REGIONAL FLOOD FREQUENCY ANALYSIS, USING L-MOMENTS, ARTIFICIAL NEURAL NETWORKS AND OLS REGRESSION, OF VARIOUS SITES OF KHYBER-PAKHTUNKHWA, PAKISTAN

KHAN, M. S. R.[1*] – HUSSAIN, Z.[2] – AHMAD, I.[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 campus, Islamabad, Pakistan*

[*]*Corresponding author*
e-mail: shafeeq.phdst02@iiu.edu.pk; phone: +92-3347-603-022

(Received 10th Aug 2020; accepted 19th Nov 2020)

Abstract. This study provides the results of flood frequency analysis adopting a regional approach using annual maxima's of peak flows (APF) of eight catchments located on various small rivers of Khyber-Pakhtunkhwa, Pakistan. Initial screening reveals that the recorded data of APF for all catchments are independent, random, free from significant trend and identically distributed. L-moments based heterogeneity measure indicates that the study region is homogeneous. The results of |Z-Dist| statistic and L-moment ratio diagram being goodness of fit measures are in favor of Generalized Pareto (GPA) distribution among five candidates of regional distribution. For the ungauged sites, flood quantiles have been estimated through OLS regression and artificial neural networks (ANN). The estimated quantiles using ANN method are relatively accurate compared to OLS regression. The historical assessment indicates that quantile estimates obtained through ANN and index flood method are close to the highest recorded APF values for shorter as well as longer return periods for each site.
Keywords: *annual maximum peaks, GPA distribution, L-moments, least squares regression, machine learning methods, ungauged sites*

Abbreviations: APF, Annual maxima's of Peak Flows; KPK, Khyber-Pakhtunkhwa; RFA, Regional frequency analysis; GPA, Generalized Pareto; GNO, Generalized Normal; GLO, Generalized Logistic, GEV, Generalized Extreme Value; PE3, Pearson Type-3; ANN, Artificial Neural Networks; OLS, Ordinary Least Square; RA, Regression Analysis; AARF, Average Annual Rainfall; long, Longitude; lat, Latitude; ele, Elevation; ARMS, Average Rainfall in Monsoon; l_1, first sample L-moment

Introduction

Frequency analysis of extreme events (like floods, rainfall, winds and droughts, etc.) is necessary for effective planning and management against these natural disasters. Moreover the estimates of frequency analysis are also useful in design and development of hydrological structures such as dams, barrages, culverts, bridges etc., to ensure public safety and the effective use of surface water resources. For this purpose, different approaches are available in literature like at-site or regional. At-site frequency analysis may not be a preferred choice with a shorter or limited span of recorded data set at any site or is inapplicable for the estimation at any specific site with no observed record (ungauged site). Moreover, 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,

http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/1901_471489

regional frequency analysis (RFA) (combining sites based on similar site characteristics) is an optimum choice. Advantages of using RFA are many folds like more reliable estimation of quantiles at gauged sites and estimation or improvement of the quantiles at ungauged or partially/poorly gauged sites within the homogeneous region(s) through interpolation/extrapolation of the T-years quantiles of gauged sites. RFA using L-moments is a popular method of estimation and has been used frequently in various case studies worldwide. For example; in Canada, (Requena et al., 2017); in Norway, (Hailegeorgis and Alfredsen, 2017); in Iran, (Mesbahzadeh et al., 2019); in India, (Alam et al., 2016); in Korea, (Lee and Kim, 2019); in China, (Yang et al., 2010); in Turkey, (Aydoğan et al., 2016). GREHYS (1996a, b) provided a detailed comparison of several regional flood estimation procedures. A brief of the development in RFA has been illustrated in Malekinezhad and Zare-Garizi (2014). RFA has been applied in few of the published studies related to Pakistan. These include Hussain and Pasha (2009), Hussain (2011), Ahmad et al. (2017), Shahzadi et al. (2013), Ahmad et al. (2013, 2016), Batool (2017), Khan et al. (2017), Fawad et al. (2018, 2019). Few highlighted points in the some published literature related to flood frequency analysis in Pakistan are:

The study of Hussain and Pasha (2009) is the first application (to the best of authors' knowledge) of L-moments based RFA in Pakistan. The study area consists of seven sites of three major rivers of Punjab namely Jhelum, Chenab and Ravi with annual maximum peak flows as the main variable for analysis. The results of various accuracy measures calculated through simulation experiments reveal that Generalized Normal (GNO) Distribution is the robust distribution for the study area.

In another study, Hussain (2011) used annual maximum peak flows of seven sites located on the mainstream of the biggest river of the river systems of Pakistan namely the Indus River. The study area was divided into two regions, the upper half containing four sites (Tarbela, Kalabagh, Chashma and Taunsa) and the lower half containing three sites (Guddu, Sukkur and Kotri). Pearson Type-3 (PE3) was identified as robust regional distribution for the upper half while Generalized Logistic (GLO) Distribution for the sites of the lower half of the region.

The study of Ahmad et al. (2017) performed L-moments based RFA using 10 days average of low flows of nine sites belong to the different rivers of river systems of Pakistan. The study area includes 6 sites of Indus River (Tarbela, Kalabagh, Chashma, Tausa, Guddu and Sukkur), 1 site of Kabul River (Nowshera), 1 site of Jhelum River (Mangla) and 1 site of Chenab River (Marala). The study area was divided into two homogeneous regions (Region 1 consisting of Tarbela, Nowshera, Kalabagh and Taunsa while Region 2 includes Chashma, Guddu, Mangla and Marala) using basin drainage area in square miles and mean annual minimum flow. The study concluded that the best fit distribution for Region 1 is GNO and for Region 2 is Generalized Pareto (GPA).

Hussain (2017) analysed various sites of the major rivers located in Punjab, Pakistan namely Jhelum, Chenab, Ravi and Sutlej using L-moments based RFA. The results showed that the best-fitted regional distribution, for the region consisting of two sites (Trimmu and Panjnad) at the confluence of the rivers is PE3, while for the region consisting of the rest of the nine sites (Mangla, Rasul, Marala, Khanki, Qadirabad, Balloki, Sidhnai, Suleimanki and Islam) is GNO. Similar results have been reported in Hussain and Pasha (2009). Moreover, a multiple linear regression model in log transformed form was developed to estimate the mean value of the APF for ungauged locations using average rainfall during monsoon and catchment area of the corresponding rivers as explanatory variables. Adequacy of the developed regression

http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/1901_471489

model was established using statistical measures and comparison of the estimated floods with historical flood information available at the gauged sites.

The aforementioned details reveal that the catchments of major rivers located in Punjab and the Indus River have been the focus of the published studies so far. Therefore, there is a need to adopt standard procedures for analyzing APF of small rivers and streams of other parts of the country, especially Khyber Pakhtunkhwa (the north-western area of Pakistan). An important feature of these rivers or streams is that most of them originate in Pakistan with natural flows and are not or less affected by man-made changes like construction of barrages and dams, etc. Furthermore, floods are increasing in frequency and intensity in Pakistan by the year 2000 and their trends are alarming from 2010 onwards (Government of Pakistan, 2016). The area of KPK is known as more vulnerable to flash flooding due to its steep geography and mountainous land. This region is badly affected from flash floods in 1992 and 2010 (Pakistan Meteorological Department, 2012). Therefore, estimation of magnitude and frequency associated to these floods is a desperate need for the area of KPK to generate flood risk maps, management of stream water and feasibilities/designing of new hydraulic structures for the rivers and streams located in the province using a standard methodology available in details in Hosking and Wallis (1997).

In RFA, various methods are used to develop and estimate the forecast equation for ungauged sites. 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. None of the adopted method(s) so far received universal acceptability; however, success depends on the availability and suitability of gauged site characteristics. Anilan et al. (2016) illustrated details of commonly used site characteristics as independent variables in different studies around the world for estimation at ungauged sites adopting regression and ANN methods. These site characteristics are drainage area, slope of stream, and mean annual rainfall. Availability and identification of the most influential site characteristics that can be used for the estimation at ungauged site within the region is an ongoing area of research. Development of an adequate model depends on the site characteristics having significant relationship with the recorded data sets at gauged sites. This study has used available or significant site characteristic(s) using ANN and regression methods, emphasizing the justification of critical assumptions associated with the estimation procedures. In addition, a comparison has been made with the historical flood data to evaluate the reliability of the given estimates using different methods.

Main objectives of the study are:

i. Estimation of flood quantiles at gauged sites of important rivers/streams of the area of KPK using L-moments based RFA.
ii. Development of a suitable relationship for the estimation of quantiles at ungauged sites.
iii. Historical comparison of the provided estimates for validity and reliability of the provided estimates.

http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/1901_471489

Materials and Methods

Study area and data descriptions

APF in cusecs, of eight sites of important rivers of KPK, Pakistan namely Naranji, Bagiari, Dallus, Shahi Bala, Garandi, Chprial, Shah Alam, and Kalpani Raisalpur, have been used to perform RFA. The observed flow record and the site characteristics are provided by Provincial Irrigation Department of KPK. Details of the site characteristics including average annual rainfall (AARF), longitude (long), latitude (lat), elevation (ele), and average rainfall during monsoon season (ARMS) are given in *Table 1*. There are few missing observations in the data at sites Naranji, Bagiari, Shahi Bala, Chprial, Jani Khwar, and Kalpani Raisalpur. *Table 1* also illustrates the percentage of these missing observations. These missing values are estimated by the averages of APF at the respective sites. Geographical locations of the eight sites are given in *Fig. 1*.

Table 1. *Site characteristics and details of missing observations of eight sites of Khyber-Pakhtunkhwa, Pakistan*

S. No.	Site Name	Latitude (North)	Longitude (East)	Elevation (meters)	AARF (mm)	ARMS (mm)	Percentage of MO
1	Naranji	34.2480	72.3427	356	639	272	6
2	Bagiari	34.2254	72.1543	313	559	227	10
3	Dallus	34.1650	71.5931	310	460	151	0
4	Shah Alam	34.1669	71.3689	397	422	204	0
5	Shahi Bala	34.1702	71.7466	300	460	151	12
6	Chprial	34.9866	72.3520	1243	478	212	6
7	Garandi	34.0869	71.4719	328	384	105	0
8	Kalpani Raisalpur	34.3303	71.9085	345	556	222	6

Note: average annual rainfall (AARF); average rainfall during monsoon (ARMS); missing observations (MO), *mm* denotes millimeter

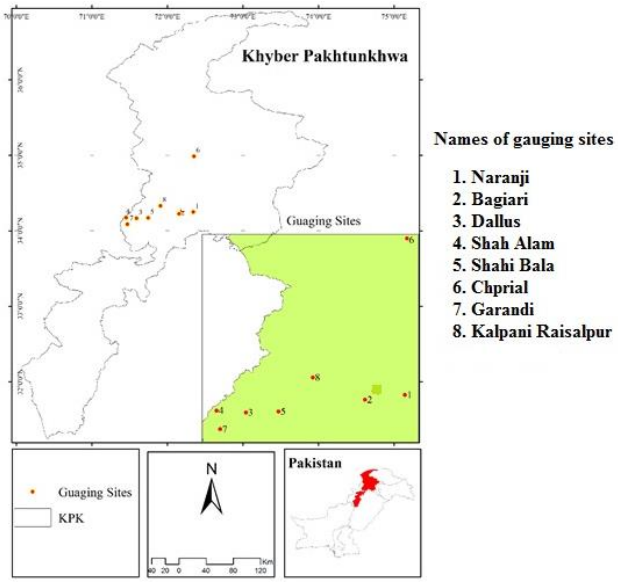

Figure 1. *Study area and geographical locations of the sites of Khyber-Pakhtunkhwa, Pakistan*

http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/1901_471489

Data preprocessing

This section provides details of measures related to data preprocessing for RFA.

Run test

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

$$Z = \frac{R-E(R)}{S.E.(R)} \quad \text{(Eq.1)}$$

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.

Rank Sum test

To validate the assumption of the identical distribution of the data of each site, Rank-Sum test 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 ten, following test statistics is used

$$W = mini(W_1 - W_2)$$

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

In case of large sample, i.e. greater than ten, the test statistic is

$$Z_w = \frac{W-\mu_w}{\sigma_w} \quad \text{(Eq.2)}$$

Here

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

Wald-Wolfowitz test

An important assumption with respect to the data that it is independent and free from significant trends is tested through the Wald-Wolfowitz test (Wald and Wolfowitz, 1943). For a sample size of less than ten, the test statistics is given as:

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

With expected mean and variance

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

http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/1901_471489

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

For large sample, i.e. greater than ten, the test statistics is:

$$Z_{ww} = \frac{K-\mu_k}{\sigma_k} \quad \text{(Eq.3)}$$

Measures of RFA

Discordancy measure (D_i) has been calculated for each site within the initial group of sites to check whether any site is discordant or not. Formula to calculate D_i is:

$$D_i = \frac{1}{3} N(t_i - \bar{t})^T S^{-1}(t_i - \bar{t}), \qquad i = 1,2,3,\ldots..N$$
$$S = \sum_{i=1}^{N} (t_i - \bar{t})\,(t_i - \bar{t})^T \quad \text{(Eq.4)}$$

where $t_i \; and \; \bar{t}$ are the ith site sample L-moments ratios and their means respectively. *N* is the total number of sites in the region.

Heterogeneity measures have been calculated to check the homogeneity of the region. The statistic to compute heterogeneity measure (*H*) is:

$$H = \frac{V-\mu_v}{\sigma_v} \quad \text{(Eq.5)}$$

where $V = \left[\frac{\sum_{i=1}^{N} n_i (\tau^i - \tau^R)^2}{\sum_{i=1}^{N} n_i}\right]^{\frac{1}{2}}$ and μ_v is average value and σ_v is the simulation-based standard deviation obtained by fitting four-parameters Kappa distribution.

|Z-Dist| statistic has been used as a goodness-of-fit criterion. The formula for |Z-Dist| is:

$$|\text{Z-Dist}| = \frac{\tau_4^{Dist} - \tau_4^R + \beta_4}{\sigma_4} \quad \text{(Eq.6)}$$

where τ_4^{Dist} is L-kurtosis of candidate probability distribution, regional L-kurtosis is τ_4^R, its bias is β_4 and the standard deviation is σ_4 calculated through simulations.

The regional quantiles using quantile function of the best-fitted regional distribution are used to estimate at-site flood quantiles using the following equation:

$$\hat{Q}_i(F) = l_1^{(i)} \hat{q}(F) \quad \text{(Eq.7)}$$

For site *i*, $\hat{Q}_i(F)$ is flood quantile at given return period, $l_1^{(i)}$ is mean of APF and $\hat{q}(F)$ is regional quantile

http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/1901_471489

Artificial neural networks

ANN is receiving increasing popularity in statistical hydrology. This study uses back propagation neural network (BPNN) because of its suitability for prediction of river flows (Maier and Dandy, 2000; Abrahart et al., 2004). The model comprises an input, hidden and output layers. Neuron layers interact via a network of feed forward weighted connection. For computations, every input of the neurons multiplied by weight which is known as connection parameter and combined output with some bias is produced. This value is managed with an activation function. 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.

$$f(x) = \frac{1}{1+e^{-x}} \quad \text{(Eq.8)}$$

The relationship between the mean of the APF of each site (l_1) and site characteristics of the region is estimated using ANN model. l_1 is the estimated or dependent variable and the site characteristics are used as an input or independent variables of the model.

Results and Discussion

Data preprocessing

The results for the validation of critical assumptions for the data at each site of the region are provided in *Table 2* including test statistics and their respective p-values.

Table 2. *Test statistics calculated and corresponding p-values of the Run Test, Rank Sum Test and Wald-Wolfowitz Test of eight sites of KPK, Pakistan*

S. No.	Site name		Run Test	Rank Sum	Wald-Wolfowitz
1	Naranji	Test statistic	-0.4460	-1.6560	0.9353
		P-value	0.6556	0.0977	0.3496
2	Bagiari	Test statistic	0.2930	1.3400	1.6385
		P-value	0.7695	0.1802	0.1013
3	Dallus	Test statistic	-1.0590	-0.6900	1.2807
		P-value	0.2892	0.4902	0.2003
4	Shah Alam	Test statistic	0.3716	-0.6640	-0.2300
		P-value	0.7102	0.5067	0.8181
5	Shahi Bala	Test statistic	-0.7675	-0.4080	1.9036
		P-value	0.4427	0.6833	0.0570
6	Chprial	Test statistic	-1.0890	1.6010	1.5120
		P-value	0.2762	0.1094	0.1305
7	Garandi	Test statistic	0.7188	1.1350	-1.0420
		P-value	0.4723	0.2564	0.2973
8	Kalpani Raisalpur	Test statistic	-0.3483	1.4670	1.0960
		P-value	0.7276	0.1424	0.2729

http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/1901_471489

The results of *Table 2* illustrate that the data of all sites fulfill important assumptions of RFA procedure at 5% significance level. Therefore, it is suitable to perform RFA.

The descriptive statistics in term of L-moments and values of D_i for each site using *Equation 4* are given in *Table 3*. The critical value of D_i for eight sites is 2.14 (Hosking and Wallis, 1997). The results of *Table 3* show that the value of D_i for the site "Garandi" is slightly greater than the critical value of 2.14. Therefore, there is a need for close examination of the observed data series at this site. A time series plot of APF of the site Garandi is illustrated in *Fig. 2*. The plot shows high variations (like observation number 25 in the year 2003) and a gradual decreasing pattern from observations number 16 to 20 (from the year 1994 to 1999). Importantly, these changes did not affect the statistical randomness of the observed APF but increased the values of L-skewness and L-kurtosis at this site, resultantly a slightly higher value of D_i.

Table 3. *Descriptive statistics and discordancy measures of eight sites of KPK, Pakistan*

S. No.	Site Names	n	l_1	t	t_3	t_4	t_5	D_i
1	Naranji	52	5447.02	0.445	0.333	0.235	0.151	1.57
2	Bagiari	31	5767.03	0.488	0.218	-0.037	0.013	1.06
3	Dallus	25	8196.84	0.474	0.252	0.066	-0.030	0.22
4	Shah Alam	30	7343.07	0.400	0.265	0.048	-0.011	0.84
5	Shahi Bala	25	2792.40	0.514	0.242	0.067	0.043	0.88
6	Chprial	34	10479.75	0.442	0.263	0.065	0.026	0.12
7	Garandi	33	1004.63	0.449	0.374	0.172	0.108	2.21
8	Kalpani Raisalpur	34	34773.34	0.368	0.336	0.159	0.075	1.10

Note: n is the number of observations at each site, l_1 is the first sample L-moment, t is the sample L-CV, t_3 is the sample L-skewness, t_4 is the sample L-kurtosis, t_5 is the sample L-moment ratio based on 5th sample L-moment and D_i is the discordancy measure

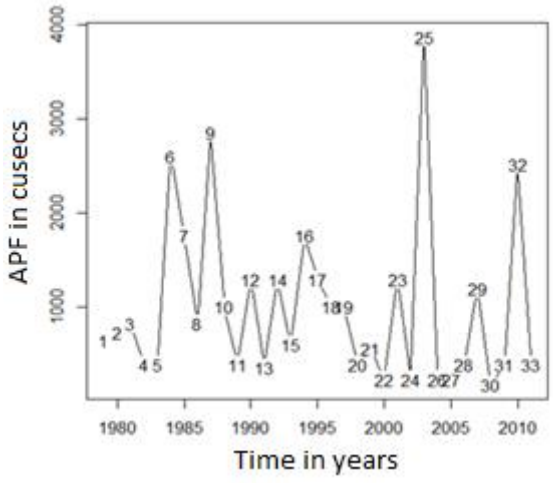

Figure 2. *Time series plot of the site Garandi*

Keeping in view that few abrupt changes are typical in extreme value analysis, the observed APF at this site have passed other preprocessing steps and less number of sites in the region, this site has been retained for further analysis. The study of Hussain and

http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/1901_471489

Pasha, 2009 had also suggested retaining a site with a D_i value less than 3 and no serious irregularities or inconsistencies in the observed data series.

Heterogeneity measures

The study area consists of eight geographically contiguous sites. Therefore, it is reasonable to assume that the factors influencing the site's flow behavior are homogenous in nature. Keeping this in view, the group of eight sites is treated as a region and heterogeneity measures are calculated using *Equation 5*. The values of *H* based on first three sample L-moment ratios are 0.52, -0.41 and 0.48, respectively; suggesting that the region is definitely homogeneous and suitable to perform RFA.

Goodness of fit measures

To find the best fit regional distribution for the study area, estimated values of |Z-Dist| statistic obtained through *Equation 6* for five standard three-parameter distributions in the family of extreme value distributions, i.e. Generalized Extreme Value (GEV), Generalized Pareto (GPA), GNO, GLO and PE3 are available in *Table 4*. These findings indicate that only one distribution, i.e. GPA distribution fulfils the criteria of best fit regional distribution (having value of $|Z\text{-Dist}| \leq 1.64$).

Table 4. *Values of | Z-Dist | statistic for each candidate distributions*

Distributions	**GLO**	**GNO**	**GEV**	**GPA**	**PE3**
| Z-dist |	5.23	3.29	4.12	1.09	1.85

L-moment ratio diagram (a graphical goodness of fit procedure), provided in *Fig. 3*, shows that average point of sample L-skewness and L-kurtosis lies closest to the theoretical curve of GPA distribution so as the tendency of the points, i.e. sample L-kurtosis and L-skewness for eight sites. Therefore, L-moment ration diagram (like |Z-Dist| statistic) favors GPA distribution as best-fit regional distribution. The results of both goodness-of-fit measures are in agreement to each other.

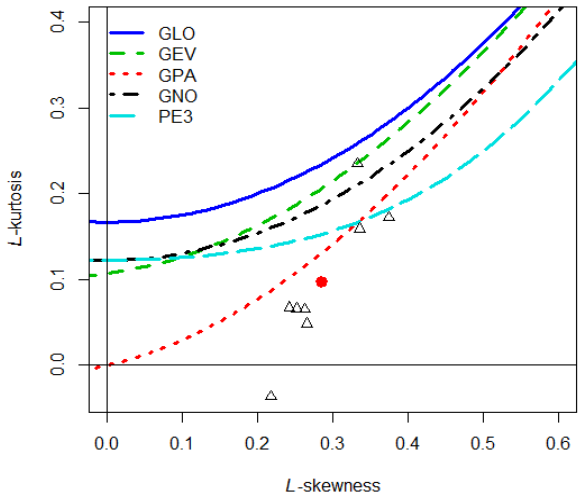

Figure 3. *L-moment ratio diagram of the region*

http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/1901_471489

Estimates of parameters of regional distribution and quantiles

After the selection of GPA as best-fit regional distribution, its parameters have been estimated by the method of L-moments. The quantile function of GPA distribution has been used to estimate the regional flood quantiles for 5, 10, 20, 50 and 100 years return periods and results are provided in *Table 5*.

Table 5. *Estimated parameters of GPA distribution and regional quantiles for various return periods in years*

Distribution	Estimated Parameters			Quantiles				
	ε	α	k	5	10	20	50	100
GPA	0.0685	1.0209	0.0960	1.5910	2.1775	2.7264	3.3980	3.8682

Estimated quantiles at each site using *Equation 7* are given in *Table 6* for 5, 10, 20, 50 and 100 years return periods.

Table 6. *Estimated quantiles at each site using Equation 4 for various return periods*

S. No.	Site names	At-site quantiles				
		5	**10**	**20**	**50**	**100**
1	Naranji	8666	11861	14851	18509	21070
2	Bagiari	9175	12558	15723	19596	22308
3	Dallus	13041	17849	22348	27853	31707
4	Shah Alam	11683	15990	20020	24952	28404
5	Shahi Bala	4443	6080	7613	9489	10802
6	Chprial	16673	228120	28572	35610	40538
7	Garandi	1598	2188	2739	3414	3886
8	Kalpani Raisalpur	55324	75719	94806	118160	134510

Estimation for ungauged sites using OLS regression and ANN

As mentioned earlier that this study has used OLS regression and ANN to develop a model with dependent variable l_1 (at-site mean of APF) and site characteristics as independent variables. The estimates of the model will be used to predict the mean of APF for an ungauged site within the homogeneous region. This estimated mean can be used to predict the quantiles of ungauged site using index flood procedure.

Development of OLS regression model

For the development of a log-transformed simple linear regression model, the most relevant site characteristic is used as an independent variable. The development of multiple linear regression model has not been considered due to the presence of strong significant correlations between the explanatory variables (i.e. the problem of multi-collinearity). The details of the procedure are:

To identify the most important site characteristic among available, i.e. having the highest degree of linear relationship with l_1, correlations between l_1 and site characteristic have been calculated. This correlation matrix is provided in *Table 7*.

http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/1901_471489

Table 7. *Correlations between l_1 and site characteristics. Parenthesis include P-values for testing the significance of correlation coefficient*

	l_1	Latitude	Longitude	Elevation	AARF	ARMS
l_1	1.0000	0.2366 (0.5727)	0.1434 (0.7348)	0.0564 (0.8945)	0.3343 (0.4143)	0.3426 (0.4061)
Latitude	0.2366	1.0000	0.6333 (0.0919)	0.9677 (0.0001)	0.1221 (0.7733)	0.3318 (0.4220)
Longitude	0.1434	0.6333	1.0000	0.4913 (0.2163)	0.7767 (0.0234)	0.7188 (0.0445)
Elevation	0.0564	0.9677	0.4913	1.0000	-0.0776 (0.8551)	0.1847 (0.6615)
AARF	0.3343	0.1221	0.7767	-0.0776	1.0000	0.8601 (0.0061)
ARMS	0.3426	0.3318	0.7188	0.1847	0.8601	1.0000

Results of *Table 7* show that the correlation between $l_1$and ARMS is highest and statistically significant. Therefore, ARMS can be chosen as the most relevant variable for the development of simple linear regression model. However, for further investigation of the choice of ARMS as the most suitable independent variable, the frequencies of recorded APF at eight sites during four seasons of a year namely summer (monsoon), autumn, winter and spring have been calculated to observe the trends of occurrence of APF in a season, if any. The percentages of these frequencies of occurrence are illustrated in *Table 8*. These values show that the highest percentage of frequencies of occurrence of APF at all sites is in the monsoon season. Therefore, ARMS is the most suitable variable for the development of a simple linear regression model.

Table 8. *Percentage (%) of frequencies of occurrence of APF in four seasons of a year.*

S. No.	Site names	n_i	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

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

Based on the above discussion, the model using OLS estimation method in log-transformed form is:

$$ln(\hat{l}_1) = 1.6702\ ln(ARMS) \quad \text{(Eq.9)}$$

The intercept term is not included in the model 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

http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/1901_471489

Pakistan are usually dependent on the monsoon rainfall (Hussain and Pasha, 2009)). For the estimated model in *Equation 9*, the value of R^2 (coefficient of determination) is 0.9931 and adjusted-R^2 is 0.9921. This show that the linear regression line fits the data of 8 sites adequately. The estimated regression coefficient, standard error of the estimate, t-calculated and its corresponding p-value are given in *Table 9*. Results of *Table 9* show that the estimated regression coefficient is statistically significant with low standard error.

Table 9. *Results of fitted regression model. Here ARMS is average rainfall in monsoon season and S.E. is standard error of the regression coefficient*

	Coefficient	**S.E.**	**t-value**	**P-value**
ln ($\boldsymbol{ARMS}$)	1.6702	0.0525	31.81	0.0000

These details show that the model is adequate, still, assumptions related to the error term (normality, zero mean and homoscedasticity) are requisite (for details, see Gujarati, 2003). To check these assumptions, normal probability plot of residuals (illustrated in *Fig. 4*) and Anderson-Darling normality test with the null hypothesis that "the residuals follow normal distribution" have been applied. The calculated value of Anderson-Darling test statistic is 0.268 with its corresponding p-value as 0.58. As the p-value exceeds 5% level of significance; therefore, we are unable to reject the null hypothesis that the error term follows normal distribution. Moreover, the probability plot of residuals in *Fig. 4* indicate that the normality assumption with respect to residuals seems appropriate as the points follow the straight line, with a standard deviation of 0.78 and mean as zero. 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. The corresponding test statistic for White's test is W = 0.0260 with the corresponding p-value as 0.88. This shows that we are unable to reject that the residuals are homoscedastic. All these details show that the estimated regression model in *Equation 9* is an adequate fit. Therefore, can be used to predict l_1 for each ungauged site within the homogeneous region.

Artificial neural networks

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 percent, 20 percent and 20 percent, respectively. This division is useful for large data. In the present study, leave one out cross-validation (LOOCV) approach is used for the training and validation of the sample data set as it is usually considered more useful for smaller data sets. In LOOCV approach, the data set is divided 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 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 test mean squared error can be obtained from n test errors as:

$$CV_{(n)} = \frac{1}{n}\sum_{i=1}^{n} MSE_i \qquad \text{(Eq.10)}$$

http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/1901_471489

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 of ANN is to reduce the error among the target output and ANN output through adjusting weights.

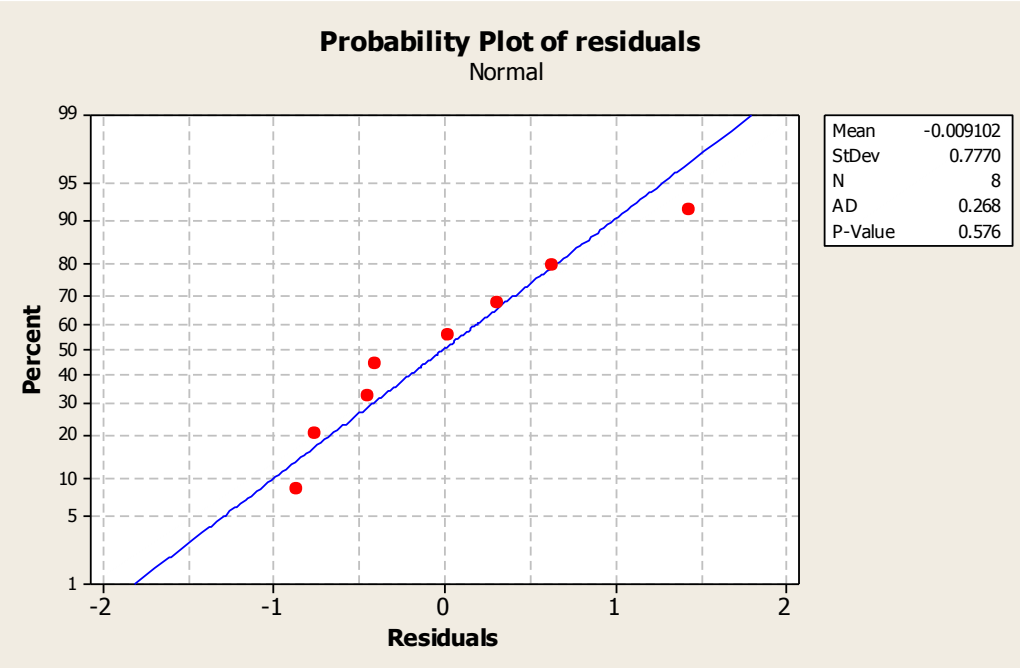

Figure 4. *Probability plot of estimated residuals, number of observations (N), arithmetic mean (mean), standard deviation (StDev) and Anderson-Darling test statistic (AD) with the corresponding p-value*

The "caret" package of R-language has been used for the training of ANN. To select the best ANN 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 entire region. The model with minimum MSE relative to other models has been selected. ANN algorithm with two hidden layers, four neurons in the first layer and two in the second layer, have been used. To avoid overfitting of the model so as ensuring the quality of the developed ANN model, testing MSE and training MSE have been compared. Training of ANN has been terminated for an observed increase in the test MSE or even decrease in the training MSE. The function of input and output variables is:

$$f(l_1) = g(lat, long, ele, AARF, ARMS) \quad \text{(Eq.11)}$$

Few of the published studies have also developed such ANN model with only two input variables and three hidden layers for the estimation of floods, for instance, Aziz et al. (2014). The graphical representation of the fitted model is shown in *Fig. 5*.

Predicted values of $l_1$by using ANN and OLS regression methods are provided in *Table 10*. The results show that predicted values of l_1 through ANN are close to their true values relative to regression estimates.

http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/1901_471489

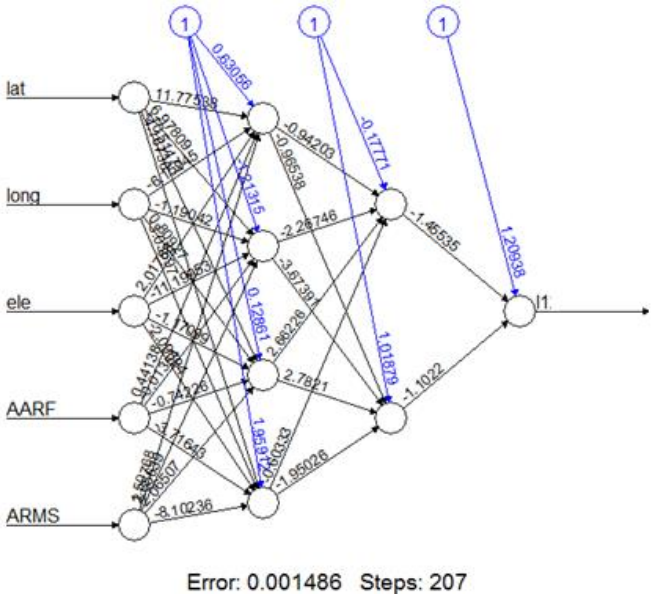

Figure 5. *Graphical representation of ANN procedure to show the convergence of the model. Average annual rainfall (AARF), longitude (long), latitude (lat), elevation (ele), average rainfall during monsoon season (ARMS) and l_1 is the mean of APF of each site*

Table 10. *Estimated values of l_1 of gauged sites through ANN and regression analysis (RA)*

S. No.	Site name	l_1 (observed)	l_1 (fitted) using ANN	l_1 (fitted) using RA
1	Naranji	5447	5306	11648
2	Bagiari	5767	6516	8611
3	Dallus	8197	8705	4359
4	Shah Alam	7343	7311	7204
5	Shahi Bala	2792	2236	4359
6	Chprial	10480	11034	7682
7	Garandi	1005	979	2376
8	Kalpani Raisalpur	34773	33383	8297

Practical validation of the estimates

The estimates of flood quantiles obtained through RFA are accurate and reliable but their practical validation is still crucial. For this assessment, a comparison is given in *Table 11* where the estimated quantiles are compared with the first and second highest values of APF for each respective site. Results of *Table 11* show that the predicted quantiles through OLS regression analysis for smaller return period (10 years) are comparable with the highest values of observed APF for Naranji, Bagiari, Shah Alam, Shahi Bala, Chprial and Garandi sites. 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 analysis for longer return periods (100 years) or outside the available span of the data, show large deviations from highest values of observed APF for all the sites. This is a usual and major disadvantage of using OLS regression analysis for estimating flood quantiles for longer return periods or beyond the span of the observed data series.

http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/1901_471489

The comparison further reveals that the estimated flood quantiles obtained using RFA are nearly close to the highest values of observed APF of all the sites for shorter and longer return periods. This shows the strength of the adopted procedure for the estimation of quantiles. Moreover, the comparison of the estimated quantiles using ANN reveals that the estimates are accurate and close to the highest values of APF for all the sites. Therefore, ANN would be a preferred method relative to OLS regression for estimation at ungauged sites within the region.

Table 11. *Comparison of estimated flood quantiles using RFA, artificial neural networks (ANN) and regression analysis (RA) with highest observed APF at various sites*

S. no.	Site name	Highest observed values of APF (year)		Estimates using RFA			Estimates using ANN			Estimates using RA		
		Highest	2nd Highest	10	20	100	10	20	100	10	20	100
1	Naranji	30000 (2010)	15704 (1997)	11861	14851	21070	11554	14466	20524	25364	31757	45057
2	Bagiari	16688 (2006)	16023 (2010)	12558	15723	22308	14188	17764	25204	18751	23478	33310
3	Dallus	21700 (2010)	19984 (2006)	17849	22348	31707	18954	23732	33671	9491	11884	16860
4	Shah Alam	20000 (2010)	18513 (2005)	15990	20020	28404	15919	19931	28278	15687	19641	27867
5	Shahi Bala	8911 (1995)	7427 (1996)	6080	7613	10802	4870	6097	8651	9491	11884	16860
6	Chprial	33836 (1993)	24639 (2003)	22820	28572	40538	24025	30082	42680	16728	20945	29716
7	Garandi	3869 (2003)	2860 (1987)	2188	2739	3886	2133	2670	3789	5173	6478	9190
8	Kalpani Raisalpur	118604 (2010)	80615 (2008)	75719	94806	134510	72691	91015	129131	18066	22621	32094

Summary and Conclusion

The results of this study contributes 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 ANN to estimate floods and so on. Few key findings are summarized below:

For preprocessing of the APF at various sites, necessary assumptions have been tested using nonparametric tests. The findings show that the recorded data sets at all sites are random, independent, identically distributed, and free from significant trends. The values of discordancy statistic D_i show that the site "Garandi" is discordant in the group of eight sites. One possible reason for the slightly higher value of D_i may be high skewness of the data due to the presence of high outliers. In flood estimation, high outliers should not be discarded from the data. Moreover, considering the fewer number of sites for the analysis and the observed APF have passed the other preprocessing steps related to RFA; the site "Garandi" is retained in the analysis.

The estimates of L-moment ratios showed that there exists 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

http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/1901_471489

monsoon rainfall. Hussain (2017) found similar results for the sites of river basins in Punjab, Pakistan.

The set of eight sites is homogenous as confirmed by the heterogeneity measure based on L-moments. After the confirmation of homogeneity in the region, L-moment ration diagram and |Z-Dist| statistic showed that GPA distribution is the best-fit regional distribution. Estimates of regional quantiles show the rising trend for smaller to longer return periods and larger than the mean values of recorded APF of each site. The results of this study reveal that the shape/distribution associated to the frequency of observed APF for the sites of north-western streams and rivers of the country is different relative to the sites of the Indus River and its major eastern tributaries. For the region of KPK, the identified regional distribution is GPA while for the sites of the Indus River and its eastern tributaries these are GNO, GLO or PE3, etc. as suggested in Hussain and Pasha (2009), Hussain (2011, 2017), and Ahmad et al. (2017). This shows that the observed APF of various sites of the current study area has low L-kurtosis values against high values of L-skewness.

OLS regression and ANN methods have been used to predict the average of APF for ungauged sites. Comparison of the estimates reveals that predictions based on ANN are more accurate relative to the OLS regression.

For applied validation of the estimates, a comparison has been demonstrated using the first and second highest of observed APF of each respective site. RFA estimates have similar tendencies like the highest recorded APF for various return periods at all sites. In addition, the estimates using ANN (although the given comparison is only for gauged locations) are very similar to the highest observed APF of each respective site for various return period quantiles. Hence, ANN is a preferred method for estimating flood quantiles at ungauged sites within the homogeneous regions (particularly for larger return periods).

Findings of this research will not only useful for officials concerned with the management of flood risk but it will be beneficial for agriculture water management and to improve design capacity of the current and proposed hydrological schemes within the study region. These results can also be useful to improve the quality of quantiles of poorly gauged sites within the homogeneous region. For future studies, the focus would be to include maximum available sites of the province to perform RFA. Secondly, the inclusion of few other site characteristics for the development of models to estimate quantiles at ungauged sites. Another important area is to perform RFA using variables other than annual maxima's like 3 days, 5 days or 7 days maxima's to add more data for the application of L-moments based RFA. Supposedly, it will further improve the quality and usefulness of the estimates for the officials dealing with disasters management.

Acknowledgements. Authors are very grateful to the Higher Education Commission, Pakistan for financial support under the project number: 5790/Federal/NRPU/R&D/HEC/2016. We are also thankful to the Irrigation Department of Khyber Pakhtunkhwa for providing flood data for the study.

http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/1901_471489

REFERENCES

[1] Abrahart, R., Kneale, P. E., See, L. M. (eds.) (2004): Neural networks for hydrological modeling. – CRC Press.
[2] 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.
[3] 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): 2255-2264.
[4] Ahmad, I., Yasin, M., Fawad, M., Saghir, A. (2017): Regional frequency analysis of low flows using L-moments for Indus Basin, in Pakistan. – Pakistan Journal of Science 69(1): 75-84.
[5] 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.
[6] 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.
[7] 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.
[8] 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.
[9] 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.
[10] Bradley, J. V. (1968): Distribution-free statistical tests. – No. 04; QA278. 8, B7.
[11] Cunnane, C. (1988): Methods and merits of regional flood frequency analysis. – Journal of Hydrology 100(1-3): 269-290.
[12] Faridzad, M., Yang, T., Hsu, K., Sorooshian, S., Xiao, C. (2018): Rainfall frequency analysis for ungauged regions using remotely sensed precipitation information. – Journal of hydrology 563: 123-142.
[13] 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.
[14] 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. https://doi.org/10.1016/j.energy.2019.05.153.
[15] Gado, T. A., Hsu, K., Sorooshian, S. (2017): Rainfall frequency analysis for ungauged sites using satellite precipitation products. – Journal of Hydrology 554: 646-655.
[16] Government of Pakistan (2016): Annual flood report 2016. – 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.
[17] GREHYS (1996a): Inter-comparison of regional flood frequency procedures for Canadian rivers. – Journal of hydrology (Amsterdam) 186: 85-103.
[18] GREHYS (1996b): Presentation and review of some methods for regional flood frequency analysis. – Journal of hydrology (Amsterdam) 186: 63-84.
[19] Griffis, V. W., Stedinger, J. R. (2007): The use of GLS regression in regional hydrologic analyses. – Journal of Hydrology 344(1-2): 82-95.

http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/1901_471489

[20] Gujarati, D. N. (2003): Basic Econometrics. – McGraw-Hill, New York.
[21] 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.
[22] 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.
[23] Hosking, J. R. M., Wallis, J. R. (1993): Some statistics useful in regional frequency analysis. – Water resources research 29(2): 271-281.
[24] Hosking, J. R. M., Wallis, J. R. (1997): Regional frequency analysis: an approach based on L-moments. – Cambridge University Press.
[25] 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.
[26] 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.
[27] 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.
[28] James, G., Witten, D., Hastie, T., Tibshirani, R. (2013): An introduction to statistical learning with Applications in R. – New York: Springer.
[29] 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, article ID: 6954902.
[30] Kuhn, M., Johnson, K. (2013): Applied predictive modeling. – Springer, New York.
[31] 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.
[32] 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.
[33] Malekinezhad, H., Zare-Garizi, A. (2014): Regional frequency analysis of daily rainfall extremes using L-moments approach. – Atmósfera 27(4): 411-427.
[34] 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.
[35] 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.
[36] 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.
[37] 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(9): 2521-2539.
[38] 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.
[39] 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 forecast combination. – Hydrology and earth system sciences 6(4): 671-684.

http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/1901_471489

[40] 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.
[41] 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.
[42] 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.

http://www.aloki.hu ● ISSN 1589 1623 (Print) ● ISSN 1785 0037 (Online)
DOI: http://dx.doi.org/10.15666/aeer/1901_471489