2007

A joint probability approach for the confluence flood frequency analysis

Cheng Wang
Iowa State University

A joint probability approach for the confluence flood frequency analysis

by

Cheng Wang

A thesis submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Major: Environmental Science

Program of Study Committee:
Ramesh S. Kanwar, Major Professor
Roy Gu, Co-Major Professor
U. Sunday Tim

Iowa State University

Ames, Iowa

2007

UMI Number: 1447555

UMI Microform 1447555

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

Table of Contents

List of Tables

List of Figures

Acknowledgements

I would like to take this opportunity to express my thanks to Dr. Ramesh Kanwar, and Dr. Roy Gu for their guidance and financial support during my graduate study. I would also like to thank Dr. Sunday Tim for his efforts and contributions to this work.

I am indebted to my family for their support and encouragement that enable me to successfully complete this work. Especially, I am grateful to my wife Lili for her love, support and encouragement during my study. I would also like to thank my sons, Daniel and Eric, who brighten my life.

Abstract

The flood frequency analysis at or nearby the confluence of two tributaries is of interest because it is necessary for the design of the highway drainage structures, which often are located near the confluence point and may be subject to inundation by high flows from either stream or both. The shortage of the hydrological data of the confluence point which are necessary to the univariate flood frequency analysis makes the flood estimation at the confluence challenging. This thesis presents a practical procedure for the flood frequency analysis at the confluence of two streams by multivariate simulation of the annual peak flow rate of the tributaries based on joint probability and Monte Carlo simulation.

Four steps are involved in the proposed approach, the distribution identification of annual peak flow rate of the tributary streams, the identification of joint probability distribution of the tributary stream flows, the generation of the synthetic annual peak flow rate at the confluent point by using Monte Carlo simulation, and identification of the flood frequency of the confluent point by the univariate flood frequency analysis.

Due to the difficulty identifying the joint probability distribution of two specified marginal distributions, an easy and practical method for the identification of joint probability distribution is needed. Copulas method is introduced and several often used copulas are employed to identify the joint probability.

Two case studies are conducted and the results are compared with the flood frequency of the confluence point obtained by the well accepted univariate flood frequency analysis based on the observation data. The results are also compared with the ones by the National Flood Frequency program developed by United State Geological Survey. It is found out that the results by the proposed model are very close to the results by the unvariate flood frequency analysis, while the National Flood Frequency program tends to underestimate the flood for a certain return period, especially when the return period is less than 50 or 100 years, and when the river basin is getting larger.

Keywords: Flood frequency analysis, goodness-of-fit, Chi-square test, Kolmogorov-Smirnov test, joint probability, Monte Carlo simulation, confluence point, copulas

Chapter 1 Introduction

1.1 Introduction

The ability to adequately define the magnitude and frequency of floods is necessary for the regulation, planning, and design of activities along rivers and streams. One of the first considerations in the safe and economical design of drainage structures is the magnitude and frequency of the design flood or the maximum peak flow that can safely pass through the structure, many of which are located at or near the confluence point of the tributaries. The most desirable basis for selection of the design discharge is a flood-frequency analysis of a long-term records of flood that have occurred at or near the site, but long-term flood records are rarely available for the site where they are needed, for example, the confluence of the tributaries.

This thesis presents a flood frequency analysis for the confluent point of the tributaries based on the joint probability distribution and Monte Carlo simulation. Copula method is introduced to obtain the joint probability distribution with specified marginal distributions, which plays a key role in the proposed model but usually very difficult to be identified since there are no general approaches available or addressed in relative detail in engineering area.

1.2 Background and Problem Identification

Highway drainage structures and water management facilities are often located near the confluence of two or more streams (see Figure 1-1), where they may be subject to inundation by high flows from one stream or all. These structures are designed to meet specified performance objectives for floods of a specified return period (e.g., the 100-year flood). Because the flooding of structures on one stream can be affected by high flows on the other stream, it is important to know the relationship between the coincident exceedence probabilities on the confluent stream pair (i.e., the joint probability of the coincident flows). Accurate estimates of the joint probability of design flows at stream confluences are a crucial element in the design of efficient and effective highway drainage structures and water management facilities. No accurate generally accepted estimation procedure for determining coincident flows currently exists for use in the design of highway structures and water management facilities at the confluence of the tributary. A practical procedure for the determination of joint probabilities of design flows at stream confluences is needed.

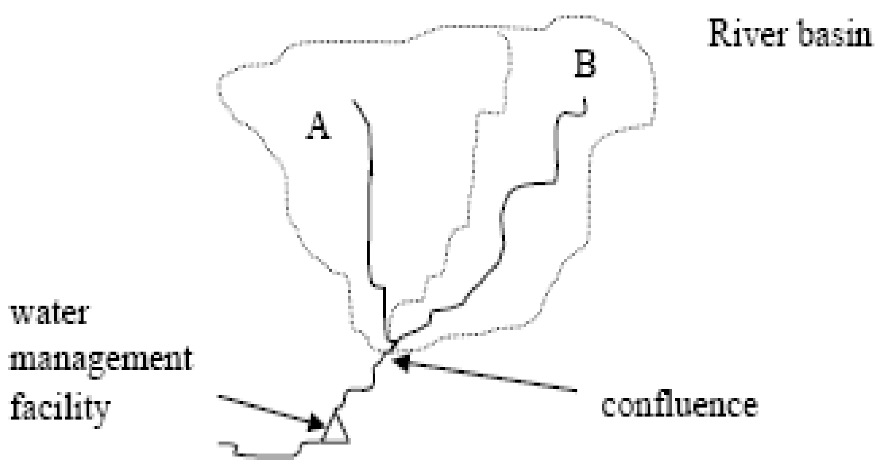

Figure 1-1 River basin illustration

1.3 Review of literature

1.2.1 Flood Frequency Analysis

Flood frequency analysis is a key issue in hydrology. The main objective of flood frequency analysis is to relate the flood magnitude of extreme events to their frequency of occurrence. The results of flood flow frequency analysis can be used for many engineering purposes: for the design of dams, bridges, culverts, and flood control structures; to determine the economic value of flood control projects; and to delineate flood plains and determine the effect of encroachments on the flood plain (Chow et al., 1988).

All the proposed flood frequency analysis methods may be roughly classified into three categories depending on the availability and the length of observed flood data for the site: regional analysis, stream-based analysis and time series analysis. Regional analysis and stream-based analysis are more often used. The well established univariate flood frequency analysis based on the annual peak flow rate distribution is employed in the case that a long enough flow records are available, while for the un-gauged stream, the regional analysis currently seems the only effective method to apply that relates the flood magnitude to the hydrologic characters of a specified region, such as rainfall, drainage area, and so on. Some researchers, i.e. Rao and Hamed (2000) consider the time series a special case of stream-based analysis, which is proposed in Flood Studies Report (1975). It is separated from the stream-based analysis in this thesis based on the time interval length of the flood observations. Annual peak flow rate is mainly used in most stream-based analysis while the daily flow rate is preferred in the time series method.

In the time series method, the flow hydrograph is considered to be a time series in which the flows are represented by a series of ordinates at equally spaced intervals of time (days). To use the time series models, relatively long records are required and the data requirements are greater than for univariate flood frequency analysis. Rao and Hamed (2000) described the time series method as follows:

"Ideally, if a hydrograph is considered to be a stochastic process in continuous time, properties of such a series can be deduced from those of the parent process. If Q(t) is the flow on day t, and time series model may be written as the sum of trend, seasonal, and stochastic components. Estimation of model formulation and parameters proceed together through the three components beginning with trend and ending with the stochastic component. "

Flood frequency analysis of a single variable has been discussed since 1950's to relate the magnitude of extreme events to their frequency of occurrence by using the probability distributions (Chow et al., 1988). It has been well established and accepted in academic and engineering field, which is called univariate flood frequency analysis (UFF) in this thesis. Many literatures about the development and application of this approach have been addressed (Tod, 1957; Burkhardt and Prakash, 1976; Linsley, 1986, Sigh and Sigh, 1985; Rossi et al, 1984; Moharran et al,1993). Rao and Hamed (2000) summarized the conventional flood frequency analysis in detail and presented many examples for different stream discharge distributions and with different parameter estimations.

The flood frequency analysis based on the distribution is preferred to use when an adequate observation record of annual flood is available, such as 30 years or more of flood records. The most commonly used model of this approach is annual maximum series model. The annual peak flow rate data are used to establish a probability distribution that is assumed to describe the flooding process, and that can be evaluated by using data to determine the flood magnitude at any frequency. This approach has many advantages and also disadvantages. All the impact factors on the flood frequency, such as rainfall, are taken into account in the procedure so it is relatively easy to use. However, this approach may miss some information. For example, the second and third peak within a year may be greater than the maximum flow in other years and yet they are ignored (Kite, 1977; Chow et al. 1988; Rao and Hamed, 2000). This means this approach may underestimate or overestimated the true flood. Another disadvantage is that sometimes not all the existing data are available for the use of this approach for some reasons. For example, due to land use changes or the watershed characters change or the construction of the water management facilities in the site or upstream, i.e., a dam, the hydrologic characteristics may change. This change may result in the change of the trend of corresponding annual peak flow rate and this may make the annual flow data prior to the hydrologic condition changes are irrelevant to the future flood prediction. This actually reduces the available data from the existing record, and may bring some estimation error if not enough attention is paid on this. So although this approach has been well established and popular in academic and engineering, sometimes the dilemma exists when it is employed. Generally, the longer stream discharge record the studied stream has, the more accuracy UFF approach brings, while there are situations sometimes that no discharge flow record available or not long enough for UFF to obtain a accurate result, i.e. near of at the confluence of stream tributaries, or in some underdeveloped area with shortage of the historical hydrologic data.

The second approach, regional analysis is based on the concept of regional homogeneity and often used for the flood frequency estimation, especially valuable at ungauged sites. It is also used to enhance the flood estimation at gauged sites where historical records are short. This approach often based on the rainfall data. The rainfall-runoff routing process may be involved to convert the rainfall into flood discharge in this case, and the rainfall-runoff model provides the link between the rainfall data and the flood frequency estimation. This approach is relatively complex and time consuming. The U.S. Geological Survey (USGS) developed a set of regression equations by statistically relating the flood characteristics to the physical and climatic characteristics of the watersheds for a group of gauging stations within a region that have virtually natural stream flow conditions, with a format of $Q_T = aX^bY^cZ^d$, for rural area flood estimation in every state of U.S., where Q_T is the T-year rural flood-peak discharge, X, Y, Z are watershed or climatic characteristics, and a, b, c, d are regression coefficients. Drainage area or contributing drainage area is used as independence variable for the regression in almost all the regression equations for the 50 states of US. The other most frequently used watershed and climatic characteristics are main-channel slope and mean annual precipitation. The nationwide urban flood estimation regression equations based on multiple regression analysis of urban flood-frequency data from 199 urbanized basins are also provided in which more variables are included, such as drainage area, main channel slope, rainfall, basin storage, and so on. In the 1990's, a computer program called the National Flood Frequency Program (NFF) was developed, which compiled all the USGS available regression equations for estimating the magnitude and frequency of floods in the United States and Puerto Rico (USGS, 2002).

NFF is probably the most often used model and one of the very few models available for the ungauged site flood frequency estimation in US from the author's knowledge. It is relatively easy to use; however, it is inconvenient most time. In this approach all the states in US are divided into multiple hydrologic regions determined by using major watershed boundary and/or some other hydrologic characteristics, i. e., the mean elevation of watershed. A series of regression equations of T-year flood (T=2 , 5, 10, 25, 50, 100, 200 and 500 year) associated with each hydrologic region are developed in terms of hydrologic characteristics based on the gauged site records. One has to determine the hydrologic region of the interest site first among all the hydrologic regions and then pick up the developed regression equations to perform the flood frequency analysis. Moreover, some equations in this approach have high errors, for example, some equations generated for the western part of the US have standard error greater than 100 percent, although the average standard error of NFF is between 30 and 60 percent (USGS, 2002).

Based on the above review, one accurate and practical approach for ungauged confluence flood estimation that can overcome the shortages of UFF and NFF model is needed. The desire approach can use the available stream discharge records around the study site, which may be obtained relatively easily. Also the desire approach should be convenient for use. A joint probability approach is proposed in this thesis that may meet the two criteria.

1.2.2 Bivariate Flood frequency analysis

The research on bivariate distribution has been of interest of statisticians for a long time and many methods have been proposed to derive the joint distribution functions with the same or different margins (Molenberghs and Lesaffre, (1997); Ronning, 1977). With the recognition that the complex hydrological events such as floods are always affected by one or more correlated events and that an accurate estimate of the joint probability of the correlated events plays an important role for hydrology analysis, much attention has been paid on the bivariate and even multivariate flood frequency analysis since 1980s.

Sackl and Bergmann (1987), Chang et al. (1994), Yue (1999), and Beersma and Buishand (2004) used the bivariate normal distribution to perform the flood frequency analysis and hydrology events analysis. Krstanovic and Singh (1987) derived the multivariate Gaussian and exponential distributions by the principle of maximum entropy and applied the bivariate distributions for the analysis of flood peak and volume. Goel et al. (1998) employed a multi-variate normal distribution to perform flood frequency analysis after normalizing the peak flow data, volume and duration. Yue (2001a) applied the bivariate lognormal distribution for multivariate flood events analysis and described the relationship of flood peaks and volumes as well as flood volumes and durations by joint distribution and the corresponding conditional distribution.

Hashino (1985), Choulakian et al. (1990), Singh and Singh (1991), Bacchi et al. (1994), and Ashkar et al. (1998) investigated and applied bivariate exponential distributions for the hydrological events analysis. Bacchi et al. (1994) proposed a numerical procedure for the estimation of parameters of a bivariate exponectial model used to simulation the storm intensity and duration simultaneously.

Buishand (1984), Yue et al. (2001b) applied bivariate extreme value distributions to analyze multivariate flood/storm events. Yue and Wang (2004) compared the performance in flood analysis between two bivariate extreme value distributions, the Gumbel mixed model and the Gumbel logistic model. Shiau et al (2007) derived a joint probability distribution with a mixture of exponential and gamma marginal distribution to simulate the relationship between drought duration and drought severity.

Some researchers used bivariate gamma distribution for the flood frequency analysis (Moran, 1970; Crovelli, 1973; Prekopa and Szantai, 1978; Clarke, 1980; Yue, 2001b, 2001c; Yue, et al. 2001). Among them, Yue (2001c) investigated the applicability of the bivariate gamma distribution model to analyze the joint distribution of two positively correlated random variables with gamma marginals. Yue (2001b) reviewed three bivaraite gamma distribution models with two gamma marginal distributions. Durrans et al (2003) presented two approximate methods for joint frequency analysis using Pearson Type III distribution to estimate the joint flood frequency analyses on seasonal and annual basis. Nadarajah and Gupta (2006) developed exact distribution of intensity-duration based on bivariate gamma distribution.

Wang (2001) developed a procedure for record augmentation of annual maximum floods by applying the bivariate extreme value distribution for annual maximum floods at gauged stations with generalized extreme value distribution. Yue and Rasmussen (2002) discussed the concepts of bivariate hydrology events and demonstrated the concepts by applying a bivariate extreme value distribution to represent the joint distribution of flood peak and volume from a basin. Johnson et al. (1999) reviewed some techniques for obtaining bivariate distributions and presented the properties of some bivariate models, such as bivariate Weibull distribution, bivariate inverse Gaussian distribution, bivariate S_{BB} distribution and bivariate normal-lognormal distribution.

Zhang and Singh (2006) derived bivariate distributions of flood peak and volume, and flood volume and duration by using copula method. In the paper, four often used one parameter Archimedean copulas are introduced, the corresponding parameter estimation is described and the criteria of copula selection are addressed.

Most of the researchers just applied bivariate or multivariate distribution with the same type of marginal distributions, either two normal distributions or two gamma marginal distributions, and so on. Only a few of them, i.e., Zhang and Singh (2006) and Wang (2001) employed bivariate distribution with two different types of distribution. Although many researchers performed flood frequency analysis with the bivariate distributions, most of them focused more on identifying the relationship of different hydrologic variables, such as flood peak and volume, and flood volume and duration. In their researches, the flow discharge records of the site of interest are usually required. A bivariate distribution approach is presented in this thesis to estimate the flood and frequency at the confluence of the tributaries without the requirement of records of the studied sites.

1.4 Objective and scope of work

This research is to develop practical procedures for the flood frequency analysis for the confluence of the tributaries where many drainage structures are located but the long-term flood records may be unavailable sometimes, and guidelines for applying the procedures. The estimation of joint probabilities of the stream peak flow of the tributary streams is the key task in the research. The scope of this research is limited to riverine areas and does not include coastal areas.

A whole procedure for the design coincident flows at stream confluences is introduced first, which comprises of the following four steps, the identification of the each of the tributary using the USGS gauge station data, the estimation of the joint probability of the two tributary flows based on the identified marginal annual peak flow distributions of the two tributaries, the synthesis of the confluence flows based on the joint probability, and the univariate flood frequency analysis based on the synthetic flows at the confluence. Then two case studies in Iowa and Georgia, respectively are conducted to demonstrate the proposed approach.

Due to the difficulty identifying the joint probability, a simply method is needed. The copula method is introduced and the application procedure is addressed. Two case studies are also presented for the demonstration.

1.5 Format and content

This thesis is organized as follows. Chapter 2 presents the practical procedures of estimating the flood coincidence of the flood at the confluence. Chapter 3 presents the concepts and application of copula method for the joint probability estimation, which is the key task in the proposed joint probability approach for the estimation of confluence flood analysis. Chapter 4 summarizes the work presented in this thesis and outlines the opportunities for the future work beyond the scope of this thesis.

References

Ashkar, F., and Rousselle, J. (1982), A multivariate statistical analysis of flood magnitude, duration and volume, in *Statistical analysis of rainfall and runoff*, edited by V. P. Singh, pp. 659–669, Water Resource publication, Fort Collins, Colo..

Ashkar, F., El-Jabi, N., and 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. and Kottegoda, N. T. (1994), Bivariate exponential model applied to intensities and durations of extreme rainfall, *Journal of hydrology,* **155,** 225-236.

Beersma, J. J. and Buishand, T. A. (2004), The joint probability of rainfall and runoff d eficits in the Netherlands. Proceedings, EWRI 2004 World Water & Environmental Resources Congress 2004, June 27 - July 1, 2004, Salt Lake City, USA (CD-ROM).

Blachnell, D. (1994), New method of the simulation of correlated K-distribution clutter, *EE Proceedings Radar, Sonar and Navigation,***141**, 53-58.

Buishand TA. (1984), Bivariate extreme-value data and the station-year method, *Journal of Hydrology*, **69**, 77–95.

Burkhardt, G. and Prakash, A. (1976), An analysis of computer graphics to extremes, *Water Research Bulletin*, **12**, 1245-1258.

Chang, C. H., Tung, Y. K., and Yang, J. C. (1994), Monte Carlo simulation for correlated variables with marginal distributions, *Journal of Hydraulics Engineering*, **120,** 313-331.

Choulakain, V., Ei-Jabi, N., and Moussi, J. (1990), On the distribution of flood volume in partial duration series analysis of flood phenomena. *Stochastic Hydrology and Hydraulics,* **4,** 217-226.

Chow, V. T., Maidment, D. R. and Mays, L. W., (1988), *Applied hydrology*. McGraw-Hill, New York.

Clarke, R. T.(1980), Bivariate gamma distributions for extending annual streamflow records from precipitation: some large-sample results, *Water Resources Research*, **16**, 863–870.

Crovelli, R. A. (1973), A bivariate precipitation model, paper presented at 3rd Conference on Probability and Statistics in Atmospheric Science, American Meteorological Society, Boulder, CO, USA; 130–134.

Durrans, S. R., Eiffe, M. A., Thomas, Jr. W. O., and Goranflo, H. M. (2003), Joint seasonal/ annual flood frequency analysis, *Journal of Hydrologic Engineering*, **8**, 181-189.

Goel NK, Seth SM, Chandra S. (1998), Multivariate modeling of flood flows, *Journal of Hydraulic Engineering*, **124,** 146–155.

Hashino, M. (1985), Formulation of the joint return period of two hydrologic variates associated with a Possion process, *Journal of Hydroscience and Hydranlic Engineering*, **3,** 73-84.

Johnson, R. A., Evans, J. W., and Green, D. W., (1999), Some Bivariate Distributions for Modeling the Strength Properties of Lumber, United States Department of Agriculture , Forest Service , Forest Products Laboratory Research Paper FPL.RP.575

Kite, G. W., and Stuart, A. (1977), Frequency and risk analysis in hydrology. *Water Resources public.*Fort Collins, Co.

Krstanovic, P. F., and Singh, V. P. (1987), A multivariate stochastic flood analysis using entropy, in *Hydrologic Frequency Modelling*, edited by V. P. Singh, PP. 515–539, Reidel, Dordrecht, The Netherlands.

Linsley, R. K. (1986), Flood estimates: how good are they? *Water Resources Research*, **22,** 1595-1645.

Michael, J. R. and Schucany, W. R., (2002), The mixture approach for simulating bivariate distributions with specified correlations. *The American statistician*, **56**, 48-54.

Moharran, S. H., Gosain, A. K., and Kapoor, P. N. (1993), A comparative study for the estimators of the generalized pareto distribution. *Journal of hydraulics*, **150,** 169-185.

Molenberghs, G. and Lesaffre, E., (1997), Non-linear integral equations to approximate bivariate densities with given marginals and dependence function, *Statistica Sinica*, **7**, 713-738.

Moran, PAP. (1970), Simulation and evaluation of complex water systems operations. *Water Resource Research*, **6**, 1737–1742.

Nadarajah, S. and Gupta, A. K., (2006), Intensity–duration models based on bivariate gamma distributions, *Hiroshima mathematical journal*, **36**, 387–395.

Natural Environmental Research Council (1975), Flood Studies Report, Volume I: Hydrological Studies, report, Natural Environmental Research Council, London.

Oklin, I. and Harshall, A. (1988), Families of multivariate distributions, *Journal of American Statistical association*, **83**, 834-841.

Prekopa A, Szantai T. (1978), A new multivariate gamma distribution and its fitting to empirical streamflow data. *Water Resource Research*, **14**, 19–24.

Rao, A. R. and Hamed, K. H., (2000a), *Flood frequency analysis.* CRC Press, New York, NY, pp. 8.

Rao, A. R. and Hamed, K. H., (2000b), *Flood frequency analysis.* CRC Press, New York, NY, pp. 10.

Ronning, G. (1977), A Simple Scheme for Generating Multivariate Gamma Distributions with Non-Negative Covariance Matrix, *Technometrics*, **19**, 179-183.

Rossi, F. M. F. and Versace, P. (1984), Two-component extreme value distribution for flood frequency analysis. *Water Resource Research*, **20,** 847-856.

Sackl, B., and Bergmann, H. (1987), A bivariate flood model and its application, in *Hydrologic frequency modeling*, edited by V. P. Singh, pp. 571–582, Dreidel, Dordrecht, The Netherlands.

Shiau, J., Feng, S., and Nadarajah, S. (2007), Assessment of hydrological droughts for the Yellow River, China, using coplus, *Hydrological Processes*, **21**, 2157-2163.

Singh, K., Singh, V. P. (1991), Derivation of bivariate probability density functions with exponential marginals. Stochastic *Hydrology and Hydraulic*s, **5**, 55–68.

Singh, V. P. and Singh, K. (1985a), Derivation of the Pearson Type (PT) III distribution by using the principle of maximum entropy (POME), *Journal of Hydraulics*, **80,** 197-214.

Singh, V. P. and Singh, K. (1985b), Derivation of gamma distribution by using the principle of maximum entropy. *Water Resources Bulletin,* **21,** 1185-1191.

Tod, D. K.(1957), Frequency analysis of streamflow data, *Journal of the Hydraulics Division*, Process ASCE HY1, pp. 1166-1, 1166-16.

U.S. Geological Survey (2002), The National Flood Frequency Program, Version 3: A Computer Program for Estimating Magnitude and Frequency of Floods for Ungaged Sites. U.S. Geological Survey Water-Resources Investgations Report 02-4168

Wang, Q. J. (2001). A Bayesian Joint probability approach for flood record augmentation. *Water Resources Research,* **37,** 1707-1712.

Yue S. (1999). Applying the bivariate normal distribution to flood frequency analysis. *Water International,* **24,** 248–252.

Yue, S. (2001a). The bivariate lognormal distribution to model a multivariate flood episode. *Hydrological Processes*, **14,** 2575-2588.

Yue, S. (2001b), A bivariate extreme value distribution applied to flood frequency analysis, *Nordic Hydrology*, **32**, 49–64.

Yue, S. (2001c), A bivariate gamma distribution for use in multivariate flood frequency analysis, *Hydrological Process*, **15**, 1033-1045.

Yue, S. and Rasmussen, P. (2002), Bivariate frequency analysis: discussion of some useful concepts in hydrological application, *Hydrological Process*, **16**, 2881–2898.

Yue, S., Ouarda, T. B. M. J., and Bobée, B. (2001), A review of bivariate gamma distributions for hydrological application, *Journal of Hydrolgoy*, **246**, 1-18.

Yue, S. and Wang, C. Y. (2004), A comparison of two bivariate extreme value distributions, *Stochastic Environmental Resea*, **16**, 61-66.

Zhang, L. and Singh, V. P. (2006), Bivariate flood frequency analysis using the copula method, *Journal of hydrologic engineering*, **11**, 150-164.

Chapter 2 A Joint Probability Approach for Confluence Flood Frequency Analysis

Abstract

This paper presents a practical procedure for the flood frequency analysis at the confluence of two streams based on the flow rate data from the upstream tributaries. Four steps are involved in the approach, the distribution identification of annual stream peak flow of the tributary streams, the identification of joint probability distribution of the tributary stream flows, the generation of the synthetic stream flow at the confluent point by using Monte Carlo simulation, and identification of the flood frequency of the confluent point by the univariate flood frequency analysis. Two case studies are conducted and the results are compared with the flood frequency obtained by the univariate flood frequency analysis based on the observation data, and with the ones by National Flood Frequency Program developed by United State Geological Survey. It shows that the results by the proposed approach are much closer to flood estimated by the univariate flood frequency analysis based on the observation data than the results by the national flood frequency program, especially when the return period is less than 50 or 100 years.

Keywords: Flood frequency analysis, goodness-of-fit, Chi-square test, Kolmogorov-Smirnov test, joint probability, Monte Carlo simulation, confluence point

2.1 Introduction

The flood frequency analysis at or nearby the confluence of two tributaries is of interest because it is necessary for the design of the highway drainage structures, which often are located near the confluence point and may be subject to inundation by high flows from either stream or both. These infrastructures are designed to meet specified performance objectives for floods of a specified return period (e.g., the 100-year flood). The shortage of the hydrological data of the confluence point which are necessary to the univariate flood frequency analysis makes the flood estimation at the confluence challenging. An accurate and practical approach for the flood frequency estimation for this situation is needed.

To estimate the flood without discharge records, the flow routing may be performed which usually involves complicated numerical scheme and tedious of computation. Currently, the National Flood Frequency Program (NFF) (US Geology Survey, 2002) developed by US Geology Survey (USGS) based on the regional analysis probability is probably the most popular method for the ungauged site flood estimation, and could be employed for the flood estimate at the confluence. Although many researchers have proposed many regional flood analysis approaches, in NFF model all the states in US are divided into multiple hydrologic regions by using major watershed boundary and/or some other hydrologic characteristics, i. e., the mean elevation of watershed. It is assumed that the hydrologic characteristics are homogeneous in each region so that the flood at the ungauged sites can be estimated by the gauged sites. A series of regression equations of T-year flood (T=2 , 5, 10, 25, 50, 100, 200 and 500 year) associated with each hydrologic region are developed in terms of hydrologic characteristics based on the gauged site records. All the sites in each region share the same regression equation for the flood estimation associated with a specified return period. However, some equations in this approach have high errors; for example, some equations generate standard errors greater than 100 percent for the western part of the US, although the average standard error of NFF is between 30 and 60 percent (USGS, 2002).

He et al. (2007) derived a time coefficient of flood discharge model and a kinetic wave routing model based on the flood events on a long cycle to evaluate the flood behaviors at a confluence of the middle Yellow River in China by considering the flood frequency, intensity and duration. This model requires relative detail historic flood events information of the river basin which is unavailable sometimes.

Because the flooding of structures on one stream could be affected by high flows on the other stream, it is important to know the relationship between the coincident exceedence probabilities on the confluent stream pair (i.e., the joint probability of the coincident flows). It is reasonable to assume that an accurate flood estimation approach may be developed based on the joint probability of the coincident flows of the tributary streams. In the proposed approach in this search, accurate estimates of the joint probability of design flows at stream confluences are a crucial element in the design of efficient and effective highway drainage structures. With the recognition that the complex hydrological events such as floods are always affected by one or more correlated events and that an accurate estimate of the joint probability of the correlated events plays an important role for hydrology analysis, much attention has been paid on the bivariate and even multivariate flood frequency analysis since 1980s.

The research on bivariate distribution has been of interest of statisticians for a long time and many methods have been proposed to derive the joint distribution functions with the same or different margins (Molenberghs and Lesaffre, (1997); Marshall and Olkin, 1988; Schucany and Michael, 2002; Blachnell, 1994; Ronning, 1977).Sackl and Bergmann (1987), Chang et al. (1994), Yue (1999), and Beersma and Buishand (2004) used the bivariate normal distribution to perform the flood frequency analysis and hydrology events analysis. Krstanovic and Singh (1987) derived the multivariate Gaussian and exponential distributions by the principle of maximum entropy and applied the bivariate distributions for the analysis of flood peak and volume. Goel et al. (1998) employed a multivariate normal distribution to perform flood frequency analysis after normalizing the data of flood peak, volume and duration. Hashino (1985), Choulakian et al. (1990), Singh and Singh (1991), Bacchi et al. (1994), and Ashkar et al. (1998) investigated and applied the bivariate exponential distributions for the hydrological events analysis. Buishand (1984), Raynal and Salas (1987), Yue (2001a) applied bivariate extreme value distributions to analysis multivariate flood/storm events. Yue and Wang (2004) compared the performance in flood analysis between two bivariate extreme value distributions, the Gumbel mixed model and the Gumbel logistic model. Many researchers used bivariate gamma distribution for the flood frequency analysis (Moran, 1970; Prekopa and Szantai, 1978; Clarke, 1980; Yue, 2001b). Among them, Yue (2001b) investigated the applicability of the bivariate gamma distribution model to analyze the joint distribution of two positively correlated random variables with gamma marginals. Yue et al (2001) reviewed three bivariate gamma distribution models with two gamma marginal distributions. Durrans et al (2003) presented two approximate methods for joint frequency analysis using Pearson Type III distribution to estimate the joint flood frequency analyses on seasonal and annual bases. Nadarajah and Gupta (2006) developed exact distribution of intensity-duration based on bivariate gamma distribution. Shiau et al (2007) derived a joint probability distribution with a mixture of exponential and gamma marginal distribution to simulate the relationship between drought duration and drought severity. Wang (2001) developed a procedure for record augmentation of annual maximum floods by applying the bivariate extreme value distribution for annual maximum floods at to gauging stations with generalized extreme value distribution. Yue and Rasmussen (2002) discussed the concepts of bivariate hydrology events and demonstrated the concepts by applying a bivariate extreme value distribution to represent the joint distribution of flood peak and volume from an actual basin. Johnson et al. (1999) reviewed the some techniques for obtaining bivariate distributions and presented the properties of some bivariate models that include bivariate Weibull distribution, bivariate inverse Gaussian distribution, bivariate S_{BB} distribution and bivariate normal-lognormal distribution.

Although many of above researchers performed flood frequency analysis with the joint probability approach, most of them focused more on the determination of the relationship of different hydrologic variables, such as flood peak and volume, and flood volume and duration, where the flow discharge records of the site of interest are usually required. No one has applied the joint probability approach for the flood estimation at the ungauged sites, especially ungauged confluence point of the tributaries. A joint probability approach is presented in this paper to estimate the flood and frequency at the confluence of the tributaries without the requirement of records of the studied site.

2.2 Methodology

2.2.1 The procedure of the approach

Four steps are involved in the approach, stream flow distribution identification of the tributary streams, identification of joint probability distribution of the tributary stream flows, identification of the synthetic stream flow at the confluent point by using Monte Carlo simulation, and identification of the flood frequency of the confluent point by the conventional flood frequency analysis. The flow chart for the procedure is seen in Figure 2-1.

Step 1. Stream flow distribution identification of the tributary streams

In the step, the historical annual stream peak flow data of the two tributary streams are collected first, the parameters associated with the assumed distributions are estimated by method of moment, method of maximum likelihood, or method of probability weighted moments, and then the test of goodness-of-fit is performed to identify the annual stream peak flow distributions of the two tributary streams. Chi-square test and Kolmogorov-Smirnov (K-S) are used in this step.

Step 2. Identification of joint probability distribution of tributary stream flows

In this step, the correlationship of the annual stream peak flow data of the two tributaries is identified first by calculating the correlation coefficient, and then the joint probability distribution of the tributary stream flow is identified based on the annual peak flow distributions of the two tributary streams identified in the first step and the correlationship of the annual peak flow data of the tributary streams. If the correlationship is small enough, say, less than 0.2, it is reasonable to assume that the two set of data are independent, in other words, the annual peak flow of the two tributary streams are independent. In this simplified case, the joint probability distribution of the stream annual peak flow of the two tributary streams is simply the multiplication of the annual peak flow distributions of the tributary streams. Otherwise the joint probability needs to be estimated by an appropriate method, such as well established empirical bivariate distributions equations. The conditional annual peak flow distribution is also identified in this step based on which the Monte Carlo simulation will be performed in the next step.

Step 3. Monte Carlo simulation

In this step, Monte Carlo simulation is performed to obtain the synthetic annual peak flow of the two tributary streams, based on the annual peak flow distributions of the tributary streams and the conditional annual peak flow distribution. The synthetic annual peak flow at the confluence point is assumed to be the summation of the annual peak flow and the two tributary steams.

Step 4. Conventional flood frequency analysis

In this step, the distribution of the synthetic annual stream peak flow is identified by the test of goodness-of-fit first, and then the peak flows corresponding to specified return periods are calculated by using frequency factors or inverse method, based on the synthetic annual peak flow at the confluence in the previous step.

2.2.2 Distribution identification of the tributary streams

The distribution identification of the tributary streams involves parameter estimation and goodness-of-fit test.

2.2.2.1 Parameter estimation

There are many methods to estimate the parameters of a distribution; however, the

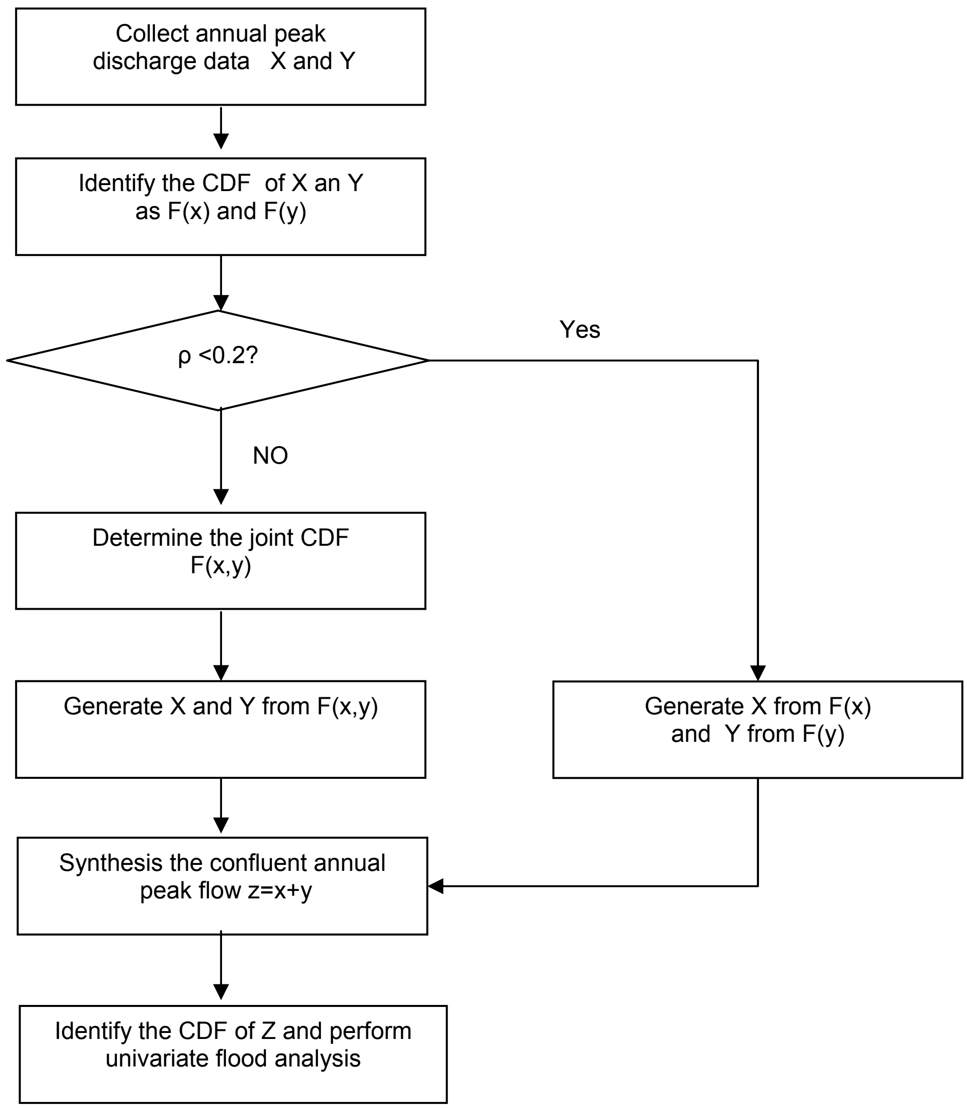

Figure 2-1 Flow chart of the procedure of proposed approach

three most often used methods are the method of moments (MOM), the method of maximum likelihood (ML) and the probability weighted moments method (PWM). The advantages and disadvantages of the three methods are addressed by Rao and Hamed (2000) as follows,

"The maximum likelihood method (ML method) is considered the most efficient method since it provides the smallest sampling variance of the estimated parameters, and hence of the estimated quantiles, compared to other methods. However, for some particular cases, such as the Pearson type III distribution, the optimality of the ML method is only asymptotic and small sample estimates may lead to estimates of inferior quality (Bobee and Ashkar, 1991). Also the ML method has the disadvantage of frequently giving biased estimates, but these biases can be corrected. Furthermore, it may not be possible to get ML estimates with small samples, especially if the number of parameters is large. The ML method requires higher computational efforts, but with the increased use of high-speed personal computers, this is no longer a significant problem.

The method of moments (MOM) is a natural and relatively easy parameter estimation method. However, MOM estimates are usually inferior in quality and generally are not as efficient as the ML estimates, especially for distributions with large number of parameters (three or more), because higher order moments are more likely to be highly biased in relatively small samples.

The PWM method (Greenwood et al,. 1979; Hosking, 1986) gives parameter estimates comparable to the ML estimates, yet in some cases the estimation procedures are much less complicated and the computations are simpler. Parameter estimates from small samples using PWM are sometimes more accurate than the ML estimates (Landwehr et al., 1979). Also, in some cases, such as the symmetric lambda and Weibull distributions, explicit expressions for the parameters can be obtained by using PWM, which is not the case with the ML or MOM methods."

For the convenience of application, the often used distributions and the associated parameter are listed in Appendix A.

2.2.2.2 Test of goodness-of-fit

The choice of distribution to be used in flood frequency analysis has been a topic of interest for a long time (Rao and Hamed, 2000). When a theoretical distribution has been assumed, the validity of the assumed distribution may be verified or disproved statistically by goodness-of-test (Ang and Tang, 1975a). Chi-square test and Kolmogorov-Smirnov (K-S) test have been typically used to identify the stream flow distributions for flood frequency analysis.

Chi-square test

In Chi-square test, the observed values of the relative frequency or the cumulative frequency function are compared with the corresponding value of the assumed theoretical distribution to test the goodness of fit of a probability. In the test, the data are divided into k class intervals (k is recommended to be more than 5). The statistic Chi-square (χ^2) is given by

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \qquad \textbf{Eq. 2.1}$$

where O_i is the observed number of events in the class interval i, E_i is the number of events that would be expected from the summed theoretical distribution and k is an arbitrary number of classes to which the observed data are divided. The above equation can also be written as follows,

$$\chi^2 = \sum_{i=1}^{k} \frac{n[(f_s(x_i) - p(x_i)]^2}{p(x_i)} \qquad \textbf{Eq. 2.2}$$

where n is total number of observations, $f_s(x_i)$ is the observation relative frequency function, which is defined as $f_s(x_i) = n_i / n$ where n_i is the number of observations in interval i, and

$p(x_i)$ is the incremental probability function, which is defined as $p(x_i) = F(x_i) - F(x_{i-1})$, where $F(x_i)$ is the cumulative probability $P(X \le x_i)$.

If $\chi^2 < c_{1-\alpha,f}$, where $c_{1-\alpha,f}$ is the value of the Chi-square distribution with f degree of freedom, at the cumulative probability $1-\alpha$, the assumed theoretical distribution is accepted at the significance level of α. Otherwise, the null hypothesis that the assumed distribution fits the data adequately is rejected at the significance level of α. A typical value for the significance level is 0.05. The values of $c_{1-\alpha,f}$ can be looked up in most statistics textbooks.

Kolmogorov-Smirnov test

Kolmogorov-Smirnov (K-S) test is another widely used goodness-of-fit besides Chi-square test. K-S test is based on the deviation of the sample distribution function from the specified continuous hypothetical distribution function, providing a comparison of a fitted distribution with the empirical distribution. The test statistic is the maximum vertical distance between the empirical and hypothetical cumulative distribution function (CDF), which is defined as follows,

$$D_n = \sup_x |F(x) - S_n(x)| \qquad \textbf{Eq. 2.3}$$

where $F(x)$ are the estimated values by the proposed theoretical distribution, and $S_n(x)$ is denoted by

$$\mathrm{S_n}(x) = \begin{cases} 0 & x < x_1 \\ k/n & x_k < x < x_{k+1} \\ 1 & x \ge x_n \end{cases} \qquad \textbf{Eq. 2.4}$$

where $x_{1,}\ x_{2,}\ ...\ x_k\ ...\ x_n$ are the values of the increasingly ordered sample data and n is the sample size.

Theoretically, D_n is a random variable whose distribution depends on n. In K-S test, the value of D_n must be less than the critical value D_n^α at a specified significance level α in

order to accept the proposed distribution at the specified significance level α ; otherwise, the assumed distribution would be rejected at the specified significance level α .
For larger sample sizes (n>30), the approximate critical value D_n^α is expressed by the equation $D_n^\alpha = \frac{c(\alpha)}{\sqrt{n}}$, where c(α) is the coefficient associated with the significant level α , which is given by Table 2-1.

Table 2-1 Critical D value for K-S test

α	0.10	0.05	0.25	0.01	0.005	0.001
Critical value D_n^α	1.22	1..36	1.48	1.63	1.73	1.95

The advantage of the K-S tests over the Chi-square test it that it is not necessary to divide the data into bins; hence the problems associated with the chi-square approximation for small number of intervals would not appear with the K-S test (Ang and Tang, 1975a).

In the proposed approach of identification of distribution, the assumed distribution is required to be appraised by both chi-square test and K-S test.

2.2.3. Joint probability identification

Many researchers have dealt with bivariate flood frequency analysis (Kite 1978; Zhang 2006, Durrans 2003;Yue,1999, 2001,2000,2001a,2001b). In the statistical literature, a few bivariate or multivariate distribution models have been developed and studied (Gumbel and Mustafi, 1967; Buishand, 1984). Unfortunately, there are currently no well established general methods to derive the joint probability from the marginal distributions directly. However, some empirical formulas for the joint distributions with certain specified margins work well. Some often used joint CDFs and/or joint PDFs are listed in this paper, which include bivariate normal distribution, bivariate exponential distribution, bevariate gamma distribution, bivariate extreme value distribution, and so on.

The bivariate normal PDF have been well developed and used in many areas for a long time, which is given by

$$f(x,y)=\frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}\exp\{-\frac{1}{2(1-\rho^2)}[(\frac{x-\mu_x}{\sigma_x})^2-2\rho(\frac{x-\mu_x}{\sigma_x})(\frac{y-\mu_y}{\sigma_y})+(\frac{y-\mu_y}{\sigma_y})^2]\}\quad -\infty<x<\infty;-\infty<y<\infty \qquad \textbf{Eq. 2.5}$$

where X and Y are two variables; σ_x and σ_y are the standard deviation of the sample data of variable X and Y, respectively; μ_x and μ_y are the mean of sample data of X and Y, respectively; ρ refers to the correlation coefficient of X and Y

Stuart and Ord (1987) presented the joint PDF and cumulative density function (CDF) for two exponential distributions. Given two exponential distribution

$$F_X(x)=1-e^{-ax},\ x\geq 0, a>0;\ F_Y(y)=1-e^{-by},\ y\geq 0, b>0 \qquad \textbf{Eq. 2.6}$$

$$f_{x,y}(x,y)=[(a+cy)(b+cx)-c]e^{-ax-by-cxy} \qquad \textbf{Eq. 2.7}$$

$$F_{X,Y}(x,y)=1-e^{-ax}-e^{-by}+e^{-ax-by-cxy} \qquad \textbf{Eq. 2.8}$$

where a and b are the parameters of the exponential distribution of variable X and Y, respectively; c denotes a parameter describing the joint variable of the variates, which is related to the correlation coefficient of X and Y, $0\leq c\leq ab$

Smith et al. (1982) pointed out the joint PDF and CDF of two positively correlated random variables X and Y with gamma marginal distributions as follows:

$$f(x,y)=\begin{cases}\frac{K_1}{K_2}\sum_{j=0}^{\infty}\sum_{k=0}^{\infty}c_{jk}(\beta_x x)^j(\eta\beta_y y)^{j+k} & \text{if } \rho>0\\ f_X(x)f_Y(y) & \text{if } \rho=0\end{cases} \qquad \textbf{Eq. 2.9}$$

$$F_{X,Y}(x,y)=\begin{cases}J\sum_{j=0}^{\infty}\sum_{k=0}^{\infty}d_{jk}H(\gamma_x+j,\frac{x}{1-\eta})H(\gamma_y+j+k,\frac{y}{1-\eta}) & \text{if } \rho>0;\\ F_X(x)F_Y(y) & \text{if } \rho=0\end{cases} \qquad \textbf{Eq. 2.10}$$

where

$$K_1=(\beta_x x)^{\gamma_x-1}(\beta_y y)^{\gamma_y-1}\exp(-\frac{\beta_x x+\beta_y y}{1-\eta}) \qquad \textbf{Eq. 2.11}$$

$$K_2=(1-\eta)^{\gamma_x}\Gamma(\gamma_x)\Gamma(\gamma_y-\gamma_x) \qquad \textbf{Eq. 2.12}$$

$$c_{jk} = \frac{\eta^{j+k}\Gamma(\gamma_y - \gamma_x + k)}{(1-\eta)^{2j+k}\Gamma(\gamma_y + j + k)\, j!k!} \quad \textbf{Eq. 2.13}$$

$$\eta = \rho\sqrt{\frac{\gamma_y}{\gamma_x}} \quad \textbf{Eq. 2.14}$$

$$J = \frac{(1-\eta)^{\gamma_y}}{\Gamma(\gamma_x)\Gamma(\gamma_y - \gamma_x)} \quad \textbf{Eq. 2.15}$$

$$d_{jk} = \frac{\eta^{j+k}\Gamma(\gamma_y - \gamma_x + k)}{\Gamma(\gamma_y + j + k)\, j!k!} \quad \textbf{Eq. 2.16}$$

$$H(a,z) = \int_0^z t^{a-1} e^{-t} dt \quad \textbf{Eq. 2.17}$$

$$\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt \quad \textbf{Eq. 2.18}$$

$$k! = k(k-1)(k-2)\cdots 3 \bullet 2 \bullet 1 \quad \textbf{Eq. 2.19}$$

where H is the incomplete gamma function, $\Gamma(\cdot)$ is the gamma function, η is the association parameter between X and Y, and ρ is the product-moment correlation coefficient of X and Y and is estimated from the sample data:

$$\rho = \frac{E[(x - M_x)(Y - M_y)}{S_x S_y} \quad \textbf{Eq. 2.20}$$

in which (M_x, S_x) and (M_y, S_y) are the sample mean add standard deviations of he variable X and Y respectively,. (β_x, γ_x) and (β_y, γ_y) are the scale and shape parameters of the single-variable gamma distributions of X and Y respectively. The PDF, $f_X(x)$ and $f_Y(y)$ of the marginal distributions of X and Y are respectively given as follow:

$$f_X(x) = \frac{1}{\Gamma(\gamma_x)} x^{\gamma_x - 1} \beta_x^{\gamma_x} e^{-\beta_x x} \quad \textbf{Eq. 2.21}$$

$$f_Y(y) = \frac{1}{\Gamma(\gamma_y)} y^{\gamma_y - 1} \beta_y^{\gamma_y} e^{-\beta_y y} \quad \textbf{Eq. 2.22}$$

Yue (2001b) applied the bivariate gamma distribution to perform the flood analysis for a river in Canada, with the data of flood volume, flood days and flood flow.

Wang (2001) derived a format of joint probability function for two extreme value distributions as follows. Given the generalized extreme value (GEV) distributions for variable X,

$$F_X(x) = \begin{cases} e\,\mathrm{xp}\{-[1-\kappa_x(x-\xi_x)/\alpha_x]^{1/\kappa_x} & \kappa_x \neq 0 \\ \exp[-(x-\xi_x)/\alpha_x] & \kappa_x = 0 \end{cases} \qquad \textbf{Eq. 2.23}$$

with a density function

$$f_X(x) = \begin{cases} F_X(x)(\alpha_x)^{-1}[1-\kappa_x(x-\xi_x)/\alpha_x]^{1/\kappa-1_x} & \kappa_x \neq 0 \\ F_X(x)(\alpha_x)^{-1} e^{-(x-\xi_x)/\alpha_x} & \kappa_x = 0 \end{cases} \qquad \textbf{Eq. 2.24}$$

where κ_x, ξ_x, and α_x are parameters. Similarly CDF and PDF for Y, then the joint PDF is derived as follows,

$$f_{X,Y}(x,y) = f_{U,V}(u,v)[(\alpha_x - \kappa_x(x-\xi_x)]^{-1}[(\alpha_y - \kappa_y(y-\xi_y)]^{-1} \qquad \textbf{Eq. 2.25}$$

Where

$$u = \begin{cases} -(\kappa_x)^{-1}\ln(1-\kappa_x(x-\xi_x)/\alpha_x] & \kappa_x \neq 0 \\ (x-\xi_x)/\alpha_x & \kappa_x = 0 \end{cases} \qquad \textbf{Eq. 2.26}$$

$$v = \begin{cases} -(\kappa_y)^{-1}\ln(1-\kappa_y(y-\xi_y)/\alpha_y] & \kappa_y \neq 0 \\ (y-\xi_y)/\alpha_y & \kappa_y = 0 \end{cases} \qquad \textbf{Eq. 2.27}$$

$$F_{U,V}(u,v) = \exp[-(e^{-mu} + e^{-mv})^{1/m}] \qquad \textbf{Eq. 2.28}$$

$$f_{U,V}(u,v) = F_{U,V}(u,v)e^{-m(u+v)}(e^{-mu} + e^{-mv})^{-2+1/m} \cdot [m-1+(e^{-mu} + e^{-mv})^{-2+1/m}] \qquad \textbf{Eq. 2.29}$$

(Gumbel and Mustafi, 1967; Johnson and Kotz, 1972)

where U and V are independent of each other when m=1 and completely dependent of each other when $m = \infty$. In general $\rho_{U,V} = 1 - m^2$, where $\rho_{U,V}$ is the correlation coefficient between U and V (Gumbel and Mustafi, 1967)

Papadimitriou et al. (2006) introduced an analytical framework for analyzing the arbitrarily correlated trivariate Weibull distribution in a very complicated formation. A joint probability function for three Weibull distributions was expressed in that paper.

2.2.4. Multivariate Monte Carlo simulation

Monte Carlo simulation is required for the problems involving random variables with assumed (or known) probability distributions (Ang and Tand, 1975b). The key task in Monte Carlo simulation procedure is to generate the appropriate values of the variables in accordance with the specified probability distribution. In the proposed confluence flood frequency analysis approach, the key task of Monte Carlo simulation is the generation of jointly distributed random numbers in accordance with the respectively prescribed joint probability distributions.

There are several commonly used methods of generating multivariate random numbers, conditional distribution approach, transformation approach, rejection approach, and Gibbs approach. Conditional distribution approach is to generate random numbers for a marginal distribution, and then to generate random numbers for a sequence of condition distributions. Another way to generate multivariate random number is to generate a vector of identically independent distribution variates, and then apply a transformation to yield a vector from the specified multivariate distribution. An example of this method for the random number generation of multivariate normal distribution was addressed by Gentle (1998). Gibbs method is an iterative method used to generate multivariate random numbers. The conditional distribution approach is adopted in this study since it is relatively easy to apply and the results by this approach seem better than those by the other approaches based on our observation.

The conditional approach reduces the problem of generating a multi-dimensional random vector into a series univariate generation problems (Johnson, 1987).

Let $X_1, X_2, ..., X_n$ be a set of n random variables. The joint PDF is

$$f_{X_1, X_2, ..., X_n}(x_1, x_2, ..., x_n) = f_{X_1}(x_1) f_{X_2|X_1}(x_2|x_1) \cdots f_{X_n|X_1, X_2 ... X_{n-1}}(x_n|x_1, x_2, ..., x_{n-1}) \quad \textbf{Eq. 2.30}$$

where $f_{X_i}(x_i)$ is the marginal PDF of X_i $(i \in [1,n])$, and $f_{X_i|X_1,X_2...X_{i-1}}(x_i|x_1, x_2, ..., x_{i-1})$ is the conditional PDF of X_i $(i \in [1,n])$ given $X_1 = x_1, X_2 = x_2, ..., X_{i-1} = x_{i-1}$. And the corresponding joint CDF is

$$F_{X_1, X_2, ..., X_i}(x_1, x_2, ..., x_i) = F_{X_1}(x_1) F_{X_2|X_1}(x_2|x_1) \cdots F_{X_i|X_1,X_2...X_{i-1}}(x_i|x_1, x_2, ..., x_{n-1}) \quad \textbf{Eq. 2.31}$$

where $F_{X_i}(x_i)$ is the marginal CDF of X_i $(i \in [1,n])$, and $F_{X_i|X_1,X_2...X_{i-1}}(x_i|x_1, x_2, ..., x_{i-1})$ is the conditional CDF of X_i $(i \in [1,n])$ given $X_1 = x_1, X_2 = x_2, ..., X_{i-1} = x_{i-1}$.

If the $X_1, X_2, ..., X_n$ random variables are statistically dependent, the conditional approach involves the following steps:

- A set of uniformly distributed random numbers between 0 and 1, $(u_1, u_2, ..., u_n)$ are generated.
- A value of x_1 is determined as $x_1 = F^{-1}{}_{X_1}(u_1)$.
- With the value of x_1 and the conditional CDF of X_2, $F_{X_2}(x_2|x_1)$, a value of x_2 may be determined from $x_2 = F^{-1}{}_{X_2}(x_2|x_1)$
- Similarly, the value of x_i can be determined from $x_i = F^{-1}{}_{X_i}(x_i|x_1, x_2, ..., x_{i-1})$

In the case that the $X_1, X_2, ..., X_n$ random variables are statistically independent, the random numbers for each variate can be generated separately and independently from the marginal PDF of each variate, $x_i = F^{-1}(x_i)$.

2.2.5. Univariate flood frequency analysis

The probability of non-exceedence $F(x_T)$ of an even for a specified return period T is defined as,

$$F(x_T) = 1 - p(x \geq x_T) = 1 - \frac{1}{T} \quad \textbf{Eq. 2.32}$$

Hence the flood of magnitude X_T for a given return period, can be solved

$$x_T = F^{-1}(1 - \frac{1}{T}) \quad \textbf{Eq. 2.33}$$

This is straight forward and the basis for estimating a flood. However, some probability distribution functions cannot be expressed directly in the inverse from the equation $F(x_T) = 1 - 1/T$. In this case, indirect methods or numerical methods are needed to estimate the flood at a specified return period corresponding to given value of probability F. Chow (1951) proposed the following equation to calculating x_T,

$$x_T = u_1^{'} + K_T \sqrt{\mu_2} \quad \textbf{Eq. 2.34}$$

where $u_1^{'}$ is the sample mean from the observation data, μ_2 is the standard deviation, and K_T is called frequency factor which is a function of the return period and the parameters of the distribution. In this method, the parameters in the equation are calculated by the MOM. K_T and the equations in the direct method for the commonly used distributions in hydrology are listed in Appendix B..

2.2.6. Evaluation

Five numeric evaluation criteria may be used, which include:

Ratio of standard deviation of predicted to observed discharges: The ratio of standard deviation of predicted and observed discharges would indicate a better model as it approaches to 1.

$$CO = \sqrt{\frac{\sum\left(y_f - \bar{y}_f\right)^2}{\sum\left(y_o - \bar{y}_o\right)^2}} \quad \textbf{Eq. 2.35}$$

Root-mean-square error (RMSE): The RMSE would indicate a better model as the value approaches zero.

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} \text{S.E.}}{N}} \quad \textbf{Eq. 2.36}$$

Ratio of the mean error to the mean observed discharge: The ratio of the mean error to the mean observed discharge would indicate a better model when it approaches zero.

$$R = \frac{\sum (y_f - y_o)}{N \cdot \bar{y}_o}$$

Eq. 2.37

Square of the Pearson product moment correlation coefficient: The square of the Pearson product moment correlation coefficient would indicate a better model as it approaches 1.

$$r^2 = \left[\frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2}} \right]^2$$

Eq. 2.38

Mean of Percent Error (PE): The PE indicates a better model when its value approaches zero.

$$P.E. = \frac{\sum_N \left[\frac{(y_f - y_o)}{y_o} * 100 \right]}{N}$$

Eq. 2.39

where, yo and yf are the observed and forecasted discharges, respectively. N is the total number of data points involved. The bar above each parameter indicates the arithmetic mean.

2.3 Application examples

2.3.1 Case study 1

The proposed joint flood frequency analysis is applied to the Des Moines River basin in Iowa. The task is to estimate the flood at the site of USGS 05481300 (Station C) by the proposed joint probability approach, assuming there is no gauge station at this site. Still the data from the downstream station C are collected and the univariate flood analysis is performed for this site for the demonstration of the proposed approach.

Two upstream USGS gauge stations, USGS 05480500 (Station A) and USGS 05471000 (Station B), are selected, and 38 years (1968-2005) annual peak flow records of the two gauge stations are collected. The gate station names and locations are shown in Table 2-2 and Figure 2-2.

Table 2-2 USGS Gauge Stations located in the Des Moines River basin

Station	A	B	C
Station Name	Des Moines River at Fort Dodge, IA	Boone River near Webster City, IA	Des Moines River near Stratford, IA
USGU Station No.	05480500	05481000	05481300

For station A, by performing the two tests of goodness-of-fit, Chi-square test and K-S test, after the parameter estimation by maximum likelihood method, among the assumed distributions of exponential distribution, normal distribution, log-normal distribution, 3-parameter log-normal distribution, and 3-parameter gamma distribution, only 3-parameter log-normal and 3-parameter gamma distributions fit the data at a 0.05 of significant level. However, there is some difficulty in the parameter estimation using maximum likelihood method for the 3-paramter gamma distribution based on the data of Station A. The Newton-Raphson iteration cannot converge even after 50 iterations. Probability weighted moment method may need to be applied for the parameter estimation. Here we choose 3-parameter log- normal distribution as the distribution for station A. The probability plot for Station A, B and C are shown in Figure 2-3, Figure 2-4 and Figure 2-5.

Similarly, we can identify the distribution for station B and C. The goodness-of-fit test shows the 3-parameter log-normal and 3-parameter gamma distributions fit the data of station B and C. Here we pick up 3-parameter log-normal for Station B and both 3-parameter log-normal and 3-parameter gamma distributions for Station C for comparison. The

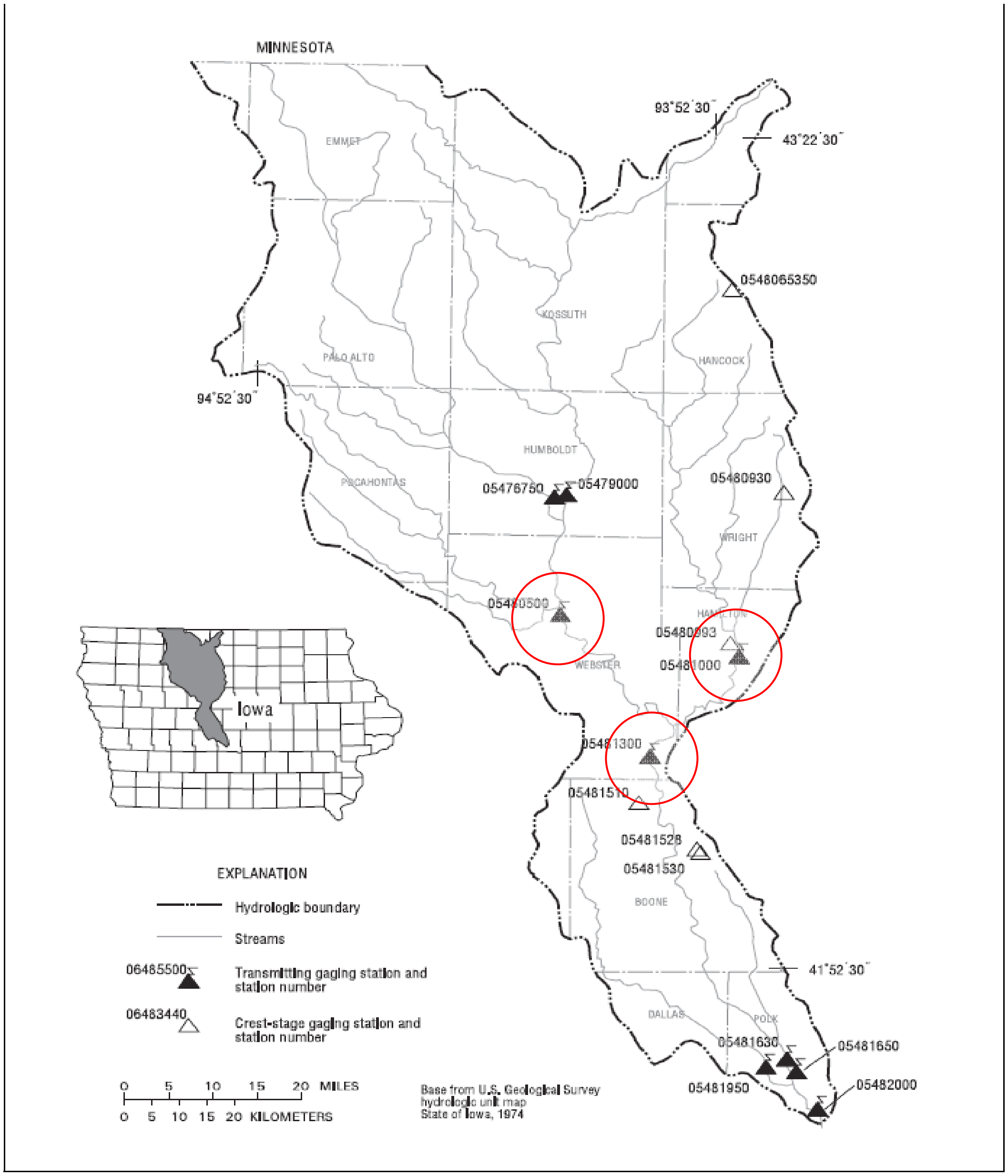

Figure 2-2   Location of USGS Gauge Stations in the Des Moines River Basin

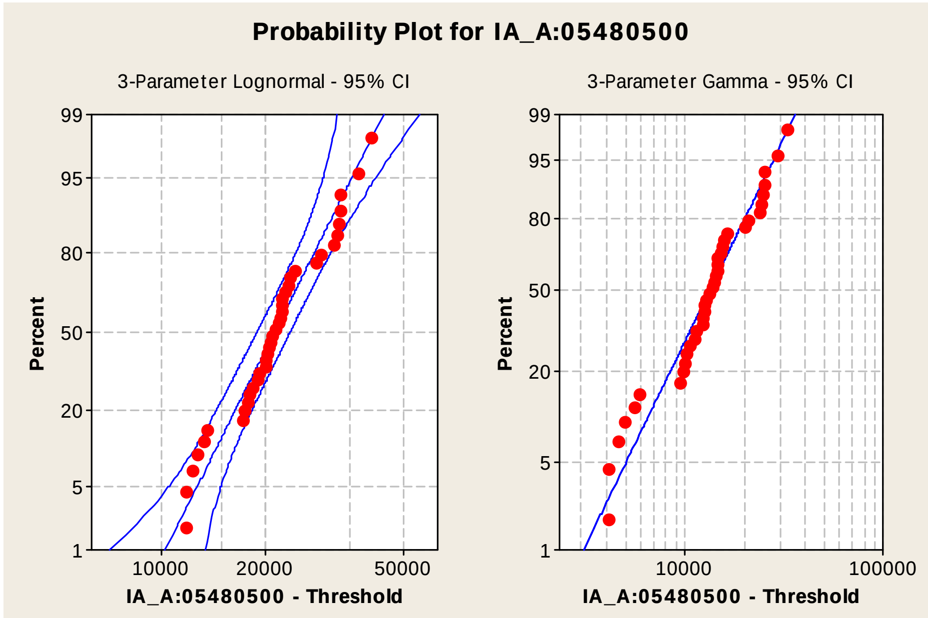

Figure 2-3 Probability plot for Station A in Des Moines River basin

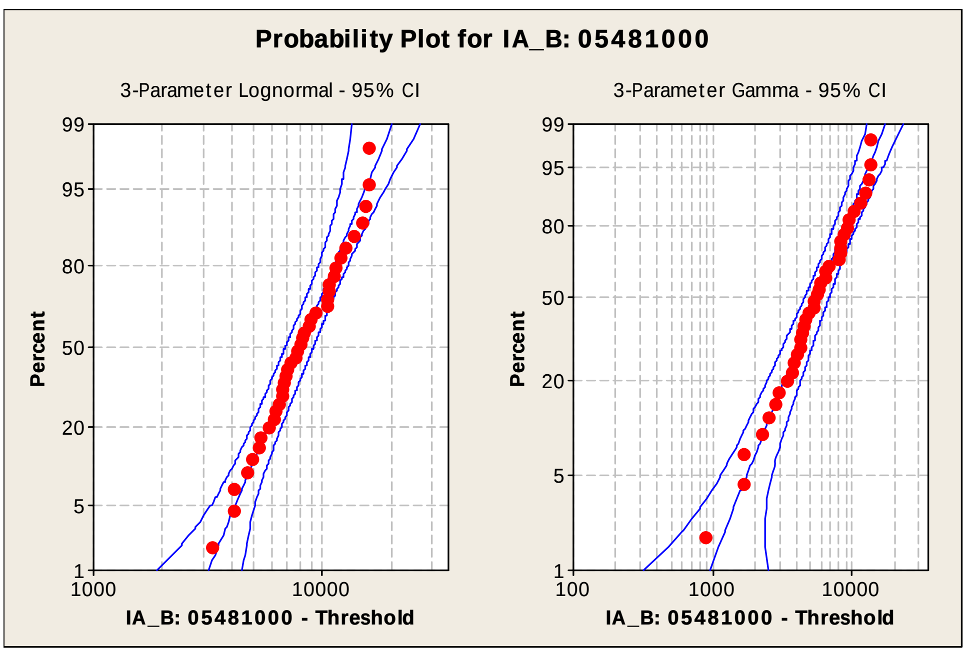

Figure 2-4 Probability plot for Station B in Des Moines river basin

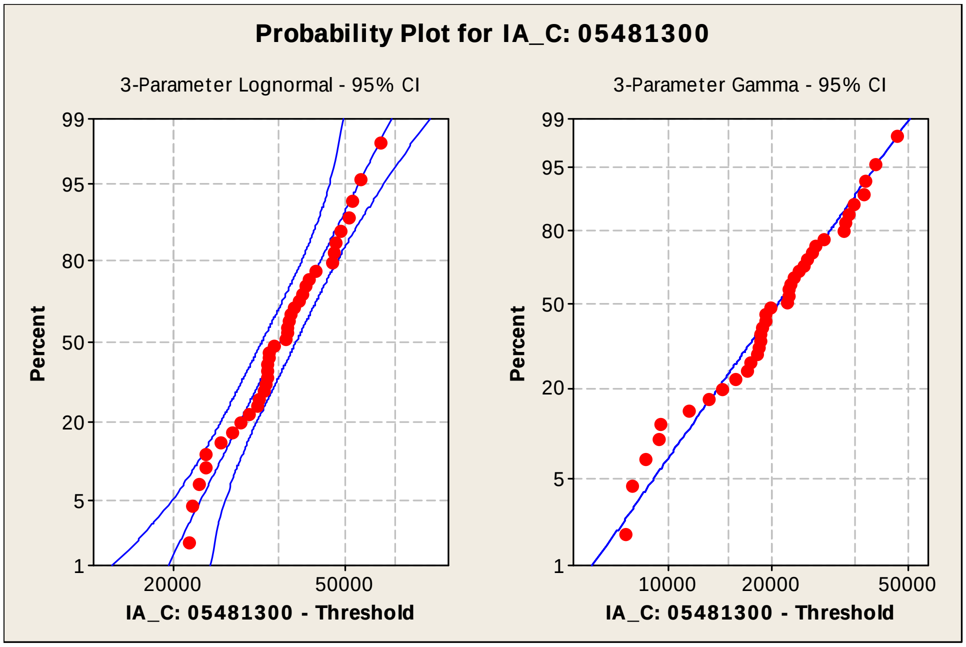

Figure 2-5 Probability plot for Station C in Des Moines river basin

Table 2-3 Gauge station data distribution information of Des Moines basin

<table>
<tr><td></td><td></td><td>Station A</td><td>Station B</td><td>Station C</td></tr>
<tr><td colspan="2">Data Mean</td><td>13087.1</td><td>6303.42</td><td>18564.7</td></tr>
<tr><td colspan="2">Data Standard deviation</td><td>7063.02</td><td>3432.53</td><td>9402.36</td></tr>
<tr><td colspan="2">Correlation coefficient of A and B</td><td colspan="2">0.804</td><td></td></tr>
<tr><td rowspan="3">3-parameter log-normal</td><td>Location</td><td>9.96384</td><td>8.98548</td><td>10.46668</td></tr>
<tr><td>Scale</td><td>0.31125</td><td>0.39657</td><td>0.25572</td></tr>
<tr><td>Threshold</td><td>-9203.6</td><td>-2323.1</td><td>-17719</td></tr>
<tr><td rowspan="3">3-parameter gamma</td><td>Shape</td><td rowspan="3"></td><td>3.09469</td><td>5.38697</td></tr>
<tr><td>Scale</td><td>2004.594</td><td>4109.859</td></tr>
<tr><td>Threshold</td><td>99.82</td><td>-3574.96</td></tr>
</table>

statistics and the parameters of the distributions for station A, B and C are shown in Table 2-3.

For the annual peak flow in Station A (random variable X), we use the following form of 3-paramter log-normal distribution.

PDF: $$f(y) = \frac{1}{\sqrt{2\pi}\sigma_y}\exp[-\frac{1}{2}(\frac{y-\mu_y}{\sigma_y})^2]$$ **Eq. 2.40**

CDF: $$F(y) = \frac{1}{\sqrt{2\pi}\sigma_y}\int_{-\infty}^{x}\int_{-\infty}^{y}\exp[-\frac{1}{2}(\frac{y-\mu_y}{\sigma_y})^2]dydx$$ **Eq. 2.41**

where $y = \ln(x-\gamma)$, μ_y, and σ_y are the mean and standard deviation of y, respectively, and γ is the threshold parameter.

For annual peak discharge data of Station C, 3-paramter gamma (Pearson III) distribution is adopted and the following format is used in this research,

PDF: $$f(x) = \frac{1}{\alpha^{\beta}\Gamma(\beta)}(x-\gamma)^{\beta-1}e^{-(x-\gamma)/\alpha} \quad x \geq \gamma$$ **Eq. 2.42**

CDF: $$F(x) = \frac{1}{\alpha^{\beta}\Gamma(\beta)}\int_{\gamma}^{x}(x-\gamma)^{\beta-1}e^{-(x-\gamma)/\alpha}dx \quad x \geq \gamma$$ **Eq. 2.43**

where α , β and γ are the shape parameter, scale parameter and threshold parameter, respectively.

The bivariate normal distribution can be expressed by,

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}\exp\{-\frac{1}{2(1-\rho^2)}[(\frac{x-\mu_x}{\sigma_x})^2 - 2\rho(\frac{x-\mu_x}{\sigma_x})(\frac{y-\mu_y}{\sigma_y}) + (\frac{y-\mu_y}{\sigma_y})^2]\} \qquad -\infty < x < \infty; -\infty < y < \infty$$ **Eq. 2.44**

where μ_x, σ_x μ_y, σ_y , and ρ the mean and standard deviation of x, the mean and standard deviation of y, and the correlation coefficient of X and Y, respectively.

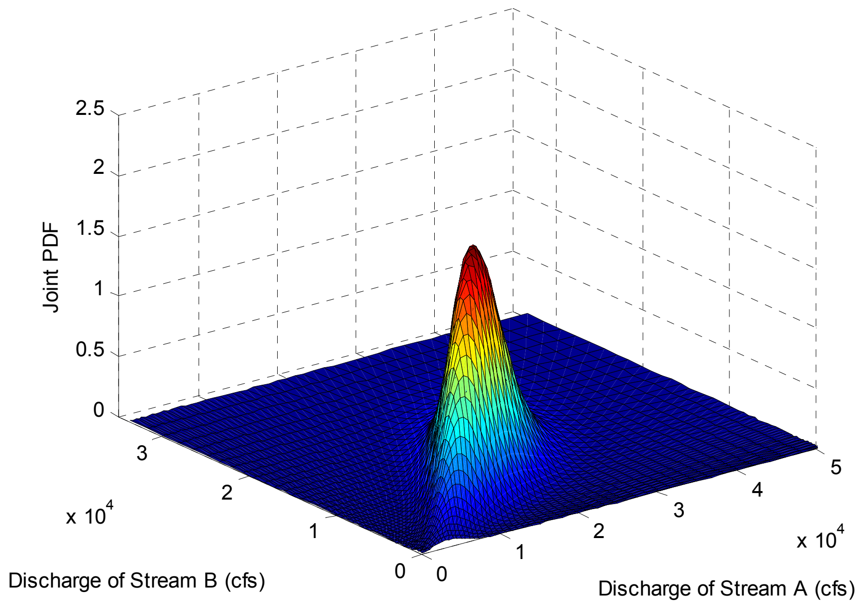

Figure 2-6 Joint PDF of the tributary discharge of Des Moines River basin

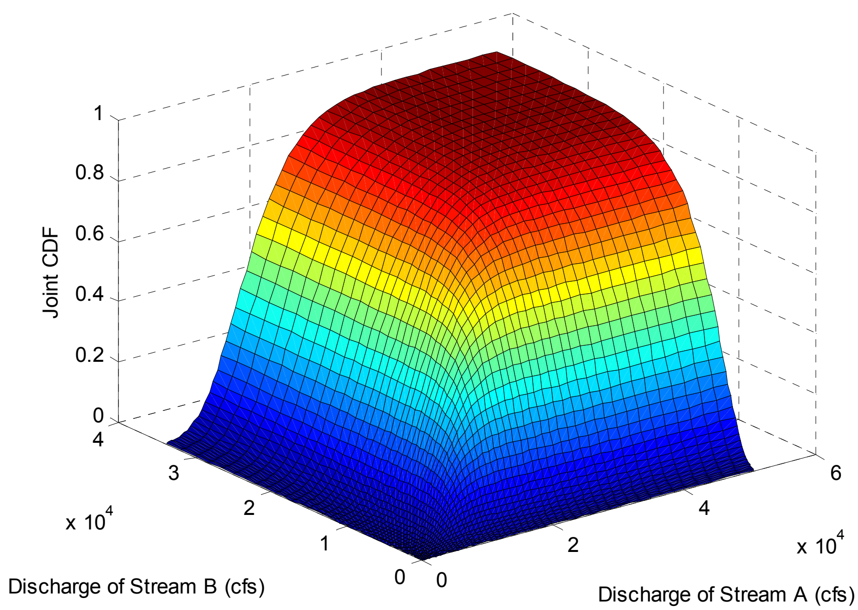

Figure 2-7 Joint CDF of the tributary discharge of Des Moines River basin

In the 3-parameter log-normal case, the following transformation can be performed in order to use the above equation for bivariate normal distribution. Introduce two random variables T and S, such that $s = \ln(x-\gamma_x)$ and $t = \ln(y-\gamma_y)$, where γ_x and γ_y donate the threshold parameters of the random variable X and Y, respectively. Therefore, the bivariate 3-parameter log-normal distribution can be expressed by the following equation,

$$f(s,t) = \frac{1}{2\pi\sigma_s\sigma_t\sqrt{1-\rho^2}}\exp\{-\frac{1}{2(1-\rho^2)}[(\frac{s-\mu_s}{\sigma_s})^2 - 2\rho(\frac{t-\mu_t}{\sigma_t})(\frac{s-\mu_s}{\sigma_s}) + (\frac{t-\mu_t}{\sigma_t})^2]\} \quad \textbf{Eq. 2.45}$$

where ρ is the correlation coefficient of S and T. The joint PDF and CDF of the annual peak flow of A and B are shown in Figure 2-6 and Figure 2-7.

After the joint probability distribution is identified, the Monte-Carlo simulation can be performed to obtain the synthetic annual peak flow at the confluence. The conditional distribution is identified as follows. Given the PDF of the bivariate normal distribution $f_{S,T}(s,t) = f_{T|S}(t \mid s) f_T(s)$, thus the conditional PDF is

$$f_{T|S}(t \mid s) = \frac{f_{S,T}(s,t)}{f_T(s)} = \frac{1}{2\pi\sigma_s\sqrt{1-\rho^2}}\exp[-\frac{1}{2}(\frac{t-\mu_t-\rho(\sigma_t/\sigma_s)(s-\mu_s)}{\sigma_t\sqrt{1-\rho^2}}] \quad \textbf{Eq. 2.46.}$$

By arranging the above equation, it is the above the conditional distribution of T given S is also a normal distribution with the conditional mean of T given S. The conditional mean of T is

$$E(T \mid s) = \mu_t + \rho\frac{\sigma_t}{\sigma_s}(s-\mu_s) \quad \textbf{Eq. 2.47}$$

and the conditional standard deviation of T is

$$\sigma_{T|S} = -\sigma_s\sqrt{1-\rho^2} \quad \textbf{Eq. 2.48}$$

An i.i.d. number is generated first from the marginal distribution of S with μ_s and σ_s, and then a value of t is generated from the above conditional distribution with the above conditional mean and standard deviation. In this research 10000 random annual peak discharge values are generated from Monte Carlo simulation for each of T and S. The synthetic downstream annual peak flow Z can calculate by $z = s + t$.

For the synthetic z, the test of goodness-of-fit indicates that 3-parameter log-normal and 3-parameter gamma distributions fit the data at the significant level of 0.05. The parameters are listed in Table 2-4. We select 3-parameter log-normal distribution in this study. It should be noted that the parameters for the synthetic flow rate of Station C may vary slightly when the model is run every time due to variation of the random numbers generated by MC simulation.

Table 2-4 Distribution parameters of synthetic annual peak flow at the confluence of Des Moines River

		Synthetic data From MC	Observation data of Station C
Data Mean		19341.0	18564.7
Data Standard deviation		10245.1	9402.36
3-parameter log-normal	Location	10.27151	10.46668
	Scale	0.32870	0.25572
	Threshold	-11156	-17719
3-parameter gamma	Shape	5.13091	5.38697
	Scale	4490.22071	4109.859
	Threshold	-3697.88	-3574.96

Perform the flood frequency analysis for the synthetic annual peak flow at the downstream station, the flood corresponding to the return period of 2, 5, 10, 25, 50, 100, and 200 year are listed in Table 2-5 and Figure 2-8.

To verify the model, the flood frequency analysis is performed based on the observation data of Station C. NFF model is also employed and the results are listed in the table for comparison with the proposed model. The regression equations of flood frequency analysis for the Des Moines River basin Stratford, IA are given by

$$Q_2 = 33.8 DA^{0.656}$$

$$Q_5 = 60.8DA^{0.658}$$

$$Q_{10} = 80.1DA^{0.660}$$

$$Q_{25} = 105DA^{0.663}$$

$$Q_{50} = 123DA^{0.666}$$

$$Q_{100} = 141DA^{0.669}$$

$$Q_{200} = 159DA^{0.672}$$

where Q_T = discharge for a return period of T-years (cfs)

DA = the drainage area corresponding to that gauge station (USGS 05481300) = 5452 sq.mi

Table 2-5 Comparison of simulation results with the observation data and NFF model results: Des Moines River

Return Period (yr)	Flood estimated by the proposed model		Flood by NFF model		Flood estimated from Station C data based on log-normal (cfs)	Flood estimated from Station C data based on Pearson III (cfs)
	Food (cfs)	Relative error to observation data (Pearson III)	Food (cfs)	Relative error to observation data (Pearson III)		
2	17741	0.031	9550	-0.445	17406	17211
5	26951	0.039	17500	-0.325	25841	25941
10	32880	0.050	23400	-0.253	31028	31328
25	40222	0.066	31500	-0.165	37241	37727
50	45603	0.080	37900	-0.103	41670	42230
100	50924	0.094	44600	-0.042	45958	46533
200	56229	0.109	51600	0.018	50152	50685
Total relative error		0.469		1.351		

Note: Relative error $= \frac{x_e - x_0}{x_0}$ where x_e is the predicted value, x_0 is the observed value.

Total relative error $= \sum_{i=1}^{n} \left| \frac{x_{e,i} - x_{0,i}}{x_{0,i}} \right|$

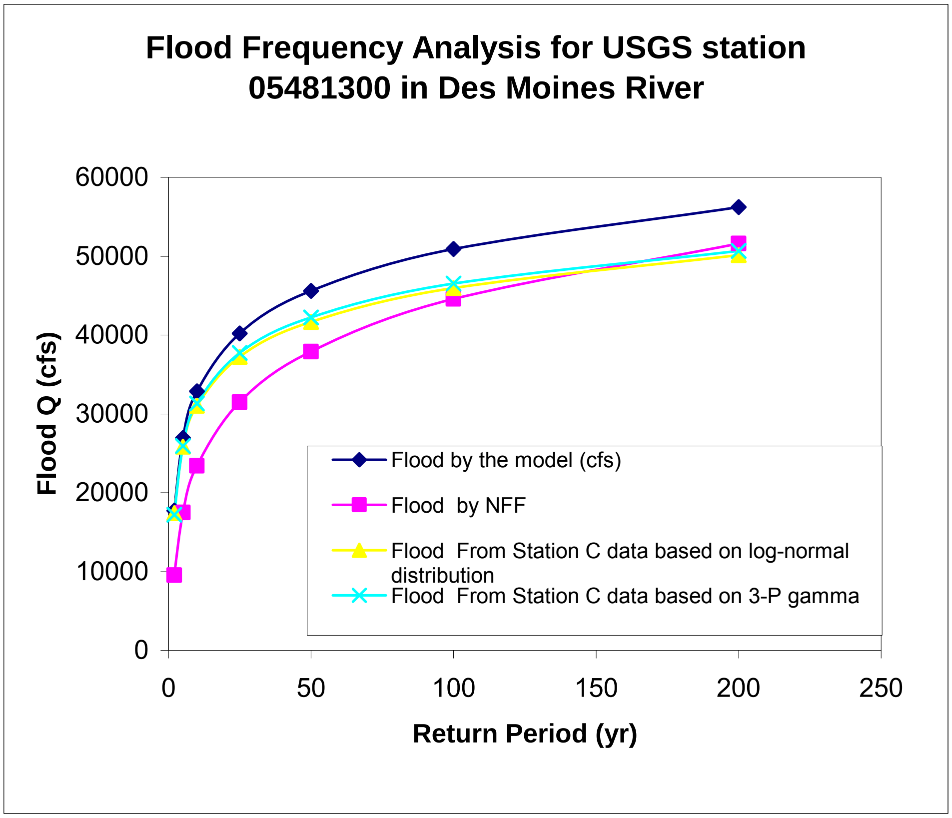

Figure 2-8 Simulation results by joint probability model, NFF model and univariate flood frequency analysis based on observation data

2.3.2 Case study 2

The proposed approach is applied for the Altamaha River basin also where Oconee Rive at Dublin, GA and Ocmulgee River at Lumber City, GA are two tributary streams and Altamaha River near Baxley, GA is the confluent stream. The gauge station information for the three streams is shown in and Figure 2-9. The annual peak discharge of the two tributaries USGS 0223500 and USGS02215500 are collected, and the task is to estimate the flood frequency for the confluence point of the tributaries by assuming the annual peak discharge data of Station C are not available. Totally 35 year data (1971 through 2005) of each of the three stations are used in this research.

Figure 2-9 Location of USGS Gauge Stations in Altamaha River basin, GA

The test of goodness of fit indicates that both Station A and Station B fit 3 parameter normal distribution (See Figure 2-10 and Figure 2-11) and 3-paramter gamma distribution at 0.05 level of significant, with the parameters as shown in Table 2-7. In this research, the 3-parameter lognormal distribution is selected for both Station A and Station B. The joint PDF and conditional distribution can be obtained by Eq. 2.45 and Eq. 2.46. The plot of the joint PDF and CDF are shown in Figure 2-13 and Figure 2-14.The synthetic annual peak discharges for the confluence are then generated from the Monte Carlo simulation based on the conditional probability of binormal distribution. The test of goodness of fit indicates that the 3-parameter gamma distribution fits the synthetic flow rate at a level of 0.05 significant level, with a shape, scale and threshold parameter of 4.89260, 13438.12 and -3634.72, respectively.

To verify the performance of the joint probability approach, the observation data of the gauge station at the confluence, USGS 02205000, are also collected and the conventional flood frequency analysis is performed. The goodness of fit test for the observation data of Station C shows that the annual peak flow follows a 3-parameter gamma distribution or a 3-parameter lognormal distribution, with parameters as shown in Table 2-7. The 3-parameter gamma distribution is selected in this research. Also the result from NFF model is employed here for the comparison with the joint probability model.

The univariate flood frequency analysis is performed for the confluence and the results are shown in Table 2-8 and Figure 2-15. The NFF model is also employed to estimate the flood at the confluence for the comparison with the joint probability model, as shown in Table 2-8 and Figure 2-15. In NFF model the Station C is located in the Region 3 in GA and the drainage area is 11600 mi^2, and the regression equations for Region 3 are given by

$$Q_2 = 76A^{0.620}$$

$$Q_5 = 133A^{0.620}$$

$$Q_{10} = 176A^{0.621}$$

Table 2-6 USGS Gage Stations located in the Altamaha River basin in GA

Station	Tributary A	Tributary B	Confluence C
Station Name	Oconee River at Dublin, GA	Ocmulgee River at Lumber City, GA	Altamaha River Near Baxley, GA
USGU Station No.	02223500	02215500	02225000

Table 2-7 Data distribution information of Altamaha River basin in GA

		Station A	Station B	Station C
Data Mean		32315.4	29612.3	58671.4
Data Standard deviation		17024.9	16956.3	25488.2
Correlation coefficient of A and B		0.6404		
3-parameter log-normal	Location	10.67145	10.32607	11.56073
	Scale	0.36144	0..46212	0.22866
	Threshold	-13667	-4353	-49010
3-parameter gamma	Shape	2.68265	2.58517	8.58058
	Scale	10755.08	10124	8528.94931
	Threshold	3463	3438	-145119

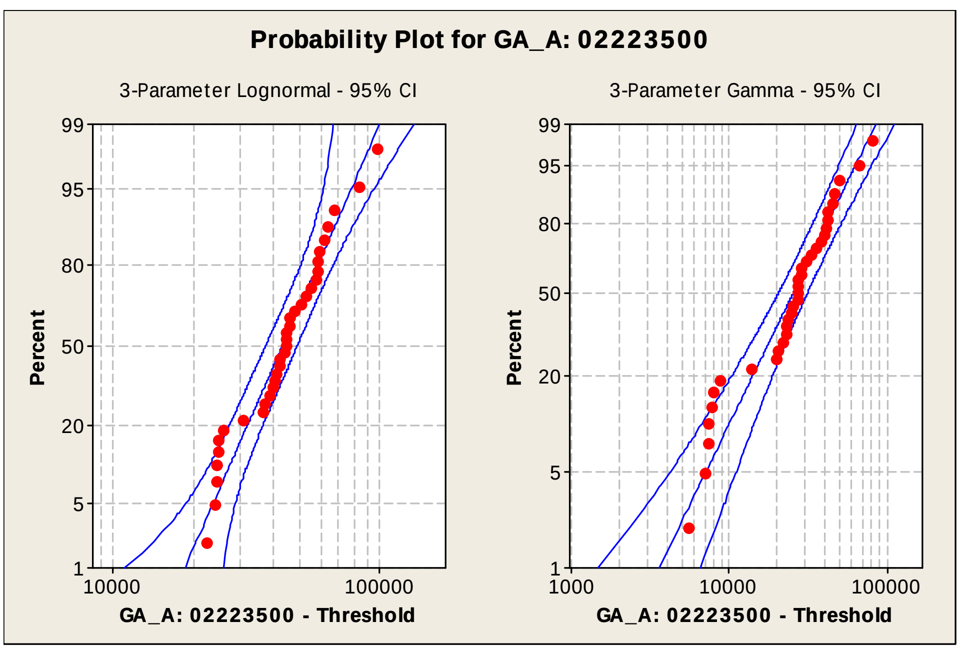

Figure 2-10 Probability plot of Station A in Altamaha River basin, GA

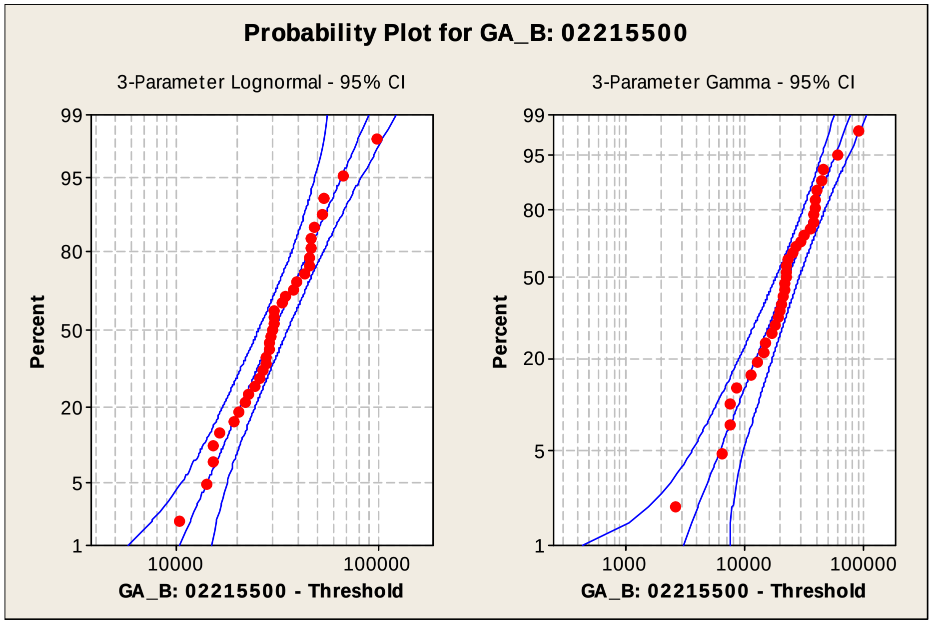

Figure 2-11 Probability plot of Station B in Altamaha River basin, GA

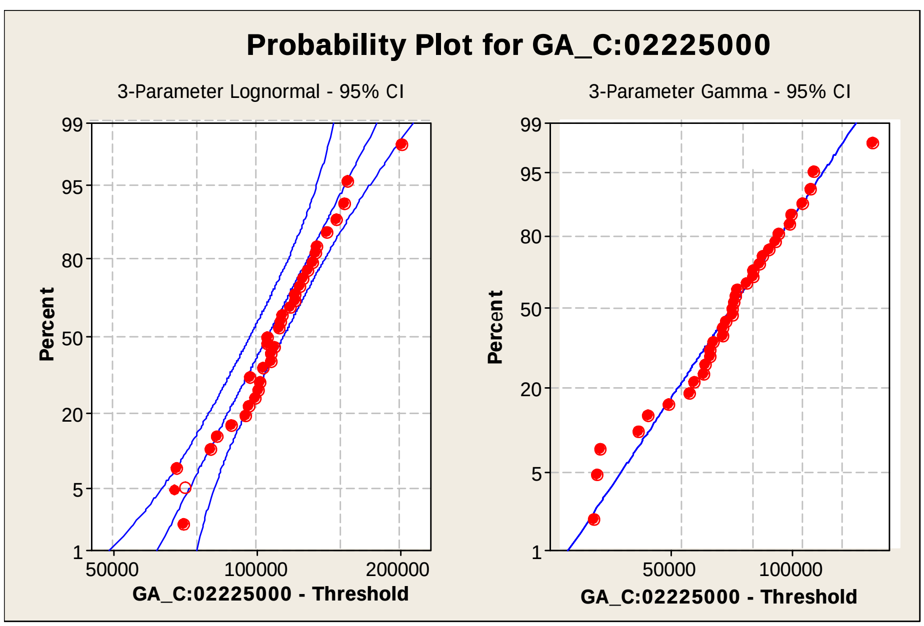

Figure 2-12 Probability plot of Station C in Altamaha River basin, GA

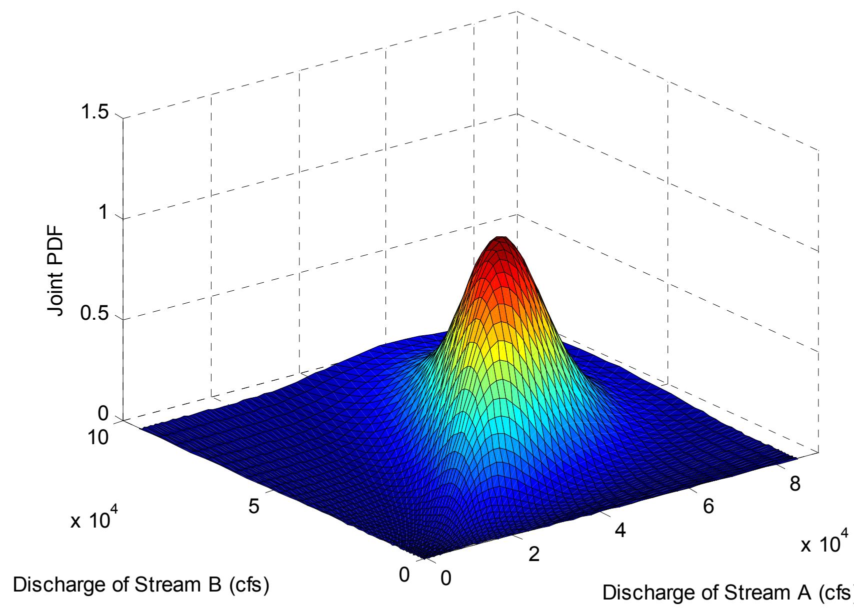

Figure 2-13 Joint PDF of the tributary streamflow of Altamaha River basin, GA

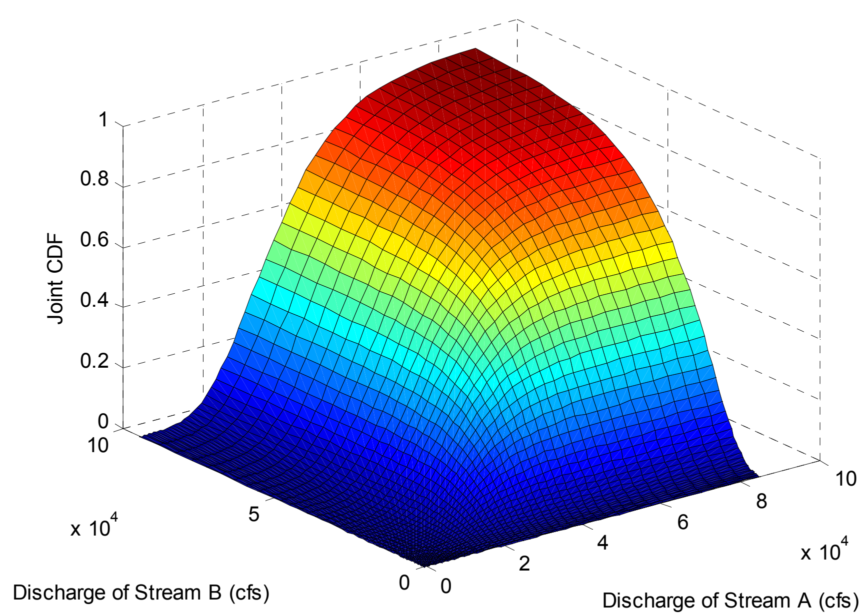

Figure 2-14 Joint PDF of the tributary streamflow of Altamaha River basin, GA

$Q_{25} = 237A^{0.623}$

$Q_{50} = 287A^{0.625}$

$Q_{100} = 340A^{0.627}$

$Q_{200} = 396A^{0.629}$

Where A is the drainage area, 11600 mi^2.

2.3.3 Discussion

Since the annual peak discharge may fit more than one distribution, as is shown in this study, the annual peak discharges in Station B and C fit both 3-parameter lognormal and Pearson III distributions based on the test of goodness of fit. From the observation of the Station C, the results from different distributions shows just slightly difference in the flood estimate. From the empirical practice, the Pearson III and log-Pearson III distribution are recommended by USGS, which agree with the observation of about 30 other streams throughout the US in this research. It is found out that a distribution with a large number of parameters always fits the data better than the distribution with a small number of parameters. However, when size of the sample data is relative small, the distribution with smaller number of parameters may be preferred if it is accepted by the test of goodness of fit.

The results of two case studies show that the proposed model can simulate the flood frequency very well, especially when the return period is getting small. Its results agree with the results by the univariate flood frequency analysis based on the observation data of Station C. The largest error occurs at the 200 year of return period, which is around 10% for the smaller river basin and 20% for the larger river basin, relative to the result by the univariate flood frequency analysis based on Pearson III distribution. The model performs best at the smallest return periods for both small river basin and large river basin. The model always tends to overestimate the flood and the overestimation is getting larger (around 20% relative to the observation data) when the river basin is getting large. However, the relative error

Table 2-8 Data distribution information of Altamaha River basin in GA

Return Period (yr)	Flood by the proposed model		Flood by NFF model		Flood estimated from Station C data based on Pearson III (cfs)
	Food (cfs)	Relative error to observation data (Pearson III)	Food (cfs)	Relative error to observation data (Pearson III)	
2	57690	0.03	25200	-0.55	55890
5	84960	0.09	44000	-0.44	78150
10	101910	0.11	58800	-0.36	91600
25	122140	0.14	80700	-0.25	107530
50	136420	0.15	99600	-0.16	118760
100	150100	0.16	120000	-0.07	129550
200	163320	0.16	143000	0.02	140030
Total relative error		0.84		1.85	

Note: Relative error $= \frac{x_e - x_0}{x_0}$ where x_e is the predicted value, x_0 is the observed value.

Total relative error $= \sum_{i=1}^{n} \left| \frac{x_{e,i} - x_{0,i}}{x_{0,i}} \right|$

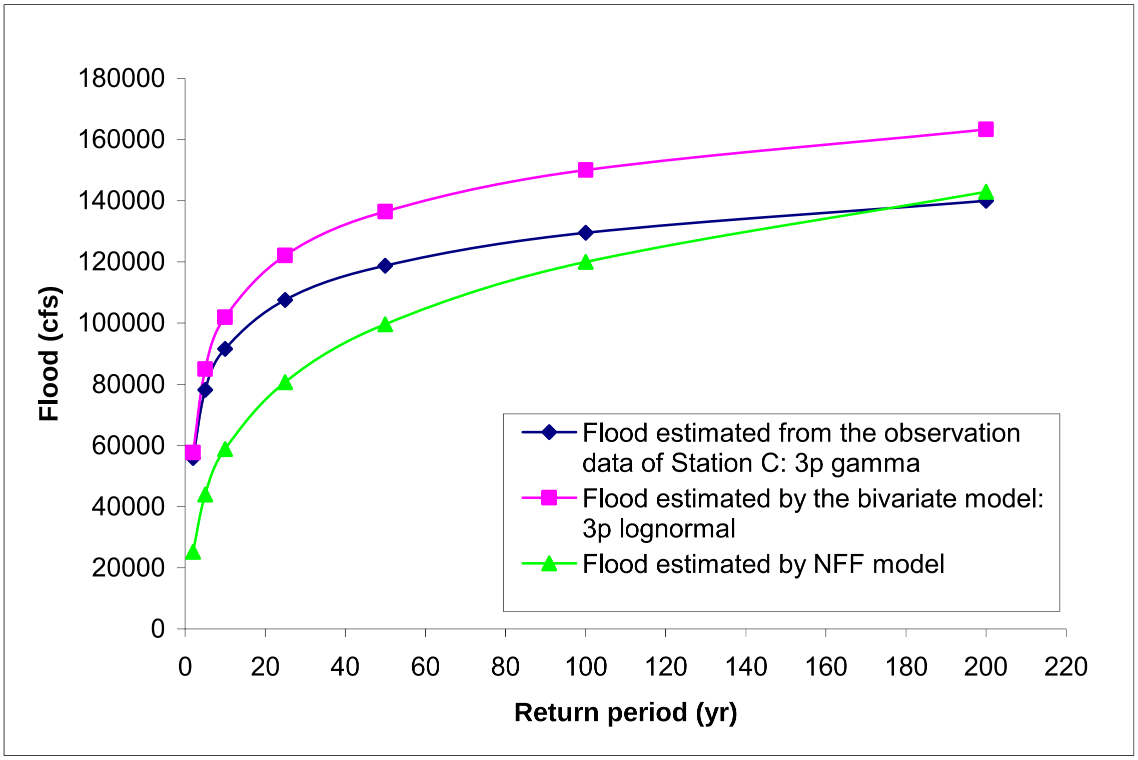

Figure 2-15 Simulation results comparison with NFF model and univairate flood frequency analysis based on observations at confluence of Altamaha River basin

tends to be stable when the return period is over 100 years. Several factors may cause the accuracy loss in Altamaha River basin in GA. The station A is very far away from the confluence point. There may be some relatively large water quantity gain or loss between Station A and the confluence while is not taken into account in the synthetic confluence annual peak flow by just summation of the generated peak flow of station A and B by MC method. Another possible reason is that the Station C is also far away from confluence, so the real difference of the flood between Station C and confluence might be a little large.

The NFF model developed by USGS also works well, especially when the return period is larger than 100 years. However, it seems that the NFF model underestimates the flood when the return period is smaller than 100 year, and the smaller the return period is, the larger the underestimation is. The relative error is also getting larger with the increase of the river basin. The largest error occurs in the smallest return period, which maybe more than 50% relative to the observation data. The NFF model works best when the return period is large enough, say no smaller than 100 year.

2.4 Conclusion and future work

This research proposed a joint probability approach that provides a practical way for the confluence flood frequency analysis with an acceptable accuracy. The approach performs better for the smaller river basin than for the larger river basin. The relative error tends to get larger with the increase of return period and river basin. This error could be reduced by using variance reduction techniques, such as control variate method, or with more accurate mass balance estimate in the synthetic confluence flow, in stead of only summation of the synthetic upstream flow rate.

The procedure is discussed in the paper which is comprised of four steps, the distribution identification of annual stream peak flow of the tributary streams, the identification of joint probability distribution of the tributary stream flows, the generation of the synthetic stream flow at the confluent point by using Monte Carlo simulation, and identification of the flood frequency of the confluent point by the conventional univariate flood frequency analysis.

It is also could be extended to the confluence flood frequency analysis with more than two tributary streams by the same token of estimating a multivariate distribution of the annual peak discharge of the statistically dependent tributaries or transferring the multivariate problem to several bivariate probability problems and then performing the Monte Carlo simulation as addressed in the paper. The proposed model also provides an alternative method for the ungauged flood frequency beside the NFF model, especially when the return period is no more than 100 year within which the proposed model can perform better than NFF model.

However, it should be noted that the most challenging part of the approach is to estimate accurately the joint probability distribution. There are no well established methods reported that can meet this requirement although there are many bivariate or multivariate distributions reported in the literature. Most of them focus on the bivariate normal distribution, bivariate exponential distribution, or biavariate gamma distribution. Given the observation that the gamma distribution or Pearson III or log-Pearson III can fit many or even most of the streams in US, as recommended by the U. S. Water Resources Council for flood flow frequency studies (U. S. Water Resources Council, 1981), the reported empirical biavariate model should be capable of handling the most cases. It is still necessary to develop a more general and more efficient method for the bavariate distribution estimate.

References

Akaike, H. (1974), A new look at the statistical model identification, *IEEE Trans. Autom. Control, AC*, **19**, 716–722.

Ang. H-S. A. and Tang, H. W. (1975a), *Probability Concepts in Engineering Planning and Design.* Volume I- Basic Principles. John Wiley & Sons, Inc., New York.

Ang. H-S. A. and Tang, H. W. (1975b), *Probability Concepts in Engineering Planning and Design*. Volume II- Decision, Risk, and Reliability. John Wiley & Sons, Inc., New York.

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.

Ashkar, F., and Rousselle, J. (1982), A multivariate statistical analysis of flood magnitude, duration and volume, in *Statistical analysis of rainfall and runoff*, edited by V. P. Singh, pp. 659–669, Water Resource publication, Fort Collins, Colo..

Ashkar, F., El-Jabi, N., and 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. and Kottegoda, N. T. (1994), Bivariate exponential model applied to intensities and durations of extreme rainfall, *Journal of hydrology,* **155,** 225-236.

Beersma, J. J. and Buishand, T. A. (2004), The joint probability of rainfall and runoff d eficits in the Netherlands. Proceedings, EWRI 2004 World Water & Environmental Resources Congress 2004, June 27 - July 1, 2004, Salt Lake City, USA (CD-ROM).

Blachnell, D. (1994), New method of the simulation of correlated K-distribution clutter, *EE Proceedings Radar, Sonar and Navigation,***141**, 53-58.

Buishand TA. (1984), Bivariate extreme-value data and the station-year method, *Journal of Hydrology*, **69**, 77–95.

Burkhardt, G. and Prakash, A. (1976), An analysis of computer graphics to extremes, *Water Research Bulletin*, **12**, 1245-1258.

Chang, C. H., Tung, Y. K., and Yang, J. C. (1994), Monte Carlo simulation for correlated variables with marginal distributions, *Journal of Hydraulics Engineering*, **120,** 313-331.

Choulakain, V., Ei-Jabi, N., and Moussi, J. (1990), On the distribution of flood volume in partial duration series analysis of flood phenomena, *Stochastic Hydrology and Hydraulics,* **4,** 217-226.

Chow, V. T., Maidment, D. R. and Mays, L. W., (1988), *Applied hydrology*. McGraw-Hill, New York.

Clarke, R. T.(1980), Bivariate gamma distributions for extending annual streamflow records from precipitation: some large-sample results, *Water Resources Research*, **16**, 863–870.

Crovelli, R. A. (1973), A bivariate precipitation model, paper presented at 3rd Conference on Probability and Statistics in Atmospheric Science, American Meteorological Society, Boulder, CO, USA; 130–134.

Durrans, S. R., Eiffe, M. A., Thomas, Jr. W. O., and Goranflo, H. M. (2003), Joint seasonal/ annual flood frequency analysis, *Journal of Hydrologic Engineering*, **8**, 181-189.

Galambos, J. (1978), *The asymptotic theory of extreme order statistics*. Wiley, New York.

Genest, C., and Rivest L-P. (1993), Statistical inference procedures for bivariate Archimedean copulas, *Journal of American Statisticians Association*, **88**, 1034-1043.

Genest, C.,and MacKay, J. (1986), The joy of copulas: Bivariate distributions with uniform marginals, *American Statisticians Assicuation*, **40,** 280-283.

Gentile, J. E. (1998), *Random Number Generation and Monte Carlo Method,*. Springer-Verlag New York, Inc.

Goel NK, Seth SM, Chandra S. (1998), Multivariate modeling of flood flows, *Journal of Hydraulic Engineering*, **124,** 146–155.

Greenwood, J. A., Landwehr, J. M., Matalas, N. C., and Wallis, J. R. (1979), Probability weighted moments: definition and relation to parameters of several distributions expressible in inverse form, *Water Resources Research,* **15,** 1049-1054.

Gumbel, E. J., and Mustafi, C. K. (1967), Some analytical properties of bivariate extreme distributions, *Journal of American Statistician Association.* **62,** 569-588.

Hashino, M. (1985), Formulation of the joint return period of two hydrologic variates associated with a Possion process, *Journal of Hydroscience and Hydranlic Engineering*, **3,** 73-84.

He. H., Zhou, J., Yu, Q., Tian, Y. Q., and Chen, R. R. (2007), Flood frequencyand routing processes at a confluence of the middle yellow river in china, *River Research and Applications*, **23,** 407-427.

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

Hosking, J. R. M. (1993), Some statistics useful in regional frequency analysis, *Water Resources Research.***29,** 271.

Joe, H. (1997), *Multivariate mold and dependence concepts*, hapman & Hall, London.

Johnson, M. E. (1987), *Multivariate Statistical Simulation*. John Willy & Sons Inc., New York.

Johnson, R. A., Evans, J. W., and Green, D. W. (1999), Some Bivariate Distributions for modeling the strength properties of lumber, U. S. Department of Agriculture, Research paper FOL-RP-575.

Kite, G. W., and Stuart, A. (1977), Frequency and risk analysis in hydrology. *Water Resources public.*Fort Collins, Co.

Krstanovic, P. F., and Singh, V. P. (1987), A multivariate stochastic flood analysis using entropy, in *Hydrologic Frequency Modelling*, edited by V. P. Singh, PP. 515–539, Reidel, Dordrecht, The Netherlands.

Linsley, R. K. (1986), Flood estimates: how good are they? *Water Resources Research*, **22,** 1595-1645.

Marshall, A. W. and Olkin, I. (1988), Families of multivariate distributions, *Journal of the American Statistical Association*, **83**, 834-841.

Michael, J. R. and Schucany, W. R., (2002), The mixture approach for simulating bivariate distributions with specified correlations. *The American statistician*, **56**, 48-54.

Moharran, S. H., Gosain, A. K., and Kapoor, P. N. (1993), A comparative study for the estimators of the generalized pareto distribution. *Journal of hydraulics*, **150,** 169-185.

Molenberghs, G. and Lesaffre, E., (1997), Non-linear integral equations to approximate bivariate densities with given marginals and dependence function, *Statistica Sinica*, **7**, 713-738.

Moran, PAP. (1970), Simulation and evaluation of complex water systems operations. *Water Resource Research*, **6**, 1737–1742.

Nadarajah, S. and Gupta, A. K. (2006), Intensity–duration models based on bivariate gamma distributions. Hiroshima mathematical journal, **36**, 387–395

Natural Environmental Research Council (1975), Flood Studies Report, Volume I: Hydrological Studies, report, Natural Environmental Research Council, London.

Nelsen RB. 2006. *An Introduction to Copulas*, Second Edition, Springer-Verlag, New York.

Oklin, I. and Harshall, A. (1988), Families of multivariate distributions, *Journal of American Statistical association*, **83**, 834-841.

Papadimitriou, Z. G., Sagias, N. C., Bithas, P. S. Mathiopoulos, P. T., and Merakos, L. (2006), The trivariate Weibull distribution with arbitrary correlation, in Satellite and Space Communications, 2006 International Workshop on, pp.249-253.

Prekopa A, Szantai T. (1978), A new multivariate gamma distribution and its fitting to empirical streamflow data. *Water Resource Research*, **14**, 19–24.

Rao, A. R. and Hamed, K. H., (2000a), *Flood frequency analysis.* CRC Press, New York, NY, pp. 8.

Rao, A. R. and Hamed, K. H., (2000b), *Flood frequency analysis.* CRC Press, New York, NY, pp. 10.

Raynal, J.A., and Salas, J.D., (1987), A probabilistic model for flooding downstream of the junction of two rivers, in *Hydrologic frequency modeling*, edited by V. P. Singh, pp.595-602, Dordrecht, The Netherlands, D. Reidel Publishing Company.

Ronning, G. (1977), A Simple Scheme for Generating Multivariate Gamma Distributions with Non-Negative Covariance Matrix, *Technometrics*, **19**, 179-183.

Rossi, F. M. F. and Versace, P. (1984), Two-component extreme value distribution for flood frequency analysis. *Water Resource Research*, **20,** 847-856.

Sackl, B., and Bergmann, H. (1987), A bivariate flood model and its application, in *Hydrologic frequency modeling*, edited by V. P. Singh, pp. 571-582, Dreidel, Dordrecht, The Netherlands.

Schucany, W. R., Michael, J. R. (2002). The mixture approach for simulating bivariate distributions with specified correlations. *The American statistician*, **56**, 48-54.

Schweizer, B. (1991), *Thirty years of copulas. In: Dall's Aglio G, Kotz S, Salinetti G(eds) advances in probability distributions with given marginals*. Kluwer, Dorderecht. pp13-50

Shiau, J., Feng, S., and Nadarajah, S. (2007), Assessment of hydrological droughts for the Yellow River, China, using coplus. *Hydrological Processes*, **21**, 2157-2163.

Shih, J. H. and Louis, T. A. (1995), Inferences on the association parameter in copulas models for bivariate survival data. *Biometrics*, **51**, 1384-1399.

Singh, K., and Singh, V. P. (1991). Derivation of bivariate probability density functions with exponential marginals. *Stochastic Hydrology & Hydraulics*, **5**, 55-68.

Singh, V. P. and Singh, K. (1985a), Derivation of the Pearson Type (PT) III distribution by using the principle of maximum entropy (POME), *Journal of Hydraulics*, **80,** 197-214.

Singh, V. P. and Singh, K. (1985b), Derivation of gamma distribution by using the principle of maximum entropy. *Water Resources Bulletin,* **21,** 1185-1191.

Smith, O.E., Adelfang, and Tubbs, JD. S. I. (1982). A bivariate Gamma probability distribution with application to gust model. NASA Technical Memorandum 82483, National Aeronautics and Space Administration.

Song, P. X. K., Y. Fan, and J. D. Kalbfleisch (2005), Maximization by parts in likelihood inference, *Journal of the American Statistical Association*, **100**, 1145–1158.

Stuart, A., and Ord, J. K. (1987). Kendall's Advanced theory of statistics, Vol.1, 5th ed.,Distribution theory, Charles Griffin and Company, London. The first of three volumes of the classic reference book in statistics and probability-not for beginners.

Tod, D. K.(1957), Frequency analysis of streamflow data, *Journal of the Hydraulics Division*, Process ASCE HY1, pp. 1166-1, 1166-16.

Trivedi, P. K. and Zimmer, D. M. (2005), Copulas modeling: an introduction for practitioners, *Foundations and trends in econometrics*, **1**, 1-111.

U.S. Geological Survey (2002). The National Flood Frequency Program, Version 3: A Computer Program for Estimating Magnitude and Frequency of Floods for Ungaged Sites. U.S. Geological Survey Water-Resources Investgations Report 02-4168

U.S. Water Resouces Council (1981) Guidelines for determining flood flow frequency, bulletin 17B, available from Office of Water Data Coordination, U. S. Geological Survey, Reston, VA 22092.

Wang, Q. J. (2001). A Bayesian Joint probability approach for flood record augmentation. *Water Resources Research,* **37,** 1707-1712.

Yue S, Ouarda TBMJ, Bob´ee B. 2001. A review of bivariate gamma distributions for hydrological application. Journal of Hydrology 246: 1–18.

Yue S. (1999). Applying the bivariate normal distribution to flood frequency analysis. *Water International,* **24,** 248-252.

Yue, S. (2001a). A bivariate extreme value distribution applied to flood frequency analysis. *Nordic Hydrology*, **32**, 49–64.

Yue, S. (2001b). The bivariate lognormal distribution to model a multivariate flood episode. *Hydrological Processes*, **14,** 2575-2588.

Yue, S. (2001c) A bivariate gamma distribution for use in multivariate flood frequency analysis. *Hydrology processes*, **15,** 1033-1045.

Yue, S. and Rasmussen, P. (2002), Bivariate frequency analysis: discussion of some useful concepts in hydrological application, *Hydrological Process*, **16**, 2881–2898.

Yue, S. and Wang, C. Y. (2004), A comparison of two bivariate extreme value distributions, *Stochastic Environmental Research*, **18**, 61-66.

Yue, S., Ouarda, T. B. M. J., and Bobée, B. (2001), A review of bivariate gamma distributions for hydrological application, *Journal of Hydrolgoy*, **246**, 1-18.

Zhang, L. and Singh, V. P. (2006), Bivariate flood frequency analysis using the copula method, Jo*urnal of hydrologic engineering*, **11**, 150-164.

Chapter 3 A copulas-based joint probability approach for confluence Flood Frequency Analysis

Abstract

Joint probability approach shows the capablity of estimating the flood frequency accurately. But it is difficult to estimate the joint probability distribution in the approach, especially when the joint probability has different type of margins, which is the key task in joint probability approach. Copulas provide a way to construct multivariate distribution functions. This paper reviews the joint probability approach for confluence flood frequency analysis, and introduces mainly the copulas which can be used to construct multivariate distribution with any type of margins. The method of constructing copulas and the often used Archimedean copulas are introduced. And the dependent parameter and the copula evaluation are also presented in the paper.

Keywords: Flood frequency analysis, copula, joint probability, Monte Carlo simulation, confluence point

3.1 Introduction

The flood frequency analysis at or nearby the confluence of two tributaries is of interest because it is necessary for the design of the highway drainage structures, which often are located near the confluence point and may be subject to inundation by high flows from either stream or both. The univariate flood frequency analysis approach is not applicable for this sometimes because of the shortage of the hydrological data at the confluence point. Currently, the National Flood Frequency Program (NFF) (US Geology Survey, 2002) developed by US Geology Survey (USGS) based on the regional analysis probably is the most popular model for the ungauged site flood estimation, and could be employed for the flood estimate at the confluence. In NFF model each state in US are divided into multiple hydrologic regions by using major watershed boundary and/or some other hydrologic characteristics, i. e., the mean elevation of watershed. It is assumed that the hydrologic characteristics are homogeneous in each region so that the flood at the ungauged sites can be estimated by the gauged sites. A series of regression equations of T-year flood (T=2 , 5, 10, 25, 50, 100, 200 and 500 year) associated with each hydrologic region are developed in terms of hydrologic characteristics based on the gauged site records. All the sites in each region share the same regression equation for the flood estimation associated with a specified return period. However, some equations in this approach have high errors; for example, some equations generate standard errors greater than 100 percent for the western part of the US, although the average standard error of NFF is between 30 and 60 percent (USGS, 2002).

To avoid the flow routing procedure which is usually time and effort consuming, a practical approach for the flood frequency estimation for this situation is needed. An approach based on the joint probability may be developed for the confluence flood estimation.

Bivariate or multivariate flood frequency analysis has received much attention recently. Sackl and Bergmann (1987), Chang et al. (1994), Goel et al. (1998), Yue (1999, 2000), and Beersma and Buishand (2004) used the bivariate normal distribution to perform the flood frequency analysis and hydrology events analysis. Krstanovic and Singh (1987) derived the multivariate Gaussian and exponential distributions by the principle of maximum entropy and applied the bivariate distributions for the analysis of flood peak and volume. Yue (2000) applied the bivariate lognormal distribution multivariate flood events analysis and described the relationship of flood peaks and volumes as well as flood volumes and durations by joint distribution and the corresponding conditional distribution.

Hashino (1985), Choulakian et al. (1990), Singh and Singh (1991), Bacchi et al. (1994), and Ashkar et al. (1998) investigated and applied the bivariate exponential distributions for the hydrological events analysis. Bacchi et al. (1994) proposed a numerical procedure for the estimation of parameters of a bivariate exponectial model used to simulation the storm intensity and duration simultaneously.

Buishand (1984), Raynal and Salas (1987), Yue et al. (1999) and Yue (2001a) applied bivariate extreme value distributions to analysis multivariate flood/storm events. Yue and Wang (2004) compared the performance in flood analysis between two bivariate extreme value distributions, the Gumbel mixed model and the Gumbel logistic model.

Bivariate gamma distribution is also widely used for the flood frequency analysis (Moran, 1970; Prekopa and Szantai, 1978; Yue, 2001). Among them, Yue (2001b) investigated the applicability of the bivariate gamma distribution model to analyze the joint distribution of two positively correlated random variables with gamma marginals. Yue et al (2001) reviewed three bivaraite gamma distribution models with two gamma marginal distributions.

Durrans et al (2003) presented two approximate methods for joint frequency analysis using Pearson Type III distribution to estimate the joint flood frequency analyses on seasonal and annual bases. Nadarajah and Gupta (2006) developed exact distribution of intensity-duration based on bivariate gamma distribution.

Yue and Rasmussen (2002) discussed the concepts of bivariate hydrology events and demonstrated the concepts by applying a bivariate extreme value distribution to represent the joint distribution of flood peak and volume from an actual basin.

Johnson et al. (1999) reviewed the some techniques for obtaining bivariate distributions and presented the properties of some bivariate models that include bivariate Weibull distribution, bivariate inverse Gaussian distribution, bivariate S_{BB} distribution and bivariate normal-lognormal distribution.

The existing techniques for estimating joint distributions of hydrology data often require some assumptions, for example, the same type of the marginal distribution are always assumed. The estimation and inference for data that are assumed to be multivariate normal distributed are highly developed, but general approaches for joint nonlinear modeling of nonnormal data are not well developed, and there is a frequent tendency to consider modeling issues on a case-by-case basis. This research explores the copula approach for hydrology modeling of joint parameter distributions, providing a way to perform flood frequency analysis at the confluence in the river basin. The copula approach involves specifying marginal distributions of each random variable along with a function (copula) that binds them together. Although theoretical foundations of copulas are complex, this paper demonstrates that practical implementation and estimation is relatively straightforward. One of the properties of copulas that are very useful in implication is that the same copula can be used for the joint distribution of $(\ln X, \ln Y)$ as the copula for the joint distribution of (X, Y).This is useful because it may be more convenient for the analysis to express the hydrology data in natural unit.

Although well known in the statistical literature for more than 40 years, applications of the copula theory in statistical modeling are a more recent phenomenon.Very few applications have been reported in hydrology area (Wang, 2001; Zhang, 2006; Shuiau; 2006).

Zhang and Singh (2006) derived bivariate distributions of flood peak and volume, and flood volume and duration by using copula method. In the paper, four often used one parameter Archimedean copulas are introduced, the corresponding parameter estimation is described and the criteria of copula selection are addressed. Wang (2001) developed a procedure for record augmentation of annual maximum floods by applying the bivariate extreme value distribution (Gumbel-Hougaard copula) for annual maximum floods at two gauging stations with generalized extreme value distribution. Shiau et al (2007) applied copulas with a mixture of exponential and gamma marginal distribution to simulate the relationship between drought duration and drought severity.

This study aims to review the joint probability approach for the confluence flood frequency analysis and introduce the copulas into the application for the joint probability approach. The concept of copula, the construction of copula, parameter estimation for the copulas and the criteria for copula selection is presented. The copula-based joint probability approach is applied to estimate the flood in two river basins to demonstrate the proposed approach.

3.2 Methodology

3.2.1 Review of the joint probability approach

In the section, the proposed copula method is applied in the joint probability approach for the confluence flood analysis. The joint probability approach for the confluence flood estimate involves the following steps, (1) stream flow distribution identification of the tributary streams, (2) identification of joint probability distribution of the tributary stream flows,(3) identification of the synthetic stream flow at the confluent point by using Monte Carlo simulation, and (4) identification of the flood frequency of the confluent point by the conventional flood frequency analysis. Figure 3-1 shows briefly the procedure of the joint probability approach for flood frequency analysis. In this approach, the accurate estimate of the joint probability of the upstream tributaries plays a key role and probably is the most challenging part.

3.2.2 Copulas

Copulas are defined by Nelson (2006) “functions that join or “copula” multivariate distribution functions to their one-dimensional marginal distribution functions”. The biggest advantage of copula method is that it is capable of determining the multivariate distribution in an easy way regardless of the marginal distributions, compared to the other methods which may involve either very complicated derivation or have some strict requirement, i.e. the

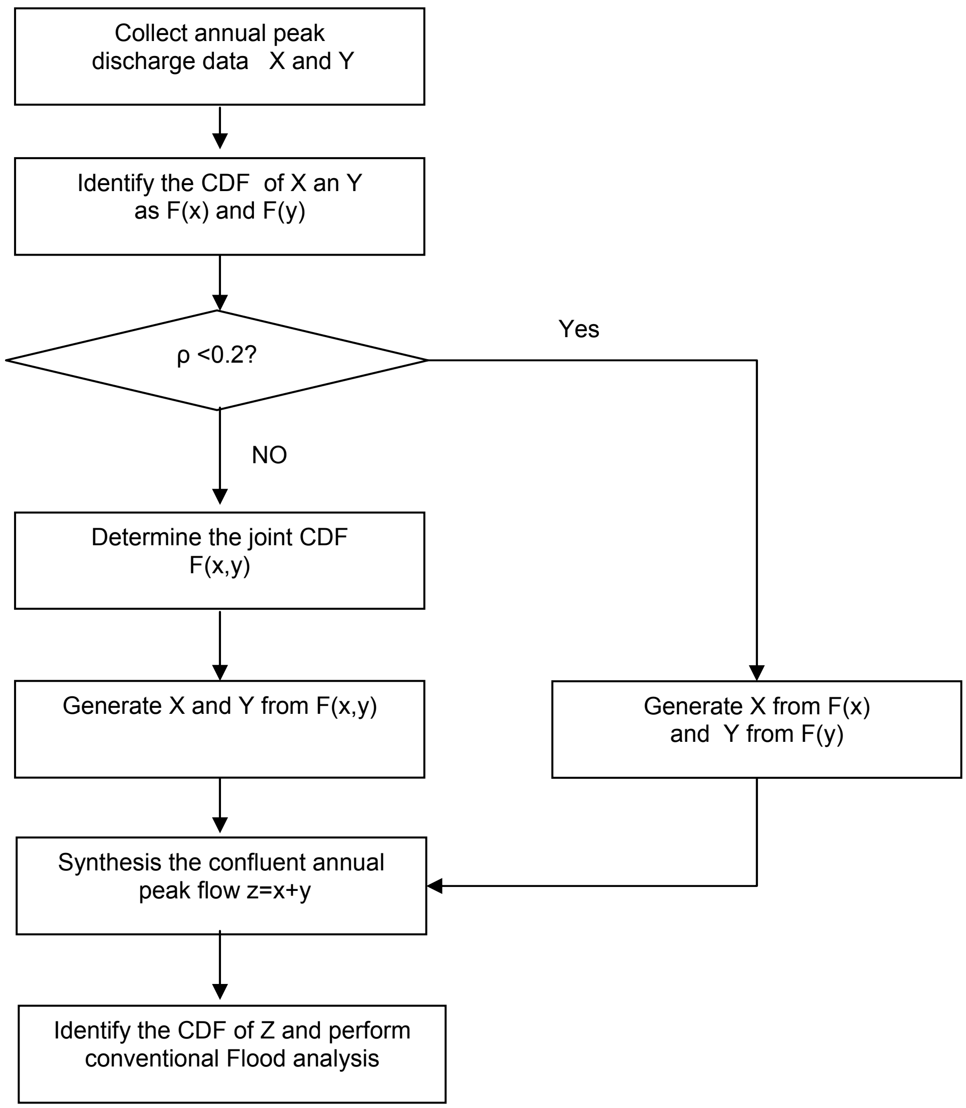

Figure 3-1 Flow chart of joint probability approach for confluence flood frequency analysis

margins are the same type of distribution.

3.2.2.1 Concept of copula

The term of “copula” was first employed by Sklar (1959), and then developed and addressed by many researchers (Galambos, 1978; Genest and Mackay, 1986; Schweizer, 1991; Genest and Rivest 1993; Joe, 1997; Shih and Louis, 1995; and Nelson, 2006).

To demonstrate the copula method of deriving multivariate distribution function with multi univariate distributions, the Sklar’s theorems, developed by Sklar (1959) are introduced as follows, by assuming the following assumption holds that a distribution function (CDF) is a function F with domain $\bar{R}\,([-\infty,+\infty])$ such that (1) F is nondecreasing and (2) $F(-\infty)=0$ and $F(+\infty)=1$.

Sklar’s theorem. Let H be a joint distribution function with margins F and G. Then there exists a copulas C such that for all x, y in $\bar{R}\,([-\infty,+\infty])$,
$H(x,y)=C(F(x),G(y))$

If F and G are continuous, then C is unique; otherwise, C is uniquely determined on RanF$\times$RanG. Conversely, if C is a copula and F and G are distribution functions, then the function H defined by the above equation is a joint distribution function with margins F and G, where RanF (=**I=**[0,1])and RanG (=**I=**[0,1] donate the range of F and the range of G, respectively.

Sklar’s theorem in n-dimensions. Let H be an n-dimensional distribution function with margins $F_1, F_2, ..., F_n$, then there exists an n-copulas C that for all x in $\bar{R}^n$,
$H(x_1,x_2,...,x_n)=C(F_1(x),F_2(x),...,F_n(x))$.
If $F_1, F_2, ..., F_n$ are all continuous, then C is unique; otherwise, C is uniquely determined on $\text{RanF}_1\times\text{RanF}_2\times\cdots\times\text{RanF}_n$. Conversely, if C is an n-copula and $F_1, F_2, ..., F_n$ are distribution functions, then the function H defined by the above equation is an n-dimensional distribution function with margins $F_1, F_2, ..., F_n$.

According to the Sklar’s theorems, the joint CDF or n-dimensional multivariate joint CDF can be determined if the marginal CDFs are know and the copula or the n-dimensional copula can be determined. By splitting the marginal behaviors from the dependence relation, the copulas method allows very flexible joint distributions.

The following theorem can be proved.

Theorem For $n \geq 2$, let $X_1, X_2, ..., X_n$ be continuous random variables with margins $F_1, F_2, ..., F_n$. Then

1. $X_1, X_2, ..., X_n$ are independent if and only if the n-copula of $X_1, X_2, ..., X_n$ is $\mathbf{\Pi}^n$, and
2. each of the random variables $X_1, X_2, ..., X_n$ is almost surely a strictly increasing function of any of the others if and only if the n-copula of $X_1, X_2, ..., X_n$ is $\mathbf{M}^{\text{n}}$, where $\mathbf{\Pi}^n = F_1 \times F_2 \times \; ... \times \; F_n$, and $\mathbf{M}^{\text{n}} = \min(F_1, F_2, ..., F_n)$.

3.2.2.2 Methods of copula generation

Method of Inversion

According to Sklar's theorem, for random variable X and Y with continuous margins $F(x)$ and $G(y)$, respectively, and the joint continuous distribution function $H(x, y)$, there exists a unique copula C which can be generated by

$$H(x, y) = C(F(x), G(y)) \qquad \textbf{Eq. 3.1}$$

Let $u = F(x)$ and $v = G(y)$, where u and v are standard uniform variables, and we can express X and Y as $x = F^{-1}(u)$ and $\mathrm{y} = G^{-1}(v)$, respectively, where F^{-1} and G^{-1} are the inverse functions of F and G, respectively. Then the above equation can be wroten as

$$C(u,v) = H(F^{-1}(u), G^{-1}(v)) \qquad \textbf{Eq. 3.2}$$

Example (Nelsen, 2006)

Consider the Gumbel's bivariate exponential distribution (Gumbel 1960), which is given by

$$H(x, y) = \begin{cases} 1 - e^{-x} - e^{-y} + e^{-(x+y-\theta xy)} & x, y \geq 0 \\ 0 & x, y < 0 \end{cases} \qquad \textbf{Eq. 3.3}$$

For given marginal distributions $u = F(x) = 1 - e^{-x}$ and $v = G(y) = 1 - e^{-y}$, where θ stands for a parameter in [0,1].Thus,

$$F^{-1}(u) = -\ln(1-u) \qquad \textbf{Eq. 3.4}$$

$$G^{-1}(y) = -\ln(1-v) \qquad \textbf{Eq. 3.5}$$

Therefore, the copula C can be obtained by

$$\begin{aligned} C(u,v) &= 1 - e^{-F^{-1}(u)} - e^{-G^{-1}(v)} + e^{-(F^{-1}(u)+y-\theta F^{-1}(u)G^{-1}(v))} \\ &= u + v - 1 + (1-u)(1-v)e^{-\theta \ln(1-u)\ln(1-v)} \end{aligned} \qquad \textbf{Eq. 3.6}$$

The inversion method is straight forward; however, in this method the joint distribution is required to derive the copula. This limits the usefulness of the method for applications in which the joint distribution is often unknown.

Geometric methods

Geometric methods is related to the following property of copula, 1) for every u, v in **I**, $C(u,0) = C(0,v) = 0$ and $C(u,1) = u$ and $C(1,v) = v$; 2) For every u_1, u_2, v_2, v_2, in **I** such that $u_1 \le u_2$ and $v_1 \le v_2$, $C(u_2,v_2) - C(u_2,v_1) - C(u_1,v_2) + C(u_1,v_1) \ge 0$.

Consider the parameter θ as an observation of a continuous random variable Z with distribution function $\Lambda_\theta(z)$. $\{C_z\}$ is a finite collection of copulas. In the geometric methods, the copula is derived from the integration

$$C_\theta(u,v) = \int_{\mathbf{R}(Z)} C_{\mathbf{Z}}(u,v) d\Lambda_\theta(z) \qquad \textbf{Eq. 3.7}$$

$C_\theta(u,v)$ are called the convex sum of $\{C_z\}$ with respect to $\Lambda_\theta(z)$. This equation can be extended by replacing the $C_{\mathbf{Z}}(u,v)$ by more general bivarition distribution functions. Set

$$H(u,v) = \int_0^\infty F^z(u) G^z(v) d\Lambda_\theta(z) \qquad \textbf{Eq. 3.8}$$

where $\Lambda_\theta(0) = 0$.

Example

Marshall and Olkin (1988) provided an example how convex sums can lead to copulas constructed from Laplace transforms of distribution functions.Let $\varphi(t)$ denote the Laplace transform of the mixing distribution $\Lambda_\theta(z)$, i.e., $\varphi(t) = \int_0^\infty e^{-zt} d\Lambda(z)$. Let F and G be the marginal distributions given by $F(u) = e^{-\varphi^{-1}(u)}$ and $G(v) = e^{-\varphi^{-1}(v)}$, then

$$
\begin{aligned}
H(x,y) &= \int_0^{\infty} [e^{-\varphi^{-1}(u)}]^z [e^{-\varphi^{-1}(v)}]^z d\Lambda_{\theta}(z) \\
&= \int_0^{\infty} \exp\left\{-z[\varphi^{-1}(u)+\varphi^{-1}(v)]\right\} d\Lambda_{\theta}(z) \\
&= \varphi[\varphi^{-1}(u)+\varphi^{-1}(v)]
\end{aligned}
\qquad \textbf{Eq. 3.9}
$$

Marshall and Olkin (1988) showed this is a joint distribution of X and Y. Thus,

$$C(u,v) = \varphi[\varphi^{-1}(u)+\varphi^{-1}(v)] \qquad \textbf{Eq. 3.10}$$

and $\varphi[\varphi^{-1}(u)+\varphi^{-1}(v)]$ is also called Archimedean copula.

Algebraic methods

Algebraic methods use the algebraic relationship between the joint probability and its univariate margins to derive the copulas. Then this relationship can be expressed in terms of a dependence parameter. Nelsen (2006) addressed two examples of copulas generation with this methods, Plackett and Ali – Mikhail – Haq distributions.

Example 1 Placekett Distributions

Consider two random variables X and Y with margins $F(x)$ and $G(y)$, respectively, and let the joint continuous distribution function be $H(x,y)$, the copula be C, and θ be the dependence parameter. Consider Table 3-1,

Table 3-1 2×2 contingency table for Algebraic methods

		Y variable	
		$y \le Y$	$y > Y$
X variable	$x \le X$	a	B
	$x > X$	c	D

It holds that

$$a = P(x \le X, y \le Y) = H(x,y) \qquad \textbf{Eq. 3.11}$$

$$b = P(x \le X, y \ge Y) = F(x) - H(x,y) \qquad \textbf{Eq. 3.12}$$

$$c = P(x > X, y \le Y) = G(y) - H(x,y) \qquad \textbf{Eq. 3.13}$$

$$d = P(x \le X, y \le Y) = 1 - F(x) - G(y) + H(x, y)$$ **Eq. 3.14**

The dependence parameter θ is defined as the cross product ratio, or odds ration,

$$\theta = \frac{ad}{bc} = \frac{H(x, y)[1 - F(x) - G(y) + H(x, y)]}{[F(x) - H(x, y)][G(y) - H(x, y)]}$$ **Eq. 3.15**

Using the relationship $u = F(x)$ and $v = G(y)$ and Skar's theorem, the equation can be rewrite as

$$\theta = \frac{C(x, y)[1 - u - v + C(x, y)]}{[u - C(x, y)][v - C(x, y)]}$$ **Eq. 3.16**

Solving the above equation for C,

$$C(u, v) = \begin{cases} \dfrac{[1 + (\theta - 1)(u + v)] - \sqrt{[1 + (\theta - 1)(u + v)]^2 - 4uv\theta(\theta - 1)}}{2(\theta - 1)} & \theta \ne 1 \\ uv & \theta = 1 \end{cases}$$ **Eq. 3.17**

Example 2 Ali – Mikhail – Haq distributions

First suppose variable X and Y are independent, then $H(x, y) = F(x)G(y)$.The odds ratio is defined as

$$\begin{aligned} \frac{1 - H(x, y)}{H(x, y)} &= \frac{1 - F(x)G(y)}{F(x)G(y)} = \frac{1}{F(x)}\frac{1}{G(y)} - 1 \\ &= [1 + \frac{1 - F(x)}{F(x)}][1 + \frac{1 - G(y)}{G(y)}] - 1 \\ &= \frac{1 - F(x)}{F(x)} + \frac{1 - G(y)}{G(y)} + \frac{1 - F(x)}{F(x)}\frac{1 - G(y)}{G(y)} \end{aligned}$$ **Eq. 3.18**

Based on the odds ratio in the independent case, Ali, Mikhail, and Haq (1978) proposed a generalized bivariate ratio with a dependence parameter θ,

$$\frac{1 - H(x, y)}{H(x, y)} = \frac{1 - F(x)}{F(x)} + \frac{1 - G(y)}{G(y)} + (1 - \theta)\frac{1 - F(x)}{F(x)}\frac{1 - G(y)}{G(y)}$$ **Eq. 3.19**

thus, using the relationship $u = F(x)$ and $v = G(y)$ and Skar's theorem, the equation can be rewrite as

$$\frac{1-C(x,y)}{C(x,y)}=\frac{1-u}{u}+\frac{1-v}{v}+(1-\theta)\frac{1-u}{u}\frac{1-v}{v}$$ **Eq. 3.20**

Solving C from the equation, the copula C can be obtained

$$C(x,y)=\frac{uv}{1-\theta(1-u)(1-v)}$$ **Eq. 3.21**

3.2.2.3 Archimedean copulas

Archimedean copulas have been used in a wide range of application because they are easily generated and are capable of capturing wide ranges of dependence.

Consider continuous function φ with the properties, 1) $\varphi(1)=0$; 2) $\varphi(0)=\infty$; 3) $\varphi^{'}(t)<0$; 4) $\varphi^{''}(t)>0$, for all $t\in(0,1]$. These properties ensure φ to be a decreasing convex function and the inverse function φ^{-1} exits. The bivariate Archimedean copulas take the form: $C_\theta(u,v)=\varphi^{-1}[\varphi(u)+\varphi(v)]$, where θ refers to the dependence parameter.

The joint density function c can be derived by differentiating with respect to the two variables U and V.

$$c(u,v)=\frac{\partial^2(u,v)}{\partial u\partial v}$$ **Eq. 3.22**

It can also be expressed in term of x and y,

$$f(x,y)=c(u,v)\frac{\partial u}{\partial x}\frac{\partial v}{\partial y}$$ **Eq. 3.23**

The conditional joint function takes the form of

$$C_{\theta,U|V=v}(u|V=v)=\frac{\partial C_\theta(u,v)}{\partial v}\Big|_{V=v}$$ **Eq. 3.24**

$$C_{\theta,U|V}(u|V\le v)=\frac{C_\theta(u,v)}{v}$$ **Eq. 3.25**

Many copulas have been reported in the literature. Nelsen (2006) summarized the bavariate copulas and their corresponding properties. Here just the often used four copulas in empirical application are discussed. See Nelsen (2006) for the description of more copulas.

Ali-Mikhail-Haq copula

Let $\varphi(t) = \ln\frac{(1-\theta t)}{t}$, then Ali-Mikhail-Haq copula and the joint density function take the form of

$$C_\theta(u,v) = \frac{uv}{1-\theta(1-u)(1-v)} \quad \theta \in [-1,1)$$ **Eq. 3.26**

$$c_\theta(u,v) = \frac{[1-\theta(1-u)(1-v)](1-\theta)+2\theta uv}{[1-\theta(1-u)(1-v)]^3} \quad \theta \in [-1,1)$$ **Eq. 3.27**

Clayton copula

Let $\varphi(t) = \frac{t^{-\theta}-1}{\theta}$, the Clayton copula and the joint density function can be expressed as

$$C_\theta(u,v) = (u^{-\theta}+v^{-\theta}-1)^{-1/\theta} \quad \theta \geq 0$$ **Eq. 3.28**

$$c_\theta(u,v) = (uv)^{-\theta-1}(\theta+1)(u^{-\theta}+v^{-\theta}-1)^{-\frac{1}{\theta}-2}$$ **Eq. 3.29**

This copula is also called Cook and Johnson family (Nelsen, 2006). As θ approaches zero, the marginals become independent. The Clayton copula cannot account for negative dependence. It shows strong left tail dependence and relatively weak right tail dependence. When correlation between two events is strongest in the left tail of the joint distribution, Clayton is an appropriate modeling choice.

Frank copula

Let $\varphi(t) = -\frac{e^{-\theta t}-1}{e^{-\theta}-1}$, the Clayton copula and the joint density function can be expressed as

$$C_\theta(u,v) = -\frac{1}{\theta}\ln\left[1+\frac{(e^{-\theta u}-1)(e^{-\theta v}-1)}{e^{-\theta}-1}\right] \quad \theta \neq 0$$ **Eq. 3.30**

$$c_\theta(u,v) = \frac{\theta e^{-\theta(u+v)}(e^{-\theta}-1)}{(e^{-\theta(u+v)}-e^{-\theta u}-e^{-\theta v}e^{-\theta})^2}$$ **Eq. 3.31**

When θ approaches zero, the marginal distributions are independent. The Frank copula permits both negative and positive dependence between the marginals and the dependence is symmetric in both tails. Because of it properties, the Frank copula have been widely used in empirical applications.

Gumbel-Hougaard copula

Let $\varphi(t) = (-\ln t)^{\theta}$, the Clayton copula can be expressed as

$$C_{\theta}(u,v) = \exp\left\{-\left[(-\ln u)^{\theta} + (-\ln v)^{\theta}\right]^{1/\theta}\right\} \quad \theta \geq 1 \qquad \textbf{Eq. 3.32}$$

Similar to the Clayton copula, Gumbel copula does not allow negative dependence, but it contrast to Clayton, Gumbel exhibits strong right tail dependence and relatively weak left tail dependence. If outcomes are known to be strongly correlated at high values but less correlated at low values, then the Gumbel copula is an appropriate choice.

3.2.2.4 Parameter estimation of copulas

Several methods for the copula parameters estimation have been proposed and applied, which include the maximum likelihood approach (ML), the sequential two-step maximum likelihood method (TSML), inference function for margins (IFM),Bayesian approach, and the approach based on the rank correlation.

Maximum likelihood approach (ML)

ML is a direct method to estimate the parameters, which involves the following steps,

1. For the two variables X and Y, pick the PDF $f_X(x;\alpha)$ and $f_Y(y;\beta)$, respectively, where α and β are the parameters of $f_X(x)$ and $f_Y(y)$, which include $\alpha_1, \alpha_2, ...\alpha_i, ...\alpha_m$, $i \in [1,m]$ and $\beta_1, \beta_2, ...\beta_j, ...\beta_n$, $j \in [1,n]$,respectively, where m and n are the number of the parameters in $f_X(x)$ and $f_Y(y)$, respectively. In this step, just the type of marginals are needed to identified but it is no need to estimate the parameters α and β in this step.
2. Select an assumed copula, and express the copula $C_{\theta}(u,v;\alpha,\beta,\theta)$ in terms of α, β, θ, u, and v, where θ is the dependence parameter in copula, u and v are the CDF of X and Y, respectively.
3. Derive the copula density function

$$f_{X,Y}(x,y;\alpha,\beta,\theta)=c_\theta(u,v:\alpha,\beta,\theta)\frac{\partial u}{\partial x}\frac{\partial v}{\partial y}=c_\theta(u,v;\alpha,\beta,\theta)\,f_X(x;\alpha)\,f_Y(y;\beta)$$ **Eq. 3.33**

where $c_\theta(u,v;\alpha,\beta,\theta)=\dfrac{\partial^2(u,v)}{\partial u\partial v}$ **Eq. 3.34**

4. Write the log-likelihood function,

$$\ln L[f_{X,Y}(x,y;\alpha,\beta,\theta)]=\sum_{k=1}^{K}\ln c_\theta(F_X(x_k;\alpha),F_Y(y_k;\beta);\theta)+\sum_{k=1}^{K}[\ln f_X(x_k;\alpha)+f_Y(y_k;\beta)]$$

Eq. 3.35

where K is the number of the observations,

5. Let $\Omega=(\alpha,\beta,\theta)$, solve

$$\frac{\partial\ln L[f_{X,Y}(x,y;\alpha,\beta,\theta)]}{\partial\Omega}=0$$ **Eq. 3.36**

for $\Omega=(\alpha,\beta,\theta)$. Therefore, the parameters are determined.

This method is efficient and consistent. However, in most time, the method involves solving of the nonlinear system and numerical algorithms need to be used.

Maximization-by-parts approach

Song et al. (2005) proposed modified maximum likelihood estimation, called maximization-by-parts (MBP) approach, by estimating the parameters for a bivariate Gaussian copula. The following steps are involved in this approach.

1. By observing the log-likelihood function,
2. $\ln L[f_{X,Y}(x,y;\alpha,\beta,\theta)]=\sum_{k=1}^{K}\ln c_\theta(F_X(x_k;\alpha),F_Y(y_k;\beta);\theta)+\sum_{k=1}^{K}[\ln f_X(x_k;\alpha)+f_Y(y_k;\beta)]$

   Eq. 3. 3-1

it can be rewrite as $\ln L[f_{X,Y}(x,y;\alpha,\beta,\theta)]=L_1+L_2$

where

$$L_1=\sum_{k=1}^{K}[\ln f_X(x_k;\alpha)+f_Y(y_k;\beta)]$$ **Eq. 3.37**

$$L_2=\sum_{k=1}^{K}\ln c_\theta(F_X(x_k;\alpha),F_Y(y_k;\beta);\theta)$$ **Eq. 3.38**

It can seen that in L1, just α and β are involved, and in L2, all α , β and θ are involved. Let $\mathrm{P}=(\alpha,\beta)$, and Solve $\frac{\partial L_1}{\partial \mathrm{P}} = 0$ for P .

3. Using the result of P as initial estimate P_0 , solve

$$\frac{\partial \ln L[f_{X,Y}(x,y;\mathrm{P}_0,\theta)}{\partial \theta} = 0 \text{ for } \theta_0 \qquad \textbf{Eq. 3.39}$$

4. Plug θ_0 into the log-likelihood function and obtain $\ln L[f_{X,Y}(x,y;\alpha,\beta,\theta_0)]$, solve for P_1 based on

$$\frac{\partial \ln L[f_{X,Y}(x,y;\mathrm{P},\theta_0)]}{\partial \mathrm{P}} = 0 \qquad \textbf{Eq. 3.40}$$

5. Plug P_1 into the log-likelihood function and solve for θ_0 .

By this pattern, the parameters can be estimated.

Bayesian approach

Wang (2001) proposed this approach for the parameter estimation for a bivairate extreme value distribution, in the inference analysis of the flood at two stations in Australia. See Wang (2001) for detail of this approach.

Two-step likelihood method

For the two variables X and Y, the empirical CDFs $F_X(x)$ and $F_Y(y)$ are computed. Let $u_1 = F_X(x_1),...,u_i = F_X(x_i),\ \mathrm{i} \in [1,\mathrm{N}],\ v_1 = F_Y(y_1),...,v_i = F_Y(y_i),\ \mathrm{i} \in [1,\mathrm{N}]$. Give u and v, for a specified copula, the estimate $\hat{\theta}$ of dependence parameter θ can be obtained by

$$\hat{\theta} = \arg\max_{\theta} \sum_{i=1}^{N} \ln C(u_i, v_i; \theta) \qquad \textbf{Eq. 3.41}$$

Inference function for margins (IFM)

Joe (1997) proposed this method and Shiau (2006) employed this method to estimate the copula with exponential and gamma margins in the drought duration and severity analysis for a gauge station in Taiwan. The basic idea of this method is to separate the estimation of the dependence parameter from the estimation of marginal parameters. So it is easy to employ and saves computation effort, especially when the ML is difficult to solve all the parameters simultaneously, for example, when there is a large dimension of parameters. Basically, the following two steps are involved,

1. Based on the log-likelihood functions of the two margins, the parameters of α and β are estimated for the PDF of PDF $f_X(x;\alpha)$ and $f_Y(y;\beta)$, respectively, where α and β are the parameters of $f_X(x)$ and $f_Y(y)$, which may include $\alpha_1, \alpha_2, ...\alpha_i, ...\alpha_m$, $i \in [1, m]$ and $\beta_1, \beta_2, ...\beta_j, ...\beta_n$, $\mathrm{j} \in [1, n]$,respectively,
2. Using the estimated α and β, solve the full log-likelihood function

$$\ln L[f_{X,Y}(x, y;\alpha,\beta,\theta)] = \sum_{k=1}^{K} \ln c_\theta(F_X(x_k;\alpha), F_Y(y_k;\beta);\theta) + \sum_{k=1}^{K}[\ln f_X(x_k;\alpha) + f_Y(y_k;\beta)]$$

for the dependence parameter θ.

Approach based on the rank correlation

This approach is on the basis of the relationship between the dependence parameter θ and the rank correlation coefficient. The two commonly used measures of correlation are Kendall's tau τ and Spearman's rho ρ_s, both of which have a range between -1 and 1.

Kendall's tau and Spearman's rho can be calculated by the following equations.

For random variable X and Y, let $\{x_1, x_2, ..., x_n\}$ and $\{y_1, y_2, ..., y_n\}$ donate n observations of X and Y, respectively, then

$$\tau = \binom{\mathrm{n}}{\mathrm{c}}^{-1} (n_c - n_d) \quad \text{or} \quad \tau = \binom{\mathrm{n}}{\mathrm{c}}^{-1} \sum_{i<j} \mathrm{sign}[(x_i - x_j)(y_i - y_j)] \quad \mathrm{i,j} \in [1,\mathrm{n}] \qquad \textbf{Eq. 3.42}$$

where n_c and n_d are the number of concordance pairs and discordance pairs, respectively, and n is the number of observations.

$$\rho_s(x, y) = \frac{\sum_{i=1}^{n} R(x_i)R(y_i) - n(\frac{n+1}{2})^2}{\left(\sum_{i=1}^{n} R(x_i)^2 - n(\frac{n+1}{2})^2\right)^{1/2} \left(\sum_{i=1}^{n} R(y_i)^2 - n(\frac{n+1}{2})^2\right)^{1/2}} \qquad \textbf{Eq. 3.43}$$

where $R(x)$ and $R(y)$ are the ranks (ascendingly ordered) of a pair of variables (x and y).

Nelsen (2006) expresses the Kendall's tau and Spearman's rho in terms of copulas as follows,

$$\tau = 4\iint_{I^2} C(u,v)dC(u,v) - 1 \qquad \textbf{Eq. 3.44}$$

Table 3-2 Kendall's tau and Spearman's rho for the often used four copulas

Copula type	Kendall's tau	Spearman's rho
Ali-Mikhail-Haq	$\tau = \frac{3\theta-2}{3\theta} - \frac{2}{3}(1-\frac{1}{\theta})^2 \ln(1-\theta)$ $\tau \in [(5-8\ln 2)/3, 1/3]$ or [-.1817,0.3333]	$\rho_s = \frac{12(1+\theta)}{\theta^2}\operatorname{dilog}(1-\theta) - \frac{24(1-\theta)}{\theta^2}\ln(1-\theta) - \frac{3(\theta+12)}{\theta}$ $\rho_s \in [33-48\ln 2, 4\pi^2-39]$ or [-0.2711, 0.4784]
Clayton	$\tau = \frac{\theta}{\theta+2}$	Complicated form
Frank	$\tau = 1 - \frac{4}{\theta}[1 - D_1(\theta)]$	$\rho_s = 1 - \frac{12}{\theta}[D_1(\theta) - D_2(\theta)]$
Gumbel-Hougaard	$\tau = \frac{\theta-1}{\theta}$	No closed form

Note: 1. $D_k(x)$ is the Debye function, for any positive integer k,

$$D_k(x) = \frac{k}{x^2}\int_0^x \frac{t^k}{e^t - 1}dt$$

2. dilog(x)=$\int_1^x \frac{\ln t}{1-t}dt$

For Archimedean Copulas, Kendall's tau takes the form of $\tau = 1 + 4\int_0^1 \frac{\varphi(t)}{\varphi'(t)}dt$, where $\varphi(t)$ is the generator function. And the Spearman's rho is given by

$$\rho_s = 12\iint_{I^2} C(u,v)dudv - 3 \qquad \textbf{Eq. 3.45}$$

Based on the relationship between Kendall's tau, Spearman's rho and the copulas, the dependence parameter can be determined. Table 3-2 shows some of the dependence parameters by Kendall's tau and Spearman's rho.

Zhang and Singh (2006) employed this approach to estimate several copulas for the bivariate flood frequency analysis for one river in Canada and one river in US.

3.2.2.5 Copula selection

How to select an appropriate copula that fits the data best among the many copulas becomes an issue when the copulas are constructed or picked up from the proposed family and the parameters are estimated. Because copulas separate marginal distributions from dependence structures, the appropriate copula for a particular application is the one which best captures dependence features of the data. To identify a copula, two steps are basically involved. First, the univariate marginal distributions need to be identified appropriately with the technique of goodness-of-fit test, and the corresponding parameter need to be estimated with a appropriate technique, i.e. maximum likelihood method. The better the fit of the marginal, the more precisely the model can fit the dependence structure (Trivedi and Zimmer, 2005). Second, a specified copula needs to be identified. Several methods have been discussed in the literature, some of which are addressed here.

Genest and Rivest method

Genest and Rivest (1993) described a procedure for selection among bivariate Archimedean copulas. For random variables X and Y of size n, with CDF $F_X(x)$ and $F_Y(y)$, respectively, the corresponding copula is C(u,v) ,where u, v are the CDF of X and Y, respectively. The following steps are involved in Genest and Rivest method:

1. Let the random variable Z=Z(x,y) which had the property $K(z) = \Pr(Z \leq z)$, where K(z) is defined as $K(z) = z - \frac{\varphi(z)}{\varphi'(z)}$

where $\varphi(z)$ is the generator function and $\varphi'(z)$ donates the derivative of $\varphi(z)$ with respect to z. The appropriate generator function needs to be identified so as to identify the appropriate copula.

2. calculate the empirical copula, $\hat{K}(z)$,

Define the variable

$$z_i = \frac{\text{number of } (x_j, y_j) \text{ such that } x_j < x_i \text{ and } y_j < y_i}{n-1} \quad i, j = 1, 2, \ldots, n \qquad \textbf{Eq. 3.46}$$

Set the estimate of K such that $\hat{K}(z) =$ the portion of z_i's $\leq z$

3. Calculate Kendall's tau by

$$\tau=\begin{pmatrix} n \\ c \end{pmatrix}^{-1} \sum_{i<j} \text{sign}[(x_i - x_j)(y_i - y_j)] \quad \text{i,j} \in [1,\text{n}] \qquad \textbf{Eq. 3.47}$$

Calculate the dependence parameter θ according to the relationship between θ and Kendall's tau, and then the generator function $\varphi(z)$ corresponding to each copula is obtained for a specified copula.

4. Using $\varphi(z)$, calculate a parametric estimate of K by $K_{\varphi}(z) = z - \frac{\varphi(z)}{\varphi^{'}(z)}$ corresponding to each generator function.

5. Compare the $K_{\varphi}(z)$ with the nonparametric estimate $\hat{K}(z)$, and choose the generator function that has the closest difference between $K_{\varphi}(z)$ and $\hat{K}(z)$ as the appropriate one. This is can be determine by employing Q-Q plot or by minimizing the distance function $\int \hat{K}_{\varphi}(z) - \hat{K}(z) dK(z)$

AKaike information criterion (AIC)

AKaike information criterion (AIC) (Akaike, 1974) is also often applied to identify the appropriate copula, which is defined as

$$AIC(m) = -2\log(\text{maximum likelihood of the model}) + 2\text{m} \qquad \textbf{Eq. 3.48}$$

where m is the number of parameter being estimated, which is determined by the type of univariate marginal distributions and the copula parameters. The AIC has another format,

$$AIC(m) = n\ln(\text{MSE}) + 2\text{m} \qquad \textbf{Eq. 3.49}$$

where n is the number of observations, m donates the number of fitted parameters, and MSE is the mean square error,

$$\text{MSE}=\text{E}(x_c - x_0) = \frac{1}{n-m}\sum_{i=1}^{n}[x_{c,i} - x_{0,i}]^2 \qquad \textbf{Eq. 3.50}$$

where $x_{c,i}$ and $x_{0,i}$ donate the i-th theoretic value by the copula and the i-th observed value, respectively.

and Zhang (2006) applied AIC for copula identification.

Quadratic distance criterion

Define the empirical copula $\hat{C}$ as

$$\hat{C}(\frac{t_x}{T},\frac{t_y}{T})=\frac{1}{T}\sum_{t=1}^{T}1_{\{x_t\le x_{t_x},y_t\le y_{t_y}\}} \quad \textbf{Eq. 3.51}$$

where $1_{\{A\}}$ is the indicator function that equals 1 if the event A occurs, x_{t_x} and y_{t_y} are the t_x-th and t_y-th order statistics of X and Y variable, and T is the number of observations. The empirical copula is the proportion of elements from the sample that satisfies $x_t \le x_{t_x}$ and $y_t \le y_{t_x}$

The quadratic distance between two copulas C_1 and C_2 in a set of bivariate points A = {a_1, a_2, …, a_m} is defined as:

$$d(C_1,C_2)=\left\{\sum_{i=1}^{m}\left[C_1(a_i)-C_2(a_i)\right]^2\right\}^{1/2} \quad \textbf{Eq. 3.52}$$

Among the estimated parametric copulas C_i, the one closest to the empirical copula is the most appropriate choice.

3.3 Application examples

Two river basins are studied in this paper, Des Moines River near Stratford, IA, and Altamaha River near Baxley, GA.

3.3.1 Application for the Des Moines River basin near Stratford, IA

Two USGS gauge stations, USGS 05480500 (Station A) and USGS 05471000 (Station B), located in the upstream of Des Moines River basin near Stratford, IA, Des Moines River at Fort Dodge, IA and Boone River near Webster City, IA, are selected and 38 years (1968-2005) annual peak flow records for the two gauge stations are collected in order to estimate the flood at the confluence using the joint probability approach. The gate station names and locations are shown in Table 3-3. Near the confluence point of the tributaries, there is a gauge station located in Des Moines River basin near Stratford, IA, USGS 05481300 (Station C). The data from Station C are also collected and the conventional flood frequency analysis is performed based on the discharge data of the Station C for the verification of copula method application in the joint probability approach.

The test of goodness of fit indicates that the annual peak discharge of Station A follows a 3-parameter lognormal distribution and the annual peak discharge of Station B follow a 3-paramter gamma (Pearson III) distributions. The parameters associated with the

Table 3-3 USGS Gauge Stations located in the Des Moines River Basin

Station	A	B	C
Station Name	Des Moines River at Fort Dodge, IA	Boone River near Webster City, IA	Des Moines River near Stratford, IA
USGU Station No.	05480500	05481000	05481300

distribution of each Station are shown in Table 3-4. The 3-parameter lognormal distribution and the 3-parameter gamma distribution are given by the following PDF and CDF.

3-paramter log-normal distribution.

PDF: $$f(y) = \frac{1}{\sqrt{2\pi}\sigma_y} \exp[-\frac{1}{2}(\frac{y-\mu_y}{\sigma_y})^2]$$ **Eq. 3.53**

CDF: $$F(y) = \frac{1}{\sqrt{2\pi}\sigma_y} \int_{-\infty}^{x}\int_{-\infty}^{y} \exp[-\frac{1}{2}(\frac{y-\mu_y}{\sigma_y})^2] dydx$$ **Eq. 3.54**

where $y = \ln(x-\gamma)$, μ_y, and σ_y are the mean and standard deviation of y, respectively, and γ is the threshold parameter.

3-paramter gamma distribution is adopted and the following format is used in this research,

PDF: $$g(x) = \frac{1}{\alpha^{\beta}\Gamma(\beta)}(x-\gamma)^{\beta-1} e^{-(x-\gamma)/\alpha} \quad x \geq \gamma$$ **Eq. 3.55**

CDF: $$G(x)=\frac{1}{\alpha^{\beta}\Gamma(\beta)}\int_{\gamma}^{x}(x-\gamma)^{\beta-1}e^{-(x-\gamma)/\alpha}dx \quad \mathrm{x}\geq\gamma$$ **Eq. 3.56**

where α , β and γ are the shape parameter, scale parameter and threshold parameter, respectively.

Table 3-4 Annual peak flow distribution information of Des Moines river basin

		Station A	Station B	Station C
Data Mean		13087.1	6303.42	18564.7
Data Standard deviation		7063.02	3432.53	9402.36
Correlation coefficient of A and B		0.804		
3-parameter log-normal	Location	9.96384		
	Scale	0.31125		
	Threshold	-9203.6		
Pearson III	Shape		3.09469	5.38697
	Scale		2004.594	4109.859
	Threshold		99.82	-3574.96

The four one-parameter copula families are assumed and the corresponding dependence parameters listed in Table 3-5. The Kendall's tau is calculated first and the Ali_Mikhail-Haq copula is denied since the Kendall 's tall of this copula need to be within the range of [-0.18, 1/3] while the Kendall's tau is 0.6885 in this research. The IFM method with the rank correlation method is applied for the parameter estimation. The parameters for the margins are estimated as shown in Table 3-4 and then the dependence parameter is estimated from Kendall's tau. And AIC criterion is adopted for the copula selection, as show in **Error! Reference source not found.**. AIC value shows that Frank copula fits the data best among Clayton copula, Frank copula and Gumbel-Hougaard copula. So the joint CDF is given by Frank copula as follows

$$C_{\theta}(u,v)=-\frac{1}{\theta}\ln\left[1+\frac{(e^{-\theta u}-1)(e^{-\theta v}-1)}{e^{-\theta}-1}\right] \qquad \theta=9.3428$$ **Eq. 3.57**

where u=F(x) and v=G(y), where F(x) is the CDF of 3-parameter lognormal distributions and G(x) is the CDF of 3-parameter gamma distribution. The corresponding conditional copula is given by Eq. 3.25.

Table 3-5 Dependence parameter for each copula and AIC values

	Ali-Mikhail-Haq	Clayton	Frank	Gumbel-Hougaard
Kendall's tau	0.6472			
Dependence parameter	Not available	3.6689	9.3428	2.8345
ML value	Not available	15.98	21.96	18.15
AIC	Not available	-3.54	.-4.18	-3.80

Based on the conditional copula, the Monte Carlo simulation is performed and 5000 pairs random number of u and v are generated first for Station A and B. The stream annual peak discharge is then calculated from $x=F^{-1}(u)$ and $y=G^{-1}(v)$. The synthetic confluent annual discharge is then given by z=x+y. The test of goodness fit implies that the synthetic confluence discharge follow a Pearson III distribution with the scale, shape and threshold parameter of 4.83202, 4553.2963 and -2703.5513, respectively. The simulation flood can be obtained by the conventional flood frequency analysis for the confluence point with the values of z.

To verify the joint probability model, the conventional flood frequency analysis is performed for the Station C based on the observation data. The goodness-of-fit test shows the Pearson III distributions fits the observation data of station C with the corresponding parameters shown in **Error! Reference source not found.**. The estimation results are given in **Error! Reference source not found.**and **Error! Reference source not found.**. The results from NFF model by USGS are also included in this study for the comparison with the joint probability model by using copula method, as shown in **Error! Reference source not found.**and **Error! Reference source not found.**.

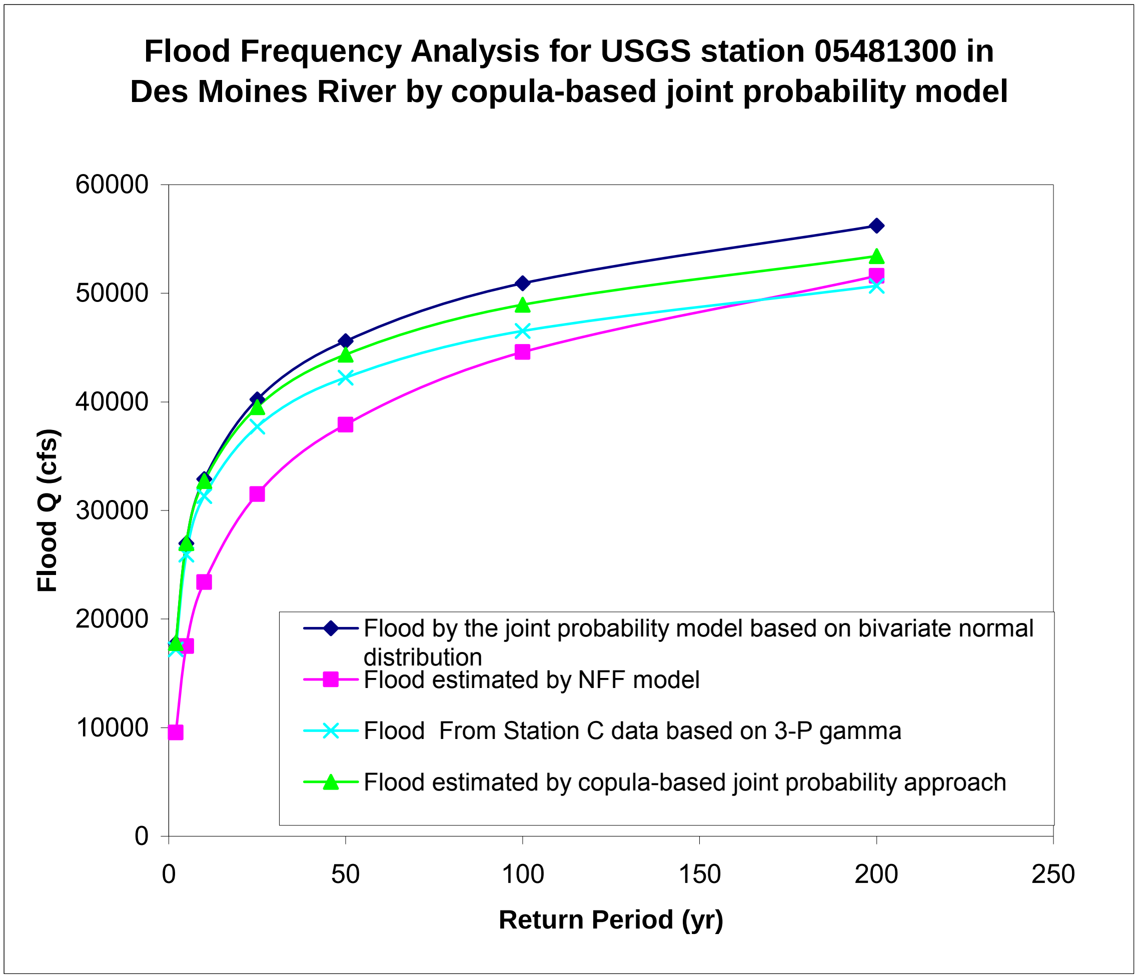

Figure 3-2 Simulation results by joint probability approach and comparison with NFF model, empirical bivariate approach and Copula: Des Moines River

Table 3-6 Simulation results comparison with NFF model and observation data

Return Period (yr)	Flood by the joint probability model based on empirical bivariate normal distribution		Flood by the copula-based joint probability model		Flood by NFF model		Flood estimated from Station C data based on Pearson III (cfs)
	Flood (cfs)	Relative error to observation data (Pearson III)	Flood (cfs)	Relative error to observation data (Pearson III)	Flood (cfs)	Relative error to observation data (Pearson III)	
2	17741	0.031	17800	0.034	9550	-0.445	17211
5	26951	0.039	26984	0.040	17500	-0.325	25941
10	32880	0.050	32701	0.044	23400	-0.253	31328
25	40222	0.066	39525	0.048	31500	-0.165	37727
50	45603	0.080	44345	0.050	37900	-0.103	42230
100	50924	0.094	48963	0.052	44600	-0.042	46533
200	56229	0.109	53426	0.054	51600	0.018	50685
Total relative error		0.469		0.322		1.351	

Note: Relative error $= \frac{x_e - x_0}{x_0}$ where x_e is the predicted value, x_0 is the observed value.

$$\text{Total relative error} = \sum_{i=1}^{n} \left| \frac{x_{e,i} - x_{0,i}}{x_{0,i}} \right|$$

3.3.2 Application for the Altamaha River basin near Baxley, GA

The second cased study is to apply the copula-based joint probability mode for the Altamaha River basin which is much larger than Des Moines River basin. In the Altamaha River basin, Oconee Rive at Dublin, GA and Ocmulgee River at Lumber City, GA are two tributary streams and Altamaha River near Baxley, GA is the confluent stream.The gauge station information for the three streams is shown in Table 3-7. The annual peak discharge of the two tributary gauge stations USGS 0223500 and USGS02215500 are collected, and the task is to estimate the flood frequency for the confluence point of the tributaries by assuming the annual peak discharge data of Station C are not available. Totally 35 year data (1971 through 2005) of each of the three stations are used in this research.

Table 3-7 USGS Gage Stations located in the Altamaha River basin

Station	Tributary A	Tributary B	Confluence C
Station Name	Oconee River at Dublin, GA	Ocmulgee River at Lumber City, GA	Altamaha River Near Baxley, GA
USGU Station No.	02223500	02215500	02225000

Table 3-8 Annual peak discharge distribution parameters: Altamaha River basin

		Station A	Station B	Station C
Data Mean		32315.4	29612.3	58671.4
Data Standard deviation		17024.9	16956.3	25488.2
Correlation coefficient of A and B		0.6404		
3-parameter log-normal	Location	10.67145		
	Scale	0.36144		
	Threshold	-13667		
Pearson III	Shape		2.58517	8.58058
	Scale		10124	8528.94931
	Threshold		3438	-145119

The test of goodness of fit indicates that both Station A and Station B fit 3 parameter normal distribution and Pearson III distribution, respectively, with the parameters as shown in Table 3-8.

Following the same token with the Des Moines River basin, four one-parameter copula families are assumed and the corresponding dependence parameters listed in Table 3-9. Ali_Mikhail-Haq copula is not applicable for this case since the Kendall's tau is out of its normal range of [-0.18, 1/3]. The IFM method with the rank correlation method is applied for the parameter estimation. The parameters for the margins are estimated as shown in Table 3-8. and then the dependence parameter is estimated from Kendall's tau. And AIC criterion is adopted for the copula selection, as show in Table 3-9. AIC value shows that Frank copula fits the data best among Clayton copula, Frank copula and Gumbel-Hougaard copula. So the joint CDF is given by Frank copula as follows

Table 3-9 Dependence parameter AIC value for each copula

	Ali-Mikhail-Haq	Clayton	Frank	Gumbel-Hougaard
Kendall's tau	0.6885			
Dependence parameter	Not available	4.3981	10.9208	3.2103
ML value	Not available	13.88	23.54	14.44
AIC	Not available	-3.26	.-4.32	-3.34

$$C_\theta(u,v) = -\frac{1}{\theta}\ln\left[1+\frac{(e^{-\theta u}-1)(e^{-\theta v}-1)}{e^{-\theta}-1}\right] \qquad \theta = 10.9208 \qquad \textbf{Eq. 3.58}$$

where u=F(x) and v=G(y), where F(x) is the CDF of 3-parameter lognormal distributions and G(x) is the CDF of Pearson III distribution. The corresponding conditional copula is given by Eq. 3.25.

Based on the conditional copula, the Monte Carlo simulation is performed and 5000 pairs random number of u and v are generated first for Station A and B. The stream annual peak discharge is then calculated from $x=F^{-1}(u)$ and $y=G^{-1}(v)$. The synthetic confluent annual discharge is then given by z=x+y. The test of goodness fit implies that the synthetic confluence discharge follow a gamma distribution with the scale and shape parameter of 3.89691 and 15709.72, respectively. The simulation flood can be obtained by the conventional flood frequency analysis for the confluence point with the values of z.

To verify the performance of the joint probability approach, the observation data of the gauge station at the confluence, USGS 02205000, are also collected and the conventional flood frequency analysis is performed for this site for the verification of the results by the joint probability model. Also the result from NFF model and from the joint probability mode based on the bivariate normal distribution is employed here for the comparison with the joint probability model, as shown in Table 3-10and Figure 3-3.

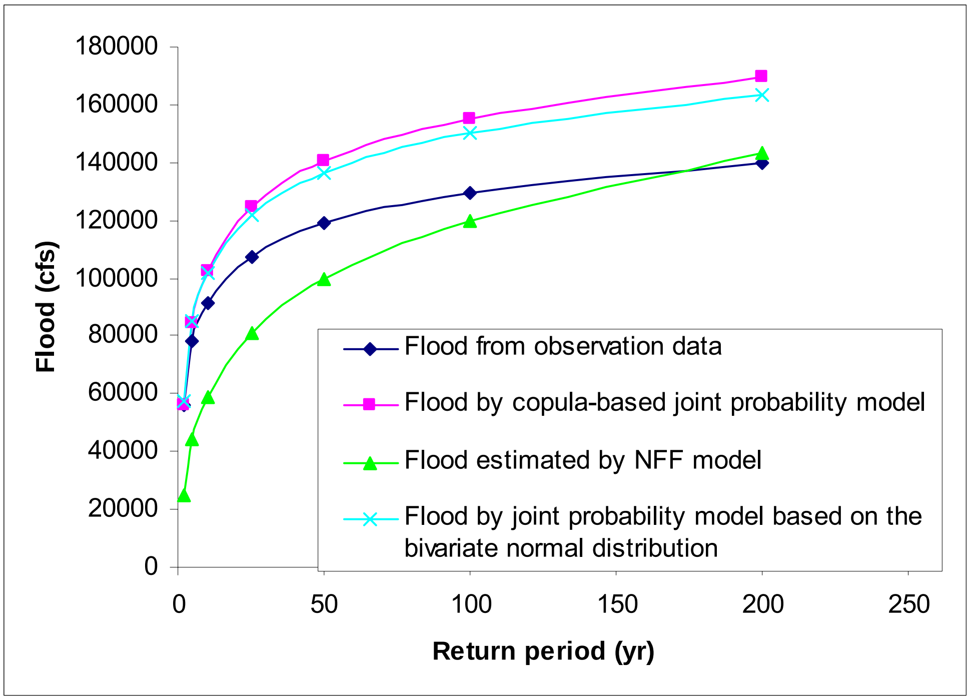

Figure 3-3 Flood simulation results comparison by the copula-based joint probability model and NFF model and the estimation from observation data for Altamaha River near Baxley, GA

Table 3-10 Flood simulation results comparison by the copula-based joint probability model and NFF model and the estimation from observation data:

Return Period (yr)	Flood by the joint probability model based on the bivariate normal		Flood by the copula-based joint probability model		Flood by NFF model		Flood estimated from Station C data based on Pearson III (cfs)
	Flood (cfs)	Relative error to observation data (Pearson III)	Flood (cfs)	Relative error to observation data (Pearson III)	Flood (cfs)	Relative error to observation data (Pearson III)	
2	57690	0.03	56070	0.003221	25200	-0.55	55890
5	84960	0.09	84660	0.083301	44000	-0.44	78150
10	101910	0.11	102790	0.122162	58800	-0.36	91600
25	122140	0.14	124660	0.159304	80700	-0.25	107530
50	136420	0.15	140230	0.180785	99600	-0.16	118760
100	150100	0.16	155220	0.198147	120000	-0.07	129550
200	163320	0.16	169770	0.212383	143000	0.02	140030
Total relative error		0.84		0.959303		1.85	

Note: Relative error $= \frac{x_e - x_0}{x_0}$ where x_e is the predicted value, x_0 is the observed value.

Total relative error $= \sum_{i=1}^{n} \left| \frac{x_{e,i} - x_{0,i}}{x_{0,i}} \right|$

3.3.3 Discussion

It is seen that the simulation results of the copula-based joint probability model for the Des Moines river basin is better than the results of the joint probability model based on the bivariate normal distribution of Station A and Station B, and the total relative error is much smaller than that of the NFF model. The largest error occurs at flood estimation for the largest return period; the accuracy is getting higher with the decrease of the return period. The performance of the copula-based joint probability model for the large river basin, such as Altamaha River basin in GA, is a little worse than that of the empirical bivariate normal distribution model. Still the relative error of the estimation for the large river basin is smaller than that of the NFF model that has a high tendency of underestimating the flood for both small river basin and large river basin when the return period is small, especially when the return period is smaller than 100 years. The accuracy of the copula-based joint probability model for large river basins could be increased by the following ways (1) carefully selecting the marginal distributions of the tributary streams, (2) selecting the best copulas from a wider range instead of from the only four often used copulas discussed in this study, (3) estimating more accurate parameters for the copulas by using a more consistent method, such as the maximum likelihood method for all the parameters including the parameters in the marginal distributions, (4) employing the variance reduction techniques.

Among the four copula families discussed in this study, Frank copula performs the best for the two cases studies of flood estimate using joint probability approach, while Ali-Mikhail-Haq is not applicable for the two cases at all due to the limit range of Kendall's tau and the high rank correlationship of the two tributaries in the two river basins.

3.4 Conclusion

A joint probability approach for the confluence flood frequency analysis is introduced briefly, and a method of multivariate distribution function estimation, Copula method is introduced in this paper which is one of the key part in the joint probability approach and could be very useful I hydrologic research. The joint probability approach provide a straightforward way to estimate the confluence flood at a acceptable accuracy without the discharge records needed for the mainstream and without tedious computation like flow routing. Copula method for the multivariate distribution function is more powerful in that it release the assumption in most empirical multivariate distribution functions that the margins follow the same type of distributions are, and it avoid the complex format of the multivariate distribution functions in many empirical formulas. The copulas are also easy to apply.

The four often used Archimedean copulas are introduced and applied in a small river basin and one large river basin for the joint probability estimation in the joint probability approach for the confluence flood frequency analysis. Frank copula and Gumbel-Hougaard copula perform better the other two, while Ali-Mikhail-Haq is not applicable in the study of the two river basins, because the high rank correlationship is beyond its Kendall's tau range. The case study shows that the copula-based joint probability approach for the confluence flood estimation performs well for the small river basin but has a relative large error when applied for the large river basin. Several techniques may be used to reduce the error, which include but not limited, (1) a more accurate estimation of the marginal distributions; (2) a more consistent and accurate estimation of the parameters in the copulas; (3) copula selection from a wider range options; and (4) variance reduction techniques.

Although several criteria have been proposed for the selection of a appropriate copula among several candidacies, there is no general criteria to verify the validity of the copulas, like the goodness-of-fit test for the univariate distributions, well developed, widely accepted, and widely applied in engineering. There is the possibility that none of the assumed the copulas perform well enough so that some other copulas need to be examined. So a general verification criterion is needed to verify the validity of the proposed copulas in application.

References

Akaike, H. (1974), A new look at the statistical model identification, *IEEE Trans. Autom. Control, AC*, **19**, 716–722.

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. and Kottegoda, N. T. (1994). Bivariate exponential model applied to intensities and durations of extreme rainfall. *Journal of hydrology,* **155,**225.

Beersma, J. J. and Buishand, T. A., (2004), The joint probability of rainfall and runoff d eficits in the Netherlands. *Proceedings, EWRI 2004 World Water & Environmental Resources Congress 2004*, June 27 - July 1, 2004, Salt Lake City, USA (CD-ROM).

Buishand, T. A. (1984), Bivariate extreme-value data and the station-year method. *Journal of Hydrology*, **69**, 77-95.

Chang, C. H., Tung, Y. K., and Yang, J. C. (1994), Monte Carlo simulation for correlated variables with marginal distributions, *Journal of Hydraulics Engineering*, **120,** 313-331.

Choulakain, V., Ei-Jabi, N., and Moussi, J. (1990), On the distribution of flood volume in partial duration series analysis of flood phenomena, *Stochastic Hydrology and Hydraulics,* **4,** 217-226.

Durrans, S. R., Eiffe, M. A., Thomas, Jr. W. O., and Goranflo, H. M. (2003), Joint seasonal/ annual flood frequency analysis, *Journal of Hydrologic Engineering*, **8**, 181-189.

Galambos, J. (1978), *The asymptotic theory of extreme order statistics*. Wiley, New York.

Gamlambos, J. (1978). *The asymptotic theory of extreme order statistics*. Wiley, New York.

Genest, C., and Rivest L-P. (1993), Statistical inference procedures for bivariate Archimedean copulas, *Journal of American Statisticians Association*, **88**, 1034-1043.

Genest, C.,and MacKay, J. (1986), The joy of copulas: Bivariate distributions with uniform marginals, *American Statisticians Assicuation*, **40,** 280-283.

Goel, N. K., Seth, S. M., and Chandra, S. (1998), Multivariate modeling of flood flows, *Journal of Hydraulic Engineering*, **124,** 146-155.

Joe, H. (1997), *Multivariate mold and dependence concepts*, hapman & Hall, London.

Johnson, R. A., Evans, J. W., and Green, D. W. (1999), Some Bivariate Distributions for modeling the strength properties of lumber, U. S. Department of Agriculture, Research paper FOL-RP-575.

Krstanovic, P. F., and Singh, V. P. (1987), A multivariate stochastic flood analysis using entropy, in *Hydrologic Frequency Modelling*, edited by V. P. Singh, PP. 515–539, Reidel, Dordrecht, The Netherlands.

Marshall, A. W. and Olkin, I. (1988), Families of multivariate distributions, *Journal of the American Statistical Association*, **83**, 834-841.

Moran, PAP. (1970), Simulation and evaluation of complex water systems operations. *Water Resource Research*, **6**, 1737–1742.

Nadarajah, S. and Gupta, A. K. (2006), Intensity–duration models based on bivariate gamma distributions. Hiroshima mathematical journal, **36**, 387–395

Nelsen RB. 2006. *An Introduction to Copulas*, Second Edition, Springer-Verlag, New York.

Prekopa A, Szantai T. (1978), A new multivariate gamma distribution and its fitting to empirical streamflow data. *Water Resource Research*, **14**, 19–24.

Raynal, J.A., and Salas, J.D., (1987), A probabilistic model for flooding downstream of the junction of two rivers, in *Hydrologic frequency modeling*, edited by V. P. Singh, pp.595-602, Dordrecht, The Netherlands, D. Reidel Publishing Company.

Sackl, B., and Bergmann, H. (1987), A bivariate flood model and its application, in *Hydrologic frequency modeling*, edited by V. P. Singh, pp. 571-582, Dreidel, Dordrecht, The Netherlands.

Schweizer, B. (1991), *Thirty years of copulas. In: Dall's Aglio G, Kotz S, Salinetti G(eds) advances in probability distributions with given marginals*. Kluwer, Dorderecht. pp13-50

Shiau, J., Feng, S., and Nadarajah, S. (2007), Assessment of hydrological droughts for the Yellow River, China, using coplus. *Hydrological Processes*, **21**, 2157-2163.

Shih, J. H. and Louis, T. A. (1995), Inferences on the association parameter in copulas models for bivariate survival data. *Biometrics*, **51**, 1384-1399.

Song, P. X. K., Y. Fan, and J. D. Kalbfleisch (2005), Maximization by parts in likelihood inference, *Journal of the American Statistical Association*, **100**, 1145–1158.

Trivedi, P. K. and Zimmer, D. M. (2005), Copulas modeling: an introduction for practitioners, *Foundations and trends in econometrics*, **1**, 1-111.

U.S. Geological Survey (2002). The National Flood Frequency Program, Version 3: A Computer Program for Estimating Magnitude and Frequency of Floods for Ungaged Sites. U.S. Geological Survey Water-Resources Investgations Report 02-4168

Wang, Q. J. (2001). A Bayesian Joint probability approach for flood record augmentation. *Water Resources Research*, **37,** 1707-1712.

Yue S. (1999). Applying the bivariate normal distribution to flood frequency analysis. *Water International*, **24,** 248-252.

Yue, S. (2000). The bivariate lognormal distribution to model a multivariate flood episode. *Hydrological Processes*, **14,** pp. 2575-2588.

Yue, S. (2001a). A bivariate extreme value distribution applied to flood frequency analysis. *Nordic Hydrology*, **32**, 49–64.

Yue, S. (2001b) A bivariate gamma distribution for use in multivariate flood frequency analysis. *Hydrology processes*, **15,** 1033-1045.

Yue, S. and Rasmussen, P. (2002), Bivariate frequency analysis: discussion of some useful concepts in hydrological application, *Hydrological Process*, **16**, 2881–2898.

Yue, S. and Wang, C. Y. (2004), A comparison of two bivariate extreme value distributions, *Stochastic Environmental Research*, **18**, 61-66.

Yue, S. Ouarda, T. B. M. J., and Bobée, B, 2001. A review of bivariate gamma distributions for hydrological application. Journal of Hydrolgoy 246, 1-18

Zhang, L. and Singh, V. P. (2006), Bivariate flood frequency analysis using the copula method, *Journal of hydrologic engineering*, **11**, 150-164.

Chapter 4 Summary and Future Work

4.1 Summary

This thesis introduces a joint probability model of the confluence flood frequency analysis and the straight forward practical procedure is presented. Due to the difficulty estimating the joint probability and the limitation of the current often used empirical formulas for the joint probability estimation, which usually need to assume the marginals follow the same type of the distribution or just can handle limited case-by-case situations, a general method is addressed in this thesis for the joint probability estimation.

The performance of the model is examined by the application studies in two river basins and the results are compared with the observation data and the NFF model. It shows that the proposed model performs well, especially for the smaller river basin. The results from the model are very close to the observation data, especially at the low return periods. It has a tendency to overestimate the flood for each of the specified return period and lose the accuracy with the return periods getting large and the river basin getting large. NFF model works very well for the large return periods situation, especially when the return period is larger than 100 years. However, it tends to underestimate the flood for the relatively low return period, say less than 50 years.

The copula method provides a way to release the restriction of the empirical approach for the joint probability estimation and a more accurate estimation if an appropriate copula is used.

4.2 Future work

As found in the application examples in the thesis, the proposed model loses the accuracy when the river basin scale is getting large. Techniques to improve the accuracy by the joint probability model, especially for the large scale river basin is needed. The possible means may include but not limited: (1) carefully selecting the marginal distributions of the tributary streams, (2) selecting the best copulas from a wider range of copulas, (3) estimating more accurate parameters for the copulas by using a more consistent method, such as the maximum likelihood method for all the parameters including the parameters in the marginal distributions, (4) employing the variance reduction techniques, (5) using a two-parameter copula.

Currently annual peak flow data from the observation gauges are used for the flood frequency analysis; however, as discussed in chapter 1 and by other researchers (Kite, 1977; Rao and Hamed, 2000), this approach may miss some peak flow information. So the performance of flood frequency analysis based on the monthly discharge peak flow may be necessary to be examined.

Copulas have been studies for more than 40 years in statistics; however, the application of copulas just has a short history. Many concept of copula, the method to structure a copula have been well established. However, the selection of a valid copula from a bunch of reported copula seems an issue. Many researchers have proposed some methods and criteria for selecting a copula that fit the observation data best, which are commonly used currently, but it is just can identify the best performed copula among the predicted one, instead of a general criteria. Some researchers extended the Chi-square test from univariate distribution to multivariate distribution which may be employed to verify the validity of copulas, but there are not enough applications to demonstrate these criteria. Criteria to verify the validity of the predicted copulas is needed to be developed.

References

Kite, G. W., and Stuart, A. (1977), *Frequency and risk analysis in hydrology*, Water Resources public.Fort Collins, CO.

Rao, A. R. and Hamed, K. H., (2000). *Flood frequency analysis.* CRC Press, New York, NY.

References

Ang. H-S. A. and Tang, H. W. (1975a), *Probability Concepts in Engineering Planning and Design.* Volume I- Basic Principles. John Wiley & Sons, Inc., New York.

Ang. H-S. A. and Tang, H. W. (1975b), *Probability Concepts in Engineering Planning and Design*. Volume II- Decision, Risk, and Reliability. John Wiley & Sons, Inc., New York.

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.

Ashkar, F., and Rousselle, J. (1982), A multivariate statistical analysis of flood magnitude, duration and volume, in *Statistical analysis of rainfall and runoff*, edited by V. P. Singh, pp. 659–669, Water Resource publication, Fort Collins, Colo..

Bacchi, B., Becciu, G. and Kottegoda, N. T. (1994), Bivariate exponential model applied to intensities and durations of extreme rainfall, *Journal of hydrology*, **155,** 225-236.

Beersma, J. J. and Buishand, T. A. (2004), The joint probability of rainfall and runoff d eficits in the Netherlands. Proceedings, EWRI 2004 World Water & Environmental Resources Congress 2004, June 27 - July 1, 2004, Salt Lake City, USA (CD-ROM).

Blachnell, D. (1994), New method of the simulation of correlated K-distribution clutter, *EE Proceedings Radar, Sonar and Navigation,***141**, 53-58.

Buishand TA. (1984), Bivariate extreme-value data and the station-year method, *Journal of Hydrology*, **69**, 77–95.

Buishand, T. A. (1984), Bivariate extreme-value data and the station-year method. *Journal of Hydrology*, **69**, 77-95.

Burkhardt, G. and Prakash, A. (1976), An analysis of computer graphics to extremes, *Water Research Bulletin*, **12**, 1245-1258.

Chang, C. H., Tung, Y. K., and Yang, J. C. (1994), Monte Carlo simulation for correlated variables with marginal distributions, *Journal of Hydraulics Engineering*, **120,** 313-331.

Choulakain, V., Ei-Jabi, N., and Moussi, J. (1990), On the distribution of flood volume in partial duration series analysis of flood phenomena. *Stochastic Hydrology and Hydraulics,* **4,** 217-226.

Chow, V. T., Maidment, D. R. and Mays, L. W., (1988), *Applied hydrology*. McGraw-Hill, New York.

Clarke, R. T.(1980), Bivariate gamma distributions for extending annual streamflow records from precipitation: some large-sample results, *Water Resources Research*, **16**, 863–870.

Crovelli, R. A. (1973), A bivariate precipitation model, paper presented at 3rd Conference on Probability and Statistics in Atmospheric Science, American Meteorological Society, Boulder, CO, USA; 130–134.

Durrans, S. R., Eiffe, M. A., Thomas, Jr. W. O., and Goranflo, H. M. (2003), Joint seasonal/ annual flood frequency analysis, *Journal of Hydrologic Engineering*, **8**, 181-189.

Gamlambos, J. (1978). *The asymptotic theory of extreme order statistics*. Wiley, New York.

Genest, C., and Rivest L-P. (1993), Statistical inference procedures for bivariate Archimedean copulas, *Journal of American Statisticians Association*, **88**, 1034-1043.

Genest, C.,and MacKay, J. (1986), The joy of copulas: Bivariate distributions with uniform marginals, *American Statisticians Assicuation*, **40,** 280-283.

Gentile, J. E. (1998), *Random Number Generation and Monte Carlo Method,*. Springer-Verlag New York, Inc.

Goel, N. K., Seth, S. M., and Chandra, S. (1998), Multivariate modeling of flood flows, *Journal of Hydraulic Engineering*, **124,** 146-155.

Greenwood, J. A., Landwehr, J. M., Matalas, N. C., and Wallis, J. R. (1979), Probability weighted moments: definition and relation to parameters of several distributions expressible in inverse form, *Water Resources Research,* **15,** 1049-1054.

Gumbel, E. J., and Mustafi, C. K. (1967), Some analytical properties of bivariate extreme distributions, *Journal of American Statistician Association.* **62,** 569-588.

Hashino, M. (1985), Formulation of the joint return period of two hydrologic variates associated with a Possion process, *Journal of Hydroscience and Hydranlic Engineering*, **3,** 73-84.

He. H., Zhou, J., Yu, Q., Tian, Y. Q., and Chen, R. R. (2007), Flood frequencyand routing processes at a confluence of the middle yellow river in china, *River Research and Applications*, **23,** 407-427.

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

Hosking, J. R. M. (1993), Some statistics useful in regional frequency analysis, *Water Resources Research.***29,** 271.

Joe, H. (1997), *Multivariate mold and dependence concepts*, hapman & Hall, London.

Johnson, M. E. (1987), *Multivariate Statistical Simulation*. John Willy & Sons Inc., New York.

Johnson, R. A., Evans, J. W., and Green, D. W. (1999), Some Bivariate Distributions for modeling the strength properties of lumber, U. S. Department of Agriculture, Research paper FOL-RP-575.

Kite, G. W., and Stuart, A. (1977), *Frequency and risk analysis in hydrology*. Water Resources public.Fort Collins, CO.

Krstanovic, P. F., and Singh, V. P. (1987), A multivariate stochastic flood analysis using entropy, in *Hydrologic Frequency Modelling*, edited by V. P. Singh, PP. 515–539, Reidel, Dordrecht, The Netherlands.

Linsley, R. K. (1986), Flood estimates: how good are they? *Water Resources Research*, **22,** 1595-1645.

Marshall, A. W. and Olkin, I. (1988), Families of multivariate distributions, *Journal of the American Statistical Association*, **83**, 834-841.

Michael, J. R. and Schucany, W. R., (2002), The mixture approach for simulating bivariate distributions with specified correlations. *The American statistician*, **56**, 48-54.

Moharran, S. H., Gosain, A. K., and Kapoor, P. N. (1993), A comparative study for the estimators of the generalized pareto distribution. *Journal of hydraulics*, **150,** 169-185.

Molenberghs, G. and Lesaffre, E., (1997), Non-linear integral equations to approximate bivariate densities with given marginals and dependence function, *Statistica Sinica*, **7**, 713-738.

Moran, PAP. (1970), Simulation and evaluation of complex water systems operations. *Water Resource Research*, **6**, 1737–1742.

Nadarajah, S. and Gupta, A. K. (2006), Intensity–duration models based on bivariate gamma distributions. Hiroshima mathematical journal, **36**, 387–395

Natural Environmental Research Council (1975), Flood Studies Report, Volume I: Hydrological Studies, report, Natural Environmental Research Council, London.

Nelsen RB. 2006. *An Introduction to Copulas*, Second Edition, Springer-Verlag, New York.

Oklin, I. and Harshall, A. (1988), Families of multivariate distributions, *Journal of American Statistical association*, **83**, 834-841.

Prekopa A, Szantai T. (1978), A new multivariate gamma distribution and its fitting to empirical streamflow data. *Water Resource Research*, **14**, 19–24.

Rao, A. R. and Hamed, K. H., (2000). *Flood frequency analysis.* CRC Press, New York, NY.

Rao, A. R. and Hamed, K. H., (2000a), *Flood frequency analysis.* CRC Press, New York, NY, pp. 8.

Rao, A. R. and Hamed, K. H., (2000b), *Flood frequency analysis.* CRC Press, New York, NY, pp. 10.

Raynal, J.A., and Salas, J.D., (1987), A probabilistic model for flooding downstream of the junction of two rivers, in *Hydrologic frequency modeling*, edited by V. P. Singh, pp.595-602, Dordrecht, The Netherlands, D. Reidel Publishing Company.

Ronning, G. (1977), A Simple Scheme for Generating Multivariate Gamma Distributions with Non-Negative Covariance Matrix, *Technometrics*, **19**, 179-183.

Rossi, F. M. F. and Versace, P. (1984), Two-component extreme value distribution for flood frequency analysis. *Water Resource Research*, **20,** 847-856.

Sackl, B., and Bergmann, H. (1987), A bivariate flood model and its application, in *Hydrologic frequency modeling*, edited by V. P. Singh, pp. 571–582, Dreidel, Dordrecht, The Netherlands.

Schucany, W. R., Michael, J. R. (2002). The mixture approach for simulating bivariate distributions with specified correlations. *The American statistician*, **56**, 48-54.

Schweizer, B. (1991), *Thirty years of copulas. In: Dall's Aglio G, Kotz S, Salinetti G(eds) advances in probability distributions with given marginals*. Kluwer, Dorderecht. pp13-50

Shiau, J., Feng, S., and Nadarajah, S. (2007), Assessment of hydrological droughts for the Yellow River, China, using coplus, *Hydrological Processes*, **21**, 2157-2163.

Shih, J. H. and Louis, T. A. (1995), Inferences on the association parameter in copulas models for bivariate survival data. *Biometrics*, **51**, 1384-1399.

Singh, K., and Singh, V. P. (1991). Derivation of bivariate probability density functions with exponential marginals. *Stochastic Hydrology & Hydraulics*, **5**, 55-68.

Singh, V. P. and Singh, K. (1985a), Derivation of the Pearson Type (PT) III distribution by using the principle of maximum entropy (POME), *Journal of Hydraulics*, **80,** 197-214.

Singh, V. P. and Singh, K. (1985b), Derivation of gamma distribution by using the principle of maximum entropy. *Water Resources Bulletin,* **21,** 1185-1191.

Smith, O.E., Adelfang, and Tubbs, JD. S. I. (1982). A bivariate Gamma probability distribution with application to gust model. NASA Technical Memorandum 82483, National Aeronautics and Space Administration.

Song, P. X. K., Y. Fan, and J. D. Kalbfleisch (2005), Maximization by parts in likelihood inference, *Journal of the American Statistical Association*, **100**, 1145–1158.

Stuart, A., and Ord, J. K. (1987). Kendall's Advanced theory of statistics, Vol.1, 5th ed.,Distribution theory, Charles Griffin and Company, London. The first of three volumes of the classic reference book in statistics and probability-not for beginners.

Tod, D. K.(1957), Frequency analysis of streamflow data, *Journal of the Hydraulics Division*, Process ASCE HY1, pp. 1166-1, 1166-16.

U.S. Geological Survey (2002). The National Flood Frequency Program, Version 3: A Computer Program for Estimating Magnitude and Frequency of Floods for Ungaged Sites. U.S. Geological Survey Water-Resources Investgations Report 02-4168

Wang, Q. J. (2001). A Bayesian Joint probability approach for flood record augmentation. *Water Resources Research,* **37,** 1707-1712.

Yue S, Ouarda TBMJ, Bob´ee B. (2001), A review of bivariate gamma distributions for hydrological application. Journal of Hydrology 246: 1–18.

Yue S. (1999). Applying the bivariate normal distribution to flood frequency analysis. *Water International,* **24,** 248–252.

Yue, S. (2000). The bivariate lognormal distribution to model a multivariate flood episode. *Hydrological Processes*, **14,** 2575-2588.

Yue, S. (2001a), A bivariate extreme value distribution applied to flood frequency analysis, *Nordic Hydrology*, **32**, 49–64.

Yue, S. (2001b) A bivariate gamma distribution for use in multivariate flood frequency analysis. *Hydrology processes*, **15,** 1033-1045.

Yue, S. and Rasmussen, P. (2002), Bivariate frequency analysis: discussion of some useful concepts in hydrological application, *Hydrological Process*, **16**, 2881–2898.

Yue, S. and Wang, C. Y. (2004), A comparison of two bivariate extreme value distributions, *Stochastic Environmental Research*, **18**, 61-66.

Yue, S., Ouarda, T. B. M. J., and Bobée, B. (2001), A review of bivariate gamma distributions for hydrological application, *Journal of Hydrolgoy*, **246**, 1-18.

Zhang, L. and Singh, V. P. (2006), Bivariate flood frequency analysis using the copula method, *Journal of hydrologic engineering*, **11**, 150-164.

Appendix A. Methods for parameter estimation

Method of moments (MOM)

The j-th sample moment of a random variable is defined as,

$$E(x^j) = \frac{1}{n}\sum_{i=1}^{n} x_i^j E(x^j) = \frac{1}{n}\sum_{i=1}^{n} x_i^j \qquad \textbf{Eq. A.1}$$

where x are the sample observation values, n is the sample size.

The j-th moment of a random variable of a distribution with probability distribution function (PDF) of f(x), is defined as,

$$E(x^j) = \int_{-\infty}^{+\infty} x^j f(x)dx E(x^j) = \int_{-\infty}^{+\infty} x^j f(x)dx \qquad \textbf{Eq. A.2}$$

If there are k parameters to be estimated for the distribution, the following system of equations can be obtained based the equation of j-th moment,

$$E(x^j) = \int_{-\infty}^{+\infty} x^j f(x)dx = \frac{1}{n}\sum_{i=1}^{n} x_i^j \quad j = 1, 2, ..., k \qquad \textbf{Eq. A.3}$$

The k unknown parameters can be obtained by solving the above system of equations simultaneously.

Method of maximum likelihood (ML)

For a distribution with a PDF given by f(x) and parameters $\alpha_1, \alpha_2, ...\alpha_k$, the likelihood function is defined as

$$L(\alpha_1, \alpha_2, ..., \alpha_k) = \prod_{i=1}^{n} f(x_i; \alpha_1, \alpha_2, ..., \alpha_k) \qquad \textbf{Eq. A.4}$$

The best value of a parameter should be the value that maximizes the likelihood L of occurrence of the observed sample; hence, the parameters $\alpha_1, \alpha_2, ..., \alpha_k$ can be estimated by solving the following system of partial differentiation equations,

$$\frac{\partial L(\alpha_1, \alpha_2, ..., \alpha_k)}{\partial \alpha_j} = 0; \text{ j= 1, 2, ..., k} \qquad \textbf{Eq. A.5}$$

In many cases, it is more convenient to work with the log-likelihood function

$$\frac{\partial \ln L(\alpha_1, \alpha_2, ..., \alpha_k)}{\partial \alpha_j} = 0; \text{ j= 1, 2, ..., k} \qquad \textbf{Eq. A.6}$$

Method of probability weighted moments (PWM) and L-moments

Greenwood et at. (1979) initiated the probability weighted moments for a random variable X with CDF F(x), or simply F, as $M_{p,r,s} = E[x^p F^r (1-F)^s]$, where p, r and s are real numbers. As it can be seen, $M_{p,0,0}$ represents the conventional p-th moment of X. When p is a non-negative number while r or s is zero, two special cases, $M_{1,0,s}$ and $M_{1,r,0}$, are often considered,

$$\begin{aligned} M_{1,0,s} &= \alpha_s = E[xF(1-F)^s] \\ M_{1,r,0} &= \beta_r = E(xF^r) \end{aligned} \qquad \textbf{Eq. A.7}$$

hence

$$\alpha_s = \sum_{i=1}^{s} \binom{s}{i} (-1)^i \beta_i, \ \beta_r = \sum_{i=1}^{r} \binom{r}{i} (-1)^i \alpha_i \qquad \textbf{Eq. A.8}$$

The L-moments are defined as $\lambda_r = E[xP^*_{r-1}(F)]$, where $P^*_r(\bullet)$ is the r-th shifted Legendre polynomials. Hosking (1990) were initially introduced the relationship between L-moments and PWMs by the equation. L-moments are related to probability weighted moments by the equation

$$\lambda_{r+1} = \sum_{k=0}^{r} p^*_{r,k} \beta_k = (-1)^r \sum_{k=0}^{r} p^*_{r,k} \alpha_k \qquad \textbf{Eq. A.9}$$

where

$$p^*_{r,k} = (-1)^{r-k} \binom{r}{k} \binom{r+k}{k} \qquad \textbf{Eq. A.10}$$

For an ordered sample $x_1 <= x_2 <= ... <= x_n$, n>r and n>s, unbiased estimators of α_r, β_r and l_{r+1} (sample L-moments) by the following equations,

$$a_r = \hat{\alpha}_r = \hat{M}_{1,0,r} = \frac{1}{n} \sum_{i=1}^{n} x_i \binom{n-i}{r} / \binom{n-1}{r} \qquad \textbf{Eq. A.11}$$

$$b_r = \hat{b}_r = \hat{M}_{1,r,0} = \frac{1}{n} \sum_{i=1}^{n} x_i \binom{i-1}{r} / \binom{n-1}{r} \qquad \textbf{Eq. A.12}$$

$$l_{r+1} = \hat{\lambda}_{r+1} = \sum_{k=0}^{r} p^*_{r,k} b_k = (-1)^r \sum_{k=0}^{r} p^*_{r,k} a_k \qquad \textbf{Eq. A.13}$$

In particular,

$$\begin{aligned} \lambda_1 &= \alpha_0 &&= \beta_0 \\ \lambda_2 &= \alpha_0 - 2\alpha_1 &&= 2\beta_1 - \beta_0 \\ \lambda_3 &= \alpha_0 - 6\alpha_1 + 6\alpha_2 &&= 6\beta_2 - 6\beta_1 + \beta_0 \\ \lambda_4 &= \alpha_0 - 12\alpha_1 + 30\alpha_2 - 20\alpha_3 &&= 20\beta_3 - 30\beta_2 + 12\beta_1 - \beta_0 \end{aligned} \qquad \textbf{Eq. A.14}$$

Estimates based on PWMs and L-moments are generally superior to standard moment-based estimates. The L-moment estimators have some desirable properties for parameter estimation. In particular, they work well with small sample and the bias tends to be small. L-moment estimators can often be used when the maximum likelihood estimates are unavailable, difficult to compute, or have undesirable properties. They may also be used as starting values for maximum likelihood estimates.

L-moment ratios are defined by Hosking (1990) as

$$\begin{aligned} \tau &= \lambda_2 / \lambda_1 \\ \tau_r &= \lambda_r / \lambda_2 \quad \mathrm{r} \geq 3 \end{aligned} \qquad \textbf{Eq. A.15}$$

where λ_1 is a measure of location, τ is a measure of scale and dispersion ($L-C_v$), τ_3 is a measure of skewness ($L-C_s$), and τ_4 is a measure of kurtosis ($L-C_k$).

L-moment ratios can be estimated by the sample L-moments

$$\begin{aligned} t &= \hat{\tau} = l_2 / l_1 \\ t_r &= \hat{\tau}_r = l_r / l_2 \quad \mathrm{r} \geq 3 \end{aligned} \qquad \textbf{Eq. A.16}$$

Some useful L-moments ratios for the commonly used distributions in hydrology are given by Hosking (1993) and some other researchers,

Normal distribution: $\tau_3 = 0$, $\tau_4 = 0.1226$

Lognormal distribution (two and three papRameters):

$\tau_4 = 0.12282 + 0.77518\tau_3^{\ 2} + 0.12279\tau_3^{\ 4} - 0.13638\tau_3^{\ 6} + 0.11368\tau_3^{\ 8}$

Exponential distribution: $\tau_3 = 1/3$, $\tau_4 = 1/6$

Gamma and Pearson III distributions:

$\tau_4 = 0.1224 + 0.30115\tau_3^{\ 2} + 0.95812\tau_3^{\ 4} - 0.57488\tau_3^{\ 6} + 0.19383\tau_3^{\ 8}$

Generalized Extreme Value distribution

$\tau_4 = 0.10701 + 0.11909\tau_3 + 0.84838\tau_3^{\ 2} - 0.06669\tau_3^{\ 3} + 0.00567\tau_3^{\ 4}$- $0.04208\tau_3^{\ 5} + 0.03763\tau_3^{\ 6}$Gumbel distribution: $\tau_3 = 0.1699$, $\tau_4 = 0.1504$

Weibull distribution: $\tau_3 = 3 - 2\dfrac{1 - 3^{\frac{\ln(1-\tau)}{\ln 2}}}{\tau}$

Normal distribution

The probability density function (PDF) of a normal distributed variable X (x_1, x_2, ... , x_n) is given by

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp[-\frac{1}{2}(\frac{x-\mu}{\sigma})^2] \qquad -\infty < x < \infty \qquad \textbf{Eq. A.17}$$

where μ and σ are the parameters of the normal distribution. The equations for the estimation for normal distribution are given by Table A-1.

Table A-1 Equations for normal distribution parameter estimation

Parameters	MOM	ML	PWM
μ	$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$	$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$	$\hat{\mu} = l_1$
σ	$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2$	$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2$	$\hat{\sigma} = \sqrt{\pi}l_2$

Two-parameter lognormal distribution

The probability density function (PDF) of a Two-parameter lognormal distributed variable X (x_1, x_2, ... , x_n) is given by

$$f(x) = \frac{1}{\sqrt{2\pi}x\sigma_y}\exp[-\frac{1}{2}(\frac{\ln x - \mu_y}{\sigma_y})^2] \qquad x > 0 \qquad \textbf{Eq. A.18}$$

where μ_y and σ_y are the mean and standard deviation of $y = \ln x$. The equations for the estimation for normal distribution are given by Table A-2.

Table A-2 Equations of lognormal distribution parameter estimation

Parameters	MOM	ML	PWM
μ	$\hat{\mu}_y = \frac{1}{n}\sum_{i=1}^{n} \ln x_i$	$\hat{\mu}_y = \frac{1}{n}\sum_{i=1}^{n} \ln x_i$	$\hat{\mu}_y = \ln l_1 - \frac{\sigma_y{}^2}{2}$
σ	$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2$	$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2$	$\hat{\sigma} = 2erf^{-1}(\frac{l_2}{l_1})$

Note: l_1 and l_2 donate the first and second L-moments.

Three-parameter lognormal distribution

The probability density function (PDF) of a Three-parameter lognormal distributed variable X (x_1, x_2, ... , x_n) is given by

$$f(x) = \frac{1}{\sqrt{2\pi}(x-a)\sigma_y} \exp\left\{-\frac{1}{2}[\frac{\ln(x-a)-\mu_y}{\sigma_y}]^2\right\} \quad x > a \qquad \textbf{Eq. A.19}$$

where μ_y and σ_y are the mean and standard deviation of y=ln(x-a), a is the shift of variable X. The equations for the estimation for 3-parameter normal distribution are given by Table A-3.

Table A-3 Equations for 3-parameter lognormal distribution parameter estimation

Parameters	MOM	ML	PWM
μ	$\hat{\mu}_y = \ln(\sqrt{m_2} / z_2) - \frac{1}{2}\ln(z_2{}^2 + 1)$	See the note	$\hat{\mu} = \ln[\frac{l_2}{erf(\sigma_y / 2)}] - \frac{\sigma_y{}^2}{2}$
σ	$\hat{\sigma}^2 = \ln(z_2{}^2 + 1)$	See the note	$\hat{\sigma} = 0.999281z - 0.006118z^3$ $+0.000127z^5$
a	$\hat{a} = m_1' - \sqrt{m_2} / z_2$	See the note	$\hat{a} = l_1 - \exp(\hat{\mu}_y + \frac{\sigma_y{}^2}{2})$

Note:

1. m_1' represents the sample mean;

2. m_2 is the second order of central sample moment $m_2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - m_1')^2$

3. $z_2 = \frac{1 - w^{2/3}}{w^{1/3}}$ where $w = \frac{-\gamma_1 + (\gamma_1{}^2 + 4)^{1/2}}{2}$ where γ_1 is the coefficient of skewness of the sample X.

4 $l_1$and l_2 donate the first and second L-moments.

5 $z = \sqrt{\frac{8}{3}}\Phi^{-1}(\frac{1+t_3}{2})$

6. for the MML estimation, numerical approach is needed to solve the following system of equations, which in some cases a solution may not exist.

$$\hat{\mu}_y = \frac{1}{n}\sum_{i=1}^{n}\ln(x_i - a) \quad \textbf{Eq. A.20}$$

$$\hat{\sigma}_y{}^2 = \frac{1}{n}\sum_{i=1}^{n}[\ln(x_i - a) - \hat{\mu}_y]^2 \quad \textbf{Eq. A.21}$$

$$\sum_{i=1}^{n}(x_i - a)^{-1}(\hat{\mu}_y - \hat{\sigma}_y{}^2) = \sum_{i=1}^{n}\frac{\ln(x_i - a)}{(x_i - a)} \quad \textbf{Eq. A.22}$$

Exponential distribution

The probability density function (PDF) of an exponential distributed variable X (x_1, x_2, ... , x_n) is given by

$$f(x) = \frac{1}{\alpha}e^{-(x-\varepsilon)/\alpha} \quad x \geq \varepsilon \quad \textbf{Eq. A.23}$$

The equations for the estimation for exponential distribution are given by Table A-4.

Table A-4 Equations for exponential distribution parameter estimation

Parameters	MOM	ML	PWM
α	$\hat{\alpha} = m_2$	$\hat{\alpha} = \frac{n(m_1^{'} - x_1)}{n-1}$	$\hat{\alpha} = 2l_2$
ε	$\hat{\varepsilon} = m_1^{'} - \hat{\alpha}$	$\hat{\varepsilon} = \frac{nx_1 - m_1^{'}}{n-1}$	$\hat{\varepsilon} = l_1 - 2l_2$

Note: x_1 represent the minimum observed value of X.

Two-parameter Gamma distribution

The probability density function (PDF) of an two-parameter Gamma distributed variable X (x_1, x_2, ... , x_n) is given by

$$f(x) = \frac{1}{\alpha^{\beta}\Gamma(\beta)}x^{\beta-1}e^{-x/\alpha} \quad x \geq 0 \quad \textbf{Eq. A.24}$$

The equations for the estimation for gamma distribution are given by Table A-5

Table A-5 Equations for gamma distributions parameter estimation

Parameters	MOM	MML	PWM
α	$\hat{\alpha}=\frac{m_2}{m_1^{'}}$	See the note	$\hat{\alpha}=l_1/\hat{\beta}$
β	$\hat{\beta}=\frac{(m_1^{'})^2}{m_2}$	See the note	$\hat{\beta}=\frac{1-0.3080\pi t^2}{\pi t^2-0.05812(\pi t^2)^2+0.01765(\pi t^2)^3}$ for $t\in(0,\frac{1}{2})$ $\hat{\beta}=\frac{0.7213\pi t^2-0.5947(\pi t^2)^2}{1-2.1817\pi t^2+1.2113(\pi t^2)^2}$ for $t\in[\frac{1}{2},1)$

Due to the difficulty of solving the system of differential equations in ML estimation, the following procedure developed by Bobee and Ashkar (1991) is often used. Let A represents the arithmetic mean of the sample and G represents the geometric mean of the sample, and set $U=\ln A-\ln G$, then

For $0\le U\le 0.5772$ $\qquad \hat{\beta}=\frac{1}{U}(0.5000876+0.1648852U-0.054427U^2)$

For $0.5772\le U\le 17.0$ $\qquad \hat{\beta}=\frac{8.898919+9.059950U+.09775373U^2}{U(17.7928+11.968477U+U^2)}$

And $\hat{\alpha}$ can be estimated by $\hat{\alpha}=A/\hat{\beta}$

Pearson III distribution

The probability density function (PDF) of an Pearson III distribution with variable X (x_1, x_2, ... , x_n) is given by

$$f(x)=\frac{1}{\alpha^{\beta}\Gamma(\beta)}\left(\frac{x-\gamma}{\alpha}\right)^{\beta-1}e^{-(x-\gamma)/\alpha} \quad x\ge\gamma$$ **Eq. A.25**

The equations for the estimation for Peason III distribution are given by Table A-6Table A-6.

Matlas and Wallis (1973) proposed an iterative numerical solution to the equation; however, they noticed a solution may not always exist, especially for very small sample skew values.

Log-Pearson III distribution

The probability density function (PDF) of an log-Pearson III distribution with variable X ($x_1, x_2, \ldots, x_n$) is given by

$$f(x) = \frac{1}{\alpha^{\beta}\Gamma(\beta)}(\frac{\ln x - \gamma}{\alpha})^{\beta-1} e^{-(\ln x-\gamma)/\alpha} \quad \ln x \geq \gamma \qquad \textbf{Eq. A.26}$$

One indirect method to estimate the parameters of the log-Pearson III distribution is to transfer the variable to $Z = \ln X$, and then estimate the parameters as for a Pearson III distribution. Another method is to direct application of MOM, MML and PWM. Due to the complication of solving the moment equations, the numerical effort required for the ML estimation, and the fact that no direct application of the PWM method to log-Pearson III distribution has been reported until now, the indirect method is recommended.

Generalized Extreme Value (GEV) distribution

The PDF of Generalized Extreme Value (GEV) distribution has the form of

$$f(x) = \frac{1}{\alpha}\exp(1 - k\frac{x-\mu}{\alpha})^{1/k-1} e^{-(1-k\frac{x-\mu}{\alpha})^{1/k}} \qquad \textbf{Eq. A.27}$$

If $k = 0$ ($C_s = 1.1396$), The GEV distribution is called Extreme Value Type I distribution (EV1), which is of the form

$$f(x) = \frac{1}{\alpha}\exp[-\frac{x-\mu}{\alpha} - \exp(\frac{x-\mu}{\alpha})] \quad -\infty < x < +\infty \qquad \textbf{Eq. A.28}$$

and

$$F(x) = 1 - e^{-(\frac{x-\mu}{\alpha})^{k}} \qquad \textbf{Eq. A.29}$$

If $k < 0$ ($C_s > 1.1396$), The GEV distribution is called Extreme Value Type II distribution (EV2). If $k > 0$ ($C_s < 1.1396$), The GEV distribution is called Extreme Value Type III distribution (EV3). EV3 is not often used in flood frequency analysis because in this case the variable x becomes upper upper bounded ($-\infty < x < \mu + \alpha / k$).When k=0 (C_s=1.1396), GEV distribution reduces to Type 1 extreme value distribution (EV1). The equations for the estimation for EV1 are given by Table A-7.

Table A-6 Equations for Pearson III distribution parameter estimation

Parameters	MOM	MML	PWM
α	$\hat{\alpha} = (m_2 / \hat{\beta})^{1/2}$	See the note	$\hat{\alpha} = \sqrt{\pi} l_2 \frac{\Gamma(\hat{\beta})}{\Gamma(\hat{\beta} + 1/2)}$
β	$\hat{\beta} = (2 / C_s)^2$	See the note	For L-moment ratio $t_3 \geq 1/3$, let $t_m = 1 - t_3$ $\hat{\beta} = \frac{0.36067 t_m - 0.5967 t_m^2 + 0.25361 t_m^3}{1 - 2.78861 t_m + 2.56096 t_m^2 - 0.77045 t_m^3}$ For L-moment ratio $t_3 < 1/3$, let $t_m = 3\pi t_3^2$ $\hat{\beta} = \frac{1 + 0.2906 t_m}{t_m + 0.1882 t_m^2 - 0.0442 t_m^3}$
γ	$\hat{\gamma} = m_1^{'} - \sqrt{m_2 \hat{\beta}}$	See the note	$\hat{\gamma} = l_1 - \hat{\alpha}\hat{\beta}$

Note: C_s is the coefficient of skewness, $C_s = \gamma_1$

The MML estimation involves solving the following system of equations simultaneously.

$$\frac{n\beta}{\alpha} - \frac{1}{\alpha^2} \sum_{i=1}^{n} x_{i-\gamma} = 0 \qquad \textbf{Eq. A.30}$$

$$-n\psi(\beta) + \sum_{i=1}^{n} \log \frac{x_i - \gamma}{\alpha} = 0 \qquad \textbf{Eq. A.31}$$

$$\frac{n}{\alpha} - (\beta - 1) \sum_{i=1}^{n} \frac{1}{x_i - \gamma} = 0 \qquad \textbf{Eq. A.32}$$

where $\psi(\beta) = \frac{\partial \ln \Gamma(\beta)}{\partial \beta} = \frac{\Gamma^{'}(\beta)}{\Gamma(\beta)}$

Table A-7 Equations for EV 1 distribution parameter estimation

Parameters	MOM	MML	PWM
α	$\hat{\alpha} = \frac{\sqrt{6}}{\pi} m_2^{1/2}$	See the note	$\hat{\alpha} = l_2 / \ln 2$
μ	$\hat{\mu} = m_1^{'} - 0.45005 m_2^{1/2}$	$\hat{\mu} = \hat{\alpha} \ln \frac{n}{\sum_{i=1}^{n} e^{-x_i / \hat{\alpha}}}$	$\hat{\mu} = l_1 - 0.5772157\hat{\alpha}$

Note: The numerical approach is involved to solve the differential equation in the ML estimation for the parameter α . By using Newton's method, the equation can be simplified as

$$\alpha_{n+1} = \alpha_n - F(\alpha_n) / F'(\alpha_n) \qquad \textbf{Eq. A.33}$$

where

$$F(\alpha_n) = \sum_{i=1}^{n} x_i e^{-x_i/\alpha_n} - (\frac{1}{n}\sum_{i=1}^{n} x_i - \alpha_n)\sum_{i=1}^{n} e^{-x_i/\alpha_n} \qquad \textbf{Eq. A.34}$$

$$F'(\alpha_n) = \frac{dF(\alpha_n)}{dF'(\alpha_n)} = \frac{1}{\alpha_n^{\ 2}}\sum_{i=1}^{n} x_i^{\ 2} e^{-x_i/\alpha_n} + \sum_{i=1}^{n} e^{-x_i/\alpha_n} + \frac{1}{\alpha_n}\sum_{i=1}^{n} e^{-x_i/\alpha_n} \qquad \textbf{Eq. A.35}$$

Weibull distribution

PDF $f(x) = \frac{\beta}{\alpha}(\frac{x-\mu}{\alpha})^{\beta-1} e^{-(\frac{x-\mu}{\alpha})^{\beta}}$ $x \geq 0$ **Eq. A.36**

CDF $F(x) = 1\text{-}e^{-(\frac{x-\mu}{\alpha})^{\beta}}$ **Eq. A.37**

The equations for the estimation for Weibull distribution are given by Table A-8

Table A-8 Equation for the Weibull distribution parameter estimation

Parameters	MOM	MML	PWM
α	$\alpha = \left\{ m_2 / [\Gamma(1+\frac{2}{\hat{\beta}}) - \Gamma^2(1+\frac{1}{\hat{\beta}})] \right\}^{1/2}$	See the note	$\hat{a} = l_2 / [\Gamma(1+1/\hat{\beta})(1-2^{-1/\hat{\beta}})]$
β	See the note		$\hat{\beta} = 1/(7.8590C + 2.95$
μ	$\hat{\mu} = m_1' - \hat{\alpha}\Gamma(1+1/\hat{\beta})$		$\hat{\mu} = l_1 - \hat{\alpha}\Gamma(1+1/\hat{\beta})$

Note: in the PWM, $C = \frac{2}{3-t_3} - \frac{\ln 2}{\ln 3}$ where t_3 is the estimated L-moment ratio

For the estimation of β by MOM, the numerical method is used to solve the equation

$$C_s = \frac{\mu_3}{\mu_2^{3/2}} = \frac{\Gamma(\frac{3}{\beta}+1) - 3\Gamma(\frac{1}{\beta}+1)\Gamma(\frac{2}{\beta}+1) + 2\Gamma^3(\frac{1}{\beta}+1)}{[\Gamma(\frac{2}{\beta}+1) - \Gamma^2(\frac{1}{\beta}+1)]^{3/2}} \qquad \textbf{Eq. A.38}$$

the following iteration equation is used.

$$(\frac{1}{\beta})_{n+1} = (\frac{1}{\beta})_n - F(\frac{1}{\beta})_n / F^{'}(\frac{1}{\beta})_n$$ **Eq. A.39**

For the ML estimation of the parameters, a numerical scheme to solve the following equations is proposed by Jenkinson (1969).

For the detail discussion on this topic, please refer to Rao and Khaled (2000) and Chow et al (1988).

Appendix B. Flood frequency factor

Table B-1 K_T and direct flood estimation equations

Distribution	Frequency factor	Direct equation
Normal	See the note	Not available
Two-parameter lognormal	$K_T = \frac{e^{\hat{\sigma}_y u - \hat{\sigma}_y^2/2} - 1}{(e^{\hat{\sigma}_y^2} - 1)^{1/2}}$	$\hat{x}_T = e^{\hat{\mu}_y + u\hat{\sigma}_y}$
Three-parameter lognormal	$K_T = \frac{e^{\hat{\sigma}_y^2 (u-1)/2} - 1}{(e^{\hat{\sigma}_y^2} - 1)^{1/2}}$	$\hat{x}_T = a + e^{\hat{\mu}_y + u\hat{\sigma}_y}$
Exponential	$K_T = \ln T - 1$	$\hat{x}_T = (\hat{\varepsilon} + \hat{a}) + \hat{a}(\ln T - 1)$
Two-parameter Gamma	$K_T = \frac{\chi^2 C_s}{4} - \frac{2}{C_s}$ See the note	Not available
Pearson Type III	$K_T = \frac{\chi^2 C_s}{4} - \frac{2}{C_s}$ $\hat{x}_T = \hat{\alpha}\hat{\beta} + \hat{\gamma} + K_T(\ln T - 1)$ See the note	$\hat{x}_T = (\hat{\varepsilon} + \hat{a}) + \hat{a}(\ln T - 1)$
Log-Pearson Type III	$K_T = \frac{\chi^2 C_s}{4} - \frac{2}{C_s}$ $\hat{x}_T = \hat{\alpha}\hat{\beta} + \hat{\gamma} + K_T \hat{a}(\ln T - 1)$ See the note	Not available
GEV	$K_T = \frac{\hat{k}\Gamma(1+\hat{k}) - [-\ln(1-\frac{1}{T})]^{\hat{k}}}{\lvert\hat{k}\rvert[\Gamma(1+2\hat{k}) - \Gamma^2(1+\hat{k})]^{1/2}}$	$x_T = \hat{u} + \frac{\hat{\alpha}}{\hat{k}}\left\{1 - [-\ln(1-\frac{1}{T})]\right\}^{\hat{k}}$
Extreme Value I	$K_T = -0.45 - .7797\ln[-\ln(1-\frac{1}{T})]$	$x_T = \hat{\beta} + \hat{\alpha}\ln[-\ln(1-\frac{1}{T})]$
Weibull	$K_T = \frac{(\ln T)^{1/b} - \Gamma(1/b+1)}{[\Gamma(2/b+1) - \Gamma^2(1/b+1)]^{1/2}}$	$x_T = \hat{m} + \hat{\alpha}(\ln T)^{1/\hat{b}}$

Note:

1. For normal distribution, K_T can be calculated by using the value of z which is given by Abramowitz and Stegun (1965),

$$z = \left| w - \frac{C_0 + C_1 w + C_2 w^2}{1 + d_1 w + d_2 w^2 + d_3 w^3} \right| + \varepsilon(p) \qquad \textbf{Eq. B.1}$$

where

$$\begin{aligned} w &= (-2\ln p)^{1/2} && \text{for } p \le 0.5 \\ &= [-2\ln(1-p)]^{1/2} && \text{for } p < 0.5 \end{aligned} \qquad \textbf{Eq. B.2}$$

and $\varepsilon(p)$ is the error which is less than 4.5×10^{-4}

2. For two-parameter gamma distribution, C_s is the skewness coefficient of the data.

$$C_s = \frac{\mu_3}{\mu_2^{3/2}} \qquad \textbf{Eq. B.3}$$

and χ^2 is calculated by

$$\chi^2 = (1 - \frac{2}{9\upsilon} + u\sqrt{\frac{2}{9\upsilon}})^3 \qquad \textbf{Eq. B.4}$$

where u is the standard normal variate corresponding to a probability of non-exceedence of F=1-1/T, and υ is the degree of freedom.It should be noted that the flood needs to be evaluated by $\hat{x}_T = \hat{\alpha}\hat{\beta} + \hat{\gamma} + K_T \hat{a}(\ln T - 1)$, instead of the equation addressed earlier.

3.For log-Pearson Type III, C_s and χ^2 are the same as those in the frequency factor equation in two parameter gamma distribution, and can be calculated by the same equations. The magnitude of the flood is estimated by $\hat{x}_T = \hat{\alpha}\hat{\beta} + \hat{\gamma} + K_T \hat{a}(\ln T - 1)$, instead of the equation addressed earlier.

Standard error of estimate

The standard error was defined by Cunnane (1989) to measure the variability of the estimated value.

$$s_T = \sqrt{E[\hat{x}_T - E(\hat{x}_T)]^2} \qquad \textbf{Eq. B.5}$$

The standard error of estimate depends in general on the method of parameter estimation (Rao and Hamed, 2000). The most efficient method gives the smallest standard error of estimate.

For a given three-parameter distribution with parameters α, β and γ , the standard error of estimate can be calculated by using the formula

$$s_T^2 = (\frac{\partial x}{\partial \alpha})^2 \operatorname{var}(\alpha) + (\frac{\partial x}{\partial \beta})^2 \operatorname{var}(\beta) + (\frac{\partial x}{\partial \gamma})^2 \operatorname{var}(\gamma) + 2(\frac{\partial x}{\partial \alpha})(\frac{\partial x}{\partial \beta}) \operatorname{cov}(\alpha, \beta) + 2(\frac{\partial x}{\partial \alpha})(\frac{\partial x}{\partial \gamma}) \operatorname{cov}(\alpha, \gamma) + 2(\frac{\partial x}{\partial \beta})(\frac{\partial x}{\partial \gamma}) \operatorname{cov}(\beta, \gamma)$$

Eq. B.6

The partial derivatives in the equation can be calculated from the relation $x_T = F^{-1}(1 - \frac{1}{T})$ or $x_T = u_1^{'} + K_T \sqrt{\mu_2}$.