doi:10.5194/piahs-368-251-2015

POME-copula for hydrological dependence analysis

DENGFENG LIU[1], DONG WANG[1], LACHUN WANG[2], YUANFANG CHEN[3], XI CHEN[3] & SHENGHUA GU[4]

1 Key Laboratory of Surficial Geochemistry, MOE, Department of Hydrosciences, School of Earth Sciences and Engineering, State Key Laboratory of Pollution Control and Resource Reuse, Nanjing University, Nanjing, China
wangdong@nju.edu.cn

2 School of Geographic and Oceanographic sciences, Nanjing University, Nanjing, China

3 School of Hydrology and Water Resources, State Key Laboratory of Hydrology-Water Resources and Hydraulic Engineering, Hohai University, Nanjing, China

4 Shanghai Hydrology Administration, Shanghai, China

Abstract Hydrological multivariate analysis has been widely studied using copula-based modelling, in which marginal distribution inference is one of the key issues. The main object of this study is to discuss the applicability of the principle of maximum entropy (POME) in marginal distribution inference, thus to develop a POME-copula framework to analyse the dependence of hydrological variables. Marginal distributions are derived with the POME approach before bivariate copulas constructed with corresponding parameters estimated by the dependence of the derived margins. The proposed POME-copula has been employed in hydrological dependence analyses, with the annual maximum streamflow and water level collected from the Yangtze River, and the monthly streamflow from the Yellow River. Results show that the POME-copula method performs well in capturing dependence patterns of various hydrological variables.

Key words the principle of maximum entropy; copula; dependence analysis; Shannon entropy; marginal distribution

1 INTRODUCTION

Hydrological multivariate analysis has attracted increasing attention in the past few years, due to limitations of single-variable analysis in hydrological applications. For instance, design of hydraulic constructions requires assessments of risks associated with more than one hydrological factor, including peak discharge, water level, etc. Therefore, multivariate dependency between all quantities defining the risk should be taken into account. In the past decades, various multivariate distributions, including multivariate normal and relevant distributions derived from extensions of Student's *t* and Fischer's *F* distributions (Johnson and Wichern, 1988), non-normal multivariate distributions including bivariate exponential (Favre *et al.*, 2002), bivariate gamma (Yue *et al.,* 2001), and bivariate extreme value distributions (Adamson *et al*., 1999), have been applied to model different dependence patterns of the hydrological variables. Although witnessed increasing applications, drawbacks in modelling these multivariate distributions are obvious, which can be summarized as: (a) the same family is needed for each marginal distribution, (b) extensions to more than just the bivariate case are not clear, and (c) parameters of the marginal distributions are also used to model the dependence between the random variables (Favre *et al.*, 2004).

The advent of copula (Sklar, 1959), a multivariate distribution constructing technique, avoiding the drawbacks mentioned above, can simplify the inference procedures and allow for splitting analyses of marginal distributions and further studies on hydrological dependence structures. Merits of copula in hydrologic applications have been discussed in Favre *et al.* (2004), Genest and Favre (2007) and Salvadori and De Michele (2007). In recent years, copula-based techniques have been widely used in hydrological dependence analysis (Vandenberghe *et al*., 2010; Gyasi-Agyei *et al.,* 2012), frequency analysis (Genest *et al*., 2007; Salvadori and Michele, 2010; Chebana *et al.,* 2012) and hydrological simulation (AghaKouchak *et al.,* 2010), etc. In copula-based modelling, marginal distribution derivation is of great importance, while in most previous researches, margins were assumed to follow certain common distributions, and problems of subjectivity might arose, accordingly. The POME, a non-parametric mathematical framework avoiding subjective bias in statistical inference, has been introduced in copula-based modelling (Hao and Singh, 2012a; 2012b). Based on previous studies, a POME-copula framework has been developed in this research, with advantages that deducted margins can have different forms and make full use of given information. The developed approach was applied to model and analyse hydrological dependence, with data collected from the Yangtze River and the Yellow River, China.

Copyright © 2015 IAHS Press

2 METHODOLOGY

2.1 The principle of maximum entropy (POME)

The principle of maximum entropy (POME) (Jaynes, 1957), a more flexible non-parametric inference framework, has been applied to derive the marginal distribution before copula modelling.

For a continuous random variable X with the probability density function (PDF) $f(x)$ ($x \in (a, b)$), the Shannon entropy H can be defined as (Shannon, 2001):

$$H = -\int_a^b f(x)\ln(f(x))dx \tag{1}$$

The POME provides a constructive criterion for setting up probability distributions on the basis of partial knowledge and one can obtain the most probable PDF with the available constraints by maximizing equation (1). Using the moments as constraints specified as:

$$\int_a^b g_i(x)f(x)dx = E(g_i), i = 0,1,2,...,m \tag{2}$$

where $E(g_i)$ is the expectation of $g_i(x)$ ($g_i(x)=x^i$). The POME-based PDF can be obtained as (Kesavan and Kapur, 1992):

$$f(x) = \exp[-\lambda_0 - \sum_{i=1}^{m} \lambda_i x^i] \tag{3}$$

where λ_i are the Lagrange multipliers. In this study, marginal distributions were derived with equation (3), using constraints of the first three moments, which can be expected to preserve the mean, variance and skewness of hydrological variables.

2.2 Copula theory

A copula is a multivariate function describing dependence of variables transformed by their margins, which can simplify inference procedures of multivariate distributions and studies on hydrological dependence.

For the continuous random vector (X, Y) with marginal distributions $F_X(x)$ and $F_Y(y)$, the joint distribution function can be expressed with its marginal distributions and copula function C as (Nelsen, 2006):

$$P(X \le x, Y \le y) = C[F_X(x), F_Y(y); \theta) = C(u, v; \theta) \tag{4}$$

where θ is the parameter of the copula that measures the dependence between margins; u and v are realizations of the random variables $U = F_X(x)$ and $V = F_Y(y)$. The density function of C is:

$$c(u, v; \theta) = \frac{d^2 C(u, v; \theta)}{dudv} \tag{5}$$

The two-dimensional copula C maps the two marginal distributions into the joint distribution as $(0, 1)^2 \rightarrow (0, 1)^2$. The value of θ can be estimated by the Spearman's correlation coefficient and the Kendall's correlation coefficient (Schweizer and Wolff, 1981).

The Archimedean copula is one of the most popular copula functions. Moreover, in the Archimedean copula the computation of measures of dependence is simplified. In this study, three ypes of Archimedean copula, the Clayton, Frank and Gumbel, were employed to model dependence patterns of different hydrological variables.

2.3 POME-copula method

The entropy-copula coupled idea has been proposed in hydrology (Hao and Singh, 2012a; 2012b), and based on relevant works, a more completed POME-copula framework is developed, which can be summarized as two steps:

(a) Derive marginal distributions $U = F_X(x)$ and $V = F_Y(y)$ using the POME.

Normalize the initial data X (Y) to (–1, 1) with the algorithm $x' = 2 \times (x - X_{min})/(X_{max} - X_{min}) + (-1)$. After that, determine the Lagrange multipliers in equation (3) using the Newton-Raphson method (Hao and Singh, 2011) with constraints of the first three moments (equation (2): $a = -1$, $b = 1$, $m = 3$) and obtain the distribution $U = F_X(x)$ ($V = F_Y(y)$). Use the Kolmogorov-Smirnov test to assess goodness-of-fit of derived distributions.

(b) Determine the best copula according to the estimation of the dependence of U and V. Calculate the Spearman's coefficient *rho* and Kendall's coefficient *tau* of U and V, so that θ can be estimated by *rho* and *tau*. Later, evaluate goodness-of-fit of different types of copulas with statistics $Sn^{(B)}$ and $Sn^{(C)}$ (Genest *et al.*, 2009), before determining the best copula to model the dependence of hydrological data.

3 CASE STUDY

The developed POME-copula framework was employed to model and analyse hydrological variables of two representative basins in China, the Yangtze River and the Yellow River. Considering streamflow and water level data as continuous random variables, dependences of the annual maxima of streamflow (denoted as S1) and water level (denoted as S2) of the Yangtze River at Yichang, for the period from 1950 to 2008; and the monthly streamflow of the Yellow River at Huayuankou (denoted as S3) and Gaocun (denoted as S4), for the period from 1998 to 2012, were analysed with the POME-copula.

3.1 Marginal distributions with the POME

The marginal PDFs $F_X(x)$ and $F_Y(y)$ derived with the POME were compared with the empirical histograms (Fig. 1). Generally, the POME-based PDFs fitted the empirical histogram well, especially that the derived POME-based PDFs of streamflows of the Yellow River are relatively accurate, illustrating the strong positive skewness of distributions.

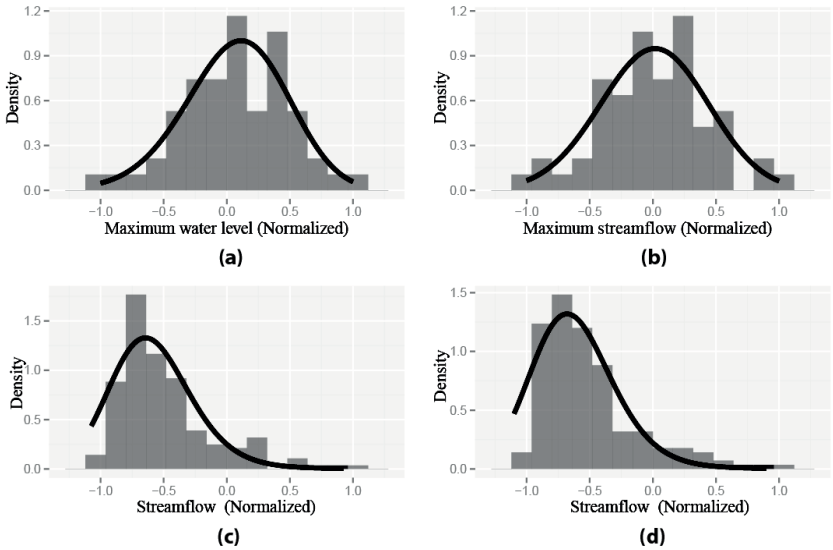

Fig. 1 Comparison of the empirical histograms and the POME-based PDFs: (a) S1, (b) S2, (c) S3 and (d) S4.

The Kolmogorov-Smirnov test was employed to assess the goodness-of-fit of POME-based marginal distributions. Results of high p value (>>0.05), especially for the data of the Yangtze River, signified that the null hypothesis that the data follow POME-based distributions should not be rejected. That is the POME approach is accurate in marginal distribution inference.

3.2 Copula selection

The Spearman's coefficient *rho* and Kendall's coefficient *tau* of U and V were calculated before parameters were estimated. The goodness-of-fit of estimated Clayton, Frank and Gumbel copulas were evaluated by statistics $Sn^{(B)}$ and $Sn^{(C)}$. From Table 1, results of low p values (<0.05) signified that the null hypothesis that the copula function is accurate should be rejected. In the two cases, the Frank copula, with higher p values than other copulas, was selected as the best copula to analyse the dependence of hydrological variables.

Table 1 Goodness-of-fit tests of different estimated copulas.

Cases	Copulas	θ		p value ($Sn^{(B)}$)		p value ($Sn^{(C)}$)	
		tau	*rho*	*tau*	*rho*	*tau*	*rho*
S1-S2	Clayton	9.41	8.75	0.06	0.07	0.07	0.07
	Frank	21.04	**17.90**	0.69	**0.73**	0.41	**0.48**
	Gumbel	5.71	5.25	0.26	0.18	0.10	0.21
S3-S4	Clayton	11.93	12.28	0.00	0.00	0.00	0.00
	Frank	26.11	**24.10**	0.62	**0.67**	0.54	**0.65**
	Gumbel	6.97	6.95	0.27	0.26	0.39	0.33

3.3 Dependence analysis using POME-Frank copula

Densities of POME-Frank copulas were compared with empirical densities (Fig. 2), and it is illustrated that densities of POME-Frank copulas fit empirical ones well. Moreover, dependence patterns of S1-S2 and S3-S4 are similar, although their marginal distributions are quite different (Fig. 1). First, densities of copulas are distributed along the main diagonals, indicating positive correlations of both two cases. Second, densities increase when they reach the upper and lower tail. Last, dependence structures of the two cases are symmetric.

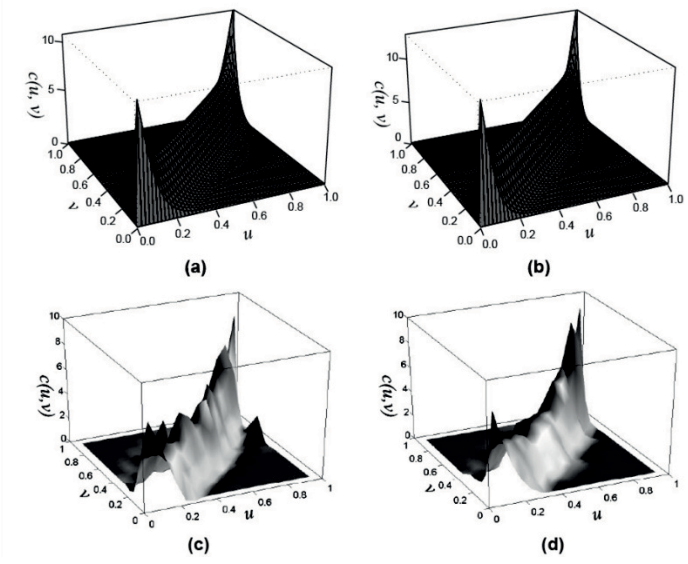

Fig. 2 Comparison of the POME-Frank copulas and empirical density distributions: (a) POME-Frank copula (S1-S2); (b) POME-Frank copula (S3-S4); (c) empirical density distribution (S1-S2); and (d) empirical density distribution (S3-S4).

One hundred pairs of series with length $N = 500$ were generated with the constructed POME-Frank copulas. Scatterplots of the observed data and the generated data with POME-Frank copulas are shown in Fig. 3. Generally, distributed patterns of the generated data matched that of the observed data. For instance, S1 and S2 shows a strong positive dependence (Pearson coefficient r =

0.95; Kendall coefficient *tau* = 0.82; Spearman coefficient *rho* = 0.95) and most of the observed data are distributed along the main diagonal. In comparison, most of the generated data are distributed along the diagonal with approximate correlations (r = 0.94; *tau* = 0.81; *rho* = 0.94).

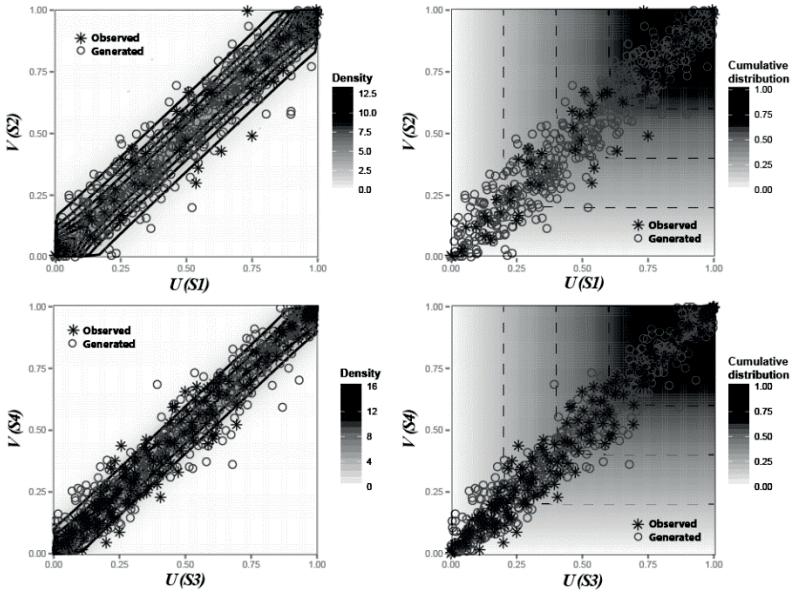

Fig. 3 Scatterplots of the observed and generated data with the POME-Frank copulas.

Boxplots were used to assess the performance of 100 generated pairs with the POME-copulas, comparing with the observed ones (Fig. 4). The performance can be judged good when the statistics calculated from the observed data fall within the ranges of boxplots drawn by simulated results. From Fig. 4, it can be shown that all correlation coefficients from the observed data are within the interval of the boxplots, and most correlation coefficients, except the Kendall coefficient in Fig. 5(a) and the Pearson coefficient in Fig. 5(b), are within the interquartile interval. Results from boxplots further verify accuracy of the POME-copulas in capturing observed dependence patterns, including linear correlation and rank correlation.

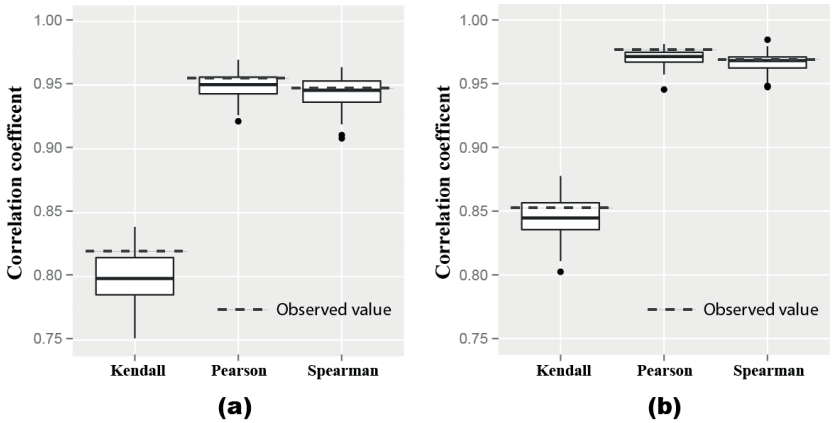

Fig. 4 Correlation coefficients of the observed and all generated data: (a) S1-S2; and (b) S3-S4.

4 CONCLUSION

A complete POME-copula framework is developed for hydrological multivariate modelling. Dependences of hydrological data from two representative basins in China have been analysed with the constructed POME-copulas, indicating obvious similarity in symmetry and tail dependence.

With their advantages of full use of providing information and objectivity in marginal distribution inference, the POME-copulas have been verified as accurate in capturing dependence patterns of various hydrological variables.

The developed POME-copula framework can also be applied in similar dependence analyses. For more complicated cases, further studies on constraints of more moments are needed, thus to reduce more potential uncertainty in marginal distribution inference.

Acknowledgements The authors gratefully acknowledge the helpful review comments and suggestions on the earlier version of the manuscript by the reviewers and Editors. This study was supported by the National Science & Technology Pillar Program (2013BAB05B01-3), National Natural Science Fund of China (No. 51190091, 41071018), Program for New Century Excellent Talents in University (NCET-12-0262), China Doctoral Program of Higher Education (20120091110026), Water Resources Public-Welfare Project (No. 201201068), Special fund of Taihu Lake, Jiangsu Province (TH2014307), Qing Lan Project, the Skeleton Young Teachers Program and Excellent Disciplines Leaders in Midlife-Youth Program of Nanjing University.

REFERENCES

Adamson, P.T., Metcalfe, A.V. and Parmentier, B. (1999) Bivariate extreme value distributions: an application of the Gibbs sampler to the analysis of floods. *Water Resources Research 35*(9), 2825–2832.

AghaKouchak, A., Bárdossy, A. and Habib, E. (2010) Copula‐based uncertainty modelling: application to multisensor precipitation estimates. *Hydrological Processes* 24(15), 2111–2124.

Chebana, F., *et al.* (2012) Exploratory functional flood frequency analysis and outlier detection. *Water Resources Research* 48(4), W04514.

Favre, A C., Musy, A. and Morgenthaler, S. (2002) Two‐site modeling of rainfall based on the Neyman-Scott process. *Water Resources Research* 38(12), 43-1.

Favre, A.C., *et al.* (2004) Multivariate hydrological frequency analysis using copulas. *Water Resources Research* 40(1), W01101.

Genest, C., *et al.* (2007) Metaelliptical copulas and their use in frequency analysis of multivariate hydrological data. *Water Resources Research* 43(9), W09401.

Genest, C. and Favre, A. C. (2007) Everything you always wanted to know about copula modeling but were afraid to ask. *Journal of Hydrologic Engineering* 12(4), 347–368.

Genest, C., Rémillard, B. and Beaudoin, D. (2009) Goodness-of-fit tests for copulas: A review and a power study. *Insurance: Mathematics and Economics* 44(2), 199–213.

Gyasi-Agyei, Y. and Melching, C. S. (2012) Modelling the dependence and internal structure of storm events for continuous rainfall simulation. *Journal of Hydrology* 464, 249–261.

Hao, Z. and Singh, V.P. (2011) Single‐site monthly streamflow simulation using entropy theory. *Water Resources Research* 47(9).

Hao, Z. and Singh, V.P. (2012a) Entropy‐copula method for single‐site monthly streamflow simulation. *Water Resources Research*, 48(6).

Hao, Z. and Singh, V.P. (2012b) Entropy-based method for bivariate drought analysis. *Journal of Hydrologic Engineering 18*(7), 780–786.

Jaynes, E.T. (1957) Information theory and statistical mechanics. *Physical Review* 106(4), 620.

Johnson, R.A. and Wichern, D.W. (1992) *Applied Multivariate Statistical Analysis* (Vol. 4). Englewood Cliffs, NJ: Prentice hall.

Kapur, J. N. and Kesavan, H. K. (1992) *Entropy Optimization Principles with Applications*. Academic Press.

Nelsen, R.B. (1999) *An Introduction to Copulas*. Springer.

Salvadori, G. and De Michele, C. (2007) On the use of copulas in hydrology: theory and practice. *Journal of Hydrologic Engineering* 12(4), 369–380.

Salvadori, G. and De Michele, C. (2010) Multivariate multiparameter extreme value models and return periods: A copula approach. *Water resources research 46*(10), W10501.

Schweizer, B. and Wolff, E. F. (1981) On nonparametric measures of dependence for random variables. *The Annals of Statistics* 9(4), 879–885.

Shannon, C. E. (2001) A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review* 5(1), 3–55.

Sklar, M. (1959) *Fonctions de répartition à n dimensions et leurs marges*. Université Paris 8.

Vandenberghe, S., Verhoest, N.E.C. and De Baets, B. (2010) Fitting bivariate copulas to the dependence structure between storm characteristics: A detailed analysis based on 105 year 10 min rainfall. *Water Resources Research 46*(1), W01512.

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