IOSR Journal Of Environmental Science, Toxicology And Food Technology (IOSR-JESTFT)
e-ISSN: 2319-2402,p- ISSN: 2319-2399. Volume 6, Issue 2 (Sep. - Oct. 2013), PP 54-60
www.Iosrjournals.Org

Application of a Mixed Gumbel Distribution to Construct Rainfall Depth-Duration-Frequency (DDF) Curves Considering Outlier Effect in Hydrologic Data

Muhammad Rizwan[1], Tae-Woong Kim, Phd[2].
[1]*MS Water Resources Engineer, Hanyang University, Seoul, Korea*
[2] *Corresponding Author, Associate Professor, Department of Civil and Environmental Engineering, Hanyang University, Ansan, Korea*

Abstract: *Values which are unusually large or small with respect to the rest of the values in dataset are called outliers. The outliers affect on estimation of parameters of distribution, evaluation of model performance, etc. The presence of outliers in rainfall data affects rainfall depth-duration-frequency (DDF) curves which provide design rainfalls for hydrologic structures. The DDF curves constructed using rainfall data containing outliers could lead to inefficient designing of hydrologic structures. This study proposed the use of a mixed Gumbel distribution for constructing rainfall DDF curves considering the effect of outliers. The results indicate that the mixed Gumbel model provides more reasonable estimates of rainfall DDF curves by reducing the considerable effect of outliers.*
Key Word : *Outliers; Rainfall depth-duration-frequency curves; mixed Gumbel distribution*

I. Introduction

One of the first steps toward obtaining a hydrologic coherent analysis is the detection and treatment of outliers in observations (Williams et al. 2002; Liu et al. 2004). The values which are unusually large or small with respect to the rest of the values in hydrologic data are called outliers (Anscombe 1960). Outliers may result from different mechanism of physical phenomenon or mistakes in measure process. Outliers may carry important information on abnormal condition of hydrologic phenomenon. However, they have unexpected influence on the distortion of parameter estimates, the inflation of error measurements, and the distortion of statistical inference, which may lead to biased conclusions and inaccurate predictions.

The study on detection and treatment of outliers has received considerable attention in the statistical literature. For example, Barnett and Lewis (1994) give nearly 1000 references. However, relatively little work has been done on outliers in hydrologic time series analysis. Outliers are quite likely to arise in hydrologic time series and may have severe effects on hydrologic modeling. Especially, the outliers in observations of hydrologic variables such as rainfall and streamflow affect the decision of design levels, since the design level is sometimes determined by the statistical frequency analysis.

Statistical frequency analysis of rainfall and flood is essential for designing hydrologic structures and establishing effective plans for water resources management. In Korea, rainfall frequency analysis is a necessary pre-requisite for designing an economic and efficient storm drainage system through providing rainfall depth-duration-frequency (DDF) or rainfall intensity-duration-frequency (IDF) curves. The rainfall DDF or IDF curves are usually used for determine design rainfalls. A design rainfall is the rainfall amount for a given duration and return period required for design of hydrologic structures. The rainfall DDF and IDF curves described annual maximum rainfalls expressed by a function of duration for given return periods or probabilities of exceedance (Overeem et al. 20008). So, through the rainfall DDF and IDF curves, the presence of outliers in the observations of annual maximum rainfall affects the decision making process for design of hydrologic structures, operation and management of water resources systems (Lee and Maeng 2003).

It is necessary to introduce an additional mathematical tool that be able to reduce the uncertainty in estimating design events due to outliers, which are needed in many water engineering studies and projects. Bulletin 17B, which is recommended for use by US Federal agencies, provides a recommended procedure for the adjustment of outliers including deleting outliers (McCuen 2004; Griffis and Stedinger 2007). However, the deletion of outliers could lead to the underestimation of the design hydrologic quantities, and the use of the detected outliers lead to over design which will not be economically feasible or inadequate for the design requirements. Such undesired situation must be taken into account in practice. Even though it has not yet been clearly proven that outliers result from the different mechanism of hydrologic process or the mistakes during observation process, hydrologists are asked to deal with outliers to construct rainfall DDF curves for hydrologic design and management.

The main objective of this study is to develop a practical method to construct rainfall DDF curves considering the effect of outliers after evaluating outlier effects on rainfall DDF curves. This paper is organized as follow. First, a mixed Gumbel distribution is described. A mixed Gumbel distribution is used in this study to estimate the probability density function incorporating the probabilities of outlier occurrences. Next, the effect of outliers on rainfall DDF curves is evaluated, and the use of a mixed Gumbel distribution is justified for the construction of rainfall DDF curves.

II. Mixed Gumbel Distribution

The objective of hydrologic frequency analysis is to estimate the extreme hydrological magnitude corresponding to any return period of occurrence based on the probability distribution of the hydrologic variable of interest. To estimate a design rainfall, a conventional univariate frequency analysis is usually applied for the annual maximum rainfall depths by being extracted for the selected storm duration from the historical rainfall records (Chow et al. 1998). The Gumbel distribution has been proposed to describe the occurrence probability of extreme rainfalls (Loaiciga and Leipnik 1999), for example, Gumbel-based rainfall DDF or IDF relationships have been used for the planning and design of stormwater drainage systems in Korea and Hong Kong (Heo et al. 2009; Xu and Tung 2009).

A mixed distribution is a combined distribution of two continuous and/or discrete distributions (Yoo et al. 2005). Owing to the intermittency of rainfall, Kedem et al. (1990) proposed to use a mixed distribution to represent the statistical behavior of rainfall. Shimizu (1993) also showed the possible application of a mixed distribution for the analysis of rainfall data, especially using the mixed log-normal distribution. Yoo et al. (2005) used a mixed Gamma distribution to evaluate the global warming effect on daily rainfall in Korea.

The conventional univariate probability distributions can hardly consider the occurrence probability of outliers. So this study employed a mixed Gumbel distribution to incorporate the occurrence probability of outliers with the probability density function of rainfall amount to overcome the limitation of univariate probability distributions. A mixed Gumbel model has the probability density function (PDF) and the cumulative distribution function (CDF) as follows

$$f(x) = \frac{p}{\alpha_1} exp^{-\left(\frac{x-\nu_1}{\alpha_1}\right)} exp^{-exp^{-\left(\frac{x-\nu_1}{\alpha_1}\right)}} + \frac{(1-p)}{\alpha_2} exp^{-\left(\frac{x-\nu_2}{\alpha_2}\right)} exp^{-exp^{-\left(\frac{x-\nu_2}{\alpha_2}\right)}} \quad (1)$$

$$\mathbf{F(x) = p\,exp^{-exp^{-\left(\frac{x-\nu_1}{\alpha_1}\right)}} + (1-p)exp^{-exp^{-\left(\frac{x-\nu_2}{\alpha_2}\right)}}} \quad (2)$$

where ν_1 and ν_2 are the location parameter of the first and second population respectively. α_1 and α_2 are the scale parameter of the first and second population respectively. The first population means a set of outliers, and the second population means annual maximum rainfall data set discarding outliers in this study. So, p is the occurrence probability of outliers.

In the rainfall frequency analysis using annual maximum data, the return period T is related to the non-exceedance probability which is the same as the CDF. The relationship between the return period and the non-exceedance probability becomes as Eq. (3) considering the occurrence probability of the outliers in the data sets.

$$\mathbf{F_{mixed} = \left(1 - \frac{1}{T}\right) - (1-p)} \quad (3)$$

The methodology proposed in this study (shown in Fig. 1) may be simple to be applied to the practice. The practical application of a mixed Gumbel distribution is addressed as follow.

III. Application And Discussions

Fig. 3 shows a graphical representation of 24-hr annual maximum rainfall observations at Gangneung gauging station which is one of sample stations located on the north eastern side of Korean peninsula as shown in Fig. 2. The highest value is 883.8 mm in 2002 resulted from the typhoon Rusa. Preliminary studies indicated that several rainfall observation stations in Korea are statistically significant for detecting outliers and the possibility of heavy rainfall occurrence larger than the design rainfall becomes higher (Kwon et al. 2008; Lee et al. 2009).

This study collected annual maximum rainfall observations from 1961 to 2006, and performed the Grubbs outlier test (Grubbs 1969) to check the existence of outliers in 24-hrs annual maximum rainfall observations for 56 rainfall observation stations (shown in Fig. 2) managed by the Korean Meteorological Administration Agency. Table 1 shows the statistics and results of Grubbs outliers test for 7 stations which has statistically significant high outliers in data set.

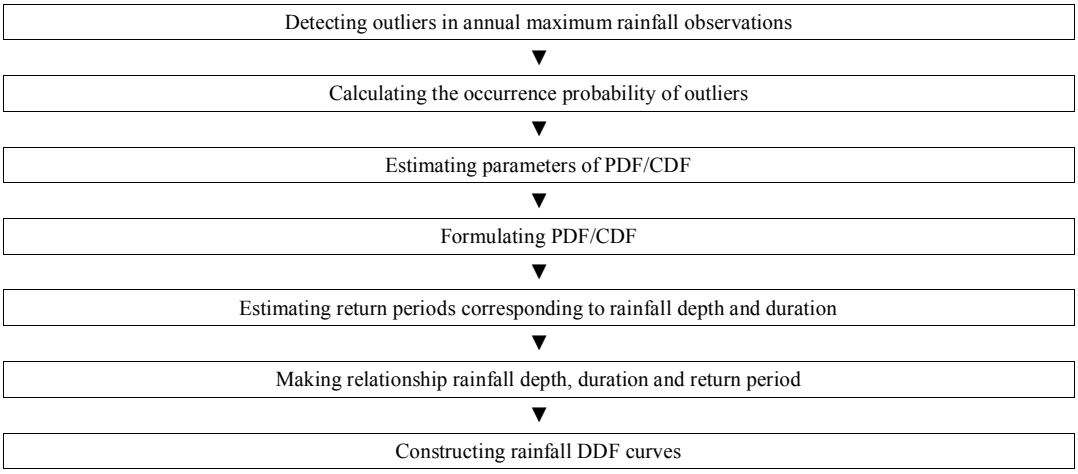

Fig. 1 Flowchart for using a mixed Gumbel distribution to construct rainfall DDF curves

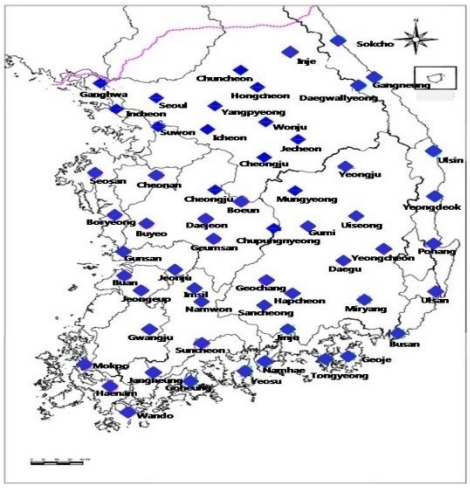

Fig. 2 Rainfall observation stations in Korea

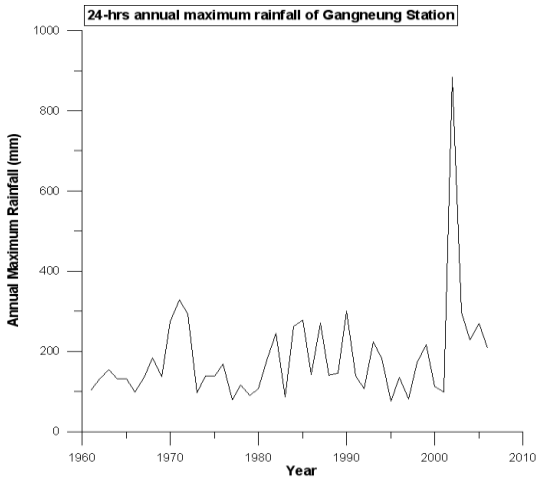

Fig. 3 Maximum annual rainfalls at Gangneung gauging station, Korea

Table 1 Statistics and Results of Grubbs test for sample stations

Gauging Station	Mean (mm)	Median (mm)	Standard Deviation (mm)	Coefficient of Skewness	Threshold of high outlier (mm)	Highest value (mm)
Gangneung	185.6	141.7	126.9	3.74	610.2	883.8
Ganghwa	201.1	169.5	105.9	2.23	549.0	622.2
Pohang	142.6	119.9	85.7	3.22	420.8	576.8
Buyeo	155.4	139.8	77.7	3.68	353.3	537.8
Gwangju	147.0	130.6	61.2	1.83	369.6	380.7
Goheung	186.3	163.2	101.1	2.34	522.8	596.2

Bulletin 17B of US and the guidelines on river design of Korea recommend that the detected outliers should be censored to improve the reliability of the flood (or rainfall) quantile estimates of interest (KWRA 2005; Griffis and Stedinger 2007). However, in order to eliminate outliers from observations, the apparent knowledge of its parent distribution is required (Cordery et al. 2007). In addition, in practice, to perform a reliable rainfall frequency analysis, it is desired to use rainfall observations as many as possible, since at most 50 years of hourly rainfall observed data are available in Korea.

Numerous studies have indicated that the characteristics of extreme storm events in Korea might follow a Gumbel distribution (Heo et al. 2006; Kwon et al. 2008). A Gumbel distribution is more tractable mathematically than other extreme probability models such as log-normal, Gamma, and GEV distribution (Loaiciga and Leipnik 1999). Therefore, in this study, a mixed Gumbel model was employed to incorporate the occurrence probability of outliers with the probability density function of rainfall amount. The Gumbel

distribution was found in this study to yield a mixed distribution function without excessive computational effort. The parameters of a mixed Gumbel distribution were estimated by the method of moments as shown in Table 2. Fig. 4 illustrates the PDF and the CDF at Gangneung station, in Korea, to give an instance.

Table 2 Parameters of mixed Gumbel distribution at sample stations

Gauging Station	Mixed Gumbel Distribution			
	ν_1	α_1	ν_2	α_2
Gangneung	128.54	98.91	137.84	55.92
Ganghwa	153.47	82.55	153.92	59.66
Pohang	104.00	66.83	107.72	43.66
Buyeo	120.51	60.52	126.37	30.29
Gwangju	119.42	47.73	118.99	39.48
Goheung	140.86	78.80	141.69	55.84

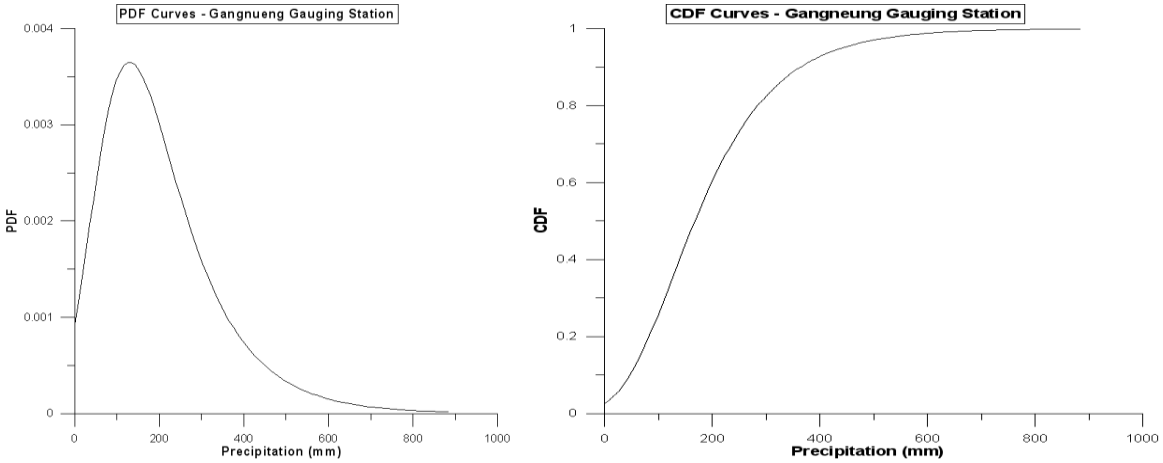

Fig. 4 Illustration of the PDF and the CDF at Gangneung station, in Korea

This study employed a Quantile-Quantile (Q-Q) plot to assess the validity of the assumption that the rainfall observations follow a mixed Gumbel distribution. A Q-Q plot compares two probability distributions, usually the sample distribution function and a theoretical distribution function. A Q-Q plot is a simple and quite general way to check whether a batch of data conforms, approximately to a particular probability model. Fig. 5 illustrates the comparison of theoretical and empirical probabilities for Gangneung rainfall gauging station, for example. The dotted lines indicate the upper and lower limits estimated from the critical values of the Kolmogorov-Smirnov test at significance level of 5% ($D_{n=46}^{\alpha=0.05} = 0.197 \cong 0.20$). For all selected rainfall gauging stations, the discrepancies are within the upper and lower limits. So, the estimated mixed Gumbel distribution was verified as an acceptable model at 5% significance level.

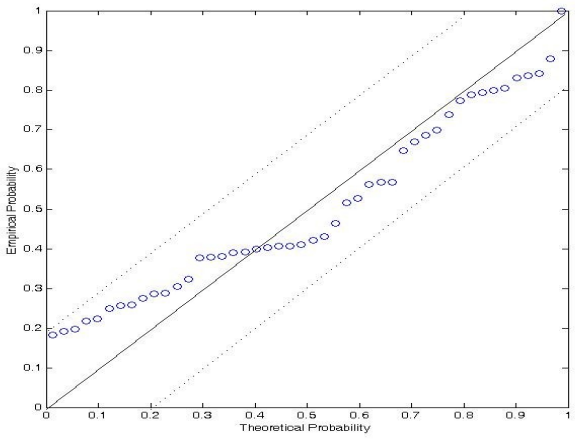

Fig. 5 Comparison of the theoretical and empirical probabilities at Gangneung station, Korea

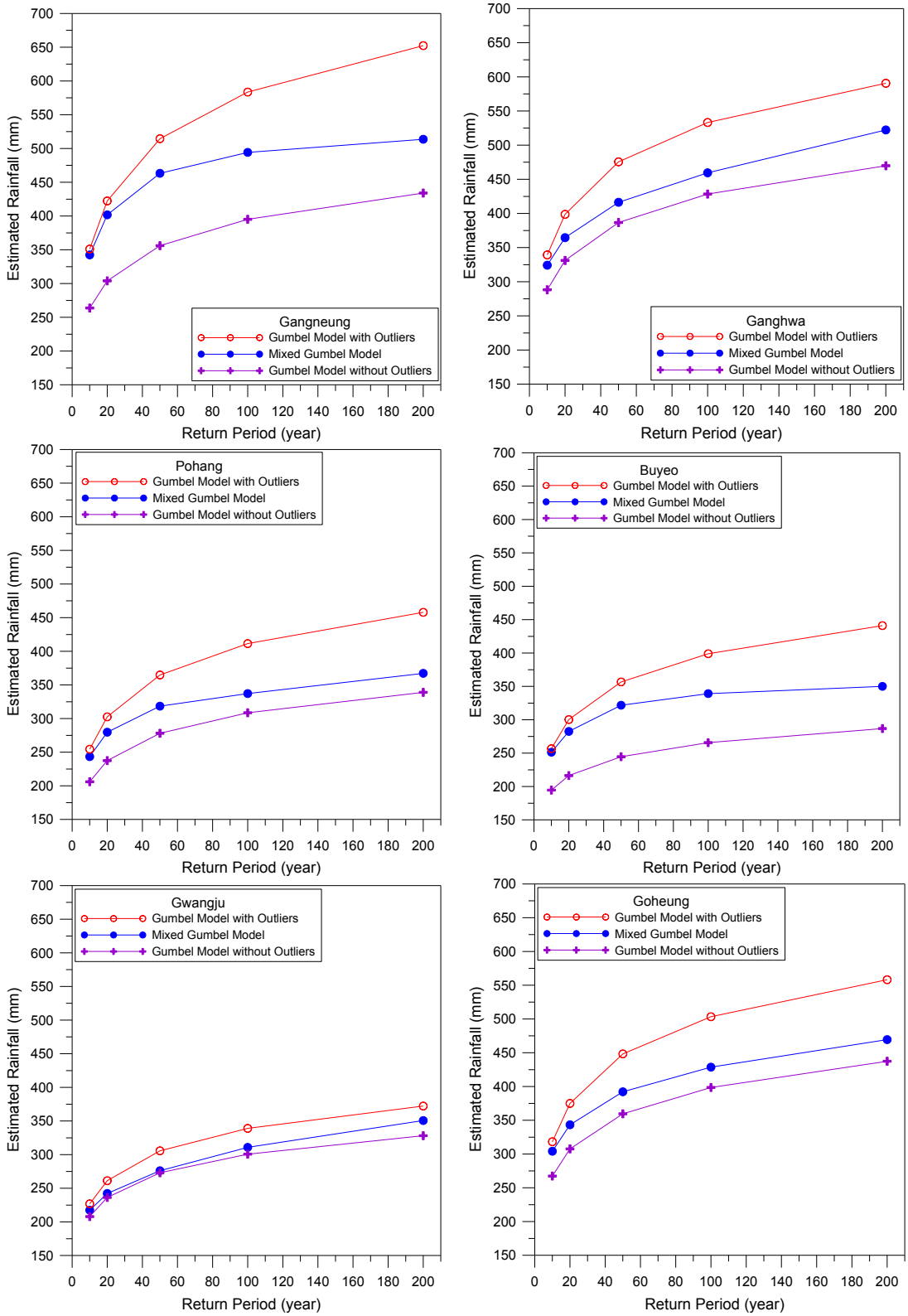

Fig. 6 Rainfall Depth-Duration-Frequency Curves for sample stations

Rizwan and Kim (2009) empirically investigated the effect of outliers on rainfall DDF curves by making various data sets, i.e., annual maximum rainfall data containing outliers, annual maximum rainfall data censoring outliers, and annual maximum rainfall data in which outliers are replaced with the maximum value among non-outlier values. Compensating the shortcomings and discussions on Rizwan and Kim (2009), this study employed the mixed Gumbel distribution to systematically evaluate the effect of hydrologic outliers on

rainfall DDF curves. Fig. 6 shows the comparison of rainfall DDF curves for sample rainfall gauging stations. The red line with open circle is the rainfall DDF curve calculated using a Gumbel distribution for data with outliers. The violet line with cross symbol is the rainfall DDF curve constructed using a Gumbel distribution for data after discarding outliers. The blue line with closed circle is the rainfall DDF curve constructed using a mixed Gumbel distribution. The Gumbel model with outliers gives very high values of design rainfall while the Gumbel model without outliers gives very low values. In both the cases, the rainfall DDF curves result in over estimated or underestimated hydrologic design quantities.

Table 3 shows the percentage reduction in 24-hr design rainfalls when the mixed Gumbel distribution was used. For example, the design rainfalls at Gangneung station were reduced 2.5 % for 10 years return period, 4.88 % for 20 years return period, 9.96 % for 50 years return period, 15.31 % for 100 years return period and 21.26 % for 200 years return period using the mixed Gumbel distribution. Fig. 6 and Table 3 confirm that the mixed Gumbel model provides more reasonable estimates of rainfall DDF curves by reducing the considerable effect of outliers.

Table 3 Percentage reduction in design rainfall using a mixed Gumbel distribution

Return Period (yrs)	%age Reduction in Design Rainfall using Mixed Gumbel Distribution					
	Gangneung	Ganghwa	Pohang	Buyeo	Gwangju	Goheung
10	2.50	4.45	4.36	2.10	4.17	4.43
20	4.88	8.56	7.55	5.92	7.25	8.44
50	9.96	12.46	12.72	9.79	9.70	12.53
100	15.31	13.82	18.06	14.98	8.32	14.84
200	21.26	11.57	19.81	20.64	5.79	15.90

IV. Concluding Remarks

Outliers in 24-hr annual maximum rainfall observations were studied in this study. A mixed Gumbel distribution was employed to incorporate the occurrence probability of outliers with the probability density function of rainfall amount. The innovative aspect of this paper is that the occurrence probability of outliers is systematically taken into account in the scheme of extreme rainfall analysis. Since conventional univariate probability distributions can hardly consider the occurrence probability of outliers, the use of a mixed Gumbel distribution was practicable for constructing feasible rainfall DDF curves for hydrologic design and management.

It has not yet been clearly proven that outliers in hydrologic observations result from the different mechanism of hydrologic process or the mistakes during observation process. In the presence of outliers, the rainfall DDF curves give very high values of design rainfalls which will be safe but not economically feasible and very high budget is required to complete hydrologic projects with high values of design rainfalls. While the use of rainfall DDF curves constructed using annual maximum rainfall observations after deleting outliers can lead to under estimation which will not be safe. The low values of design rainfall lead to under design, the under design will damage the structure.

The frequency analysis model proposed in this study can be adapted for the regional rainfall frequency analysis if the spatial variation of extreme rainfall occurrence is generalized for the large basin. The rainfall DDF curves constructed using the mixed Gumbel distribution, as proposed in this study, will be a good choice for safe and economical design and management of hydrological systems.

Acknowledgement

This research was supported by Basic Science Research Program through the National Research Foundation (NRF) of Korea funded by the Ministry of Education (2013R1A1A2013160).

References

[1] Anscombe FJ (1960) Discussion on rejection of outliers. *Technometrics 2:165-166*
[2] Barnett V, Lewis T (1994) Outliers in statistical data. Wiley, New York
[3] Chow VT, Maidment DR, Mays LW (1998) Applied Hydrology. McGraw-Hill, New York
[4] Cordery I, Mehrotra R., Nazemosadat MJ (2007) How reliable are standard indicators of stationary? *Stochastic Environmental Research and Risk Assessment 21:765-771*
[5] Griffis VW, Stedinger JR (2007) Evolution of flood frequency analysis with bulletin 17. *Journal of Hydrologic Engineering 12:283-297*
[6] Grubbs F (1969) Procedures for detecting outlying observations in samples. *Technometrics 11:1-21*
[7] Heo JH, Lee DJ, Shin H, Nam W (2006) A study on uncertainty of risk of failure based on Gumbel distribution. *Journal of Korea Water Resource Association 39:659-668*
[8] Kedem B, Chiu LS, Karni Z (1990) An analysis of the threshold method for measuring area- average rainfall. *Journal of Applied Meteorology 29:3–20*

[9] Kwon YM, Han JW, Kim T-W (2008) Estimation of design rainfalls considering nonstationarity in observations. In: Proceedings of the 5th Annual Meeting Asia Oceania Geosciences Society, Busan, Korea
[10] KWRA (2005) Guidelines on river design. Korea Water Resources Association, Korea
[11] Lee CH, Ahn JH, Kim T-W (2009) Nonstationary rainfall frequency analysis considering climate variability. In: Proceedings of the 6th Annual Meeting Asia Oceania Geosciences Society, Suntec, Singapore
[12] Lee S, Maeng S (2003) Frequency analysis of extreme rainfall using L-moment. *Irrigation and Drainage 52:219–230*
[13] Liu H, Shah S, Jiang W (2004) On-line outlier detection and data cleaning. *Computers and Chemical Engineering 28:1635–1647*
[14] Loaiciga HA, Leipnik RB (1999) Analysis of extreme hydrologic events with Gumbel distributions: marginal and additive cases. *Stochastic Environmental Research and Risk Assessment 13:251-259*
[15] McCuen RH (2004) Hydrologic analysis and design. Pearson Prentice Hall, New Jersey
[16] Overeem A, Buishand A, Holleman I (2008) Rainfall depth-duration-frequency curves and their uncertainties. *Journal of Hydrology 348:124-134*
[17] Rizwan M, Kim T-W (2009) Treatment of outliers in rainfall observations for constructing IDF curves. In: Proceedings of the 6th Annual Meeting Asia Oceania Geosciences Society, Suntec, Singapore
[18] Shimizu K (1993) A bivariate mixed lognormal distribution with an analysis rainfall data. *Journal of Applied Meteorology 32:161–171*
[19] Williams GJ, Baxter RA, He HX, Hawkins S, Gu L (2002) A comparative study of RNN for outlier detection in data mining. In: Proceedings of the 2nd IEEE International Conference on Data-mining (ICDM'02), Maebashi City, Japan
[20] Xu Y-P, Tung Y-K (2009) Constrained scaling approach for design rainfall estimation. *Stochastic Environmental Research and Risk Assessment 23:697-705*
[21] Yoo C, Jung K-S, Kim T-W (2005) Rainfall frequency analysis using a mixed Gamma distribution: evaluation of the global warming effect on daily rainfall. *Stochastic Environmental Research and Risk Assessment 19:3851-3861*