 

# Regional Flood Frequency Analysis of the Red River Basin Using L-moments Approach

Y.H. Lim[1]

[1]Civil Engineering Department, School of Engineering & Mines, University of North Dakota, 243 Centennial Drive Stop 8115, Grand Forks, ND 58202-8115; PH (701) 775-3998; FAX (701) 777-3782; email: howelim@mail.und.nodak.edu

## Abstract

A basin-wide flood frequency analysis using Index flood and L-moments approach offers an attractive solution to provide flood quantile estimations at hundreds of ungauged sites within the Red River Basin of the North. L-moments diagrams and hierarchical clustering techniques were used initially to delineate hydrologic regions within the basin. Although the northern plain is relatively flat and almost monotonous, the analyses have shown that four homogeneous regions could be established for the basin. The whole basin as a region is also established for the purpose of flood estimations along the main stem. Appropriate probability distributions are fitted to the pooled regional flood peaks in each region. Monte Carlo simulations are performed to decide the best distribution for each region, and the dominant distributions found are the Log-Pearson Type III and the generalized Pareto distributions. The regression of index flood with the physical factors of drainage area and basin slopes for two of the five regions are not very satisfactory. However, the flood quantile estimates are sufficient for preliminary planning purposes.

## Basis for Regional Flood Frequency Analysis and L-moments

The Red River of the North Basin has 91 gauged sites but that number was not sufficient for economic planning purposes as flood quantile estimations at hundreds of ungauged sites within the basin were required. Regional flood frequency analysis is a good alternative approach to flood quantile estimation using a single station approach. It is well accepted that using a regional approach in flood frequency analysis is effective in extending the flood information at a site to sites within a homogeneous region. Recent papers by Burn and Goel (2000), Cunderlik and Burn (2002), Pfister et al. (2002), and Lim and Lye (2003) are examples of such undertakings. The extension enables flood quantile estimates for any site in a region to be expressed or inferred in terms of flood data recorded at all gauging sites in that region. Design flood estimations using a regional approach can often be carried out using methods such as the index-flood method and the direct regression on quantiles method.

*L-Moment Ratios*

A good summary on L-moments and frequency analysis of extreme events can be found in Stedinger *et al*. (1992). The plot of L-moment ratios (L-CV with L-skewness) for all the sites is shown in Figure1 and the plot of L-kurtosis versus L-skewness is shown in Figure 2. It can be seen that there are a few possible clusters based on casual inspection. On the whole, the plots show somewhat consistent correlation of L-CV with L-skewness for the sample data derived from the subbasins within the Red River Basin. The same observation can be made about the plot of L-kurtosis versus L-skewness.

*Identification of Regions – Single Region*

The first natural approach would be to assume the whole basin as one hydrologic region. A set of checking procedures is then applied to validate the assumptions.

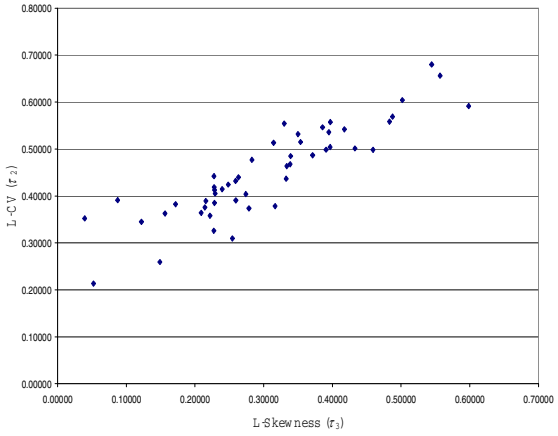


**Figure 1. L-moments ratios for all the available sites in and around the Red River Basin**

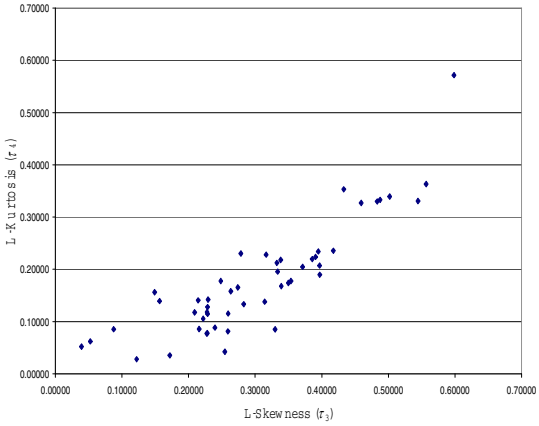


**Figure 2. L-kurtosis versus L-skewness for all available sites in and around the Red River Basin**

*Measure of Discordance*

A discordancy measure $D_i$ for a region with $N$ basins is computed based on the procedures lain down by Hosking and Wallis (1993, 1997). It is defined as:

$$D_i = \frac{1}{3} N(\mathbf{u_i} - \bar{\mathbf{u}})^T \mathbf{A}^{-1}(\mathbf{u_i} - \bar{\mathbf{u}})$$

where $u_i$ is a vector containing the L-moment ratios for basin $i$, namely the L-CV($t$), L-skewness($t_2$), and L-kurtosis($t_3$), $\bar{\mathbf{u}}$ is the unweighted regional average for $u_i$ and A is the matrix of sums of squares and cross products defined as:

$$\mathbf{A} = \sum_{i=1}^{N} (\mathbf{u_i} - \bar{\mathbf{u}})(\mathbf{u_i} - \bar{\mathbf{u}})^T$$

Any basin with the value of D greater than 2.49 would be considered as grossly discordant and would warrant removal from the defined region or cluster. For the whole basin as one region case, a few of the sites are found with D greater than 2.49.

*Heterogeneity Measure*

The L-moment ratios $t^R$, $t^R_3$ and $t^R_4$ of the proposed region are calculated as the sample means weighted proportionally to the record length $l$ of $i$ sites. The weighted standard deviation of the at-site sample L-CVs ($t_i$), $V$ is given by:

$$V = \left[ \sum_{i=1}^{N} l_i (t_i - t^R)^2 \Big/ \sum_{i=1}^{N} l_i \right]^{1/2}.$$

A homogeneity statistic, $H$, is a measure of the departure of $V$ from similar statistic obtained from simulation of some large number of realizations of a region, with $\mu_v$ and $\sigma_v$ as the mean and standard deviation of simulated $V$s:

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

Regions selected are reasonably homogeneous if H < 4 according to Hosking and Wallis (1997). Robson and Reed (1999) relaxed it to 4 as the limit instead of 2. For all the stations considered as in a homogeneous region, a heterogeneity measure $V$ is calculated. An example for the whole basin is $V$=0.08845. A simulation using the Monte Carlo method is used to generate random data based on the observed data. In this case, pairs of $t_3$ versus $t_4$ data are used. The heteogeneity measure 12.4 was obtained for the whole region indicating that it is a heterogeneous region. Further division of the sites is required. Because of the different topography and the presence of lakes, it is apparent that the basin as a region is not logical.

## Identification of Regions

*Cluster Observations*

A clustering technique is employed as a preliminary step to further divide these sites. The technique considers the factors of basin slope, contributing area, latitude, longitude, and specific peak discharge (by area); five regions are identified. The Euclidean distance is used to determine how the distance between two clusters is defined, and the average weighted linkage is the algorithm used to create the hierarchical cluster tree. The physical factors of the gauged stations form a matrix on which the hierarchical clustering of observations is performed.

*Measure of Discordance*

Tables 1 through 5 shows the major statistics and discordancy measure of each site. None of the basin is found to exhibit $D_i$ greater than the critical value except for the case of Region D.

**Table 1. Major Statistics and Discordancy ($D_i$) of Sites in Region A**

| ID | River | Name | N | Median | Mean $l_1$ | $t$ | $t_3$ | $t_4$ | $D_i$ |
|---|---|---|---|---|---|---|---|---|---|
| 8 | Red LakeRiver | Crookston | 102 | 7660.0 | 9421.2 | 0.3644 | 0.2091 | 0.1179 | 0.60 |
| 9 | Red Lake River | Red Lake | 67 | 955.0 | 936.6 | 0.3910 | 0.0875 | 0.0854 | 0.85 |
| 10 | Red Lake River | High Land. | 74 | 1590.0 | 1643.2 | 0.3522 | 0.0393 | 0.0524 | 1.30 |
| 11 | Thief River | Fall | 93 | 1630.0 | 1767.9 | 0.3627 | 0.1563 | 0.1395 | 0.73 |
| 12 | Clearwater | Plummer | 66 | 1415.0 | 1658.2 | 0.3261 | 0.2274 | 0.0773 | 0.38 |
| 13 | Lost River | Oklee | 45 | 1270.0 | 1407.1 | 0.3452 | 0.1220 | 0.0282 | 0.81 |
| 14 | Clearwater | Red L. Fall | 79 | 3200.0 | 3958.6 | 0.3580 | 0.2219 | 0.1057 | 0.49 |
| 17 | Pelican River | Fergus Falls | 40 | 262.5 | 347.8 | 0.3096 | 0.2545 | 0.0426 | 1.38 |
| 50 | Roseau River | Caribou | 86 | 1655.0 | 1769.2 | 0.2593 | 0.1490 | 0.1565 | 2.47 |

**Table 2. Major Statistics and Discordancy ($D_i$) of Sites in Region B**

| ID | River | Name | N | Median | Mean $l_1$ | $t$ | $t_3$ | $t_4$ | $D_i$ |
|---|---|---|---|---|---|---|---|---|---|
| 28 | Rush River | Amenia | 58 | 432.5 | 630.2 | 0.5360 | 0.3949 | 0.2342 | 0.48 |
| 30 | Maple River | Enderlin | 49 | 840.0 | 1555.6 | 0.5544 | 0.3300 | 0.0852 | 2.14 |
| 31 | Goose River | Portland | 44 | 665.0 | 1188.7 | 0.5582 | 0.4832 | 0.3298 | 1.05 |
| 32 | Goose River | Hillsboro | 78 | 1405.0 | 2468.7 | 0.5320 | 0.3500 | 0.1740 | 0.68 |
| 33 | Marsh River | Shelly | 61 | 1160.0 | 1434.7 | 0.4399 | 0.2634 | 0.1580 | 1.89 |
| 34 | Wild Rice River | Twin Valley | 83 | 1550.0 | 2259.7 | 0.4983 | 0.4592 | 0.3271 | 1.31 |
| 35 | Wild Rice River | Hendrum | 61 | 2860.0 | 3624.3 | 0.3906 | 0.2595 | 0.1156 | 1.15 |
| 36 | Buffalo River | Hawley | 61 | 799.0 | 911.0 | 0.3853 | 0.2285 | 0.1149 | 1.04 |
| 37 | Buffalo River | Sabin | 60 | 1250.0 | 1782.2 | 0.4869 | 0.3713 | 0.2049 | 0.20 |
| 38 | Buffalo River | Dilworth | 74 | 1615.0 | 2326.0 | 0.4985 | 0.3908 | 0.2239 | 0.25 |
| 39 | Sand Hill River | Climax | 62 | 1285.0 | 1580.5 | 0.4320 | 0.2591 | 0.0814 | 0.80 |

**Table 3. Major Statistics and Discordancy ($D_i$) of Sites in Region C**

| ID | River | Name | N | Median | Mean $l_1$ | $t$ | $t_3$ | $t_4$ | $D_i$ |
|---|---|---|---|---|---|---|---|---|---|
| 40 | Forest River | Fordville | 65 | 1080.0 | 1880.3 | 0.6042 | 0.5019 | 0.3392 | 0.73 |
| 41 | Forest River | Minto | 65 | 1100.0 | 1914.2 | 0.5689 | 0.4877 | 0.3329 | 0.83 |
| 42 | Middle River | Argyle | 65 | 939.0 | 1355.3 | 0.4676 | 0.3385 | 0.2180 | 0.29 |
| 43 | Park River | Grafton | 75 | 1420.0 | 2085.9 | 0.5465 | 0.3858 | 0.2198 | 0.94 |
| 44 | Two River | Lake Bronson | 66 | 1490.0 | 1730.1 | 0.3894 | 0.2157 | 0.0860 | 1.40 |
| 45 | Pembina River | Walhalla | 58 | 2325.0 | 3878.2 | 0.5419 | 0.4175 | 0.2358 | 0.77 |
| 46 | Tongue River | Akra | 62 | 347.5 | 559.3 | 0.5914 | 0.5985 | 0.5716 | 2.31 |
| 47 | Sprague Creek | Sprague | 58 | 603.5 | 855.1 | 0.5014 | 0.4330 | 0.3531 | 0.35 |
| 48 | Roseau River | Malung | 75 | 1700.0 | 2370.5 | 0.4635 | 0.3337 | 0.1958 | 0.15 |
| 49 | Roseau River | Ross | 75 | 1600.0 | 2091.7 | 0.3784 | 0.3166 | 0.2282 | 2.22 |

**Table 4. Major Statistics and Discordancy ($D_i$) of Sites in Region D**

| ID | River | Name | N | Median | Mean $l_1$ | $t$ | $t_3$ | $t_4$ | $D_i$ |
|---|---|---|---|---|---|---|---|---|---|
| 23 | Sheyenne River | Valley City | 64 | 1400.0 | 1678.2 | 0.4053 | 0.2293 | 0.1423 | 1.00 |
| 24 | Sheyenne River | Lisbon | 48 | 1715.0 | 1992.4 | 0.4189 | 0.2280 | 0.0780 | 1.00 |
| 25 | Sheyenne River | Kindred | 56 | 1469.0 | 1859.4 | 0.4145 | 0.2396 | 0.0887 | 1.00 |
| 26 | Sheyenne River | West Fargo | 79 | 1610.0 | 1581.5 | 0.3826 | 0.1717 | 0.0356 | 1.00 |

**Table 5. Major Statistics and Discordancy ($D_i$) of Sites in Region E (main stem)**

| ID | River | Name | n | Median | Mean $l_1$ | $t$ | $t_3$ | $t_4$ | $D_i$ |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Red River | Wahpeton | 64 | 3220.0 | 3752.5 | 0.3756 | 0.2145 | 0.1409 | 0.96 |
| 3 | Red River | Fargo | 105 | 3870.0 | 5644.0 | 0.5042 | 0.3968 | 0.2074 | 1.27 |
| 4 | Red River | Halstad | 65 | 12900.0 | 15272.5 | 0.4040 | 0.2740 | 0.1653 | 0.55 |
| 5 | Red River | Grand Forks | 123 | 17000.0 | 21941.7 | 0.4369 | 0.3328 | 0.2125 | 0.24 |
| 6 | Red River | Drayton | 66 | 27600.0 | 30165.5 | 0.3735 | 0.2785 | 0.2305 | 1.39 |
| 7 | Red River | Oslo | 45 | 24000.0 | 27266.2 | 0.4243 | 0.2484 | 0.1778 | 1.58 |

*Homogeneity Measure*

The results of the Monte Carlo (MC) simulation for the five regions are shown in Table 6 below. It can be seen that the regions selected are reasonably homogeneous as H < 4. Two are less than 2.0, which are considered as homogeneous according to Hosking and Wallis (1997). Robson and Reed (1999) relaxed it to 4 as the limit instead of 2. A plot of L-moment ratios for various regions is shown in Figure 3. The plot of moment ratios for the theoretical distributions is also superimposed on the same plot. The option for choosing the appropriate distribution by eye can be done based on the plotted location of the weighted values. However, a statistical approach is employed to confirm the appropriateness of the distribution chosen and to give a certain degree of confidence in the selected distribution. A test based on Monte Carlo simulation by Hosking and Wallis (1993, 1997) is used.

**Table 6. Kappa Distribution Fitted for Regions A–E Using MC Simulation**

| Region | A | B | C | D | E |
|---|---|---|---|---|---|
| $V$ | 0.03771 | 0.05722 | 0.07640 | 0.04585 | 0.04687 |
| $x_i$ | 0.4991 | 0.169 | 0.4461 | 0.0641 | 0.4647 |
| $a$ | 0.7584 | 0.8157 | 0.5136 | 1.1061 | 0.603 |
| $k$ | 0.2114 | –0.0777 | –0.3045 | 0.172 | –0.1235 |
| $h$ | 0.5463 | 0.8257 | 0.168 | 0.9786 | 0.3759 |
| $H$ | 2.45 | 2.69 | 2.41 | 1.96 | 1.89 |

## Selection of Appropriate Distributions

*Goodness-of-Fit Test*

For each of the proposed region, a Kappa distribution with its parameters derived from the fitting of the distribution to the regional average L-moment ratios is used to simulate some large numbers ($N_{sim}$) of the same region. For each of the $m^{th}$-simulated region, the regional average L-kurtosis $t_4^m$ is calculated. Typical three-parameter distributions are fitted to the sample regional L-moment ratios. For each of the fitted distributions, the corresponding L-kurtosis, $\tau_4^{DIST}$, is found. The goodness-of-fit measure for each distribution is given by:

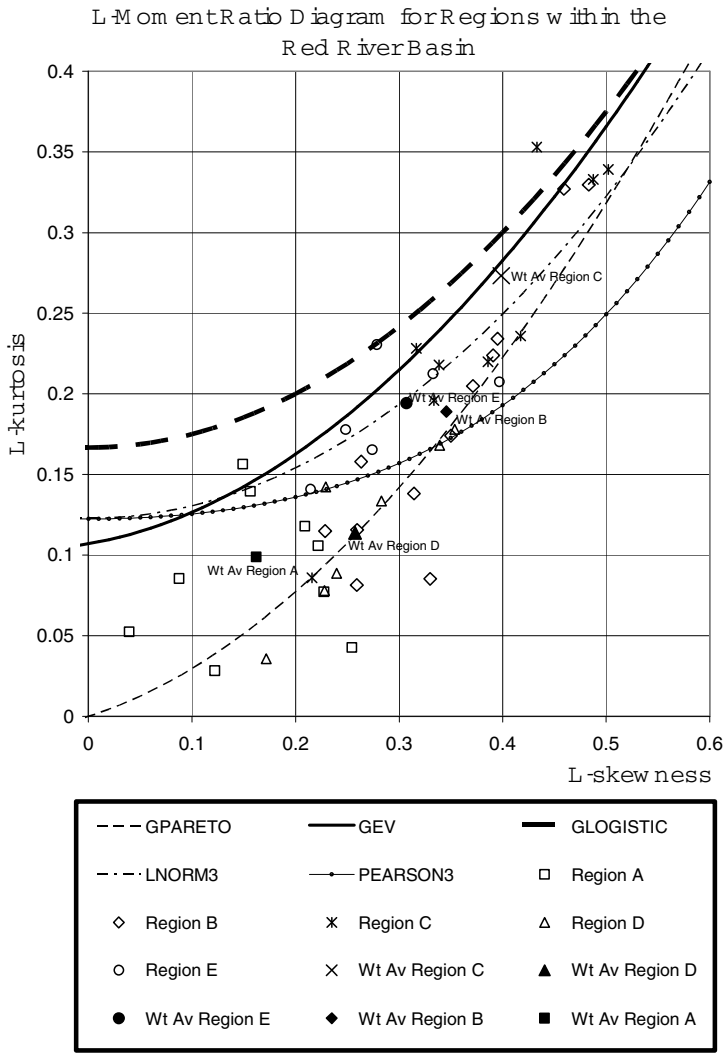


**Figure 3. L-moment Ratio Diagram with Theoretical Distribution Plots**

$$Z^{DIST} = \left(\tau_4^{DIST} - t_4^R + B_4\right) / \sigma_4$$

where the bias of $t_4^R$ is:

$$B_4 = \frac{\sum_{m=1}^{N_{Sim}} \left(t_4^m - t_4^R\right)}{N_{Sim}}$$

and the standard deviation of $t_4^R$ is given by:

$$\sigma_4 = \left[ \frac{\sum_{m=1}^{N_{Sim}} \left( t_4^m - t_4^R \right)^2 - N_{Sim} B_4^2}{N_{Sim} - 1} \right]^{1/2}$$

Any of the distributions could be declared as fitting satisfactorily if $\left| Z^{DIST} \right| \leq 1.64$ (Hosking and Wallis, 1993). From simulation tests for the Red River Basin, the regions are found to have varying distributions that fit the data well. The distributions are as follows:

GEV – Generalized Extreme Value Distribution
GPA – Generalized Pareto Distribution
GLO – Generalized Logistics Distribution
GNO – Generalized Normal Distribution
PE3 – Pearson Type III Distribution

The following distributions are found suitable for each region:

| Region | Distribution |
|---|---|
| A | PE III |
| B | GPA, PE III, LNO |
| C | GEV, GLO, LNO |
| D | GPA * |
| E | PE III, LNO, GPA |

* Based on regulated flow conditions.

**Index Flood Method**

Recent developments in statistical methods (Hosking, 1990; Hosking and Wallis, 1997; Robson and Reed, 1999) has further consolidated the usage of the index flood method. Recent published papers on the use of index flood are Fill and Stedinger (1998), Burn and Goel (2000), and Brath et al. (2001).

The index flood method is used to determine the magnitude and frequency of flood quantiles for basins of any size, gauged or ungauged, as long as it is located within a hydrologically homogeneous region. At least two regressions are required. The first is the regression of mean annual flood with some physical parameters, e.g., basin area, basin slopes, and river length, and the second relates peak flow ratio with frequency of exceedence. The peak flow ratio is the ratio of peak flow for a given frequency of exceedence to the at-site mean annual flood. A flood frequency curve can then be developed for any basin in the homogeneous region.

Typically the index flood, $Q_m$, is taken as the mean of the at-site annual maximum peak discharge series. Robson and Reed (1999) recommended using the median instead of the mean. A relationship can be established between the flood quantile, $Q_T$, of a site and $Q_m$ with the introduction of a regional growth factor, $X_T$,

that defines the dimensionless frequency distribution common to all sites within a homogeneous region. The relationship is:

$$Q_T = X_T Q_m$$

A regression of basin characteristic(s) on the index flood can be established based on available information gathered from the gauged sites. Regional growth curves showing the relationship between $X_T$ and the return period, $T$, can be derived once an appropriate probabilistic distribution has been found within a region with N sites that fits all the gauged flood series, $Q_{ij}$, where $i$ = 1, 2, 3, …, $N$, j = 1, 2, 3, …, $L_i$, and $L_i$ is the record length at site $i$. The standardized flood peak:

$$X_{ij} = Q_{ij} / Q_{im}$$

is used in the estimation of $X_T$, where $Q_{im}$ is the observed mean or median annual flood at site $i$.

## Regional Flood Frequency Curves

### *Slopes and Areas*

The slope of basins is the prime basin characteristic beside the contributing drainage area that influences the peak flood quantiles. The slopes are generally very mild but there are some differences within the basin which, in turn, provide some distinctions among the Regression Equations. Regression fits are obtained for the median peak floods, $Q_{median}$ (cfs), of each site with contributing area A (square miles) and the slope S (in decimals). The regression fits for region D and E are not very satisfactory judging from the p-values obtained.

The regression equations for the regions are as follows:

| | |
|---|---|
| Region A | $Q_{median} = 1.26\ A + 16673\ S$ |
| Region B | $Q_{median} = 1.64\ A + 5422\ S$ |
| Region C | $Q_{median} = 0.638\ A + 11462\ S$ |
| Region D | $Q_{median} = 0.031\ A + 54009\ S$ |
| Region E | $Q_{median} = 0.768\ A + 302024\ S$ |

### *Regional Growth Curves*

Based on the simulation results and the index flood procedure, five regional growth curves are derived. Figure 4 and 5 show two of the five curves. The curves are fitting reasonably well to their respective set of pooled flood peaks series in each region. The respective curves are fitted statistically based on the distribution chosen by the goodness-of-fit test.

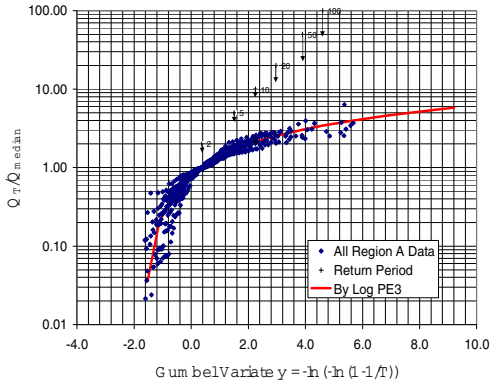


Figure 4. Region A growth curve

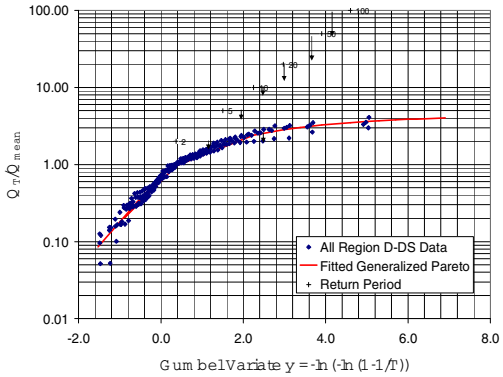


Figure 5. Region D growth curve

The equations for all the five curves are:

Region E (LPE3) $Q_T/Q_{median} = 10^{(1.153 + kp * 0.731)}$
where $kp=(2/1.108)[(1+(1.108*z/6)-(1.108^2)/36)^3]-(2/1.108)$
and $z = 5.0633[(1/T)^{0.135}-(1-1/T)^{0.135}]$

Region B (GPA) $Q_T/Q_{median}= 0.57220 + 1.36548 [1-(1-1/T)^{0.54397}]$

Region C (GEV) $Q_T/Q_{median} =0.7132 + 1.011 [1-(-\log(1/T))^{-0.3297}]$

Region D (GPA) $Q_T/Q_{median} =0.57220+1.36548 [1-(1-1/T)^{0.54397}]$

Region E (LPE3) $Q_T/Q_{median} = 10^{(1.2334 + kp * 0.9228)}$
where $kp=(2/1.772)[(1+(1.772*z/6)-(1.772^2)/36)^3]-(2/1.772)$
and $z = 5.0633[(1/T)^{0.135}-(1-1/T)^{0.135}]$

## CONCLUSION

Although the northern plain is relatively flat and almost monotonous, four homogeneous regions are found within the Red River Basin. The whole basin can also be treated as one region and flood quantiles can be estimated along the main stem. Appropriate probability distributions are fitted to the pooled regional flood peaks in each region. Monte Carlo simulations are performed to decide the best distribution for each region, and the dominant distributions found are the Log-Pearson Type III and the generalized Pareto distributions. Reasonably good regression equations of the median peak discharge with the basin area and slope are derived for three region while there is limited success for that of region D and E. Further research to include other physical factors many be required. However, the regional growth curves are fitting fairly well. With due consideration for simplicity of getting data for the ungauged sites, the method may suit the economic analysis for a basin-scale flood mitigation project.

## ACKNOWLEDGEMENT

The research was funded by Waffle Project, Environmental Energy & Environmental Research Center, University of North Dakota, Grand Forks.

## REFERENCES AND BIBLIOGRAPHY

Brath, A., Castellarin, A., Franchini, M., and Galeati, G., 2001, Estimating the index flood using indirect methods: Hydrological Sciences Journal, v. 46, no. 3, p. 399–418.

Burn, D.H., and Goel, N.K., 2000, The formation of groups for regional flood frequency analysis: Hydrological Sciences Journal, v. 45, no. 1, p. 97–112.

Cunderlik, J.M., and Burn, D.H., 2002, The use of flood regime information in regional flood frequency analysis: Hydrological Sciences Journal, v. 47, no. 1, p. 77–92.

Fill, H.D., and Stedinger, J.R., 1998, Using regional regression within index flood procedure and an empirical Bayesian estimator: Journal of Hydrology, v. 210, p. 128–145.

Hosking, J.R.M., and Wallis, J.R., 1993, Some statistics useful in regional frequency analysis: Water Resource Research, v. 29, no. 2, p. 271–281.

Hosking, J.R.M., 1990, L-moments—analysis and estimation of distributions using linear combinations of order statistics: J. Roy. Statist. Soc. Sec. B, v. 52, no. 1, p. 105–124.

Hosking, J.R.M., and Wallis, J.R., 1997, Regional frequency analysis—an approach based on L-moments: Cambridge, United Kingdom, Cambridge University Press.

Lim Y.H., and Lye L.M., 2003, Regional flood estimation for ungauged basins in Sarawak, Malaysia: Hydrological Sciences Journal, v. 48, no. 1, p, 79–94.

Pfister L., Iffly, J.-F., and Hoffmann, L., 2002, Use of regionalized stormflow coefficients with a view to hydroclimatological hazard mapping. Hydrological Sciency Journal, v. 47, no. 3, p. 479–491.

Robson, A.J., and Reed, D.W., 1999, Flood estimation handbook, v. 3—statistical procedures for flood frequency estimation: Wallingford, United Kingdom, Institute of Hydrology.

Stedinger, J.R., Vogel, R.M., and Foufoula-Georgiou, E., 1992, Frequency analysis of extreme events, in Maidment, D.R. (ed.), Handbook of Hydrology, chap. 18: New York, McGraw-Hill.