

At Site and Regional Frequency Analysis for Sudan Annual Rainfall by Using the L-Moments and Nonlinear Regression Techniques

Badreldin G. H. Hassan¹, Isameldin A. Atiem², Li Jianzhu³, Feng Ping⁴

^{1,3}Civil Engineering Department, Tianjin University, Tianjin-CHINA

²Faculty of Engineering Science, University of Nyala, Nyala, SUDAN

Abstract—In this study, a regional rainfall frequency analysis has been carried out, using the index flood L-moments approach. Mean annual rainfall data observed at 15 gauged sites on the Sudan are investigated. The aim of the study is to investigate and derive hydrological homogenous region or regions, to identify and establish the regional statistical distribution and to extend the methodologies to the case of ungauged site. For this purpose, five distribution functions are used, namely: (generalized logistic, generalized extreme-value, generalized normal, Pearson type-3 and generalized pareto distributions.) Analyses have shown that only region 3 is form a hydrological homogenous region, and this region follows a generalized Pareto (GPA) distribution. Furthermore, the other remaining regions (possibly heterogeneous and definitely heterogeneous) are also defined. Regional dimensionless growth curves for the identified regions are derived. Results are assessed on the basis of RMSE through the use of Monte Carlo simulation and a nonlinear regression model is derived and the results are investigated.

Keywords—Index flood, L-moments, Monte Carlo simulation, Nonlinear regression, Regional rainfall frequency.

I. INTRODUCTION

Rainfall frequency analysis plays an important role in hydrologic and economic evaluation of water resources projects. It helps to estimate the return periods and their corresponding event magnitudes, thereby creating reasonable design criteria. The basic problem in rainfall studies is an information problem which can be approached through frequency analysis. The classical approach to rainfall frequency analysis is hampered by insufficient gauging network and insufficient data, especially when the interest is in estimating events of large return periods. At-site rainfall frequency analysis is the analysis in which only rainfall records from the subject site are used. More commonly, it will be necessary to carry out a regional analysis where rainfall records from a group of similar catchments are used. Regionalization or regional analyses are thought to compensate for the lack of temporal data.

Regional rainfall frequency analysis, RRFA, is a probabilistic method applied to rainfall records. It attempts to respond to the need for rainfall estimation in ungauged basins and for improving the at-site estimate by using the available rainfall data within a region. Thus, it enables rainfall quantile estimates for any site in a region to be expressed in terms of rainfall data recorded at all gauging sites in that region, including those at the specific site.

The index-flood procedures are a convenient way of pooling summary statistic from different data samples. The term “index-flood” arose because early applications of the procedure were to flood data in hydrology [1], but the method can be used with any kind of data. The index flood method assumes that a region is a set of gauging sites whose flood frequency behavior is homogeneous in some quantifiable manner, and it is expected that the more homogeneous a region is, the greater is the gain in using regional instead of at-site estimation. Hosking et al. [2], Wallis & Wood [10], Lettenmaier & Potter [8], and Jin & Stedinger [5] have demonstrated that index flood procedures yield suitably robust and accurate quantile estimates. This study uses the mean annual rainfall data at 15 sites in the Sudan and explores various rainfall frequency models, using index flood L-moments. Included models are GEV, GLO, PE-3, GPA, and GNO. The study reveals that GPA model provides a good fit to rainfall data in the region.

II. MATERIALS AND METHODS

The Study Area Description. The data used have been collected from 15 raingauging stations, the records of which are published by the Sudan National Meteorology Department. Altitude and latitude have been assumed as initial statistics of hydrologic homogeneity, and station selection criteria were based on these characteristics. All sites used in this procedure are located between 31°50' and 33°25' N latitudes, and 1530 and 2300 m altitudes. Records used for the analysis have ended the same year, and there are no gaps in the records (See Table 1 & Fig. 1).

Diagnostic Statistical Tests. Hosking & Wallis [3] addressed that regional frequency analysis involves four stages, the first three of which involve subjective judgment: (a) screening of data by means of the discordance measure, D_i , which provides an initial screening of the data and identifies unusual sites in a region; (b) identification of homogeneous regions which, is the assignment of the sites to regions by the means of heterogeneity measure, H, that performs the test by calculating summary statistics (sample L-moments) of the at-site data and compares the between-site variability of these

statistics with what would be expected of a homogeneous region; (c) choice of a regional frequency distribution by means of the goodness-of-fit measure, Z, which assesses whether a candidate distribution provides an adequate fit to the data; and (d) the estimation of regional frequency distribution.

Discordancy Measure, D_i , Test for Site Data. When a single site does not appear to belong to the cloud of (τ_3, τ_4) points on the L-moment diagram, a test of discordance can be used to determine whether it should be removed from the region. Discordance measure, D_i , is used to screen out the data from unusual sites; and the test is applied by calculating the D-statistic, which is defined in terms of L-moments. Let $u_i = [\tau_i, \tau_{3i}, \tau_{4i}]^T$ be a vector containing the L-ratios for site i [3]-[7]. If the group averages \bar{u} and sample covariance matrix S are defined as:

$$\bar{u} = \frac{1}{n} \sum_{i=1}^n u_i \quad (1)$$

$$S = \frac{1}{(N-1)} \sum_{i=1}^n (u_i - \bar{u})(u_i - \bar{u})^T \quad (2)$$

then the discordancy measure for site i is:

$$D_i = 3^{-1} (u_i - \bar{u})^T S^{-1} (u_i - \bar{u}) \quad (3)$$

Where n is the total number of sites. Large values of D_i indicate sites that are the most discordant from the group as a whole and are most worthy of investigation for the presence of data errors [3].

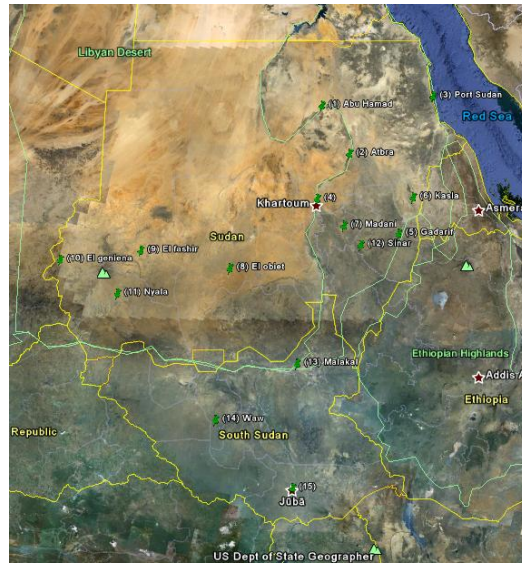


Figure 1. Satalite map of Sudan shows the study area

Heterogeneity Test. L-moment heterogeneity tests assess whether a group of sites might reasonably be treated as a homogenous region. The heterogeneity measure compares the between-site variations in sample L-moments for the group of sites with what would be expected for a homogeneous region [3].

The homogeneity test used in this study is the test that is proposed by Hosking & Wallis [3] and is based on various orders of sample L-moment ratios. It is particularly based on the variability of three different levels of tests: a test based on the L-CV only; a test based on the L-CV and L-skewness; and a test based on the L-skewness and L-kurtosis. Then, a region can be declared homogeneous with a corresponding order of L-moment if $H < 1$; the region is possibly homogeneous if $1 \leq H < 2$; and it is definitely declared heterogeneous if $H \geq 2$.

III. THE RESULTS AND DISCUSSION

Identification of Homogeneous Regions using Cluster Analysis. In this study, determination of homogeneous regions is done by Ward's clustering method as a much known approach for determination of homogeneous regions in regional frequency analysis. The physical characteristics such as the area, longitude, latitude, and the elevations of selected stations in the basin, was subjected to hierarchical clustering based on the Ward's method using Euclidean distance.

Estimation of Heterogeneity Measures. To determine whether each region in Sudan is heterogeneous, the samples L-moment ratio of data were calculated. The results are shown in Table 1.

Table 1. The hydrological and geographical characteristic of the rainfall sites selected in Sudan

Site name	Code	Data length	Mean ARF	L-cv	L-skew.	L-kurt.	lat.	Long.	Area.	Elev.
1.Abu Hamad	02ABH640	58	10.6034	0.7029	0.5437	0.3206	19.533	33.333	122.1	299.6
2.Atbra	02ATB680	58	59.3793	0.4231	0.1992	0.1345	17.667	33.967	30.4	316.3
3.Port Sudan	03PSD641	15	63.8933	0.42	0.0645	0.0767	19.583	37.217	218.8	525.3
Khartuom	04KHA721	102	150.8023	0.2835	0.1752	0.168	15.6	32.55	22.1	39.5
Gadarif	05GDF752	41	603.5854	0.1053	-0.0233	0.205	14.033	35.4	75.2	363.5
Kasala	05KSL730	28	240.7354	0.198	-0.023	0.1379	15.467	36.4	36.7	522.6
Madani	06WMD751	58	305.9828	0.1762	0.0403	0.096	14.383	33.483	27.5	416.7
Elobiet	08OBT771	90	359.3592	0.171	0.086	0.1399	13.183	30.217	185.3	492.8
Elfashir	09FSH760	84	262.332	0.2223	0.193	0.1969	13.617	25.333	296.4	405.7
Elgeniena	09GEN770	30	447.322	0.1901	0.0509	0.1614	13.483	22.45	79.4	526.2
Nyala	10NYL790	43	393.3309	0.1376	0.0674	0.1113	12.05	24.883	127.3	775.8
Senar	12SNR762	60	458.7983	0.1173	-0.0342	0.1466	13.55	33.617	37.8	692
Malakal	13MLK840	58	742.5498	0.1068	0.0463	0.1436	9.55	31.65	77.7	392.4
Waw	14AWE852	28	1078.941	0.1003	0.042	0.0979	8.767	27.4	93.9	487.8
Juba	15JUB941	43	961.4919	0.1105	0.0177	0.1987	4.867	31.6	22.9	468.2

Fig. 2. are the L-moment ratio diagrams of sites in Sudan which depict that Sudan may not be classified by L-moment ratio. The variation of L-CV is small, but that of L-skewness and L-kurtosis are relatively large. Although the low L-CV value endows Regional Shape Estimation superiority [9], it is well-admitted that the small variation of L-CV allows Index Flood method to be superior [4]. It is found that small variation of L-moment makes the heterogeneity measure low.

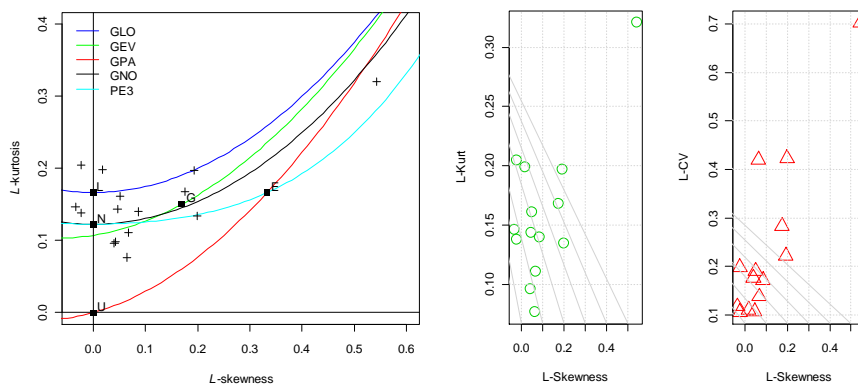


Figure 2. L-moments Ratio Diagram of sites in Sudan

Table 2. shows the heterogeneity measure H_1 , H_2 , and H_3 of each region. These measures were calculated by conducting Monte Carlo simulation after estimating parameters of kappa distribution, using sample L-moment ratio.

Table 2. The Heterogeneity measure of the Regions according to Ward's cluster

Regions	H1		H2		H3	
	Value	Judgment	Value	Judgment	Value	Judgment
Region 1	28.50	He	9.53	He	3.72	He
Region 2	1.83	P.He	-0.58	Ho	-1.04	Ho
Region 3	-1.42	Ho	-0.80	Ho	-0.76	Ho
Region 4	11.87	He	5.11	He	1.63	P. He
Region 5	2.31	He	1.02	P. He	0.66	Ho
Region 6	2.97	He	3.56	He	2.32	He

(Note) Ho. : Homogeneous, P. He. : Possibly heterogeneous, He. : Heterogeneous

Where region 1 is the whole stations, region 2 includes two stations (Nyala and Geneina), region 3 includes three stations (Gadarif, Malakal and Wau), region 4 includes four stations (Atbra, Kasala, Medani and Sennar), region 5 includes two stations (El obiet and El Fashir) and reion 6 includes two stations (Abu hamad and Portsudan).

H_1 is the measure estimated by using the sample L-CV, H_2 is calculated from the sample L-skewness, and H_3 is the sample L-kurtosis derived from the sample L-kurtosis. According to the standard by Hosking and Wallis [3,4], all regions except region (3 and 2) were judged to be heterogeneous. Region 2 seemed to be "possibly" heterogeneous. In most of the regions, H_1 was positive. This implies that the dependence of each site is not very strong except region 3. In other words, it means that one rainfall event may not effect on many sites. In addition, small variances of means and L-CV produce the small heterogeneity of measures.

Goodness-of-fit Test. The goodness-of-fit measure, Z, judges how well the simulated L-skewness and L-kurtosis of a fitted distribution matches the regional average L-skewness and L-kurtosis values obtained from the observed data. For a selected distribution, Z is defined by Hosking & Wallis [3] as:

$$Z^{dist} = (\tau_4^{dist} - \bar{\tau}_4 + \beta_4) / \sigma_4 \tag{4}$$

where $\bar{\tau}_4$ is the average L-kurtosis value computed from the data of a given region; τ_4^{dist} is the average L-kurtosis value computed from simulation for a fitted distribution; β_4 is the bias of τ_4 ; and σ_4 is the standard deviation of L-kurtosis values obtained from simulation. Thus:

$$\beta_4 = (N_{sim})^{-1} \sum_{m=1}^{N_{sim}} (\bar{\tau}_4^m - \bar{\tau}_4) \tag{5}$$

$$\sigma_4 = \left[(N_{sim})^{-1} \sum_{m=1}^{N_{sim}} [(\bar{\tau}_4^m - \bar{\tau}_4)^2 - (N_{sim}\beta_4)^2] \right]^{1/2} \tag{6}$$

Where $\bar{\tau}_4^m$ is the regional average L-kurtosis and is to be calculated for the mth simulated region. Then, a given distribution is declared to be of adequate fit if Zdist is sufficiently close to zero. A reasonable criterion is $|Z^{dist}| \leq 1.64$. This criterion corresponds to acceptance of the hypothesized distribution at a confidence level of 90% and shows approximately a standard normal distribution if the at-site L-kurtosis statistics have independent identical normal distributions.

According to Table 3. and Fig. 2. the GPA distribution is the most powerful distribution. There is no case that GPA distribution is rejected and other distribution is accepted. Although all distributions are rejected, GPA distribution produces the smallest $|Z|$ value for all regions. Therefore, GPA distribution can be the most appropriate distribution in estimating rainfall quantiles of Sudan.

Table 3. The result of goodness of fit test

Region number	GLO		GEV		GNO		PEIII		GPA	
	Z	Judg.	Z	Judg.	Z	Judg.	Z	Judg.	Z	Judgment
1	1.06	NG	-2.42	OK	-2.31	OK	-2.82	OK	-9.64	OK
2	1.02	NG	-0.35	NG	-0.13	NG	-0.18	NG	-2.99	OK
3	0.47	NG	-1.33	NG	-0.93	NG	-0.94	NG	-4.66	OK
4	1.78	OK	-0.47	NG	-0.09	NG	-0.15	NG	-4.79	OK
5	0.49	NG	-1.01	NG	-1.04	NG	-1.35	NG	-4.16	OK
6	0.50	NG	0.34	NG	-0.11	NG	-0.88	NG	-0.31	NG
Accepted (%)	1 (16%)		1 (16%)		1 (16%)		1 (16%)		5 (83%)	

Quantile Estimates. GPA distribution is the most appropriate distribution in estimating rainfall quantiles of Sudan according to the goodness of fit test. Quantiles of each site calculated by two different methods (At-site, and index-flood methods). These methods respectively led to different results in accordance with record length and L-moment ratio. 10 sites were chosen for comparison. Site 08OBT771, 09FSH760, and 12SNR762 have a large record length, but site 05KSL730 and 09GEN770 hold a small record length.

Additionally, the L-CV of site 02ATB680 and 09FSH760 are large, but site 05GDF752, 13MLK840, and 12SNR762 have a small L-CV. Also the L-skewness of site 02ATB680 and 09FSH760 are large, otherwise site 12SNR762, 05GDF752 and 05KSL730 have a small L-skewness. Table 4. Shows site statistics of chosen sites.

Table 4. Site statistics of chosen 10 sites

Region number	Site number	Region length	Mean (mm)	L-CV	L-skewness	L-kurtosis
1	02ATB680	58	59.3793	0.4231	0.1992	0.1345
2	05GDF752	41	603.5854	0.1053	-0.0233	0.205
3	05KSL730	28	240.7354	0.198	-0.023	0.1379
4	06WMD751	58	305.9828	0.1762	0.0403	0.096

5	08OBT771	90	359.3592	0.171	0.086	0.1399
6	09FSH760	84	262.332	0.2223	0.193	0.1969
7	09GEN770	30	447.322	0.1901	0.0509	0.1614
8	10NYL790	43	393.3309	0.1376	0.0674	0.1113
9	12SNR762	60	458.7983	0.1173	-0.0342	0.1466
10	13MLK840	58	742.5498	0.1068	0.0463	0.1436

Table 5. shows/depicts quantiles of each method by using the GPA distribution. Results are divided into three cases: (1) the quantiles of regional frequency analysis are larger than at-site quantiles. (2) the quantiles of regional frequency analysis are less than at-site quantiles. (3) The quantiles of regional frequency analysis are similar to at-site quantiles.

Table 5. Quantiles of each method (the GPA distribution,) (Unit: mm)

Case	Site	Records length	Methods	Return period (year)			
				10	100	1000	
$Q_i > Q_A$	02ATB680	58	At-site	95.664	114.715	119.736	
			Index flood	95.665	114.717	119.738	
	08OBT771	90	At-site	578.953	694.248	724.637	
			Index flood	578.954	694.250	724.639	
	12SNR762	60	At-site	739.156	886.355	925.153	
			Index flood	739.159	886.358	925.157	
05GDF752	41	At-site	972.419	1166.07	1217.112		
		Index flood	972.410	1166.06	1217.102		
$Q_i < Q_A$	05KSL730	28	At-site	387.842	465.078	485.436	
			Index flood	387.833	465.068	485.425	
	06WMD751	58	At-site	492.960	591.130	617.005	
			Index flood	492.955	591.124	617.000	
	09FSH760	84	At-site	422.635	506.800	528.985	
			Index flood	422.632	506.797	528.981	
	09GEN770	30	At-site	720.667	864.184	902.012	
			Index flood	720.6639	864.180	902.008	
	10NYL790	43	At-site	720.6672	864.184	902.012	
			Index flood	720.664	864.180	902.008	
	$Q_i \approx Q_A$	13MLK840	58	At-site	1196.3	1434.536	1497.33
				Index flood	1196.3	1434.536	1497.33

Q_R : quantiles of regional frequency analysis,
 Q_A : quantiles of at-site frequency analysis

In case of site 02ABH640 and 03PSD641, where GPA distribution is rejected according to goodness of fit test, the quantile of Wakeby distribution is also displayed in Table 6.

Table 6. Quantiles of each method (the Wakeby distribution) (Unit: mm)

Site	Methods	Return period in years		
		10	100	1000
02ABH640	At-site	17.08282	20.4848	21.3815
	Index flood	17.0773	20.4782	21.3746
03PSD641	At-site	102.9366	123.4358	128.839
	Index flood	102.9313	123.4294	128.832

The sites for the cases one and two shows small differences in estimated quantiles by each method, due to their long record length. Therefore, it is judged that the variances of quantiles in the two methods are stable.

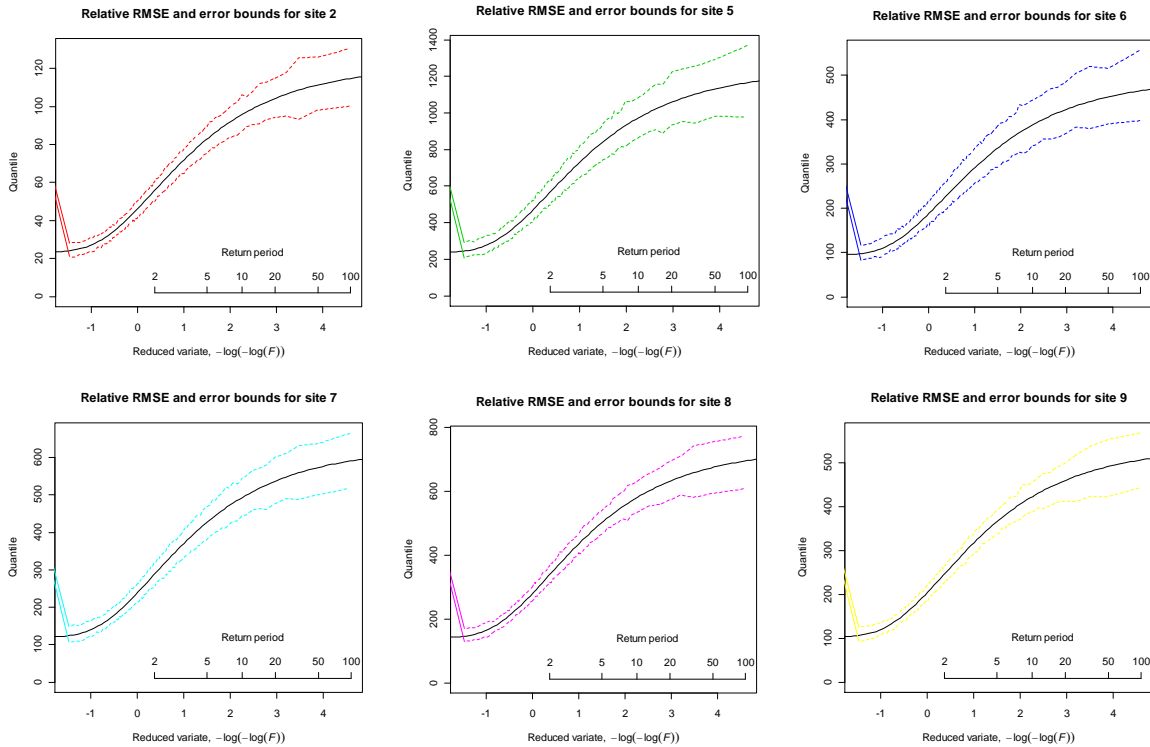
Calculation of Relative Root Mean Square Error. After the simulation, the relative root mean square errors (RMSE) of the quantiles of real data and generated data were calculated using two methods (Index flood and At-site). In addition, the mean of relative RMSE of each homogeneous region were calculated. The mean of relative RMSE of each region in this study were summarized in Table 7 and Figs. 3.

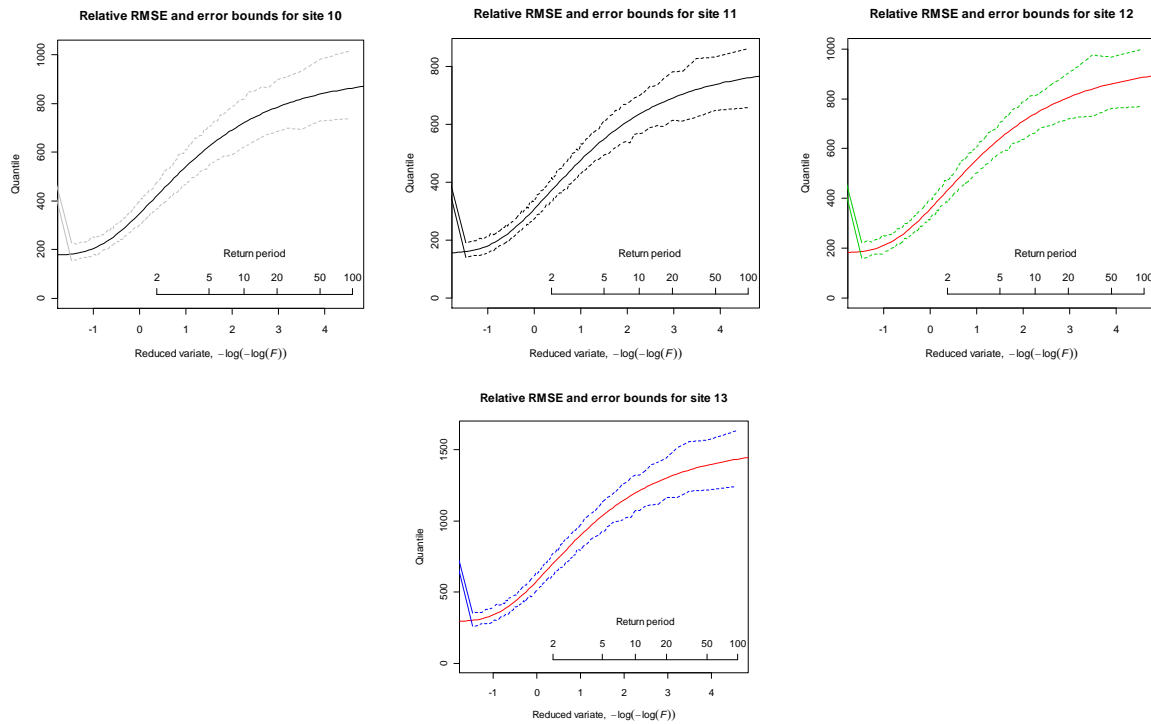
From Table 7 and Figs. 3, it is found that regional frequency analysis is more powerful than at-site frequency analysis for all regions and all nonexceedance probability. And Index Flood method is superior to Regional at-site estimation method for 9 sites in return period of 100 year, but Regional at-site estimation method is also superior to Index Flood method for only one site. The excellence of Index Flood method may be caused by all homogeneous regions in this study, and by sites with short record length (about 10 ~30) in 2 regions. However, the result from Index Flood method approaches that of At-site method as nonexceedance probability increase in all cases. These results agree Hosking's previous research. Also, in case of regional frequency analysis, the sites with small record length don't always show small relative root mean square error. The relative RMSE of 05KSL730 and 09GEN770 have 28 and 30 years are bigger to the relative RMSE of sites

08OBT771 and 09FSH760 with recrd length 90 & 84 respectively. It means that sites containing high record length always make an effect on accuracy of quantile estimation.

Table (7) The simulation result of each method for the selected sites (10 sites)

Site	Methods	Nonexcedence Probability and Relative RMSE					
		0.01	0.1	0.5	0.9	0.99	0.999
02ATB680	At-site	0.0795	0.0637	0.0557	0.0556	0.0742	0.0795
	Index flood	0.0925	0.0676	0.0739	0.0584	0.0926	0.0901
05GDF752	At-site	0.0946	0.0722	0.0618	0.0664	0.0785	0.0894
	Index flood	0.1052	0.0805	0.0685	0.0631	0.0875	0.1030
05KSL730	At-site	0.0977	0.0969	0.0768	0.0893	0.0993	0.1031
	Index flood	0.1051	0.1083	0.0773	0.1008	0.0943	0.1272
06WMD751	At-site	0.0864	0.0728	0.0497	0.0571	0.0776	0.0795
	Index flood	0.0873	0.0734	0.0660	0.0678	0.0868	0.0893
08OBT771	At-site	0.0733	0.0581	0.0404	0.0478	0.0680	0.0811
	Index flood	0.0816	0.0598	0.0479	0.0499	0.0687	0.0854
09FSH760	At-site	0.0858	0.0575	0.0505	0.0464	0.0738	0.0766
	Index flood	0.0784	0.0571	0.0513	0.0509	0.0764	0.0895
09GEN770	At-site	0.1073	0.0956	0.073	0.0809	0.0899	0.0934
	Index flood	0.1058	0.0907	0.0822	0.0773	0.1065	0.1109
10NYL790	At-site	0.0903	0.0784	0.0625	0.0690	0.0795	0.0933
	Index flood	0.1010	0.0771	0.0770	0.0775	0.0915	0.1084
12SNR762	At-site	0.0928	0.0693	0.0541	0.0558	0.0809	0.0827
	Index flood	0.0815	0.0678	0.0618	0.0578	0.0811	0.0898
13MLK840	At-site	0.0946	0.0643	0.0560	0.0507	0.0767	0.0817
	Index flood	0.0863	0.0672	0.0704	0.0608	0.0861	0.0852





Figures 3. The relative RMSE and error bound for 10 selected stations. The solid line shows theoretical distribution and the dashed lines show 5%&95% uncertainty bounds.

IV. CONCLUSION

In an aim to extend the methodologies to the case of ungauged site, a nonlinear regression model is derived and the results are investigated. Diagnostics of model results, it compares the estimated quantile for 100 years and corresponding drainage area (A km²), longitude (X_{dist}) and latitude (Y_{dist}) at different selected sites (10 sites) is carried out, which respectively results in the following derived equations:

$$M_0 = 2001.71A^{-0.52}X_{dist}^{29.76}Y_{dist}^{-158.16} \quad (7)$$

$$M_0 = 1408.51A^{-0.46}X_{dist}^{7.33}Y_{dist}^{-87.70} \quad (8)$$

With coefficient of determination R^2 of 0.6003 and 0.7969, the root mean squared error (RMSE) of 241.1631 and 82.3325, the mean absolute error (MAE) of 177.8233 and 60.8757, the percentual root mean squared error (RMSEP) of 68.5% and 25.8%, and the percentual mean absolute error (MAEP) of 41.9% and 19.6% for the defined sites.

REFERENCES

- [1]. Dalrymple, T. (1960). Flood frequency methods, U.S. Geological Survey. Water supply paper 1543A, 11-51.
- [2]. Hosking, J. R. M., Wallis, J. R., & Wood, E. F. (1985). Estimation of the generalized extreme value distribution by the method of probability weighted moments. *Technometrics*, 27(3), 251-261.
- [3]. Hosking, J. R. M. and Wallis, J. R. (1993). "Some statistics useful in regional frequency analysis", *Water Resources Research*, Vol. 29, No. 2, pp. 271-81, (Correction: *Water Resources Research*, 31(1995), 251.).
- [4]. Hosking, J.R.M. and Wallis, J.R. (1997). *Regional Frequency Analysis: An Approach based on L-Moments*. Cambridge University Press, Cambridge, UK.
- [5]. Jin, M. & Stedinger, J. R. (1989). Flood frequency analysis with regional and historical information. *Water Resources Research*, 25(5), 925-936.
- [6]. Katz, R. W., M. B. Parlange, and P. Naveau. 2002. Statistics of extremes in hydrology. *Advances in Water Resources* 25:1287-1304.
- [7]. Lee, D. J. and Heo, J. H. (2001). "Frequency Analysis of Daily Rainfall in Han River Basin Based on Regional L-moments Algorithm", *Korea Water Resources Association*. Vol. 34, No. 2, pp. 119-130 (in Korean).
- [8]. Lettenmaier, D. P. & Potter, K.W. (1985). Testing flood frequency estimation methods using a regional flood generation model. *Water Resources Research*, 21(12), 1903-1914.
- [9]. Stedinger, J. R. and Lu, L. H. (1995). "Appraisal of regional and index flood quantile estimators", *Stochastic Hydrology and Hydraulics*, Vol. 9, pp. 49-75.
- [10]. Wallis, J. R. and Wood, E. F. (1985). "Relative Accuracy of Log-Pearson-III procedures", *Journal of Hydraulic Engineering*, Vol. 111, No. 7, pp. 1043-1056.