

Modeling multisite streamflow dependence with maximum entropy copula

Z. Hao¹ and V. P. Singh^{1,2}

Received 12 April 2013; revised 9 September 2013; accepted 10 September 2013; published 4 October 2013.

[1] Synthetic streamflows at different sites in a river basin are needed for planning, operation, and management of water resources projects. Modeling the temporal and spatial dependence structure of monthly streamflow at different sites is generally required. In this study, the maximum entropy copula method is proposed for multisite monthly streamflow simulation, in which the temporal and spatial dependence structure is imposed as constraints to derive the maximum entropy copula. The monthly streamflows at different sites are then generated by sampling from the conditional distribution. A case study for the generation of monthly streamflow at three sites in the Colorado River basin illustrates the application of the proposed method. Simulated streamflow from the maximum entropy copula is in satisfactory agreement with observed streamflow.

Citation: Hao, Z., and V. P. Singh (2013), Modeling multisite streamflow dependence with maximum entropy copula, *Water Resour. Res.*, 49, 7139–7143, doi:10.1002/wrcr.20523.

1. Introduction

[2] For multisite streamflow simulation in a river basin, it is desired that statistical properties of streamflow at individual sites and dependence structure among different sites are preserved. The autoregressive moving average (ARMA) framework has been commonly used for multisite streamflow simulation [Salas and Delleur, 1980]. Parametric disaggregation method is another type of parametric methods for multisite streamflow simulation that generally consists of two steps. It first generates aggregated streamflow (e.g., annual flow) and then disaggregates or divides it into lower-level variables (e.g., monthly flow) [Valencia and Schaake, 1973; Stedinger and Vogel, 1984; Koutsoyiannis and Manetas, 1996].

[3] There are several disadvantages of the parametric ARMA and disaggregation methods, including the limitation in representing nonlinear dependences and nonstandard probability distribution forms [Sharma and O'Neill, 2002]. The nonparametric model, such as kernel density method, moving block bootstrapping method, or K-nearest neighbor resampling method, does not make assumptions about the probability distribution or dependence forms and provides an alternative for stochastic simulation [Vogel and Shallcross, 1996; Sharma et al., 1997; Prairie et al., 2007; Nowak et al., 2010].

[4] Recently, the copula method has been commonly used for modeling the dependence structure of multivariate random variables and also for the multisite stochastic simulation [Bárdossy and Pegram, 2009]. However, the ability of the commonly used parametric copulas to model dependences in higher dimensions is rather restricted [Kao and Govindaraju, 2008; Chui and Wu, 2009]. In this study, we propose the maximum entropy copula for multisite monthly streamflow simulation in which the rank correlation in higher dimensions among monthly streamflows at different sites can be modeled. The proposed method is applied to monthly streamflow simulation at three sites in the Colorado River basin.

2. Methodology

2.1. Entropy Concepts

[5] Let the joint probability density function (PDF) of two random variable X and Y on the interval $[a_1, b_1] \times [a_2, b_2]$ be $f(x, y)$. The entropy H of the joint PDF $f(x, y)$ can be defined as [Shannon, 1948; Shannon and Weaver, 1949]

$$H = - \int_{a_2}^{b_2} \int_{a_1}^{b_1} f(x, y) \ln f(x, y) dx dy \quad (1)$$

[6] The principle of maximum entropy developed by Jaynes [1957] can be employed to derive the joint probability density function $f(x, y)$ in that the joint PDF with the maximum entropy should be selected subject to the given constraints (or known information).

2.2. Maximum Entropy Copula

[7] The maximum entropy copula has been developed based on the entropy theory [Chui and Wu, 2009; Chu, 2011]. Let U and V be the marginal probabilities of the random variables X and Y with u and v denoting realizations of

¹Department of Biological and Agricultural Engineering, Texas A&M University, College Station, Texas, USA.

²Department of Civil and Environmental Engineering, Texas A&M University, College Station, Texas, USA.

Corresponding author: Z. Hao, Department of Biological and Agricultural Engineering, Texas A&M University, 2117 TAMU, College Station, TX 77843-2117, USA. (hzc07@tamu.edu)

U and V . For a copula density function $c(u, v)$, the entropy can be expressed as

$$W = - \int_0^1 \int_0^1 c(u, v) \log c(u, v) du dv \quad (2)$$

[8] The constraints can be expressed as

$$\int_0^1 \int_0^1 c(u, v) g_i(u, v) du dv = g_i \quad i = 1, 2, \dots, n \quad (3)$$

where g_i is the expectation of the function $g_i(u, v)$, i.e., $E(g_i(u, v))$. To ensure the integration of the copula density function over all the space equates one, $g_1(u, v)$ can be specified as 1. To ensure that the marginal of $c(u, v)$ is the uniform $[0, 1]$, the moments of u and v can be specified as constraints (i.e., u, u^2, u^3, v, v^2 and $v^3 \dots$) to approximate the marginal properties numerically [Chu, 2011]. To model the dependence structure, the function $g(u, v)$ can be specified in the form that is related to an association measure such that the expectation $E(g(u, v))$ becomes some linear form of rank correlation. For example, when the pairwise product constraint $g(u, v) = uv$ is used, the commonly used Spearman rank correlation (ρ) can be linked to the constraint [Chu, 2011]:

$$\int_0^1 \int_0^1 uv c(u, v) du dv = \frac{\rho + 3}{12} \quad (4)$$

[9] With the moment constraints up to order m and pairwise product constraint in equation (4), the maximum entropy copula density function can be obtained as [Chui and Wu, 2009; Chu, 2011]

$$c(u, v) = \exp \left[-\lambda_0 - \sum_{r=1}^m (\lambda_r u^r + \lambda_{r+m} v^r) - \lambda_{2m+1} uv \right] \quad (5)$$

where m is the maximum order of moments ($m=3$ in this study) and $\lambda_0, \dots, \lambda_{2m+1}$ are the Lagrange parameters. Parameter λ_0 can be expressed as a function of other parameters as

$$\lambda_0 = \int_0^1 \int_0^1 \exp \left[-\sum_{r=1}^m (\lambda_r u^r + \lambda_{r+m} v^r) - \lambda_{2m+1} uv \right] du dv \quad (6)$$

[10] The dependence structure in terms of the Spearman rank correlation can be modeled through the joint probability density function in equation (5). Note that other measures of the dependence structure, such as Blest's measure and Gini's gamma, can also be modeled through the maximum entropy copula [Chu, 2011].

[11] The joint distribution in the higher dimension is of particular interest when the multivariate dependence structure has to be modeled. In this case, a multivariate entropy in equation (2) can be defined and then copula density function with the maximum entropy can be derived straightforward. It can be seen that the derivation of the maximum entropy copula is separate from that of the marginal probability distributions. Suitable marginal distributions, such as

kernel density, can be selected to model the properties of streamflow of each month, such as skewness and bimodal properties, which have been well documented [Sharma et al., 1997; Prairie et al., 2007; Salas and Lee, 2010; Hao and Singh, 2012]. Thus, we omit the discussion of the marginal distributions but focus on the dependence structure modeling of multisite monthly streamflow through the maximum entropy copula.

2.3. Parameter Estimation

[12] For the maximum entropy copula, the Lagrange multipliers λ_i ($i=1, \dots, 2m+1$) in equation (5) have to be estimated. It has been shown that these Lagrange multipliers can be solved by finding the minimum of a convex function Γ expressed as [Kapoor, 1989]

$$\Gamma = \lambda_0 + \sum_{i=1}^{2m+1} \lambda_i g_i \quad (7)$$

[13] These parameters can be estimated using the Newton Raphson iteration method [Wu, 2003; Hao and Singh, 2011]. However, a high-dimensional integration is involved in the parameter estimation for the multisite simulation to obtain the value of λ_0 in equation (6), which makes it even more complicated than the single-site streamflow simulation. In this study, an adaptive algorithm for numerical integration over hyperrectangular region was employed for the high-dimensional integration (programmed as a MATLAB function ADAPT available from www.math.wsu.edu/faculty/genz/homepage) [Genz and Malik, 1980; Berntsen et al., 1991].

2.4. Simulation Methodology

[14] Suppose three sites from upstream to downstream are denoted as site 1, 2, and 3 and denote the marginal probability of the monthly streamflow at each site as (U_1, U_2, \dots) , (V_1, V_2, \dots) and (W_1, W_2, \dots) and the realizations as (u_1, u_2, \dots) , (v_1, v_2, \dots) and (w_1, w_2, \dots) . For site 1, the joint distribution $C(u_s, u_{s-1})$ of monthly streamflow for two adjacent months s and $s-1$ must be estimated and the conditional distribution $C(u_s | u_{s-1})$ can be used to generate the monthly streamflow (marginal) U_s given the previous monthly streamflow (marginal) U_{s-1} . For monthly streamflow of site 2, the joint distribution $C(u_s, v_{s-1}, v_s)$ has to be estimated, and the conditional distribution $C(v_s | v_{s-1}, u_s)$ can be used to generate the monthly streamflow V_s given the streamflow V_{s-1} of site 2 and the monthly streamflow U_s of site 1. Similarly, for the monthly streamflow W_s , the conditional distribution $C(w_s | w_{s-1}, u_s, v_s)$ can be used to generate the monthly streamflow W_s given the streamflow W_{s-1} of site 3, U_s of site 1 and V_s of site 2.

[15] The simulation methodology to generate the monthly streamflow (marginal) at each site can be summarized as follows:

[16] (1) Initialize monthly streamflow at sites 1, 2, and 3 of the first month, i.e., u_1, v_1 , and w_1 , by assigning random values from historical records.

[17] (2) With the initialized u_1 , generate monthly streamflow at site 1 for the second month u_2 from the conditional distribution $C(u_s | u_{s-1})$. With the generated u_2 , and initialized value v_1 , the monthly streamflow at site 2 for the second month v_2 can be generated from the distribution

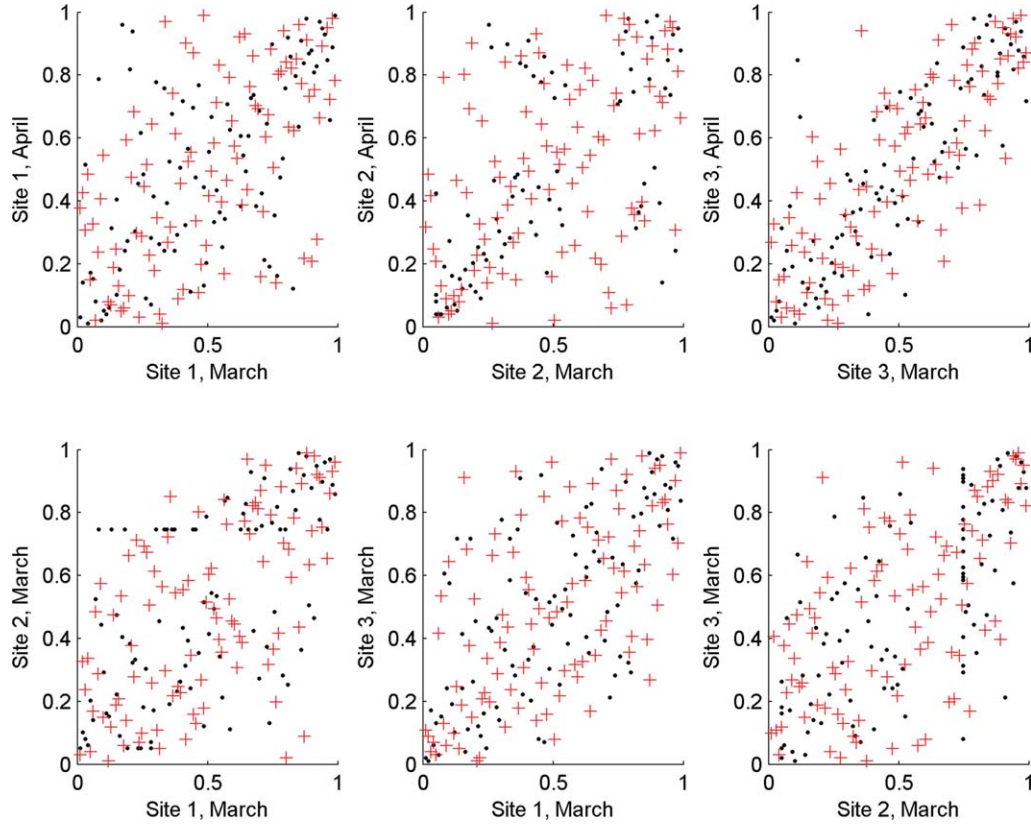


Figure 1. Scatterplots of the observed (closed circle) and simulated (plus symbol) monthly streamflow (marginal) for March and April at different sites.

$C(v_s|v_{s-1}, u_s)$. With the generated u_2, v_2 and the initialized w_1 , the monthly streamflow at site 3 for the second month w_2 can be generated from the distribution $C(w_s|w_{s-1}, u_s, v_s)$.

[18] (3) With the generated u_2, v_2 , and w_2 , repeat step (2) to generate the monthly streamflow for the next month u_3, v_3 and w_3 for sites 1, 2, and 3, respectively.

[19] (4) Repeat step (3) to generate a sequence of monthly streamflows $u_4, \dots, u_t, v_4, \dots, v_t$ and w_4, \dots, w_t up to time t .

3. Application

[20] Monthly streamflow from 1906 to 2003 of three sites in the Colorado River basin, namely Paria River at

Lees Ferry, Arizona (AZ) (denoted as site 1), Little Colorado River near Cameron, AZ (denoted as site 2), and Virgin River at Littlefield, AZ (denoted as site 3), were used for illustrating the proposed method. The monthly streamflow at each site can be downloaded from the website (<http://www.usbr.gov/lc/region/g4000/NaturalFlow/previous.html>).

[21] We illustrate the derivation of the joint probability density function for monthly streamflow at site 1 and 2 as an example. Denote the marginal probabilities of monthly streamflow for the month s at sites 1 and 2 as U_s and V_s . From equation (5), the maximum entropy copula density function $c(u_s, v_{s-1}, v_s)$ with the moment constraints up to order 3 and pairwise product constraint can be expressed as

$$c(u_s, v_{s-1}, v_s) = \exp \left(-\lambda_0 - \sum_{i=1}^3 \lambda_i u_s^i - \sum_{i=1}^3 \lambda_{i+3} v_{s-1}^i - \sum_{i=1}^3 \lambda_{i+6} v_s^i - \lambda_{10} u_s v_{s-1} - \lambda_{11} u_s v_s - \lambda_{12} v_{s-1} v_s \right) \quad (8)$$

[22] The joint distribution $C(u_s, v_{s-1}, v_s)$ and conditional distribution $C(v_s|u_s, v_{s-1})$ can be obtained from the density function accordingly.

[23] One hundred sequences of monthly streamflow (marginal) with the same length as the historical record (98 years) were generated for each site with the simulation methodology. The scatterplots of the rank of observed streamflow pairs and one sequence of simulated streamflow

pairs (marginal) from the copula at different sites for March and April are shown in Figure 1 (top). The spread pattern of simulated streamflow pairs generally matched that of observed streamflow pairs of the 2 months well. As an example, the monthly streamflow of March and April for site 3 shows a strong dependence (Spearman correlation: 0.83) and most of the streamflow pairs spread along the diagonal. The simulated streamflow pairs are also spreading

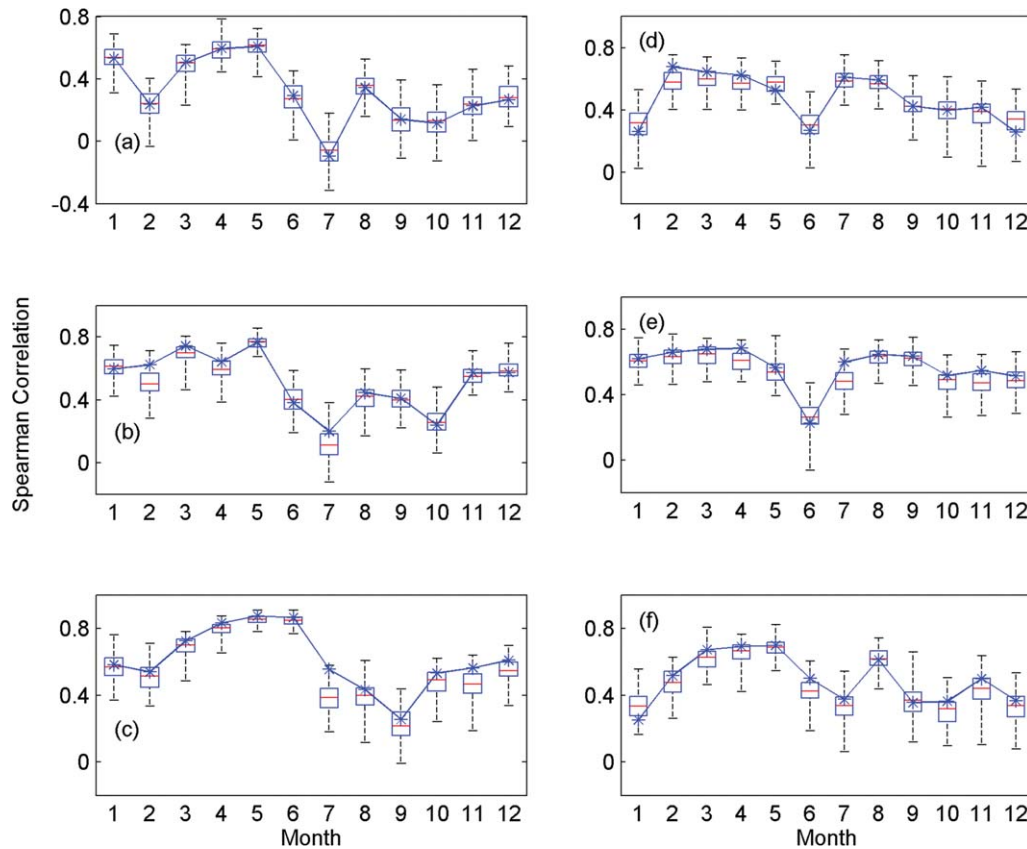


Figure 2. The observed and simulated Spearman correlations of monthly streamflows for adjacent months at site 1 (a), site 2 (b), and site 3 (c) and for the same month at different sites: (d) 1–2, (e) 1–3, and (f) 2–3.

near the diagonal with Spearman correlation 0.77. The scatterplots of the rank of observed monthly streamflow and one sequence of simulated streamflow pairs from the copula at different sites for the same month of March are also shown in Figure 1 (bottom). The simulated Spearman correlations are 0.59, 0.59, and 0.58, which are relatively close to the observed Spearman correlation (i.e., 0.65, 0.68, and 0.67).

[24] Boxplots were used to display the observed and simulated statistics, and the performance was judged to be good when a statistic fell within the boxplot [Nowak *et al.*, 2010; Salas and Lee, 2010]. Boxplots of the Spearman correlation of the observed and simulated monthly streamflows for three sites 1, 2, and 3 are shown in Figure 2 (left column), which display the temporal dependence between the adjacent months of a specific site. From Figure 2, it can be seen that for most months, the median of simulated statistics is within the boxplot. Box plots of the spatial dependence of the observed and simulated monthly streamflow of the same month between different sites are shown in Figure 2 (right column). All these simulations show good results since the observed Spearman correlation falls within the boxplots for most months. These results show that the dependence structure of the monthly streamflow at each site and between different sites can be preserved relatively well.

4. Conclusions

[25] The maximum entropy copula method is proposed for the multisite monthly streamflow simulation and shown

to be capable of modeling the rank correlation of monthly streamflows at different sites. The joint distribution (copula) is derived by specifying functions of the marginal probability as constraints having maximum entropy and its extension to higher dimensions for dependence modeling is straightforward. The proposed methodology can also be applied to similar topics, such as rainfall simulation and geostatistical interpolation. The potential drawbacks would be that the marginal properties of the copula are approximated numerically and the sum of tributary flows adding up to the downstream flow cannot be ensured with the current framework.

References

- Bárdossy, A., and G. Pegram (2009), Copula based multisite model for daily precipitation simulation, *Hydrol. Earth Syst. Sci.*, 13(12), 2299–2314.
- Berntsen, J., et al. (1991), An adaptive algorithm for the approximate calculation of multiple integrals, *ACM Trans. Math. Software*, 17(4), 437–451.
- Chu, B. (2011), Recovering copulas from limited information and an application to asset allocation, *J. Bank. Finan.*, 35(7), 1824–1842.
- Chui, C., and X. Wu (2009), Exponential series estimation of empirical copulas with application to financial returns, *Adv. Econometrics*, 25, 263–290.
- Genz, A. C., and A. Malik (1980), Remarks on algorithm 006: An adaptive algorithm for numerical integration over an N-dimensional rectangular region, *J. Comput. Appl. Math.*, 6(4), 295–302.
- Hao, Z., and V. Singh (2011), Single-site monthly streamflow simulation using entropy theory, *Water Resour. Res.*, 47, W09528, doi:10.1029/2010WR010208.

- Hao, Z., and V. Singh (2012), Entropy-copula method for single-site monthly streamflow simulation, *Water Resour. Res.*, 48, W06604, doi:10.1029/2011WR011419.
- Jaynes, E. (1957), Information theory and statistical mechanics, *Phys. Rev.*, 106(4), 620–630.
- Kao, S. C., and R. S. Govindaraju (2008), Trivariate statistical analysis of extreme rainfall events via the Plackett family of copulas, *Water Resour. Res.*, 44, W02415, doi:10.1029/2007WR006261.
- Kapur, J. (1989), *Maximum-Entropy Models in Science and Engineering*, John Wiley, New York.
- Koutsoyiannis, D., and A. Manetas (1996), Simple disaggregation by accurate adjusting procedures, *Water Resour. Res.*, 32(7), 2105–2117.
- Nowak, K., et al. (2010), A non-parametric stochastic approach for multi-site disaggregation of annual to daily streamflow, *Water Resour. Res.*, 46, W08529, doi:10.1029/2009WR008530.
- Prairie, J., et al. (2007), A stochastic nonparametric technique for space-time disaggregation of streamflows, *Water Resour. Res.*, 43, W03432, doi:10.1029/2005WR004721.
- Salas, J., and J. Delleur (1980), *Applied Modeling of Hydrologic Time Series*. Water Resour. Publ., Littleton, Colo.
- Salas, J., and T. Lee (2010), Nonparametric simulation of single site seasonal streamflows, *J. Hydrol. Eng.*, 15(4), 284–296.
- Shannon, C., and W. Weaver (1949), *The Mathematical Theory of Communication*, Univ. Illinois Press, Urbana, Ill.
- Shannon, C. E. (1948), A mathematical theory of communications, *Bell Syst. Tech. J.*, 27(7), 379–423.
- Sharma, A., and R. O'Neill (2002), A nonparametric approach for representing interannual dependence in monthly streamflow sequences, *Water Resour. Res.*, 38(7), 1100, doi:10.1029/2001WR000953.
- Sharma, A., et al. (1997), Streamflow simulation: A nonparametric approach, *Water Resour. Res.*, 33(2), 291–308, doi:10.1029/1096WR02839.
- Stedinger, J. R., and R. M. Vogel (1984), Disaggregation procedures for generating serially correlated flow vectors, *Water Resour. Res.*, 20(1), 47–56.
- Valencia, R. D., and J. J. Schaake (1973), Disaggregation processes in stochastic hydrology, *Water Resour. Res.*, 9(3), 580–585, doi:10.1029/WR1009i1003p00580.
- Vogel, R., and A. Shallcross (1996), The moving blocks bootstrap versus parametric time series models, *Water Resour. Res.*, 32(6), 1875–1882, doi:10.1029/1896WR00928.
- Wu, X. (2003), Calculation of maximum entropy densities with application to income distribution, *J. Econometrics*, 115(2), 347–354.