad, M., & Saghir, A. (2017a). Regional frequency analysis of annual total rainfall in Pakistan using L-moments. *NUST Journal of Engineering Sciences*, *10*, 19–29.

Ahmad, I., Abbass, A., Saghir, A., & Fawad, M. (2016a). Finding probability distributions for annual daily maximum rainfall in Pakistan using linear moments and variants. *Polish Journal of Environmental Studies*, *25*, 925–937.

Ahmad, I., Fawad, M., Akbar, M., Abbas, A., & Zafar, H. (2016b). Regional frequency analysis of annual peak flows in Pakistan using linear combination of order statistics. *Polish Journal of Environmental Studies*, *25*, 2255–2264.

Ahmad, I., Shah, S. F., Mahmood, I., & Ahmad, Z. (2013). Modeling of monsoon rainfall in Pakistan based on kappa distribution. *Science International (Lahore)*, *25*, 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*, 75–84.

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*, 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*, 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*, 2082–2092.

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. *Atmosfera*, *27*, 47–60.

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*, 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*, 541–554.

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.

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

Cook, N. J. (1985). *The designer's guide to wind loading of building structures part 1: Background*. Damage survey, wind data and structural classification building research establishment.

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

Development Advocate Pakistan. (2016). Water Security in Pakistan: Issues and Challenges, Volume 3, Issue 4. http://www.pk.undp.org/content/pakistan/en/home/library/hiv_aids/development-advocate-pakistan-volume-3-issue-4.html.

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*, 2289–2315.

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*, 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*, *153*, 724–737.

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

Girosi, F., & Poggio, T. (1990). Networks and the best approximation property. *Biological Cybernetics*, *63*, 169–176.

Government of Pakistan, Annual flood report 2016. *Ministry of Water and Power, Office of the Chief Engineer Advisor and Chairman*. Islamabad, Pakistan: Federal Flood Commission. https://www.ffc.gov.pk/download/AFR/Annual%20Flood%20Report%202016.pdf.

GREHYS, G. D. R. E. H. S. (1996a). Inter-comparison of regional flood frequency procedures for Canadian rivers. *Journal of Hydrology (Amsterdam)*, *186*, 85–103.

GREHYS, G. D. R. E. S. (1996b). Presentation and review of some methods for regional flood frequency analysis. *Journal of Hydrology (Amsterdam)*, *186*, 63–84.

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

Haddad, K., & Rahman, A. (2020). Regional flood frequency analysis: Evaluation of regions in cluster space using support vector regression. *Natural Hazards*, *102*, 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. In *Principles of neuro computing for science and engineering*. McGraw-Hill.

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*, 1054–1067.

Hirsch, R. M., Helsel, D. R., Cohn, T. A., & Gilroy, E. J. (1992). Statistical analysis of hydrologic data. Chapter 17. In D. R. Maidment (Ed.), *Handbook of hydrology*. McGraw-Hill.

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

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

Hussain, Z., & Pasha, G. R. (2009). Regional flood frequency analysis of the seven sites of Punjab, Pakistan, using L-moments. *Water Resources Management*, *23*, 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*, 1812–1819.

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*, 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*, 107–123.

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.

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*, 1–20.

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*, 1–15.

Lin, G. F., & Chen, L. H. (2004). A non-linear rainfall-runoff model using radial basis function network. *Journal of Hydrology*, *289*, 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*, 3598–3609.

Malekinezhad, H., & Zare-Garizi, A. (2014). Regional frequency analysis of daily rainfall extremes using L-moments approach. *Atmosfera*, *27*, 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*, e12208.

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*, 1292–1306.

Pakistan Meteorological Department. *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*. (2012). Pakistan Meteorological Department. http://www.pmd.gov.pk/reports/flood_diagnostic_2010_2011.pdf.

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*, 119–132.

Rai, R. K., Upadhyay, A., Ojha, C. S. P., & Lye, L. M. (2013). Statistical analysis of hydro-climatic variables. In R. Y. Surampalli, T. C. Zhang, C. S. P. Ojha, B. R. Gurjar, R. D. Tyagi, & C. M. Kao (Eds.), *Climate change modelling, mitigation, and adaptation* (pp. 378–388). ASCE.

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

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*, 570.

Requena, A. I., Ouarda, T. B., & Chebana, F. (2017). Flood frequency analysis at ungauged sites based on regionally estimated stream flows. *Journal of Hydrometeorology*, *18*, 2521–2539.

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.

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*, 111–136.

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

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.

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*, 378–388.

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*, 165–182.

How to cite this article: Khan, M. S. 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. https://doi.org/10.1111/jfr3.12751