Establishing acceptance region for L-moment-based goodness-of-fit test of 3-parameter distributions – Pearson Type III distribution

Professor Ke-Sheng Cheng
Department of Bioenvironmental Systems Engineering, National Taiwan University

1. Introduction

Hydrology is a subject taught in various engineering disciplines such as hydraulic and hydrological engineering, civil engineering, agricultural engineering, and environmental engineering. It is also a topic studied in geoscience-related disciplines including geography, geology, soil and water conservation, forestry, etc. The study of hydrology is thus necessary not only for engineering practice but also for scientific advancement. Hydrological processes occur in the atmosphere, land surface and subsurface of the earth. It involves numerous sub-processes and parameters which exhibit various degrees of spatial and temporal variability. Many hydrological processes are not fully understood, and simplified physical and conceptual models are developed for purpose of practical applications. In addition, observation of hydrological variables can only be conducted in limited spatial and temporal points. The inability to correctly model the hydrological processes and collect sufficient data to characterize the spatial and temporal variabilities of hydrological processes results in significant uncertainties in hydrological modeling and forecasting.

In recent years, frequent occurrences of natural hazards (e.g., droughts, flood inundation, flash floods, and storm-triggered debris flows) and increasing concern about the hydrological consequences of climate changes have led the research communities and administrative agencies to become aware of or take into account the uncertainties involved in hydrological modeling and forecasting in their decision-making process.

On the one hand, hydrology, as a branch of geoscience, is heavily dependent on available data for making inference and forecasting. On the other hand, like many other fields in geoscience, it suffers from the difficulty of quantifying uncertainties involved in modeling and forecasting. Recently, there have seen many new models like artificial neural network (ANN) and support vector machine (SVM) introduced for hydrological modeling and forecasting, partly due to the fast advancement in artificial intelligence, machine learning, and data mining. Even though theses models have gain interests and attentions of many researchers, their practical implementation in real world problems are few. The main obstacles for practical application of these models are two-fold – (1) their inability to quantify the modeling and forecasting

uncertainties, (2) the nature of data-dependent model structure embedded in these data-driven models.

In light of the importance of uncertainties in decision making, techniques of stochastic simulation will be applied to quantitatively assess the uncertainties involved in hydrological modeling and forecasting.

2. Objective

The specific objective of this study is:

Establishing sample-size-dependent acceptance regions of *L*-moments for goodness-of-fit (GOF) test of three-parameter distributions commonly used in hydrological frequency analysis.

3. Significance

In a previous study (Liou, et al., 2008) we have established 95% acceptance regions for sample *L*-moments of the normal and Gumbel distributions using stochastic simulation. The two distributions are two-parameter distributions with a location and a scale parameter, and their 3rd and 4th *L*-moments occupy a single unique point on the skewness and kurtosis *L*-moment-ratio diagram (LMRD). In contrast, a three-parameter distribution (for examples, the Pearson-Type-III (PT3) distribution and the General Extreme Value (GEV) distribution) with location, scale and shape parameters plots as a curve on the LMRD. Practically speaking, *L*-moments acceptance regions of three-parameter distributions are to be sought after since the PT3, log-PT3 and GEV distributions have been demonstrated most useful in hydrological frequency analysis. Although similar approach appears applicable for establishing *L*-moments acceptance regions of three-parameter distributions, it is actually much more complicated since the intensity of computation is several order higher and, most importantly, the acceptance regions not only vary with sample size but also the shape parameter. Unlike the normal and Gumbel distributions for which unique elliptical acceptance regions can be established, PT3 and GEV distributions have infinite number of acceptance regions with respect to various values of the shape parameter. Thus, LMRD-based goodness-of-fit test of the PT3 and GEV distributions must also consider the uncertainty of shape parameter estimation. In addition, the PT3 and GEV distributions are most commonly used for frequency analysis of both rainfall and flood. Therefore, **establishing sample-size-dependent acceptance regions for three-parameter distributions such as the PT3 and GEV distributions, with consideration of the uncertainty in parameter estimation of shape factor, will provide a complete set of tools for LMRD-based GOF test of probability distributions commonly used in hydrological frequency analysis**.

4. State of Knowledge

4.1 L-moment-ratio diagram

In recent years there have been many applications of *L*-moments for frequency analysis and the skewness and kurtosis *L*-moment-ratio diagram (LMRD) was suggested as a useful tool for discrimination between candidate distributions (Hosking, 1990; Hosking and Wallis, 1993, Vogel and Fennesset, 1993, Hosking and Wallis, 1997). The *L*-moments uniquely define the distribution if the mean of the distribution exists, and the *L*-skewness and *L*-kurtosis are much less biased than the ordinary skewness and kurtosis (Hosking and Wallis, 1997). As demonstrated in Figure 1, a two-parameter distribution with a location and a scale parameter plots as a single point on the LMRD, whereas a three-parameter distribution with location, scale and shape parameters plots as a curve on the LMRD, and distributions with more than one shape parameter generally are associated with regions on the diagram (Hosking and Wallis, 1997). However, theoretical points or curves of various probability distributions on the LMRD cannot accommodate for uncertainties induced by parameter estimation using random samples. Therefore, the effect of record length on estimates of parameters should be considered in determining the best-fit distribution for regional frequency analysis.

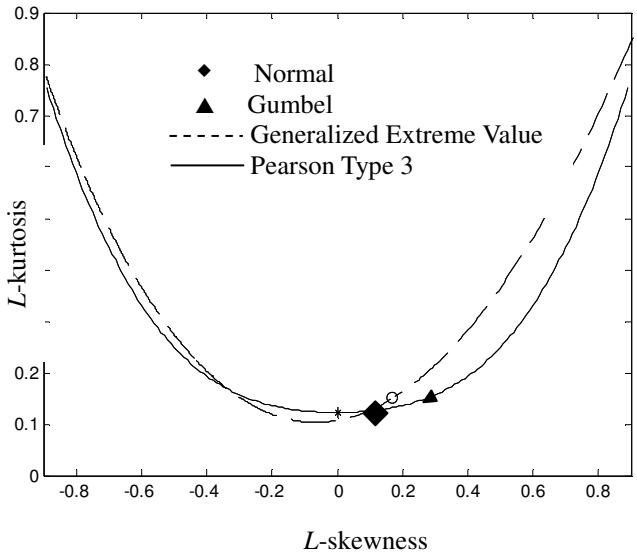


Figure 1. *L*-moment-ratio diagram of various distributions.

4.2 Bivariate distribution of sample L-skewness and L-kurtosis

L-moments are an alternative system of describing the shapes of probability distributions. For a random variable X with quantile function $x(u)$, Hosking and Wallis (1997) defined the *L*-moments $(\lambda_r, r = 1,2,\cdots)$ as

$$\lambda_r = \int_0^1 x(u) P_{r-1}^*(u) du \tag{1}$$

where

$$P_r^*(u) = \sum_{k=0}^{r} p_{r,k}^* u^k \ , \quad r = 0,1,2,\cdots \tag{2}$$

$$p_{r,k}^* = (-1)^{r-k} \binom{r}{k} \binom{r+k}{k} = \frac{(-1)^{r-k}(r+k)!}{(k!)^2 (r-k)!} \, . \tag{3}$$

The *L*-moments can also be expressed in terms of the probability weighted moments defined by Greenwood et al. (1979), and the first four *L*-moments are given by

$$\lambda_1 = \beta_0 \tag{4}$$

$$\lambda_2 = 2\beta_1 - \beta_0 \tag{5}$$

$$\lambda_3 = 6\beta_2 - 6\beta_1 + \beta_0 \tag{6}$$

$$\lambda_4 = 20\beta_3 - 30\beta_2 + 12\beta_1 - \beta_0 \tag{7}$$

where β_r, $r = 0,1,2,\cdots$, are probability weighted moments defined by

$$\beta_r = \int_0^1 x(u) u^r du \tag{8}$$

In terms of linear combination of order statistics, the *L*-moments can also be expressed by

$$\lambda_1 = E\left(X_{1:1}\right) \tag{9}$$

$$\lambda_2 = \frac{1}{2} E\left(X_{2:2} - X_{1:2}\right) \tag{10}$$

$$\lambda_3 = \frac{1}{3} E\left(X_{3:3} - 2X_{2:3} + X_{1:3}\right) \tag{11}$$

$$\lambda_4 = \frac{1}{4} E\left(X_{4:4} - 3X_{3:4} + 3X_{2:4} - X_{1:4}\right) \tag{12}$$

where $X_{k:n}$ is the *k*-th order statistic from a random sample of size *n*.

Similar to the ordinary moment ratios, the *L*-moment ratios are defined by

$$\tau_r = \frac{\lambda_r}{\lambda_2} \, , \quad r = 3,4,\cdots \tag{13}$$

Theoretical relationships between *L*-skewness (τ_3) and *L*-kurtosis (τ_4), i.e. the *L*-moment ratio diagram, of several probability distributions have been given by Hosking (1990) and can be used to distinguish different probability distributions (see Fig. 3).

Given a random sample $\{x_1, x_2, \cdots, x_n\}$, an unbiased estimator of the probability weighted moment β_r is given by

$$b_r = \frac{1}{n} \sum_{j=r+1}^{n} \frac{(j-1)(j-2)\cdots(j-r)}{(n-1)(n-2)\cdots(n-r)} x_{j:n} \tag{14}$$

The sample *L*-moments (ℓ_r) and sample *L*-moment ratios (t_r) can then be calculated by

$$\ell_1 = b_0 \tag{15}$$

$$\ell_2 = 2b_1 - b_0 \tag{16}$$

$$\ell_3 = 6b_2 - 6b_1 + b_0 \tag{17}$$

$$\ell_4 = 20b_3 - 30b_2 + 12b_1 - b_0 \tag{18}$$

$$t_r = \ell_r / \ell_2 \tag{19}$$

The sample *L*-moments ℓ_r is an unbiased estimator of λ_r; however, the estimator t_r is not an unbiased estimator of τ_r (Hosking and Wallis, 1997), even though for most distributions the biases are negligible for sample sizes of 20 or more. Also, the sample *L*-skewness (t_3) and *L*-kurtosis (t_4) are found to have a joint distribution close to bivariate normal. However, the exact distributions of the sample *L*-moment ratios are difficult to be derived.

In addition to the above mentioned sample *L*-moments and sample *L*-moment ratios, Hosking and Wallis (1997) also defined the plotting-position estimators of λ_r and τ_r as

$$\tilde{\lambda}_r = \frac{1}{n}\sum_{j=1}^{n} P_{r-1}^{*}\left(p_{j:n}\right) x_{j:n} \tag{20}$$

$$\tilde{\tau}_r = \tilde{\lambda}_r / \tilde{\lambda}_2 \tag{21}$$

where $p_{j:n}$ is a plotting-position estimator and was chosen to be

$$p_{j:n} = (j - 0.35)/n\,. \tag{22}$$

Hosking and Wallis (1997) indicated that $\tilde{\lambda}_r$ is not an unbiased estimator of λ_r, but its bias tends to zero in large samples. Hereafter t_r and $\tilde{\tau}_r$ will be respectively referred to as the probability-weighted-moment estimator and the plotting-position estimator of the *L*-moment ratio τ_r. It is generally advised to use the probability-weighted-moment estimators since they are inferior to the plotting-position estimators only for some instances of estimation of extreme quantiles in regional frequency analysis and have generally lower bias as estimators of the *L*-moment ratios (Hosking and Wallis, 1997).

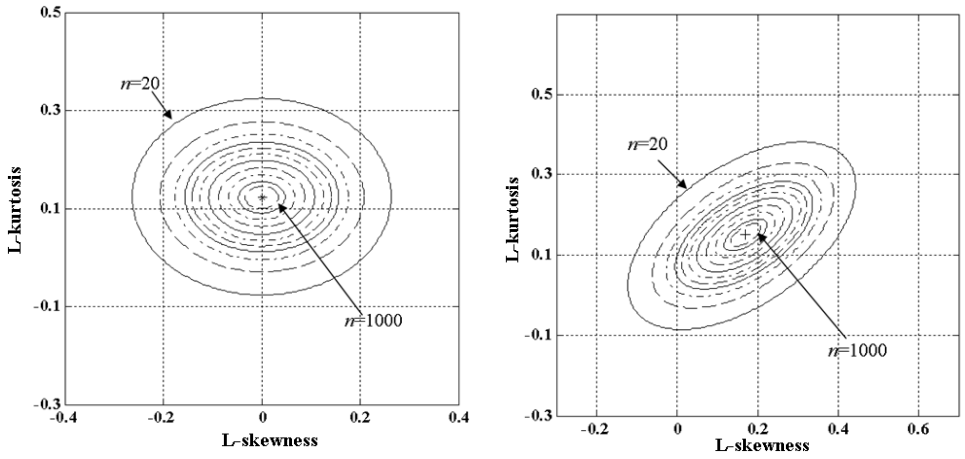
It should also be addressed that the plotting-position *L*-moment estimators are non-invariant estimators (Hosking and Wallis, 1997), and thus their statistical properties vary with changes in the location and scale parameters of the population from which random samples are drawn. In order to derive statistical properties of the plotting-position *L*-moment estimators which are generally applicable with respect to location and scale parameters, it is necessary to have the random samples preprocessed (sample mean subtraction followed by division by sample standard deviation) for normalization (zero mean and unit standard deviation), and use the

standardized data for calculation of the plotting-position *L*-moment estimates using Eqs. (20) and (21).

Through stochastic simulation of the normal and Gumbel random variables, Liou et al. (2008) demonstrated that, for both distribution types, the joint distribution of sample *L*-skewness and *L*-kurtosis seem to resemble a bivariate normal distribution for a larger sample size (n = 100).

4.3 Sample-size dependent acceptance regions of LMRD

Liou et al. (2008) established sample-size-dependent 95% acceptance regions of sample *L*-skewness and *L*-kurtosis for both normal and Gumbel distributions (Figure 2). The acceptance regions are determined by a set of equiprobable ellipses. The equiprobable ellipses are in turn characterized by the mean vector and covariance matrix of sample *L*-skewness and *L*-kurtosis which can be estimated with very high accuracy using a set of sample-size-dependent empirical relationships.



(Normal distribution) (Gumbel distribution)

Figure 2. 95% acceptance regions of *L*-moments-based GOF test for the normal and Gumbel distributions. Acceptance ellipses correspond to various sample sizes (n = 20, 30, 40, 50, 60, 75, 100, 150, 250, 500, and 1,000). (Liou et al., 2008)

5. Approach

5.1 Establishing the sample-size-dependent 95% acceptance regions of sample *L*-moments of the PT3 and GEV distributions

5.1.1 Stochastic simulation of the PT3 and GEV distributions

From the view point of random number generation, the frequency factor can be considered as a random variable K, and K_T is a value of K with exceedence probability $1/T$. For example, frequency factor of the PT3 distribution can be approximated by (Kite, 1988)

$$K_T \approx z + \left(z^2-1\right)\frac{\gamma_X}{6} + \frac{1}{3}\left(z^3-6z\right)\left(\frac{\gamma_X}{6}\right)^2 - \left(z^2-1\right)\left(\frac{\gamma_X}{6}\right)^3 + z\left(\frac{\gamma_X}{6}\right)^4 - \frac{1}{3}\left(\frac{\gamma_X}{6}\right)^5 \qquad (23)$$

where z is the standard normal deviate and γ_X is the coefficient of skewness of X. Given μ_X, σ_X and γ_X, if we can generate a set of random numbers of K, say $k_1, k_2, \cdots, k_n$, then a random sample of X, say $x_1, x_2, \cdots, x_n$, can be obtained by

$$x_i = \mu_X + k_i \sigma_X \tag{24}$$

Note that each $k_i, i = 1,2,\cdots n,$ corresponds to its own exceedence probability $1/T_i$ and, given the coefficient of skewness, the frequency factor K is only dependent on the standard normal deviate z. Thus, simulation of the PT3 distribution can be achieved by transferring random samples of the standard normal deviate to random samples of the PT3 distribution. Stochastic simulation of GEV distribution can be achieved in a similar manner.

For either of the PT3 and GEV distribution, a total of 100,000 random samples will be generated with respect to the specified sample size $n = 20$, 30, 40, 50, 60, 75, 100, 150, 250, 500, and 1,000. For each of the 100,000 samples, sample L-skewness and L-kurtosis will be calculated using the probability-weighted-moment estimator and the plotting-position estimator.

5.1.2 *Evaluating the bivariate normality assumption for sample L-skewness and L-kurtosis using the Mardia test*

The following Mardia test statistic $M_{2,p}$ has an asymptotic standard normal distribution

$$M_{2,p} = \frac{b_{2,p} - [p(p+2)(N-1)/(N+1)]}{\sqrt{8p(p+2)/N}} \tag{25}$$

where N is the number of random samples. At level of significance α, the null hypothesis of multivariate normality is rejected if

$$| M_{2,p} | > z_{1-\alpha/2} \tag{26}$$

where $z_{1-\alpha/2}$ is the 100(1-α/2)% quantile of the standard normal distribution. Hence, the null hypothesis is rejected if the absolute value of the Mardia statistic $M_{2,p}$ exceeds 1.96.

5.1.3 *Establishing acceptance regions for GOF tests*

Using this bivariate normality property, the $100(1-\alpha)\%$ acceptance region of a GOF test based on sample L-skewness and L-kurtosis can be determined by the equiprobable density contour of the bivariate normal distribution with its encompassing area equivalent to $1-\alpha$. The well-known Hotelling's T^2 statistic

$$T^2 = (X - \bar{x})^T S^{-1} (X - \bar{x})$$

$$= \frac{1}{1-r^2}\left[\frac{(X_1 - \bar{x}_1)^2}{s_1^2} - \frac{2r(X_1 - \bar{x}_1)(X_2 - \bar{x}_2)}{s_1 s_2} + \frac{(X_2 - \bar{x}_2)^2}{s_2^2}\right] \tag{27}$$

Where X_1 and X_2 are respectively *L*-skewness and *L*-kurtosis, s_1^2 and s_2^2 represent the unbiased sample variances of *L*-skewness and *L*-kurtosis, and r is the correlation coefficient of sample *L*-skewness and *L*-kurtosis. The Hotelling's T^2 is distributed as a multiple of an *F*-distribution, i.e.,

$$T^2 \sim \frac{2(N^2-1)}{N(N-2)} F_{(2,N-2)}. \tag{28}$$

where *N* is the size of a random sample of the bivariate vector *X*. At *N* = 100,000, we have

$$\frac{2(N^2-1)}{N(N-2)} F_{2,N-2}(\alpha) \approx 2F_{2,N-2}(\alpha) = \chi^2_{2,\alpha} \tag{29}$$

Therefore, the distribution of the Hotelling's T^2 can be well approximated by the chi-square distribution with degree of freedom 2. Thus, if the sample *L*-moments pair of a random sample of size *n* falls outside of the corresponding ellipse, i.e.

$$T^2 = \left(X - \bar{x}_n\right)^T S_n^{-1} \left(X - \bar{x}_n\right) > \chi^2_{2,\alpha} \tag{30}$$

the null hypothesis that the random sample is originated from a PT3 or GEV distribution is rejected.

A significant difference between the *L*-moment-ratios of the normal (or Gumbel) and PT3 (or GEV) distributions is that the former occupies a unique point on LMRD whereas the latter plots as a curve. Thus, acceptance regions of the PT3 and GEV distributions cannot be determined by a single set of sample-size-dependent equiprobable ellipses. Instead, many sets of sample-size-dependent equiprobable ellipses with respect to various values of shape factor will be established. The 95% acceptance bands as illustrated in Figure 3 will be constructed by considering the conditional distribution of the sample *L*-kurtosis given *L*-skewness.

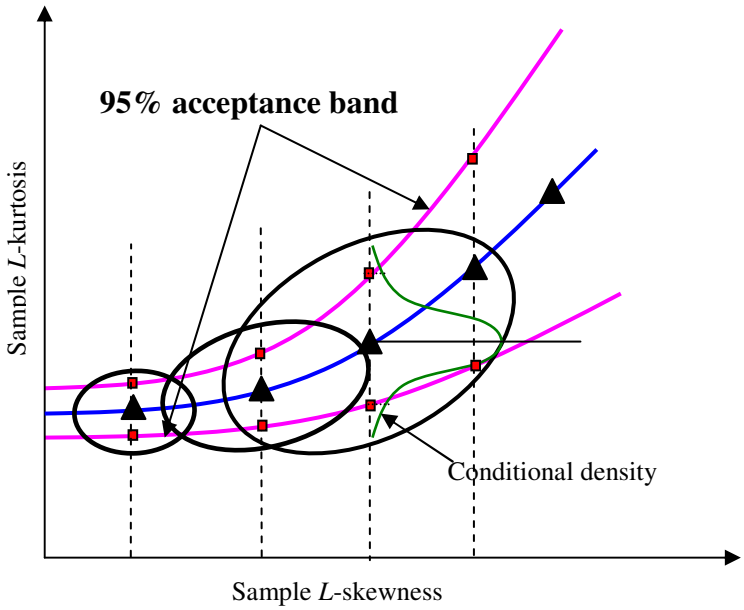


Figure 3. Illustration of 95% acceptance band using conditional density.

6. References

許介維，2004。*序率模擬應用於機率分布適合度檢定之評估*。國立台灣大學生物環境系統工程學系碩士論文。

徐宏瑋，2004。*降雨量變遷趨勢檢定與分析*。國立台灣大學生物環境系統工程學系碩士論文。

劉俊志、吳宜珍、江介倫、鄭克聲，2008。線性動差適合度檢定法之檢定力測試。*中國農業工程學報*。

Ashkar F, El-Jabi N, Issa M. 1998. A bivariate analysis of the volume and duration of low-flow events. *Stochastic Hydrology and Hydraulics* 12: 97-116.

Bacchi B, Becciu G, Kottegoda NT. 1994. Bivariate exponential model applied to intensities and durations of extreme rainfall. *Journal of Hydrology* 155: 225-236.

Cheng KS, Chiang JL, Hsu CW. 2007. Simulation of probability distributions commonly used in hydrological frequency analysis. *Hydrological Processes* 21: 51–60.

Cheng KS, Hou JC, Liou JJ. 2008. Stochastic simulation of bivariate gamma distribution – A frequency-factor based approach. Submitted to *Environmetrics*, in review.

Cherian KC. 1941. A bi-variate correlated gamma-type distribution function. *Journal of Indian Mathematical Society* 5: 133-144.

Chow, V.T., 1951. A general formula for hydrologic frequency analysis. Transaction of American Geophysics Union, 32, 231-237.

Clark RT. 1980. Bivariate gamma distribution for extending annual streamflow

records from precipitation: Some large sample results. *Water Resources Research*, 16: 863-870.

Cowpertwait PSP, O'Connell PE, Metcalfe AV, Mawdsley JA. 1996. Stochastic point process modelling of rainfall. I. Single-site fitting and validation. *Journal of Hydrology* 175: 17-46.

D'este GM. 1981. A Morgenstern-type bivariate gamma distribution. *Biometrika* 68(1): 339-340.

Goel NK, Kurothe RS, Mathur BS, Vogel RM. 2000. A derived flood frequency distribution for correlated rainfall intensity and duration. *Journal of Hydrology* 228: 56-67.

Grimaldi S, Serinaldi F, 2006. Asymmetric copula in multivariate flood frequency analysis. *Advances in Water Resources*, 29:1155-1167.

Hosking, J.R.M., 1990. *L*-moments: analysis and estimation of distributions using linear combinations of order statistics. Journal of the Royal Statistical Society Series B, 52(1), 105-124.

Hosking, J.R.M. and Wallis, J.R., 1993. Some statistics useful in regional frequency analysis. Water Resources Research, 29(2), 271-281.

Hosking, J.R.M. and Wallis, J.R., 1995. A comparison of unbiased and plotting-position estimators of *L*-Moments. Water Resources Research, 31(8), 2019-2025.

Hosking, J.R.M. and Wallis, J.R., 1997. Regional frequency analysis: an approach based on *L*-moments. Cambridge University Press, Cambridge, U.K.

Kite GW. 1988. *Frequency and Risk Analysis in Hydrology*. Water Resources Publications.

Law AM. 2007. *Simulation Modeling and Analysis*. McGraw Hill: Singapore.

Liou JJ, Wu YC, Cheng KS. 2008. Establishing acceptance regions for L-moments based goodness-of-fit tests by stochastic simulation. Journal of Hydrology doi:10.1016/j.jhydrol.2008.02.023.

Loaiciga HA, Leipnik RB. 2005. Correlated gamma variables in the analysis of microbial densities in water. *Advances in Water Resources* 28: 329-335.

Loganathan GV, Kuo CY, Yannaccone J. 1987. Joint probability distribution of streamflows and tides in estuaries. *Nordic Hydrology* 18: 237-246.

Mardia, K.V., 1970. Measures of multivariate skewness and kurtosis with applications. Biometrika, 57(3), 519-530.

Mardia, K.V., Kent, J.T., and Bibby, J.M., 1979. Multivariate Analysis. Academic Press, New York.

Moran PAP. 1969. Statistical inference with bivariate gamma distributions. *Biometrika* 56(3): 627-634.

Schmeiser BW, Lal R. 1982. Bivariate gamma random vectors. *Operation Research* 30(2): 355-374.

Scholz K. 1997. Stochastic simulation of urban hydrological processes. *Water Science and Technology* 36: 25-31.

Shiau, JT, 2003. Return period of bivariate distributed extreme hydrological events. *Stochastic environmental research and risk assessment*, 17:42-57.

Shiau, JT, Wang, HY, Tsai, CT, 2006. Bivariate frequency analysis of floods using copulas, *Journal of the American Water Resources Association*, 42(6): 1549-1564.

Singh K and Singh VP, 1991. Derivation of bivariate probability density functions with exponential marginals. *Journal of Stochastic Hydrology and Hydraulics* 5: 55-68.

Tu, JY, Chou, C and Chu, PS. 2008. Abrupt shift of typhoon activity in the vicinity of Taiwan and its association with the western North Pacific-East Asian climate change. *J. Climate*, in revision.

Werner TM, Kadlec RH. 2000. Stochastic simulation of partially-mixed, event-driven treatment wetlands. *Ecological Engineering* 14(3): 253-267.

Wu, Y.C., 2005. Establishing acceptance regions for goodness-of-fit test by stochastic simulation. Master Thesis (in Chinese), Department of Bioenvironmental Systems Engineering, National Taiwan University.

Yue, S. 1999. Applying bivariate normal distribution to flood frequency analysis. *Water International* 24(3): 248-254.

Yue, S. 2000. Joint probability distribution of annual maximum storm peaks and amounts as represented by daily rainfalls. *Hydrological Science Journal* 45(2): 315-326.

Yue, S. 2001. A bivariate gamma distribution for use in multivariate flood frequency analysis. *Hydrological Processes* 15: 1033-1045.

Yue S, Ouarda TBMJ, Bobee B. 2001. A review of bivariate gamma distributions for hydrological application. *Journal of Hydrology* 246: 1-18.

Zhang L, Singh VP, 2007. Bivariate rainfall frequency distributions using Archimedean copulas. *Journal of Hydrology*, 332: 93-109.

Zhao X, Chu PS. 2006. Bayesian Multiple Change point Analysis of Hurricane Activity in the Eastern North Pacific: A Markov Chain Monte Carlo Approach. *Journal of Climate* 19: 564-578.