Check for updates

Copula-based stochastic simulation of hydrological data applied to Nile River flows

Taesam Lee and Jose D. Salas

ABSTRACT

Modelling a multivariate distribution is a classical issue in statistics. Copula functions offer a useful solution to this issue by modelling the multivariate distribution as a function of its marginal distributions. They have been used in various problems in hydrology and water management such as flood frequency analysis and drought or rainfall intensity-duration frequency analysis. However, to the knowledge of the author, they have not been applied for stochastic simulation of hydrologic data. In this study we explore the applicability of the copula concept for stochastic streamflow simulation. Parametric and non-parametric functions are applied for fitting the distribution of the original observed data and the serial dependence structure is then modelled with alternative copula functions. The pros and cons of different copula models are investigated by comparing the statistics of the generated data. Two major features of the copula models include: (1) portraying the heteroscedasticity embedded in the serial correlation of the observed data and (2) the flexibility of applicable marginal distributions. The suggested copula models are applied to simulate synthetic annual streamflow data of the Nile River. The results showed that the benefits of using these copula models are somewhat marginal with respect to the well-known modelling procedures.

Key words | copula, drought, nonparametric, stochastic simulation, streamflow, time series

Taesam Lee (corresponding author)
Engineering Research Institute,
Department of Civil Engineering,
Gyeongsang National University,
501 Jinju-daero, Jinju-si, Gyeongsangnam-do,
660-701 Republic of Korea
E-mail: *tae3lee@gnu.ac.kr*

Jose D. Salas
Department of Civil and Environmental Engineering,
B208 Engineering Building,
Colorado State University Fort Collins,
Colorado 80523-1372,
USA

INTRODUCTION

A copula is a multivariate distribution function with standard uniform marginals. Copulas allow the modelling of any multivariate distribution from its marginal distributions and its copula function. Because of this flexibility, copulas have become popular in several fields such as economics (e.g. Chen & Fan 2006a, b; Gagliardini & Gourieroux 2007; Chen *et al.* 2009; Lee & Long 2009) and hydrology for flood and drought analysis (e.g. De Michele & Salvadori 2003; Salvadori & De Michele 2004; Shiau 2006; De Michele *et al.* 2007; Shiau *et al.* 2007; Laux *et al.* 2009; Serinaldi *et al.* 2009; Shiau & Modarres 2009). An application of copulas for time series modelling and generation of annual streamflows is presented here.

Time series generation is useful for analyzing water resources systems such as for determining the required storage capacity of a reservoir and estimating the severity of drought, etc. Conventional time series generation models used in hydrology and water resources management such as autoregressive moving average (ARMA) (Salas *et al.* 1980; Salas & Obeysekera 1982; Salas 1993; Brockwell & Davis 2003; Salas *et al.* 2006) and fractional Gaussian noise (Mandelbrot & Wallis 1969; Mandelbrot 1971; Koutsoyiannis 2002) have some drawbacks such as the assumption of Gaussian marginal distribution. However, hydrologic data such as precipitation and streamflow usually have skewed distributions and sometimes show bi- or multimodal characteristics (Lall & Sharma 1996; Sharma *et al.* 1997; Prairie *et al.* 2006) where there is more than one generating mechanism (e.g. runoff resulting from convective rainfall and snowmelt). Data transformation methods such as logarithmic, Box-Cox, gamma and power transformations are common ways of tackling with the problem of skewness. However, there still exist some limitations that may affect the reproduction of

doi: 10.2166/nh.2011.085

Downloaded from http://iwaponline.com/hr/article-pdf/42/4/318/371177/318.pdf
by guest

historical statistics. The putative multiple mechanism explanation for bimodality cannot be presented within a simple transformation model. Alternatively, some non-Gaussian models have been proposed such as lag-1 Gamma Autoregressive (GAR-1) model which has a gamma marginal distribution (Fernandez & Salas 1990). However, the GAR-1 model is restricted to time series with short memory characteristics. Moreover, this model would not be useful where the marginal distribution of target data does not follow a gamma distribution, for example a multimodal distribution or heavy tail distribution. In contrast, a copula-based model can accommodate any feasible distribution for the marginal distribution of data. This salient and flexible feature of copula functions motivates its application to the time series generation of hydrological data.

Non-parametric models such as bootstrapping, k-nearest neighbour method (Lall & Sharma 1996) and first-order non-parametric model (Sharma *et al*. 1997; Sharma 2000; Sharma and O'Neill 2002; Harrold *et al*. 2003) have been suggested to overcome some of the limitations mentioned above. However, these non-parametric models also have their own drawbacks. For example, a bootstrapping or k-nearest neighbour model generates only the observed values available in the sample data. A first-order non-parametric model reproduces only the lag-1 persistency. In addition, semi-parametric models have been developed by combining parametric and non-parametric models to mitigate the drawbacks in both type of models (e.g. Srinivas & Srinivasan 2001, 2005a, b, 2006; Kim & Valdes 2005). In these semi-parametric models, the Autoregressive (AR) model component is extracted first from the historical data as

$$\varepsilon_t = (Y_t - \mu_y) - \sum_{i=1}^{p} \varphi_i (Y_{t-i} - \mu_y) \quad (1)$$

where Y_t is the serially correlated random variable, φ_i is the ith autoregressive coefficient, ε_t is the random component which is independent of Y_{t-i} and normally distributed and p is the order of the AR model. The residuals are then block-bootstrapped (e.g. Srinivas & Srinivasan 2001, 2005a, b, 2006) or modelled with non-parametric density estimator (NPD) as in Kim & Valdes (2005). However, the generated values obtained using the technique of Srinivas & Srinivasan (2001, 2005a, b, 2006) are still limited to the variability obtained from Equation (1) and may not produce expected extremes in the lower and upper tails. On the other hand, the marginal distribution of Y_t obtained following the NPD (Kim & Valdes 2005) may not reproduce the marginal distribution of Y_t.

In the current study, the use of copulas for modelling and simulating streamflow generation is illustrated. This demonstrates the flexibility of copula functions such as matching any marginal distributions and various different types of associations. The paper is organized as follows. The fundamental definitions and the methodology of copulas used for time series simulation are discussed first. The validation with a Monte Carlo simulation is then described followed by the application of the procedure. Finally, a summary and some conclusions are given.

DEFINITIONS AND METHODOLOGY

Basic concepts of copula

A copula is a function that links feasible marginal distributions to form a multivariate distribution function. This implies a substantial freedom in choosing the univariate marginal distributions once the desired dependence framework is established for relating the variables involved. The copula concept makes it easier to formulate multivariate models, compared to other complex and limited multivariate models.

Assuming two random variables (X and Y), Sklar's theorem (Sklar 1959) states that if $F_{X,Y}(x,y)$ is a bivariate distribution function with marginal cumulative distribution functions $F_X(x)$ and $F_Y(y)$, a copula C can be denoted

$$F_{X,Y}(x,y) = C(F_X(x), F_Y(y)) \quad (2)$$

Various copula functions are shown in Table 1. Under the assumption that the marginal distributions are continuous with probability density functions (PDFs) $f(x)$ and $f(y)$,

Downloaded from http://iwaponline.com/hr/article-pdf/42/4/318/371177/318.pdf
by guest

Table 1 | Commonly used one-parameter copulas (note that $\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp(-z^2)dz$ and $\boldsymbol{\Phi}_\theta$ is bivariate normal distribution function with the correlation parameter θ)

Type	$C(u, v)$	Parameter range
Frank	$C(u,v) = -\frac{1}{\theta}\ln\left[1 + \frac{(e^{-\theta u} - 1)(e^{-\theta v} - 1)}{e^{-\theta} - 1}\right]$	$\theta \neq 0$
Clayton	$C(u,v) = (u^{-\theta} + v^{-\theta} - 1)^{-1/\theta}$	$\theta \geq 0$
Gumbel	$C(u,v) = \exp[-\{(-\log u)^{\theta} + (-\log v)^{\theta}\}^{1/\theta}]$	$\theta \geq 1$
Ali-Mikhail-Haq	$C(u,v) = \frac{uv}{1 - (1-u)(1-v)\theta}$	$-1 \leq \theta \leq 1$
Bivariate Normal (BVN)	$C(u,v) = \boldsymbol{\Phi}_\theta(\Phi^{-1}(u), \Phi^{-1}(v))$	$-1 \leq \theta \leq 1$

the joint PDF becomes

$$f_{X,Y}(x,y) = f_X(x) f_Y(y) c(F_X(x), F_Y(y)) \tag{3}$$

where $c(u,v) = \partial^2 C(u,v)/\partial u \partial v$.

Copula-based stochastic simulation procedure

As mentioned above, the copula of a bivariate distribution may be defined with any univariate marginal distribution F_Y for the time series random variables Y_t and Y_{t-1} as

$$\begin{aligned} P[Y_{t-1} \leq y_{t-1}, Y_t \leq y_t] &= F(y_{t-1}, y_t) \\ &= C[F_Y(y_{t-1}), F_Y(y_t)] = C(u,v) \end{aligned} \tag{4}$$

where the marginal distribution of Y does not change over time (i.e. the process of generating Y is temporally stationary). $F_Y(y_{t-1}) = u$ and $F_Y(y_t) = v$ are uniformly distributed random variables between zero and one. The conditional bivariate copula is also defined as

$$F(y_t|y_{t-1}) = C_{2|1}(v|u) = \frac{\partial C(u,v)}{\partial u} \tag{5}$$

where the subscript 2|1 indicates the current time variable Y_t conditioned on the previous time variable Y_{t-1}. The analytical expression of the three copulas (i.e. Clayton, Frank and Gumbel) is shown in Appendix A (Equations (A2), (A6) and (A11), respectively). By generating an initial value y_1 from the fitted marginal distribution of data, the subsequent generating values are estimated by replacing $F(y_t|y_{t-1})$ as a uniform random number and then solving Equation (5). Moreover, in the case of a bivariate normal copula, a time series can be simulated differently by describing the copula as a first-order Markov processes (Chen & Fan 2006b) such that

$$\Phi^{-1}[F_Y(y_t)] = \theta \Phi^{-1}[F_Y(y_{t-1})] + \varepsilon_t, \tag{6}$$

or

$$N_t = \theta N_{t-1} + \varepsilon_t \tag{7}$$

where $\varepsilon_t \sim N(0,\ \sigma_\varepsilon^2)$, $N_t = \Phi^{-1}[F_Y(y_t)]$, Φ is the standard normal cumulative distribution function and θ is the bivariate normal (BVN) copula parameter (Table 1). If Y_t is non-normally distributed, then Equation (6) represents a first-order Markov processes determined by the BVN copula and non-Gaussian marginal distribution (Chen & Fan 2006b). The variance and the copula parameter are related to each other by $\sigma_\varepsilon^2 = (1 - \theta^2)\, \sigma_N^2 = (1 - \theta^2)$ since N_t is a standard normal variable.

The generating procedure with the conditional copula of Equation (5) is summarized in the following steps:

(i) Fit the marginal distribution of target data with a parametric or non-parametric model, say F_Y.
(ii) Estimate the parameter θ for the Clayton, Frank, Gumbel or BVN copula (Table 1). For this purpose we may apply the inference function for margins (IFM) approach (Joe 1997; Nelsen 1999). The detailed procedures of the copula parameter estimation are presented later.
(iii) The first value y_1 is generated randomly from the fitted marginal distribution F_Y (step i).
(iv) The next value is drawn from the conditional copula equation $C_{2|1}[F(y_t)|F(y_{t-1})]$. The four conditional copula functions (Clayton, Frank, Gumbel and BVN)

Downloaded from http://iwaponline.com/hr/article-pdf/42/4/318/371177/318.pdf
by guest

are used in application. As mentioned, $C_{2|1}$ is also a random variable which is uniformly distributed in the range [0, 1].

(v) A uniform random number w is generated so that $C_{2|1}[F(y_t)|F(y_{t-1})] = w$. By solving Equation (5) the subsequent value v is obtained since $u = F_Y(y_{t-1})$ is already known. For example, in the case of Clayton copula, Equation (A2) is solved to find v.

(vi) The generated data in real domain is obtained from the inverse transformation $y_t = F_Y^{-1}(v)$ where F_Y is the marginal distribution of Y_t obtained from (i).

(vii) Steps (iv) to (v) are repeated until all the desired values are simulated.

We can extend the order of the model described in Equation (6) to higher order autoregressive models. For example, using a trivariate normal copula (TNC), a second-order autoregressive can be described similar to Equations (4) and (5). The generating procedure in the case of the TNC model, assuming that the initial values y_1 and y_2 are known, is summarized as follows.

(i) Estimate the marginal cumulative distribution, say F_Y, from the available historical data.

(ii) The TNC parameters θ_1, θ_2 and σ_ε are estimated from the normally transformed data, i.e. $N_t = \Phi^{-1}[F_Y(y_t)]$.

(iii) The data can be generated from the AR(2) model (Salas 1993) as

$$N_t = \theta_1 N_{t-1} + \theta_2 N_{t-1} + \varepsilon_t$$

(iv) The generated data will be inversed back to the original domain as $F_Y^{-1}(\Phi(N_t))$.

(v) Steps (ii) to (iii) are repeated until all the desired values are simulated.

We used the kernel density estimator (KDE) and gamma distribution as the marginal distributions. The KDE is quite flexible and may accommodate complex frequency distributions; no assumption to any particular distribution for the observed data is required. The gamma distribution is commonly used for hydrologic data because of its flexibility and lower bound characteristic.

Estimation of copula parameters and marginal distributions

The parameters of the copula functions can be estimated by maximum likelihood (ML) and inference function for margins (IFM) (Joe 1997; Nelsen 1999). The exact maximum likelihood involves the log likelihood of the observed data as a function of all the parameters including the marginal distribution parameters. However, the IFM method splits the parameters for marginal distributions and the parameters for dependence structure. We employ IFM in the current study as it is rather simpler with less computational requirements than THE ML method (Favre *et al.* 2004). After fitting the marginal distribution of the target time series (using parameter estimation by maximum likelihood), the copula parameter (θ) is estimated from maximizing the log-likelihood function of the copula such that

$$\log L(x, y; \theta) = \log L_C(x, y; \theta) + \log L(y) + \log L(x) \tag{8}$$

where $\log L(\cdot)$ is the log-likelihood of the fitted PDF to data and $\log L_C(x, y; \theta)$ is the log likelihood function of $\log L_C(x, y; \theta)$; $\log L_C(x, y; \theta)$ can therefore serve as a copula model selection criteria. From fitting the copulas in Table 1, the copula with the largest value of $\log L_C(x, y; \theta)$ is selected as suggested by Klugman & Parsa (1999). $\log L_C(y_t, y_{t-1}; \theta)$ for the time series variables is

$$\log L_C(y_t, y_{t-1}; \theta) = \sum_{t=2}^{N} \log c(F_Y(y_t), F_Y(y_{t-1}); \theta) \tag{9}$$

To estimate the copula parameter, find a value θ which maximizes $\log L_C(y_t, y_{t-1}; \theta)$. This is done by obtaining θ such that

$$\begin{aligned} dLL_c &= \frac{\partial \log L_C(y_t, y_{t-1}; \theta)}{\partial \theta} \\ &= \sum_{t=2}^{N} \frac{\partial \log c(F_Y(y_t), F_Y(y_t); \theta)}{\partial \theta} \\ &= \sum_{t=2}^{N} dl_c(F_Y(y_t), F_Y(y_{t-1}); \theta) = 0 \end{aligned} \tag{10}$$

where the functions $dl_c(u, v; \theta) = \partial \log(c(u, v))/\partial\theta$ for each copula (Clayton, Frank, Gumbel and Gaussian) are given

Downloaded from http://iwaponline.com/hr/article-pdf/42/4/318/371177/318.pdf
by guest

Table 2 | Calculated parameters of Monte Carlo simulation, estimated parameters (third row) and log-likelihood in Equation (9) for the Nile River station

		Clayton	Frank	Gumbel	Gaussian
Simulated ($\tau = 0.7$)	Parameter	4.67	11.41	3.33	0.89
	Parameter	0.47	2.62	1.39	0.38
Nile	Log-likelihood	3.06	8.31	11.05	9.08

by Equations (A.3), (A.8), (A.14) and (A.18). To solve this equation, we use the Newton–Raphson method. After setting the initial value θ_0, θ_{i+1} is calculated iteratively by

$$\theta_{i+1} = \theta_i - \frac{dLL_c}{ddLL_c} \tag{11}$$

until $\theta_{i+1} = \theta_i$ or $ddLL_c = \partial^2 \log L_c(y;\theta)/\partial\theta^2 = \sum_{t=2}^{N} ddl_c \times (F_Y(y_t), F_Y(y_{t-1});\theta)$. The functions $ddl_c(y_t, y_{t-1};\theta)$ for each copula are given by Equations (A.4), (A.9), (A.15) and (A.19). Alternatively, an optimization technique might also be employed for the copula parameter estimation finding a value that optimizes the quantity in Equation (8).

The copula parameter estimation procedure above is rather intricate for the case of trivariate normal (TVN) copula. Alternatively, the following procedure is used for TVN copula parameters as well as BVN copula.

(i) Estimate $\hat{F}(y_i)$ with a mathematical equation or the empirical formulation described in Equation (12) below.
(ii) Transform $\hat{F}(y_i)$ into normal variate $z_i = \Phi^{-1}[\hat{F}(y_i)]$, $i = 1, \ldots, n$.
(iii) Fit the normal variate (z_i) series to AR(1) for the BVN copula or AR(2) for the TVN copula and estimate the copula parameters (BVN-$\theta, \sigma_\varepsilon^2$ or TVN-$\theta_1, \theta_2, \sigma_\varepsilon^2$) as described by Salas (1993).

The marginal cumulative distribution of data required in step (2) can be obtained by either (a) using a parametric or non-parametric distribution fitting method or (b) using an

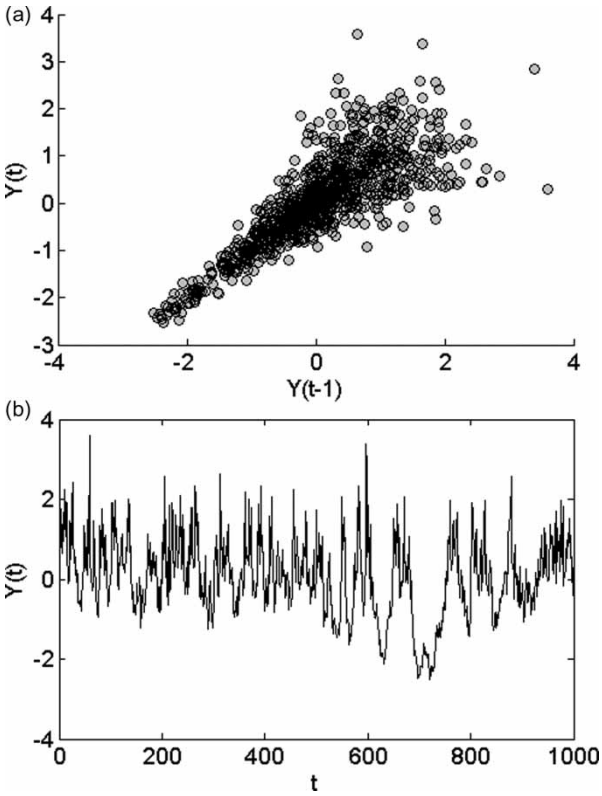

Figure 1 | Scatter plot (a) between y_t and y_{t-1} and time series plot (b) of synthetic data sampled from Clayton copula ($\theta = 4.67$) and normal marginal with 1000 record length.

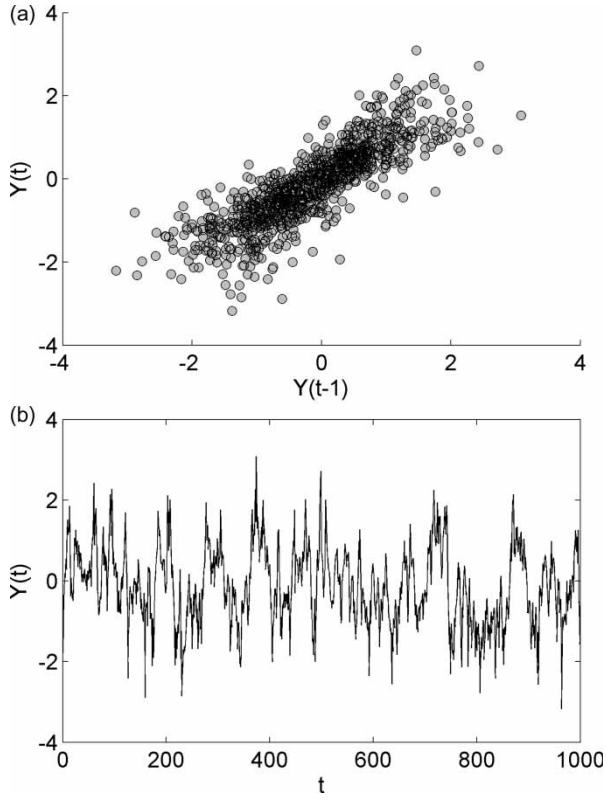

Figure 2 | As Figure 1 except with Frank copula ($\theta = 11.4$).

Downloaded from http://iwaponline.com/hr/article-pdf/42/4/318/371177/318.pdf
by guest

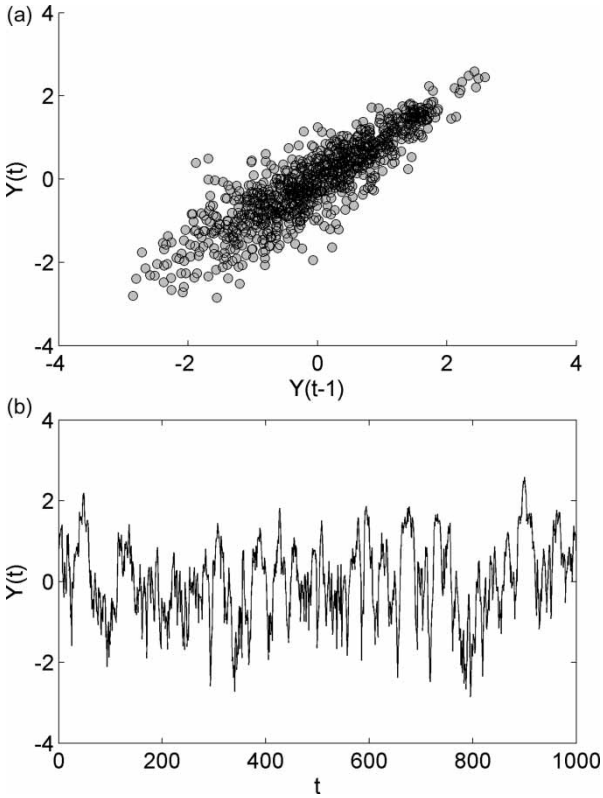

Figure 3 | As Figure 1 except with Frank copula ($\theta = 3.33$).

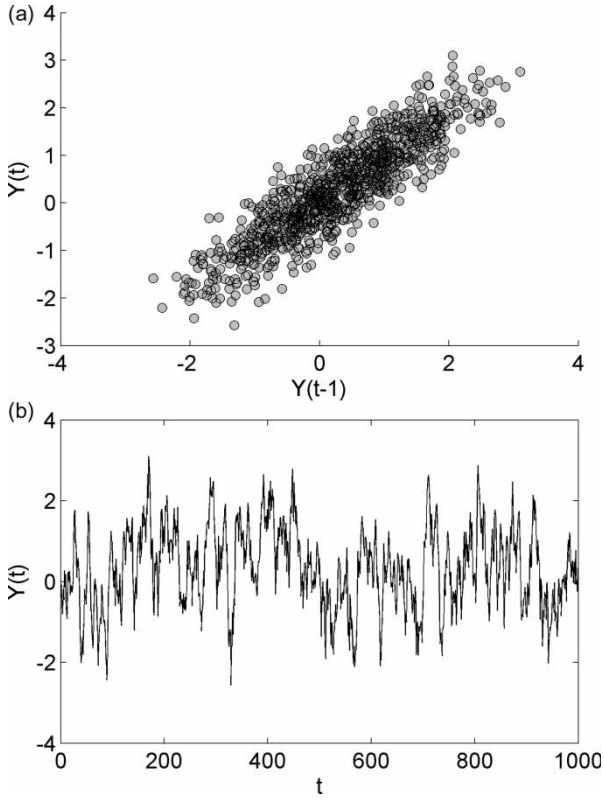

Figure 4 | As Figure 1 except with Frank copula ($\theta = 3.33$).

empirical distribution formulation such that

$$\hat{F}(y_i) = \frac{1}{n+1}\sum_{t=1}^{n} I\{Y_t \le y_i\} \tag{12}$$

$I\{a\}$ is an indicator function such that if the condition a is met then it is 1; otherwise it is 0, n is the number of observed data and $i = 1, \ldots, n$.

Validation with simulation

In order to demonstrate the performance of the copula-based time series model, a Monte Carlo simulation is performed. With a fixed association (Kendall's tau rank correlation coefficient, $\tau = 0.7$), the parameters of different copulas are calculated employing the relationship between τ and these parameters. Table 2 presents the values of the calculated parameters (Nelsen 1999). For each copula, we generate a 1000-record series. The marginal distribution employed is a standard normal distribution for all cases.

RESULTS

The relationship between y_t and y_{t-1} is shown graphically in the upper panels of Figures 1–4 (Clayton, Frank, Gumbel and Gaussian, respectively) along with the time series in the lower panels of these figures. The Clayton, Frank and Gumbel copulas are a family of Archimedean copulas. The detailed properties of Archimedean copulas are explained in Nelsen (1999). The dependence structure presented with the Clayton copula is illustrated in Figure 1(a). In the Clayton copula, a variable is strongly dependent in the case of lower values while there is more scatter for larger values. The time series generated with the Clayton copula has a peculiar behaviour reflecting the strong dependence over the lower values in Figure 1(b). When a value is small, the next sequence is also very likely to be small; when a value is large, the next sequence is more random. This behaviour might be quite useful as streamflow is very dependent in low flows and less dependent in high flows. The time series of the synthesized 1000 values reflects this behaviour:

Downloaded from http://iwaponline.com/hr/article-pdf/42/4/318/371177/318.pdf
by guest

strong persistency can be seen while values are low quantities and weak dependency can be seen while values are high. We could compare a target data time series with a synthesized one by checking whether they have similar behaviours. We can also check if the series have the same heteroscedascity in which variance depends on the current condition.

The Frank copula is quite different from the Clayton copula. In this copula, the dependencies near both extremes are lessened compared to that in the centre of the distribution as seen in the middle part of Figure 2(a). This is shown in the time series (Figure 2(b)) where there is higher variability at the high and low values.

The Gumbel copula behaves in an opposite way to the Clayton copula. The larger values have higher dependency (Figure 3(a)) than that for lower values. This feature is well presented in the realization of the copula in Figure 3(b). The higher the value, the more persistent it is. This is opposite to the Clayton copula where the lower the value, the more persistent it is. The results for the Gaussian copula are shown in Figures 4(a) and (b), demonstrating equal dependence across all values.

CASE STUDY

To test the model, the annual streamflow data of the Nile River at Aswan is used (units are in milliard cubic metres, 1 milliard $= 10^9$). The flow records are available for the period 1871–1989. The streamflow records show significant persistence and the suggested copula model is compared with the ARMA(1,1) model with log-transformation. The application involves two parts. The first part examines the model performance of different copulas. The second part checks the behaviour of different marginal distributions.

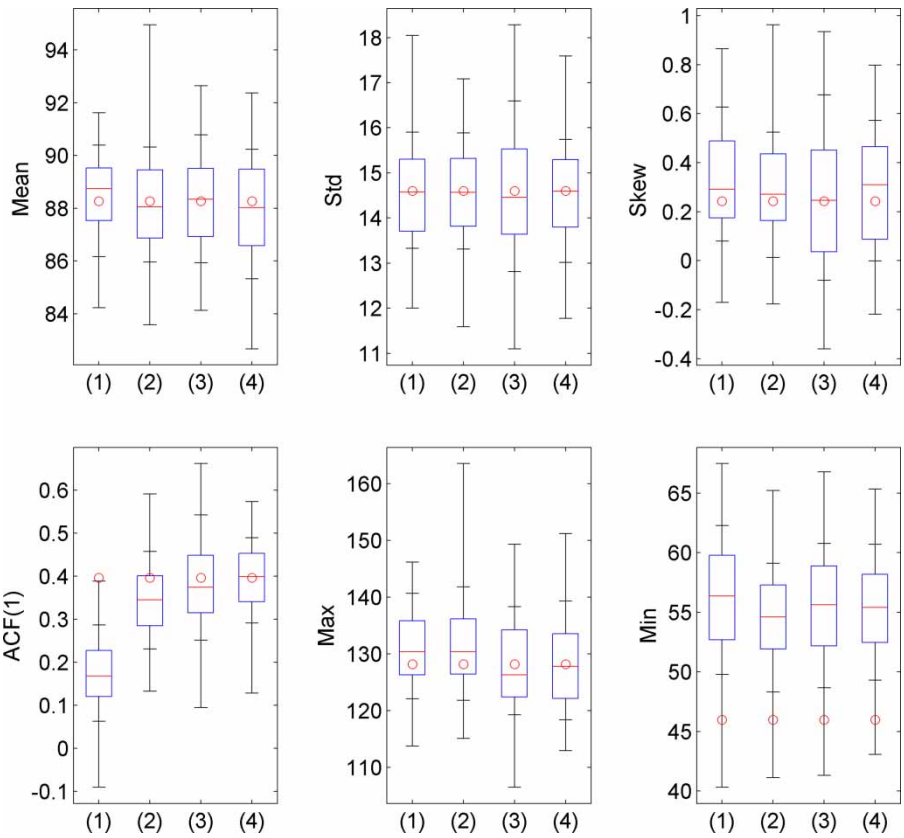

Figure 5 | Comparison of statistics of historical and generated data based on (1) Clayton copula, (2) Frank copula, (3) Gumbel copula and (4) Gaussian copula with gamma marginal distribution fitted to the Nile River flows. The 'o' symbol represents the historical statistics and the boxplot does the statistics of the generataed data from each copula model (100 sets).

Downloaded from http://iwaponline.com/hr/article-pdf/42/4/318/371177/318.pdf
by guest

Results with different copulas

The performance of different copulas such as (a) Clayton, (b) Frank, (c) Gumbel and (d) Gaussian has been compared according to various statistics. For all four copulas, the gamma marginal distribution is fitted to the Nile River annual flow data. The estimated parameters of the four copulas are presented in the third row of Table 2.

The basic statistics estimated from the historical and generated data are presented in Figure 5. The mean, standard deviation, skewness and maximum are well preserved by all models but the minimum values are underestimated in all cases. The lag-1 serial correlation for the Clayton copula is also underestimated. The log-likelihood in Equation (8) is estimated in order to help identify a preferred copula model. The result is shown in the fourth row of Table 2, and reveals that the Gumbel copula shows a better fit among the four copulas.

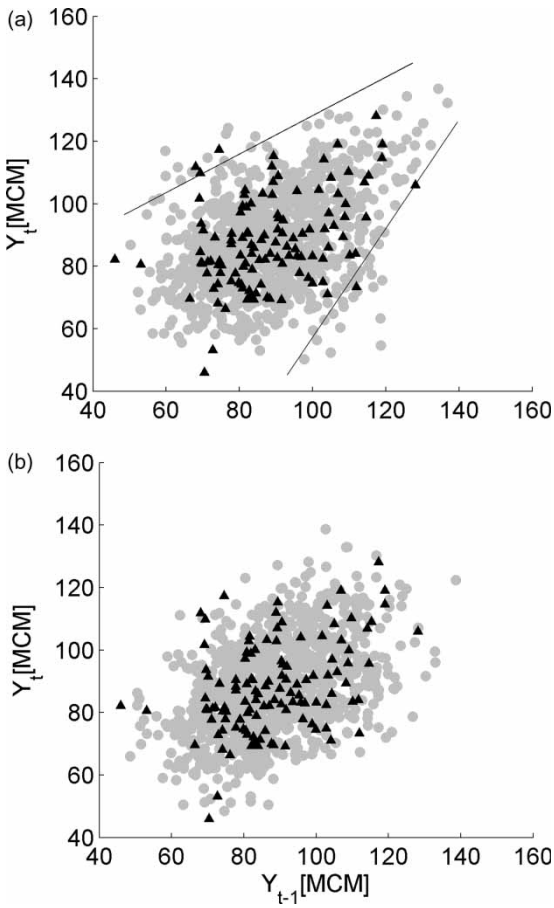

Figure 6 | Scatterplot of Y_t and Y_{t-1} based on 50 samples generated using (a) Gumbel and (b) Gaussian copulas with Gamma Marginal (filled triangle) and historical (grey-filled circle).

To illustrate the relationship between successive flow values, scatter plots of y_t and y_{t-1} from the simulated data are shown in Figure 6 for the Gumbel copula (upper panel) and the Gaussian copula (lower panel), overlaid with those obtained from the historical data. The figure shows the results based on the historical sample (filled triangles) and 50 generated samples (filled circles). The behaviours obtained from the two examples are quite different. The noticeable feature is that for the Gumbel copula, the values y_t which follow small values y_{t-1} are spread widely; the spread of values following larger values is much smaller. Two lines are added to the figure to enhance this feature. For the Gaussian copula, the relationship is not remarkable.

Figure 7 presents the state-dependent correlation coefficients defined as in Sharma *et al.* (1997). The historical values of the above-median forward (af) and backward (ab) are higher than for the below-median forward (bf) and backward (bb). 'Forwards, above median' correlation (af) is defined as the correlation between above-median flows and flows in the subsequent time step; 'forwards, below median' correlation (bf) is the correlation between all below-median flows and the flow in the subsequent time step; 'backwards, above median' correlation (ab) is the correlation between above-median flows and the preceding time steps flow; and 'backwards, below median' correlation (bb) is the correlation between below-median flows and the preceding time steps flow. Figure 7 implies that the historical flow data (presented with circles) have a higher persistency for larger values than for lower values. The results shown in Figure 7 indicate that only the Gumbel copula reproduces the heteroscedasticity feature found in the historical data. The Frank and Gaussian copulas have the same degree of the dependency regardless of the state while the Clayton copula has an opposite relation compared to the historical. This feature is well reflected in the estimated log-likelihood function (i.e. a high likelihood for the Gumbel copula and a low likelihood for the Clayton copula in the fourth row of Table 2).

Results with different marginal distributions

In addition to the comparison of the four copula functions, three other models, i.e. (a) the log-transformed ARMA(1,1)

Downloaded from http://iwaponline.com/hr/article-pdf/42/4/318/371177/318.pdf
by guest

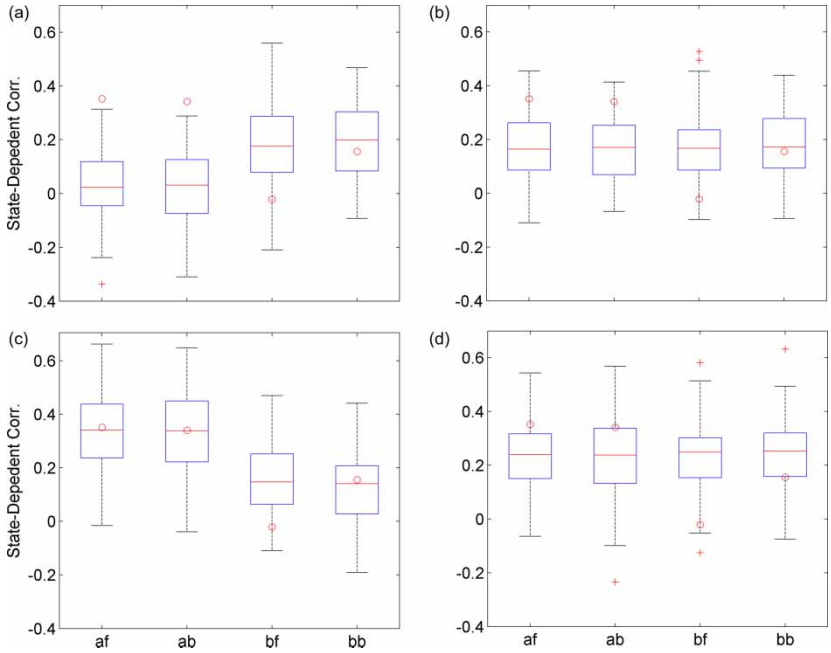

Figure 7 | State-dependent correlation of the historical (circle) and generated (boxplot) with gamma marginal and different copulas such as (a) Clayton, (b) Frank, (c) Gumbel and (d) Gaussian.

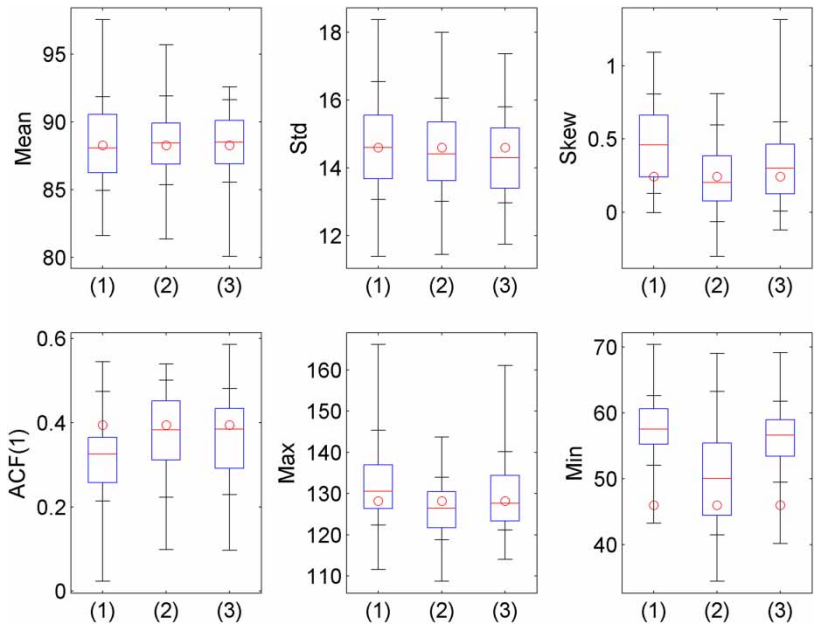

Figure 8 | Statistics of historical and generated data (1) LTARMA(1,1); (2) CTVN-KDE; and (3) CTVN-Gamma fitted to the Nile River flows. The boxplots are based on 100 generated samples and the circles represent the historical quantite.

Downloaded from http://iwaponline.com/hr/article-pdf/42/4/318/371177/318.pdf
by guest

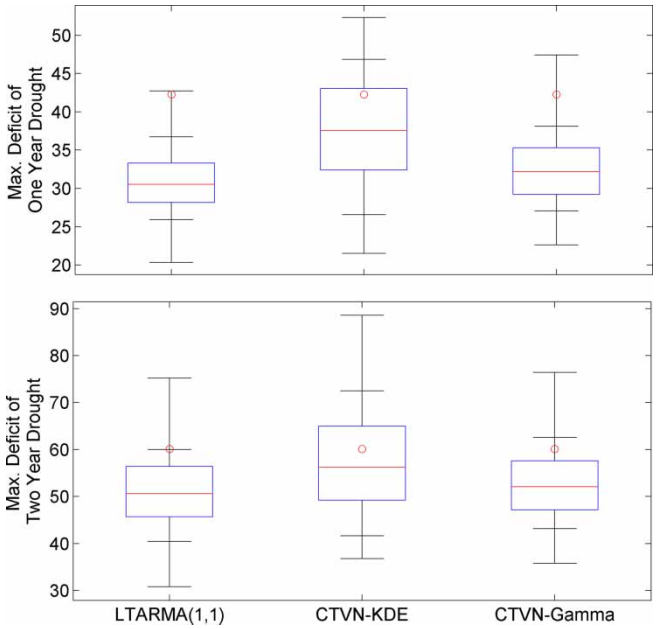

Figure 9 | Maximum deficit of 1 year (upper panel) and 2 year (lower panel) droughts for historical data (circle) and generated data (boxplot) based on 100 sample (boxplot) with the same record length as historical employing three models: LTARMA(1,1), CTVN-KDE and CTVN-Gamma.

(LTARMA(1,1)); (b) the trivariate normal copula with Kernel Density Estimate (TNC-KDE); and (c) the trivariate normal copula with gamma marginal (TNC-Gamma), have been compared. The estimated key statistics from the historical and generated (100 traces) data are presented in Figure 8. This shows that all statistics except the minimum are well preserved in all three models (in the sense that the historical statistics fall within the 25–75% whiskers). In this regard, one may also consider that the minimum is also well preserved by the TNC-KDE model, although there is a large spread which is remarkable, but the minimum is definitely overestimated by the other two models.

In Figure 9, 1 year and 2 year droughts are compared. The drought event is defined as the consecutive shortage relative to the water demand represented by the historical mean. This type of drought is relatively important when water supply systems rely primarily on direct river diversions and short-term storage. From the figure, it is revealed that the TNC-KDE model is superior in preserving the historical maximum deficit of 1 and 2 year droughts while the other two models significantly underestimate these statistics. However, for long-term droughts and storage capacity statistics (not shown) the results for all models are similar.

SUMMARY AND CONCLUSIONS

Hydrological stochastic simulation models employing the Clayton, Frank, Gumbel and Gaussian copulas have been examined in this study. The Gumbel copula has a higher association for large values and lower association for small values while the opposite is true for the Clayton copula. The Frank copula has a high association in the middle part and low in the extremes, while the Gaussian copula shows the same association strength regardless of the value. These characteristics are reflected in the results obtained by simulation where the relationship between successive flows of the Nile River appears to be reasonably well represented by the Gumbel copula model.

An interesting feature of the copula models is that any distribution can be employed for representing the marginal distribution. We also examined the applicability of Gamma distribution and Kernel density estimate as marginal distributions of the trivariate normal copula and compared their results against the LTARMA(1,1) model. The results showed that the benefits of using the copula models are somewhat marginal with respect to the well-known modelling procedures (e.g. the LTARMA). Perhaps the only significant difference is in better reproducing short-term droughts (e.g. 1 or 2 year droughts), but as far as longer term droughts and storage capacity statistics no significant difference has been found.

The weakness of the copula-based model is that a higher order of this time series model is difficult to apply since it would require higher dimensional copulas, which are considerably more difficult to apply except in the case of the Gaussian copula. However, the Gaussian copula is not different from the traditional ARMA time series models applied to normalized data.

ACKNOWLEDGEMENTS

Comments and suggestions from Reza Modarres, Fasbender Dominique and two anonymous reviewers as well as the

Downloaded from http://iwaponline.com/hr/article-pdf/42/4/318/371177/318.pdf
by guest

editor, Ian Littlewood, were helpful in preparing this manuscript.

REFERENCES

Brockwell, P. J. & Davis, R. A. 2003 *Introduction to Time Series and Forecasting*. Springer, New York.

Chen, X. H. & Fan, Y. Q. 2006a Estimation and model selection of semiparametric copula-based multivariate dynamic models under copula misspecification. *Journal of Econometrics* **135** (1–2), 125–154.

Chen, X. H. & Fan, Y. Q. 2006b Estimation of copula-based semiparametric time series models. *Journal of Econometrics* **130** (2), 307–335.

Chen, X., Koenker, R. & Xiao, Z. 2009 Copula-based nonlinear quantile autoregression. *Econometrics Journal* **12**, S50–S67.

De Michele, C. & Salvadori, G. 2003 A Generalized Pareto intensity-duration model of storm rainfall exploiting 2-Copulas. *Journal of Geophysical Research-Atmospheres* **108** (D2), 4067.

De Michele, C., Salvadori, G., Passoni, G. & Vezzoli, R. 2007 A multivariate model of sea storms using copulas. *Coastal Engineering* **54** (10), 734–751.

Favre, A. C., El Adlouni, S., Perreault, L., Thiemonge, N. & Bobee, B. 2004 Multivariate hydrological frequency analysis using copulas. *Water Resources Research* **40** (1), W01101.

Fernandez, B. & Salas, J. D. 1990 Gamma-Autoregressive Models for Stream-Flow Simulation. *Journal of Hydraulic Engineering-ASCE* **116** (11), 1403–1414.

Gagliardini, P. & Gourieroux, C. 2007 An efficient nonparametric estimator for models with nonlinear dependence. *Journal of Econometrics* **137** (1), 189–229.

Harrold, T. I., Sharma, A. & Sheather, S. J. 2003 A nonparametric model for stochastic generation of daily rainfall amounts. *Water Resources Research* **39** (12), 1343.

Joe, H. 1997 *Multivariate Models and Multivariate Dependence Concepts*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability, London, New York.

Kim, T. W. & Valdes, J. B. 2005 Synthetic generation of hydrologic time series based on nonparametric random generation. *Journal of Hydrologic Engineering* **10** (5), 395–404.

Klugman, S. A. & Parsa, R. 1999 Fitting bivariate loss distributions with copulas. *Insurance: Mathematics & Economics* **24** (1–2), 139–148.

Koutsoyiannis, D. 2002 The Hurst phenomenon and fractional Gaussian noise made easy. *Hydrological Sciences Journal-Journal Des Sciences Hydrologiques* **47** (4), 573–595.

Lall, U. & Sharma, A. 1996 A nearest neighbor bootstrap for resampling hydrologic time series. *Water Resources Research* **32** (3), 679–693.

Laux, P., Wagner, S., Wagner, A., Jacobeit, J., Bárdossy, A. & Kunstmann, H. 2009 Modelling daily precipitation features in the Volta Basin of West Africa. *International Journal of Climatology* **29** (7), 937–954.

Lee, T. H. & Long, X. D. 2009 Copula-based multivariate GARCH model with uncorrelated dependent errors. *Journal of Econometrics* **150** (2), 207–218.

Mandelbrot, B. 1971 Fast fractional Gaussian noise generator. *Water Resources Research* **7** (3), 543–553.

Mandelbrot, B. & Wallis, J. R. 1969 Computer experiments with fractional Gaussian noises.1. Averages and variances. *Water Resources Research* **5** (1), 228–241.

Nelsen, R. B. 1999 *An Introduction to Copulas*. Springer-Verlag, New York.

Prairie, J. R., Rajagopalan, B., Fulp, T. J. & Zagona, E. A. 2006 Modified K-NN model for stochastic streamflow simulation. *Journal of Hydrologic Engineering* **11** (4), 371–378.

Salas, J. D. 1993 Analysis and modeling of hydrologic time series. In: *Handbook of Hydrology* (D. R. Maidment ed.), McGraw-Hill, New York.

Salas, J. D., Delleur, J. W., Yevjevich, V. & Lane, W. L. 1980 *Applied Modeling of Hydrologic Time Series*. Water Resources Publications, Littleton, Colorado.

Salas, J. D. & Obeysekera, J. T. B. 1982 Arma Model Identification of Hydrologic Time-Series. *Water Resources Research* **18** (4), 1011–1021.

Salas, J. D., Sveinsson, O. G., Lane, W. L. & Frevert, D. K. 2006 Stochastic streamflow simulation using SAMS-2003. *Journal of Irrigation and Drainage Engineering-ASCE* **132** (2), 112–122.

Salvadori, G. & De Michele, C. 2004 Analytical calculation of storm volume statistics involving Pareto-like intensity-duration marginals. *Geophysical Research Letters* **31** (4), L04502.

Serinaldi, F., Bonaccorso, B., Cancelliere, A. & Grimaldi, S. 2009 Probabilistic characterization of drought properties through copulas. *Physics and Chemistry of the Earth* **34** (10–12), 596–605.

Sharma, A. 2000 Seasonal to interannual rainfall probabilistic forecasts for improved water supply management: Part 3 – A nonparametric probabilistic forecast model. *Journal of Hydrology* **239** (1–4), 249–258.

Sharma, A. & O'Neill, R. 2002 A nonparametric approach for representing interannual dependence in monthly streamflow sequences. *Water Resources Research* **38** (7), 5.1–5.10.

Sharma, A., Tarboton, D. G. & Lall, U. 1997 Streamflow simulation: A nonparametric approach. *Water Resources Research* **33** (2), 291–308.

Shiau, J. T. 2006 Fitting drought duration and severity with two-dimensional copulas. *Water Resources Management* **20** (5), 795–815.

Shiau, J. T., Feng, S. & Nadaraiah, S. 2007 Assessment of hydrological droughts for the Yellow River, China, using copulas. *Hydrological Processes* **21** (16), 2157–2163.

Shiau, J. T. & Modarres, R. 2009 Copula-based drought severity-duration-frequency analysis in Iran. *Meteorological Applications* **16** (4), 481–489.

Downloaded from http://iwaponline.com/hr/article-pdf/42/4/318/371177/318.pdf
by guest

Sklar, M. 1959 Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de Statistique de L'Université de Paris* **8**, 229–231.

Srinivas, V. V. & Srinivasan, K. 2001 A hybrid stochastic model for multiseason streamflow simulation. *Water Resources Research* **37** (10), 2537–2549.

Srinivas, V. V. & Srinivasan, K. 2005a Hybrid moving block bootstrap for stochastic simulation of multi-site multi-season streamflows. *Journal of Hydrology* **302** (1–4), 307–330.

Srinivas, V. V. & Srinivasan, K. 2005b Matched block bootstrap for resampling multiseason hydrologic time series. *Hydrological Processes* **19** (18), 3659–3682.

Srinivas, V. V. & Srinivasan, K. 2006 Hybrid matched-block bootstrap for stochastic simulation of multiseason streamflows. *Journal of Hydrology* **329** (1–2), 1–15.

First received 16 July 2009; accepted in revised form 13 July 2010

APPENDIX A

Mathematical derivation for generation and parameter estimation of selected copulas

$C(u,v)$ is a bivariate copula function with two variables, such as u and v, and it has one parameter θ. The followings are the derivatives of four copula functions and their log-likelihood functions.

Clayton copula

$$C(u,v) = (u^{-\theta} + v^{-\theta} - 1)^{-1/\theta} = A_c^{-1/\theta} \tag{A.1}$$

where $A_c = u^{-\theta} + v^{-\theta} - 1$.

$$C_{v|u} = \frac{\partial C(u,v)}{\partial u} = (u^{-\theta} + v^{-\theta} - 1)u^{-\theta-1} = A_c u^{-\theta-1} \tag{A.2}$$

$$\frac{\partial C(u,v)}{\partial u \partial v} = c(u,v) = (1+\theta)u^{-(\theta+1)}v^{-(\theta+1)}A_c^{-\frac{1}{\theta}-2}$$

$$\begin{aligned} dl_c(u,v;\theta) &= \frac{\partial \log(c(u,v))}{\partial \theta} \\ &= (1+\theta)^{-1} - \log(uv) + \frac{\log(A_c)}{\theta^2} \\ &\quad + \frac{(\theta^{-1}+2)(u^{-\theta}\log u + v^{-\theta}\log v)}{A_c} \end{aligned} \tag{A.3}$$

$$\begin{aligned} ddl_c(u,v;\theta) &= \frac{\partial^2 \log(c(u,v))}{(\partial\theta)^2} \\ &= -(1+\theta)^{-2} - 2\theta^{-3}\log(A_c) - 2\theta^{-2}B/A_c \\ &\quad + (1/\theta+2)(B_c^2 A_c^{-2} - A_c^{-1}(u^{-\theta}(\log u)^2 \\ &\quad + v^{-\theta}(\log v)^2)) \end{aligned} \tag{A.4}$$

where $B_c = u^{-\theta}\log u + v^{-\theta}\log v$.

Frank copula

$$C(u,v) = -\frac{1}{\theta}\log\left[1 + \frac{(e^{-\theta u}-1)(e^{-\theta v}-1)}{(e^{-\theta}-1)}\right] \tag{A.5}$$

$$C_{v|u} = \frac{\partial C(u,v)}{\partial u} = -\frac{1}{\theta}\log\left[1 + \frac{(e^{-\theta u}-1)(e^{-\theta v}-1)}{(e^{-\theta}-1)}\right] \tag{A.6}$$

$$\frac{\partial C(u,v)}{\partial u \partial v} = c(u,v) = -\frac{\theta e^{-\theta(u+v)}(e^{-\theta}-1)}{\{e^{-\theta(u+v)} - e^{-\theta u} - e^{-\theta v} + e^{-\theta}\}^2} \tag{A.7}$$

$$\begin{aligned} dl_c(u,v;\theta) &= \frac{\partial \log(c(u,v))}{\partial \theta} = \frac{1}{\theta\log\theta} + \frac{1}{C_F} \\ &\quad + \frac{u+v}{\theta} - \frac{2\{C_F(B_F\theta^u u + A_F\theta^v) - \theta A_F B_F\}}{\theta C_F(C_F - A_F B_F)} \end{aligned} \tag{A.8}$$

where $A_F = 1-\theta^u$, $B_F = 1-\theta^v$ and $C_F = 1-\theta$

$$\begin{aligned} &ddl_c(u,v;\theta) \\ &= \frac{\partial^2 \log(c(u,v))}{(\partial\theta)^2} = \frac{\{C(B_F\theta^\mu u + A_F\theta^\nu) - \theta A_F B_F\}^2}{\theta^2 C_F^2 (C_F - A_F B_F)^2} - \frac{1+\log\theta}{(\theta\log\theta)^2} \\ &+ \frac{1}{C_F^2} + \frac{u+v}{\theta^2} - \frac{2C_F[\{2\theta v C_F + \theta^2 - C_F^2 v(1-v)\}\theta^\nu - \theta^2]}{\theta^2 C_F^2 (C_F - A_F B_F)} \\ &- \frac{2B_F[\{2\theta u C_F + \theta^2 - C_F^2 u(1-u)\}\theta^\mu - \theta^2] - 4C_F^2 uv\theta^\mu\theta^\nu}{\theta^2 C_F^2 (C_F - A_F B_F)} \end{aligned} \tag{A.9}$$

Gumbel copula

$$\begin{aligned} C(u,v) &= \exp[-\{(-\log u)^\theta + (-\log v)^\theta\}^{1/\theta}] \\ &= \exp[-A^{1/\theta}] \end{aligned} \tag{A.10}$$

where, $A = (-\log u)^\theta + (-\log v)^\theta$

$$\begin{aligned} C_{v|u} &= \frac{\partial C(u,v)}{\partial u} \\ &= \exp(-A^{1/\theta})A^{-1+1/\theta}(-\log u)^{\theta-1}u^{-1} \end{aligned} \tag{A.11}$$

Downloaded from http://iwaponline.com/hr/article-pdf/42/4/318/371177/318.pdf
by guest

$$\frac{\partial C(u,v)}{\partial u \partial v} = c(u,v) = \exp(-A^{1/\theta}(uv)^{-1}A^{-2+2/\theta}(\log u \log v)^{\theta-1} \times \{1+(\theta-1)A^{-1/\theta}\} \quad (A.12)$$

$$l_c(u,v;\theta) = \log(c(u,v)) = -A^{1/\theta} - \log(uv) + (-2+2/\theta)\log A + (\theta-1)\log(\log u \log v) + \log\{1+(\theta-1)A^{-1/\theta}\} \quad (A.13)$$

$$dl_c(u,v;\theta) = \frac{\partial \log(c(u,v))}{\partial \theta} = -dA^{1/\theta} - \frac{2}{\theta^2}\log A + (-2+2/\theta)\frac{dA}{A} + \log(\log u \log v) + \frac{A^{-1/\theta} + (\theta-1)dA^{-1/\theta}}{1+(\theta-1)A^{-1/\theta}} \quad (A.14)$$

where

$$dA = \frac{\partial A}{\partial \theta} = (-\log u)^{\theta}\log(-\log u) + (-\log v)^{\theta}\log(-\log v)$$

$$dA^{1/\theta} = \frac{\partial A^{1/\theta}}{\partial \theta} = -A^{1/\theta}\theta^{-2}\log(A) + \theta^{-1}A^{1/\theta-1}dA$$

$$dA^{1/\theta-1} = \frac{\partial A^{1/\theta-1}}{\partial \theta} = \frac{dA^{1/\theta}A - A^{1/\theta}dA}{A^2}$$

$$dA^{-1/\theta} = \frac{\partial A^{-1/\theta}}{\partial \theta} = A^{-1/\theta}\theta^{-2}\log(A) - \theta^{-1}A^{-1/\theta-1}dA$$

$$ddl_c(u,v;\theta) = \frac{\partial^2 \log(c(u,v))}{(\partial\theta)^2} = -d^2A^{1/\theta} - \left(-\frac{4}{\theta^3}\log A + \frac{2}{\theta^2}\frac{dA}{A}\right) + \left(-\frac{2}{\theta^2}\right)\frac{dA}{A} + (-2+2/\theta)\left(\frac{d^2A}{A} - \frac{(dA)^2}{A^2}\right) + \frac{dA^{-1/\theta} + dA^{1/\theta} + (\theta-1)d^2A^{-1/\theta}}{1+(\theta-1)A^{-1/\theta}} - \frac{(A^{-1/\theta} + (\theta-1)dA^{1/\theta})(A^{-1/\theta} + (\theta-1)dA^{-1/\theta})}{\{1+(\theta-1)A^{-1/\theta}\}^2} \quad (A.15)$$

where

$$d^2A^{1/\theta} = -[dA^{1/\theta}\theta^{-2}\log(A) - 2A^{1/\theta}\theta^{-3}\log(A) + A^{1/\theta-1}\theta^{-2}dA] + [-\theta^{-2}A^{1/\theta-1}dA + \theta^{-1}dA^{1/\theta-1}dA + \theta^{-1}A^{1/\theta-1}d^2A]$$

$$d^2A = (-\log u)^{\theta}\{\log(-\log u)\}^2 + (-\log v)^{\theta}\{\log(-\log v)\}^2$$

$$d^2A^{-1/\theta} = [dA^{-1/\theta}\theta^{-2}\log(A) - 2A^{-1/\theta}\theta^{-3}\log(A) + A^{-1/\theta-1}\theta^{-2}dA] - [-\theta^{-2}A^{1/\theta-1}dA + \theta^{-1}dA^{-1/\theta-1}dA + \theta^{-1}A^{-1/\theta-1}d^2A]$$

$$dA^{-1/\theta-1} = \frac{\partial A^{-1/\theta-1}}{\partial \theta} = \frac{dA^{-1/\theta}A - A^{-1/\theta}dA}{A^2}$$

Gaussian copula

$$C(u,v) = \mathbf{\Phi}_{\theta}(\Phi^{-1}(u), \Phi^{-1}(v)) \quad (A.16)$$

$$\frac{\partial C(u,v)}{\partial u \partial v} = c(u,v) = \frac{\phi_{\theta}(\Phi^{-1}(u), \Phi^{-1}(v))}{\phi(\Phi^{-1}(u))\ \phi(\Phi^{-1}(v))} \quad (A.17)$$

$$dl_c(u,v;\theta) = \frac{\partial \log(c(u,v))}{\partial \theta} = [\theta(1-\theta^2) - \theta\{\Phi^{-1}(u)^2 + \Phi^{-1}(v)^2 + (1+\theta^2)\Phi^{-1}(u)\Phi^{-1}(v)\}](1-\theta^2)^2 \quad (A.18)$$

$$ddl_c(u,v;\theta) = \frac{\partial^2 \log(c(u,v))}{\partial \theta^2} = (1+\theta^2)(1-\theta^2)^{-2} + (1-\theta^2)^{-3}(6\theta + 2\theta^3) \times \Phi^{-1}(u)\Phi^{-1}(v) - (1+3\theta^2)\{\Phi^{-1}(u)^2 + \Phi^{-1}(v)^2\}(1-\theta^2)^{-3} \quad (A.19)$$

Downloaded from http://iwaponline.com/hr/article-pdf/42/4/318/371177/318.pdf
by guest