Sampling rainfall events: a novel approach to generate large correlated samples

Siao Sun, Soon-Thiam Khu and Slobodan Djordjević

ABSTRACT

It is essential that the correlation between variables is considered properly when using sampling-based methods. Modeling rainfall events is of great interest because the rainfall is usually the major driving force of hydrosystems. A novel method for generating correlated samples is introduced providing that the marginal distributions of variables as well as their correlations between them are known. The basic idea of the method is to adjust the correlations between samples by rearranging the positions inside marginal samples after each marginal sample is generated according to its distribution. The group method is developed in order to facilitate an efficient generation of correlated samples of large sizes. The theoretical precision associated with the group method is derived. There is a trade off between the computational efficiency of the algorithm and the precision that can be achieved when using different numbers of groups. The method is successfully applied to two cases of rainfall sample generation problems. The effectiveness of the group method is studied. Large group numbers are recommended in practical use as the samples distribute more broadly regardless of computational efficiency.

Key words | correlation coefficient, dependence, hydrosystem modeling, rainfall events, sampling

Siao Sun (corresponding author)
LGCIE,
INSA de Lyon,
34 avenue des Arts,
69621 Villeurbanne cedex,
France
E-mail: *siao.sun@insa-lyon.fr*

Soon-Thiam Khu
Civil Engineering,
Faculty of Engineering and Physical Sciences,
University of Surrey,
GU2 7XH,
UK

Slobodan Djordjević
College of Engineering,
Mathematics and Physical Sciences,
University of Exeter,
Exeter,
EX4 4QF,
UK

INTRODUCTION

When modeling hydrosystems, it is common to assume that the variables or inputs are independent of each other. However, such assumptions are normally not true in practice, and the variables may be correlated with each other. For example, the intensity and duration of rainfall events are usually observed to be negatively correlated; the peak, volume and duration of runoff are probably dependent on each other; a regional cross-correlation among the precipitations of different regions is possibly present when regional effects are considered.

If these variables are generated via sampling, it is important to ensure that the dependency of the samples is either incorporated in the sampling process, or maintained in the resultant samples through adjustments. Any dependencies among the variables must be considered when solving such problems as substantial biases can result if correlations are neglected (Smith *et al.* 1992). Kapelan *et al.* (2005) asserted that neglecting demand correlation in water distribution system design under uncertainty may lead to the under-design of such systems. Douglas *et al.* (2000) believed that a dramatically different interpretation would have been achieved if regional cross-correlation had been ignored when analyzing the trends in flood and low flows in the USA. Grimaldi & Serinaldi (2006) stated that for a complete analysis of the three main characteristics of a flood event, i.e. peak, volume and duration, full understanding of these variables and relationships is necessary. Kanso *et al.* (2006) found a clear correlation between the parameters in urban runoff quality modeling. Yue (2000) pointed out that the severity of the damage caused by a storm is in fact a function of the correlated storm peak and the total amount of storm.

Rainfall is usually a major driving force in hydrosystems modeling, such as sewer system design, combined sewer system overflow assessment, flood risk assessment, river discharge evaluation, reservoir design, etc. One of the most

doi: 10.2166/nh.2012.113

frequently used approaches to generate synthetic rainfall data as model input is to use a sampling based technique. Rainfall is commonly characterized by specifying two or more dependent variables such as rainfall duration, total rainfall depth (or the average rainfall intensity) and dry period (Muzik 2002; Rahman *et al*. 2002; Thorndahl & Willems 2008). A traditional way of considering the dependence between variables is to use classical families of multivariate probability distributions such as the normal, log-normal, and exponential distributions. However, such an approach suffers from the limitation that the behaviors of the multiple variables must be characterized by the same parametric family of univariate distributions. Another possibility to consider dependence between variables is via copula, which is a joint distribution function that can capture relationships between variables. Copula models are just beginning to make their way into the hydrological area. Salvadori & Michele (2006, 2007), Serinaldi & Grimaldi (2007) and Zhang & Singh (2007) used copulas to construct correlation structures of rainfall variables. When working with copulas the choice of a good fitting dependence structure is important. The experience of choosing a suitable copula to describe certain rainfall data is limited, though this is still an area of active research.

Pearson correlation coefficient (CC) and the Spearman rank correlation coefficient (RCC) (Helton & Davis 2003) are another two possibilities widely used to measure the dependence between variables. The CC r_{XY} between two random variables X and Y with expected values μ_X and μ_Y and standard deviations σ_X and σ_Y is defined as:

$$r_{XY} = \frac{\mathrm{cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{E((X - \mu_X)(Y - \mu_Y))}{\sigma_X \sigma_Y} \tag{1}$$

where cov is the covariance and E is the expected value operator. The CC r_{xy} of samples x_i and y_i, $i = 1,2,\ldots n$, which are series of measurements or samples from variables X and Y, can be calculated by:

$$r_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}} \tag{2}$$

where $\bar{x}$ and $\bar{y}$ are the sample means of X and Y, s_x and s_y are the sample standard deviations. CC is a value between -1 and 1, providing a measure of the strength of the linear relationship between two variables, with CC $= 1$ denoting the case of an increasing linear relationship, and CC $= -1$ the case of a decreasing linear relationship. The values in between in all other cases indicate the degree of linear dependence between the variables. The closer the coefficient is to either -1 or 1, the stronger the linear correlation between the variables. The RCC is defined similarly to the CC as Equations (1) and (2) but with rank-transformed data. Rank-transforming is a step to convert the original values of the samples according to the orders of the samples. More specifically, the smallest value of samples of a variable is given a rank of 1; the next smallest value is given a rank of 2; and so on up to the largest value which is given a rank equal to the sample size n. After rank-transforming, RCC is calculated by Equation (2) with the transformed ranks instead of the original values in CC. The numerical values of RCC also fall between -1 and 1 but it is a measurement of the strength of the monotonic relationship between two variables.

Henceforth, CC is a measure of linearity of the relationship between variables; while RCC is a measure of the monotonicity in the relationship between variables (Conover & Iman 1981). They are both useful in describing the dependency of variables. However, in sampling-based simulations, CC value is often used for generating correlated samples when variables are normally distributed. Otherwise, RCC is predominantly used due to the general difficulty in maintaining a specified CC value when the required variables to be generated are not normally distributed (Morgan & Henrion 1990).

This paper aims to present a method for generating correlated samples of large size with given correlation coefficients (either CC or RCC) and provided that the marginal distributions of these variables are known. The paper is organized as follows: the following section gives a brief overview of the methods for generating correlated samples; then the methodology is presented where the basic principle of the proposed approach is explained and the group method is introduced; afterwards the proposed method is applied to two cases of rainfall samples generation; at last conclusions are drawn.

CORRELATED SAMPLES GENERATION

Methods for generating correlated samples with some specific marginal distributions such as normal distributions (Cheng 1985) and Pearson family distributions (Parrish 1990) are available. Iman & Conover (1982) proposed a distribution-free and simple method for generating correlated samples and it became widely used afterwards. However, this method suffers from two drawbacks: (1) it can only generate samples with given RCC, but not CC; and (2) it is difficult to guarantee the accuracy of the resultant RCC of generated samples. Li & Hammond (1975) and Lurie & Goldberg (1998) presented some two-step methods for generating correlated samples by (1) generating an intermediate normal sample of multivariables; and (2) transforming underlying correlated normal sample into the target non-normal sample. The intermediate normal sample should have appropriate correlations which are determined by an inversion of a double integral. This is a computationally intensive procedure and a feasible solution may not be available. The correlated samples can also be generated through copulas. As stated by Genest & Rivest (1993), a natural way of specifying the distribution function is to examine the copula and marginal distributions separately. Schweizer & Wolff (1981) established that the copula accounts for all the dependence between two random variables X and Y: if g_1 and g_2 are strictly increasing functions over the range of X and Y, the transformed variables $g_1(X)$ and $g_2(Y)$ have the same copula as X and Y. Regardless of the scale in which each variable is measured, the copula is able to capture the synchronized fluctuations between X and Y. Therefore, it is possible to express RCC solely in terms of the copula function. However, as CC is affected by changes of (nonlinear) scale, specifying the copula alone is not sufficient and it requires the marginal distributions to be known (Frees & Valdez 1997). Hence, the method using copula to generate samples is generally constrained to those for which the required correlation given by RCC instead of CC.

Charmpis & Panteli (2004) and Vořechovský & Novák (2009) proposed a heuristic approach for generating correlated samples. They considered the case of sampling from a multivariate distribution with correlated variables where the specified marginal probability distributions as well as their CCs are known. The approach includes two distinct steps: the first step generates univariate random samples independently from their own specified marginal probability distributions; in the second step the generated univariate samples are rearranged in a way that the values of samples generated in the first step do not change but the positions of them change, thus the correlations between variables are adjusted and the desired correlations can be obtained. The simulated annealing (SA) algorithm is used to rearrange the positions of univariate samples in order to find a suitable arrangement. Chakraborty (2006) provided some theoretical results about 'how close to the target correlation' by rearranging univariate samples and proposed a deterministic initialization algorithm based on the theoretical results. This deterministic algorithm is claimed to speed up convergence of stochastic optimization algorithms such as SA.

METHODOLOGY

Basic principle

In this paper, the discussion focuses on generating correlated samples with known marginal distributions and dependences given by CC or RCC. The basic idea of the methodology is to approximate the sample correlations (either CC or RCC) to the target correlations by rearranging the positions of samples after each marginal sample has been independently generated according to its own marginal distribution (Charmpis & Panteli 2004; Chakraborty 2006; Vořechovský & Novák 2009). For ease of referencing, if an m-variable sample $\boldsymbol{x}$ of size n is required, the elements of $\boldsymbol{x}$ can be denoted by x_{ij}:

$$x = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix} \tag{3}$$

A column of the matrix $\boldsymbol{x}_i$ is generated from a known marginal distribution and presents a marginal sample. The correlations between these m marginal samples form a

matrix $\boldsymbol{c}$ that is required to approximate the target matrix $\boldsymbol{c}^*$:$\mathbf{c} \cong \mathbf{c}^*$. The elements of the two correlation matrices $\boldsymbol{c}$ and $\boldsymbol{c}^*$ are respectively denoted as c_{ij} and c_{ij}^* (i, $j = 1,2,\ldots m$). As $\boldsymbol{c}$ and $\boldsymbol{c}^*$ are symmetric matrices and their diagonal elements are equal to unity, only the elements below the leading diagonal need to be considered in the rearrangement procedure. The objective of the procedure is to minimize the difference between c_{ij} and c_{ij}^* ($i > j$). The positions of $\boldsymbol{x}_1$ can be kept unchanged because of the symmetry of the problem. The permutations of $\boldsymbol{x}_2$, $\boldsymbol{x}_3$,…, $\boldsymbol{x}_m$ are then determined one at a time, e.g.:

- $\boldsymbol{x}_2$ is permuted letting $c_{12} = c_{12}^*$;
- $\boldsymbol{x}_3$ is permuted letting $c_{13} = c_{13}^*$, $c_{23} = c_{23}^*$;
- …
- $\boldsymbol{x}_m$ is permuted letting $c_{1m} = c_{1m}^*, c_{2m} = c_{2m}^*, c_{m-1m} = c_{m-1m}^*$.

The root mean square error (RMSE) is used to estimate the closeness of the resultant and required correlation vector. The size of the solution space for this problem (the number of the possible permutations of $\boldsymbol{x}_2, \ldots, \boldsymbol{x}_m$) is $(n!)^{m-1}$. Charmpis & Panteli (2004) and Vořechovský & Novák (2009) used a stochastic optimization algorithm (the SA) to find a good permutation for each marginal sample that makes $\mathbf{c}_i \cong \mathbf{c}_i^*$. Random pairwise positions in a marginal sample were interchanged according to the SA scheme. As the solution space of the problem increases dramatically when the size of samples increases, Vořechovský & Novák (2009) emphasized the method is designed for generating small samples for Monte Carlo simulations (MCS). Chakraborty (2006) believed that stochastic algorithms such as SA are not efficient near local optima and proposed a deterministic algorithm. The difference between the correlations of two neighbor permutations of two samples $\boldsymbol{x}$ and $\boldsymbol{y}$ ($\boldsymbol{x}$ and $\boldsymbol{y}$ are both a marginal sample) by swapping the values at positions i and j in sample $\boldsymbol{y}$ is:

$$\mathrm{corr}(x, y) - \mathrm{corr}(x, \pi(y)) = \frac{(x_i - x_j)}{\sigma_x}\frac{(y_i - y_j)}{\sigma_y}\Big/ n \tag{4}$$

where $\pi(y)$ is a permutation of sample $\boldsymbol{y}$. In the deterministic algorithm, if a random pairwise transposition enables the derived correlation $\boldsymbol{c}$ to move towards the target correlation $\boldsymbol{c}^*$, the transposition is accepted. Such deterministic transpositions are continuously made until $\boldsymbol{c}$ cannot be improved anymore or until it achieves certain required precision. This deterministic algorithm is used in this paper as the basis of the proposed methodology.

The group method and the theoretical basis

In the application of sampling rainfall events, the required size of the generated samples depends on the range of return periods of interest and on the required accuracy. Rahman *et al.* (2002) used 10,000 samples in order to produce relatively stable estimation of derived flood frequency curves with return periods ranging from 1 to 100 years. They stated that the number of sample events should increase by orders of magnitude if the purpose of the MCS was to estimate flood events in the extreme range or if the random variables are more independent. Hence there is a potential need for a method for generating correlated samples of large sizes.

The group method proposed in this paper is to facilitate an efficient generation of correlated samples of large sizes. The method can be outlined as follows: (1) generate samples of variables from their marginal distributions; (2) rearrange the samples from each variable in increasing (or decreasing) order; (3) bunch every k samples of each variable to form groups (the reason for doing this is given afterwards), where k is a number of samples in a group and is a divisor of n. Thus (n/k) groups are formed; (4) adjust the positions of the groups while keeping the interior positions of the samples within a group to adjust the correlations between variables.

By using the group method, the dimension of the solution space of the problem is reduced. When k equals 1, the algorithm is equivalent to the method described by Chakraborty (2006) as the positions of all the elements need to be determined.

If a bivariable problem (two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ of length n are known) is considered, as Chakraborty (2006) has demonstrated, $\mathrm{corr}(x, \pi(y))$ is maximized as $c_{\max}$ when x and $\pi(y)$ are concordant (i.e. x and $\pi(y)$ are both arranged with increasing (or decreasing) orders), while it is minimized to be $c_{\min}$ when they are discordant (e.g. x is arranged with increasing orders while $\pi(y)$ with decreasing orders). When the group method is introduced, the resultant correlation range is unable to cover the whole range of $[c_{\min}, c_{\max}]$. The achievable correlation range depends

on how the samples are grouped and they fall into two categories: (1) both $\boldsymbol{x}$ and $\boldsymbol{y}$ are arranged in an increasing order, and groups are then formed; (2) $\boldsymbol{x}$ is increasingly arranged while $\boldsymbol{y}$ is decreasingly arranged and they are then divided into groups. Let $\boldsymbol{x}'$ and $\boldsymbol{y}'$ present the formed groups, $[c'_{\min}, c'_{\max}]$ denote the range of correlations that can be achieved with the group method. In both cases the correlation is maximized when $\boldsymbol{x}'$ and $\pi(y')$ are concordant, while it is minimized when they are discordant. However, there is $c'_{\min} > c_{\min}$ in case (1) because the elements inside groups of $\boldsymbol{x}'$ and $\pi(y')$ keep the monotonic relationship though groups are discordant. For the same reason, $c'_{\max} < c_{\max}$ in case (2). Therefore when forming groups, attention should be paid to make the target correlation c^* in the achievable interval $[c'_{\min}, c'_{\max}]$.

If the samples are grouped with random orders, it is difficult to identify the concordant or discordant order of the groups, thus difficult to identify $[c'_{\min}, c'_{\max}]$. Moreover, as will be shown later, it is difficult to obtain the achievable precision of the resultant correlation if the samples are reordered randomly when forming the groups.

After forming the groups, the group positions are rearranged using the deterministic algorithm. Pairwise groups are randomly chosen and their positions are exchanged only if the derived correlation is improved. The upper bound of the precision that can be achieved by the group method is determined by the distance of any permutation of groups to its nearest neighbor (as used by Chakraborty (2006)). Let τ be the permutation obtained from π by swapping positions of the groups i and j in sample $\boldsymbol{y}$, the difference between the two neighbor permutations is:

$$\mathrm{corr}(x, \pi(y)) - \mathrm{corr}(x, \tau(y)) = \sum_{l=1}^{k} \frac{(x_i^{(l)} - x_j^{(l)})}{\sigma_x} \frac{(y_i^{(l)} - y_j^{(l)})}{\sigma_y} \Big/ n \tag{5}$$

where l denotes the interior positions of samples within a group. If a permutation $\varepsilon(\boldsymbol{x})$ of x is consecutively ordered: $x_{(1)} \le \ldots \le x_{(n)}$, let $\delta_{x,k}$ denote the largest difference between $x_{(i)}$ and $x_{(i+k)}$ scaled by σ_x, define $\delta_{x,k} = 1/\sigma_x \max_{k+1 \le i < n}(x_{(i)} - x_{(i-k)})$, and the ith value belongs to the tth group. Similarly define $\delta_{y,k} = 1/\sigma_y \max_{k+1 \le i < n}(y_{(i)} - y_{(i-k)})$. Thus an upper bound of the achievable precision by the group method is:

$$\begin{aligned}
\delta &= \max_{\pi} \min_{i,j} \left| \sum_{l=1}^{k} \frac{(x_i^{(l)} - x_j^{(l)})}{\sigma_x} \frac{(y_{\pi(i)}^{(l)} - y_{\pi(j)}^{(l)})}{\sigma_y} \right| \Big/ n \\
&\le \max_{\pi} \min_{t} \left| \sum_{l=1}^{k} \frac{\max_{k+1 \le m \le n} (x_{(m)} - x_{(m-k)})}{\sigma_x} \frac{(y_{\pi(t)}^{(l)} - y_{\pi(t-1)}^{(l)})}{\sigma_y} \right| \Big/ n \\
&= \delta_x \max_{\pi} \min_{t} \left| \sum_{l=1}^{k} \frac{(y_{\pi(t)}^{(l)} - y_{\pi(t-1)}^{(l)})}{\sigma_y} \right| \Big/ n \\
&\le \delta_{x,k} \left| \sum_{l=1}^{k} \frac{\max_{k+1 \le i \le n} (y_{(i)} - y_{(i-k)})}{\sigma_y} \right| \Big/ n \\
&\le \delta_{x,k} \delta_{y,k} / (n/k)
\end{aligned} \tag{6}$$

The following rule shows how close c can get to c^* for a bivariable problem with the group method.

1. If $c^* \ge c_{\max}$ then $c = c_{\max}$.
2. If $c^* \le c_{\min}$ then $c = c_{\min}$.
3. If $c^* \in [c_{\min}, c_{\max}]$, then $|c - c^*| < \frac{1}{2}\delta$.

While $c* \in [c_{\min}, c_{\max}]$, the precision of the achievable sample correlations is associated with the number of the groups (n/k) whose positions need to be determined from Equation (6). The precision decreases as the number of samples in a group increases (as the number of groups decrease). However, the dimension of the solution space decreases from $n!$ to $(n/k)!$ when the group method is introduced. Thus there is a trade off between the achievable precision and the computational efficiency when using the group method. In practical engineering use, the precision of the obtained CCs is usually not required to be as precise as possible. For instance, one usually provides limited accuracy about the target CC (say two- or three-decimal accuracy) with good confidence. In the application section, it can be observed that the required precision (0.001) is easy to achieve even with a small number of groups n/k such as 50. The achievable precision of correlation by the developed method can also be intuitively considered as follows: there are $(n/k)!$ possible arrangements of samples by changing the positions of samples in $\boldsymbol{y}$ with each arrangement having a CC in the interval of $[-1,1]$. Such a large number of values of possible CCs distribute in a relatively

small possible range. Hence the distance between neighbor CCs must be tiny and the precision of correlation by this method can be generally high.

Figure 1 shows the procedure of the group method for generating correlated bivariable samples. It is not necessary to compute each correlation after a pairwise exchange of group positions using the correlation definition in Equation (2). The new correlation can be derived by adding Equation (5) to the current correlation. The algorithm for more than two variables is very similar, only the comparison is executed between more CCs. As mentioned before, the RMSE is used to determine the acceptance of a transposition.

APPLICATIONS

In this paper, the proposed method is applied to two applications of sampling rainfall events. In the first example, rainfall events are represented by two correlated variables – rainfall depth and duration, and in the second example, rainfall events are characterized by three correlated variables, i.e. antecedent dry period, wet period and average intensity.

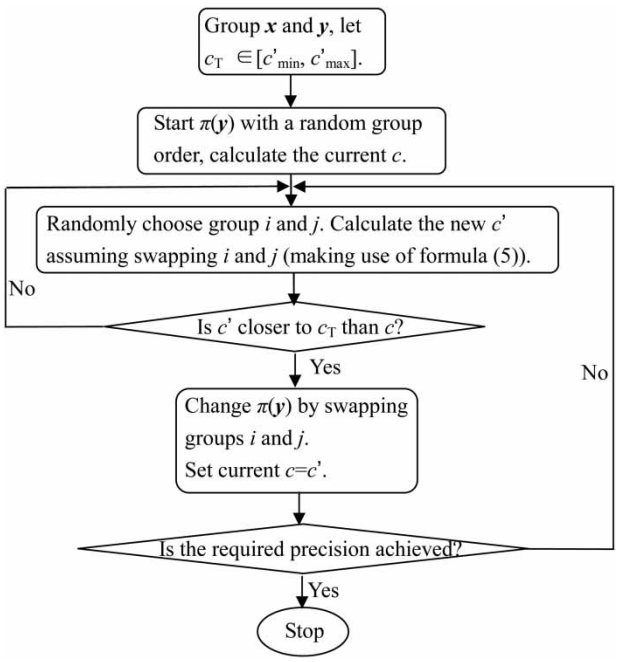

Figure 1 | The procedure of the group method for generating bivariate samples.

Rainfall events of correlated rainfall depth and duration

As the focus of this paper is on the generation of correlated samples, the marginal probability distributions of variables representing rainfall as well as their CCs are assumed to be known *a priori*. Such distributions and CCs between the variables can be obtained from statistical analysis of the real rainfall events in an actual design. The most commonly used distributions for describing rainfall variables are exponential distributions (Bacchi *et al.* 1994; Kurothe *et al.* 1997; Goel *et al.* 2000), Gumbel distributions (Koutsoyiannis & Baloursos 2000; Coles *et al.* 2003), Gamma distribution (Segond *et al.* 2006) and general Pareto distributions (Salvadori & Michele 2006). For simplicity, the distribution of rainfall depth and duration are both characterized by the Gumbel distribution in this case but such simplification has no impact on the results of the analysis. The Gumbel distribution can be expressed as:

$$F(x) = \exp\left(-\exp\left(-\frac{x-\mu}{\alpha}\right)\right) \tag{7}$$

where μ and α are the parameters of the Gumbel distribution.

In this example, the parameters are set as follows: $\mu =$ 28 mm and $\alpha = 8$ for the variable of rainfall depth and $\mu =$ 60 min and $\alpha = 12$ for the variable of rainfall duration. The CC between rainfall depth and duration is 0.78 (the same as Thorndahl & Willems (2008) obtained from the analysis of 18 years' rainfall data).

A total of 100,000 samples are drawn for rainfall depth and duration variables from their marginal distributions. The required precision on the CC is set as 0.001, i.e. the algorithm stops when the difference between the obtained correlation and the target correlation is less than 0.001. Groups of 50, 100, 500, 1,000, 5,000, 10,000, 50,000 and 100,000 are applied. Due to the stochastic nature contained in part of the algorithm, the proposed method was run 10 times for each number of groups and the average number of steps of 10 runs are listed in Table 1. It is observed that the number of steps towards the resultant samples increases as the number of groups increase. This agrees with the previous analysis on computational efficiency with the group method.

Table 1 | Average computational steps to obtain the required precision of sampling rainfall events represented by two variables with group method

Number of groups (n/k)	Average number of steps
50	326
100	490
500	2,326
1,000	4,700
5,000	23,529
10,000	47,991
50,000	239,249
100,000	477,650

The typical scatter plots of samples obtained by different number of groups are presented in Figure 2. As in this example, many plots distribute in a relatively small area, it is impossible to present the trend of samples showing all of the plots. Figure 2 endeavors to reveal the trend by the degree of grayness of areas. All figures show an obvious linear correlated relationship between the two variables. The samples obtained from small numbers of groups tend to cluster. It can be explained by the fact that the samples in a group always keep their relative close positions in the searching algorithm when samples are grouped. Thus the figure also shows that the marginal distributions and CCs

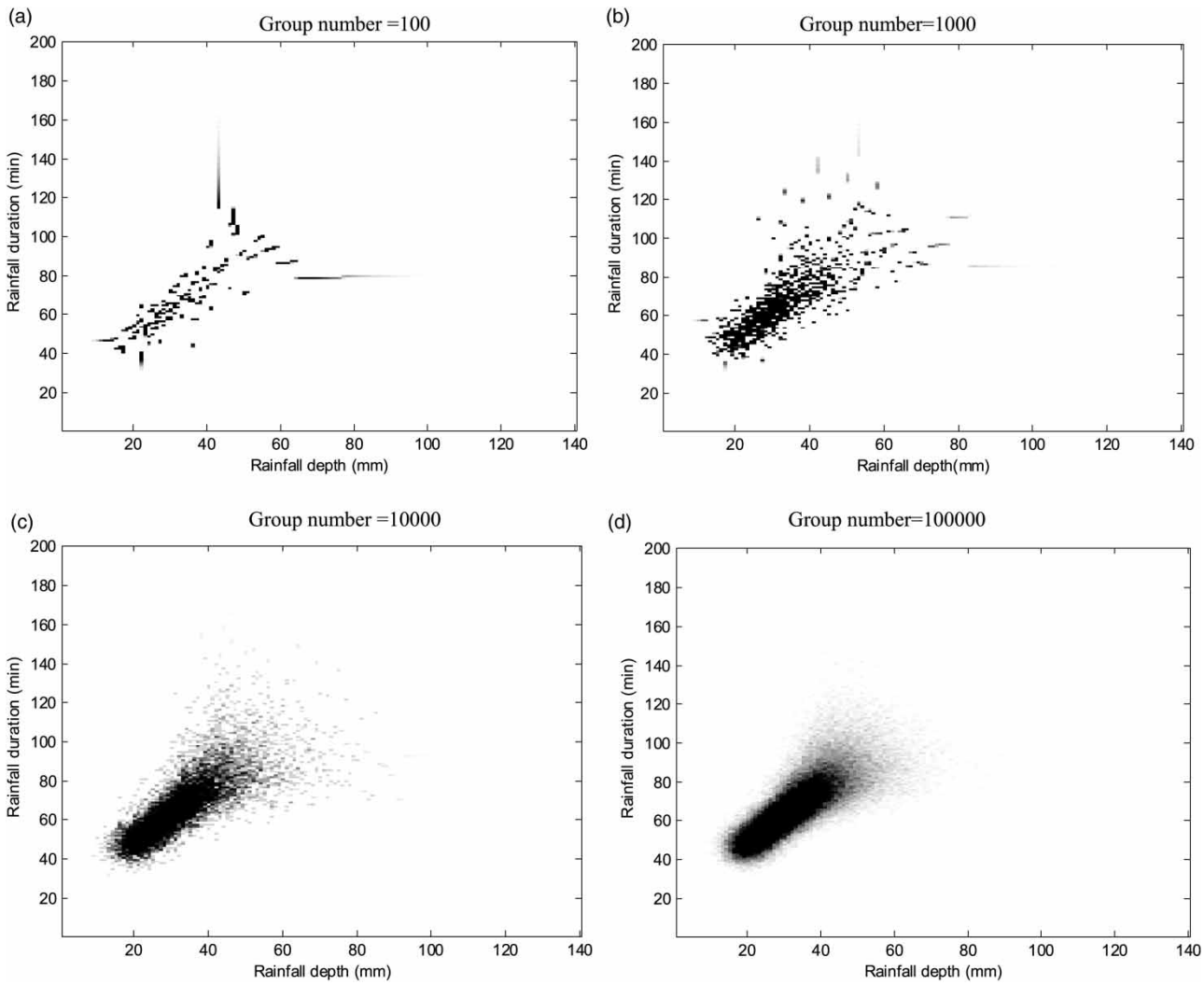

Figure 2 | Rainfall samples generated with different number of groups when rainfall is characterized by two variables (CC = 0.78).

only determine a rough trend of the samples, i.e. there is still some freedom to adjust how samples distribute under the trend. As the additional subjective constraints (the way of forming groups) are introduced to the grouping procedure, samples generated from more groups tend to be distributed in a more dispersed fashion than those from fewer groups. Therefore, it is recommended that the group number of large sizes should be used in practical use provided the computational efficiency is not an issue.

Figure 3 considers the case where the rainfall depth and duration are sampled from the same marginal distributions with a desired CC being 0.12. It is expected that the resultant samples in Figure 3 (with lower CC values) are distributed more uniformly or scattered than samples in Figure 2.

In order to examine the superiority of the proposed method over the traditional and still widely used method of Iman & Conover (1982), samples of the same distributions with RCC being 0.78 and 0.12 are also generated using Iman and Conover's method (noting that this method is incapable of generating correlations given by CC). The sample size is adopted as 10,000. As the result of Iman and Conover's method relates to the initial generated samples, 10 runs of each case are carried out and the error between targeted and obtained RCC is presented. In the case of targeted TCC being 0.78, the obtained RCC is in [0.781, 0.783] with errors possibly being 0.003, and in the case of targeted TCC being 0.12, the obtained RCC is between [0.116, 0.117] with errors greater than 0.003, in comparison with

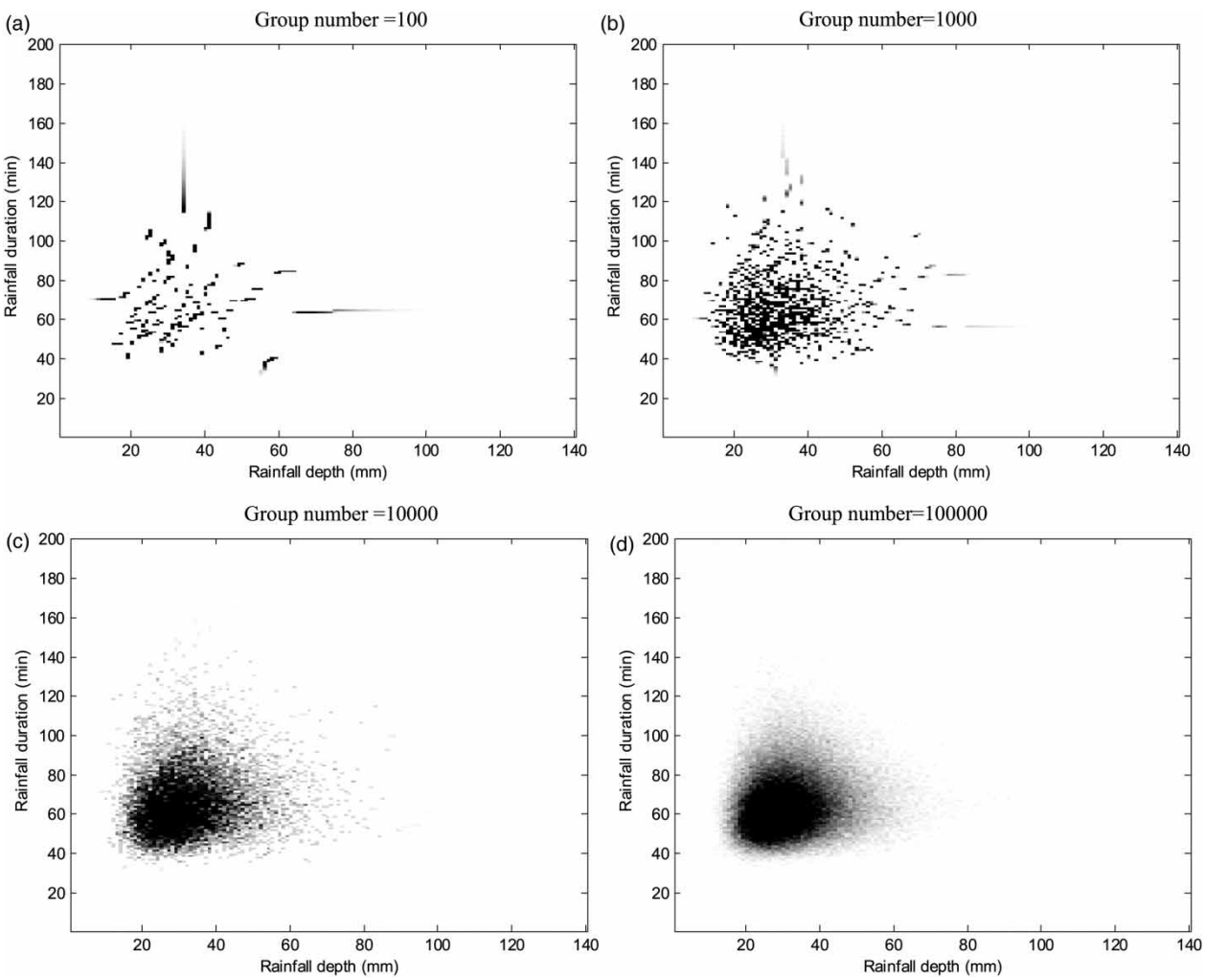

Figure 3 | Rainfall samples generated with different number of groups when rainfall is characterized by two variables (CC = 0.12).

the proposed method that can easily control errors below 0.001 and even much smaller.

Rainfall events of correlated antecedent dry duration, wet duration and intensity

In some cases, the dry period before rainfall is also important aside from the information about rainfall itself. For instance, a flood may happen under a relatively small rainfall event shortly after another event but may not happen under a large rainfall event with a long time dry period. In addition, the dry period is a crucial factor to determine the contaminated conditions of water after rains. Thus a rainfall characterizing a dry period (modeled as a duration D reporting no rainfall) followed by a wet period (modeled as a rectangular pulse having average intensity I and duration W) is frequently used as a coarse representation of rainfall. In this section, samples of rainfall events characterized by these three variables are generated.

The distributions of D, I and W are assumed to be represented by the generalized Pareto (GP) distribution:

$$F(x) = \begin{cases} 1 - \left(1 - \frac{k}{c}(x - x^*)\right)^{1/k}, & x \geq x^* \\ 0, & x < x^* \end{cases} \quad (8)$$

The parameters for each variable of the cumulative distribution function are listed in Table 2 (Salvadori & Michele (2006) from summer season rainfalls of Scoffera station, Italy). The CCs between the variables are assumed as Table 3 shows.

Ten thousand and 100,000 rainfall events are separately generated. Different group numbers are studied. The program runs 10 times for each group size. The required precision is set to be 0.001. In this case, the positions of samples from I are firstly rearranged while positions of samples from D are kept. The positions of samples from W are then rearranged in order to achieve the required correlations. The average number of steps taken (including steps of arranging both marginal samples of I and W) to achieve the required precision for different group sizes are listed in Table 4. The average number of steps needed for the generation of 10,000 samples is generally less than the generation of 100,000 samples when they use the same number of groups. This is due to the fact that a random step of samples of a smaller size generally walks more towards the required correlations than that of samples of a larger size. The average number of steps increases as the number of the groups increase. Figure 4 shows the 10,000 samples generated without grouping (i.e. when group number = 10,000) as an example. From this figure, it is clear that the dry period and rainfall intensity have positive correlation relationship whereas the rainfall duration is negatively related to the other two variables, as the target CCs indicate.

Table 2 | Values of the marginal GP parameters

Parameters	Dry duration (h)	Intensity (mm/h)	Wet duration (h)
K	−0.46	−0.49	−0.05
C	55.88	0.92	6.46
x^*	7	0	0

Table 3 | Correlated coefficients among variables

Parameters	D&I	I&W	D&W
Correlation coefficient	0.3	−0.15	−0.2

Table 4 | Average computational steps to obtain the required precision of sampling rainfall events represented by three variables with group method

	Average number of steps	
Number of groups (n/k)	10,000 samples	100,000 samples
50	461	694
100	543	846
500	2,040	3,099
1,000	3,549	5,998
5,000	17,667	28,140
10,000	35,957	53,329
50,000	–	260,572
100,000	–	478,777

CONCLUSIONS

When modeling hydrosystems, it is common that the model is driven by several variables which may be correlated with

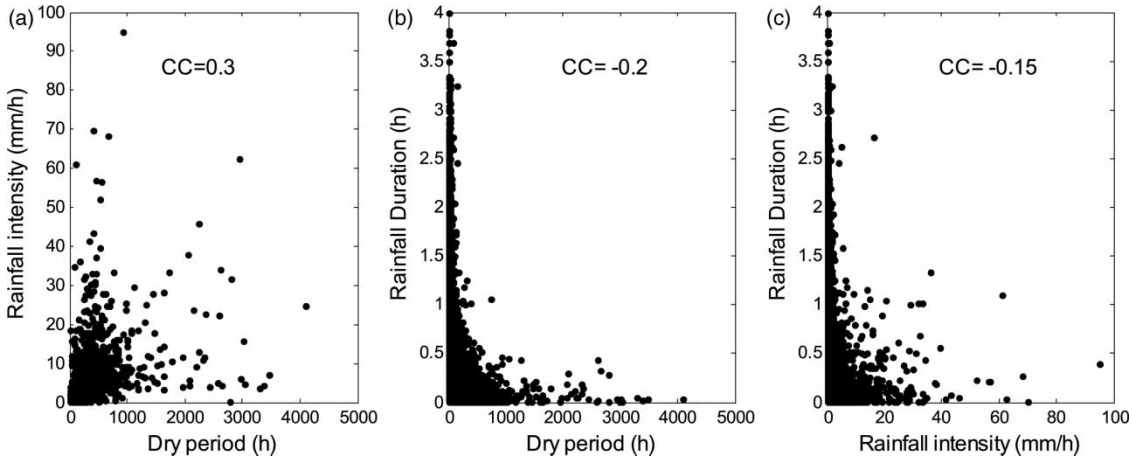

Figure 4 | Rainfall samples generated without grouping when rainfall is characterized by dry period duration, rainfall intensity and rainfall duration.

each other. If these variables are generated using sampling methods, it is important to ensure that the dependency of samples is incorporated in the resultant samples. A novel method of generating correlated samples of large sizes with known marginal distributions and desired correlations (either CC or RCC) has been introduced. Based on the idea of adjusting the correlations of samples by rearranging the positions of samples, the group method was developed to facilitate efficient generation of correlated samples of large sizes. The approach was successfully applied to two cases of generating rainfall events. The following conclusions can be drawn from the present study:

1) The theoretically achievable precision by the group method was derived. In engineering practice, the requirement of precision is usually not very high (say two or three decimal digits), a small number of groups, such as 50, can generally satisfy the required precision.
2) Due to the fact that the solution space can be dramatically decreased when the group method is introduced, the group method speeds up the searching process in adjusting sample positions for approximating required correlations. The group method works more efficiently when the number of groups decreases.
3) The correlated samples are more likely to cluster when the number of groups is small, though it still reveals the correlated relationship of samples according to the target CC. This phenomenon also shows that the profile of how samples distribute is not solely determined by marginal distributions and CCs.
4) Provided the computational efficiency is not a constraint, the group number of large sizes is recommended in practical use as the sample distributions become more dispersed.
5) The proposed method successfully generates samples of rainfall events represented by variables with known marginal distributions and correlated coefficient. However, its application to the generation of samples of rainfall events will be tested in future studies for real applications.

REFERENCES

Bacchi, B., Becciu, G. & Kottegoda, N. T. 1994 Bivariate exponential model applied to intensities and durations of extreme rainfall. *J. Hydrol.* **155** (1–2), 225–236.

Chakraborty, A. 2006 Generating multivariate correlated samples. *Computation. Stat.* **21** (1), 103–119.

Charmpis, C. D. & Panteli, L. P. 2004 A heuristic approach for the generation of multivariate random samples with specified marginal distributions and correlation matrix. *Computation. Stat.* **19** (2), 283–300.

Cheng, R. C. H. 1985 Generation of multivariate normal samples with given sample mean and covariance matrix. *J. Stat. Comput. Simul.* **21** (1), 39–49.

Coles, S., Pericchi, L. R. & Sisson, S. 2003 A fully probabilistic approach to extreme rainfall modeling. *J. Hydrol.* **273** (1–4), 35–50.

Conover, W. J. & Iman, R. L. 1981 Rank transformations as a bridge between parametric and nonparametric statistics. *Am. Stat.* **35** (3), 124–129.

Douglas, E. M., Vogel, R. M. & Kroll, C. N. 2000 Trends in floods and low flows in the United States: impact of spatial correlation. *J. Hydrol.* **240** (1–2), 90–105.

Frees, W. E. & Valdez, A. E. 1997 Understanding Relationships Using Copulas. *32nd Actuarial Research Conference*. University of Calgary, Canada.

Genest, C. & Rivest, L. P. 1993 Statistical inference procedures for bivariate archimedean Copulas. *J. Am. Stat. Assoc.* **88** (423), 1034–1043.

Goel, N. K., Kurothe, R. S., Mathur, B. S. & Vogel, R. M. 2000 A derived flood frequency distribution for correlated rainfall intensity and duration. *J. Hydrol.* **228** (1–2), 56–67.

Grimaldi, S. & Serinaldi, F. 2006 Asymmetric copula in multivariate flood frequency analysis. *Adv. Water Resour.* **29** (8), 1155–1167.

Helton, J. C. & Davis, F. J. 2003 Latin hypercube sampling and the propagation of uncertainty in analyses of complex systems. *Reliab. Eng. Syst. Safe.* **81** (1), 23–69.

Iman, R. & Conover, W. J. 1982 A distribution-free approach to inducing rank correlation among input variables. *Commun. Stat. B-Simulat.* **11** (3), 311–34.

Kanso, A., Chebbo, G. & Tassin, B. 2006 Application of MCMC–GSA model calibration method tourban runoff quality modeling. *Reliab. Eng. Syst. Safe.* **91** (10–11), 1398–1405.

Kapelan, Z., Savic, D. A. & Walters, G. A. 2005 Multiobjective design of water distribution systems under uncertainty. *Water Resour. Res.* **41** (11), W11407.

Koutsoyiannis, D. & Baloursos, G. 2000 Analysis of a long record of annual maximum rainfall in Athens, Greece, and design rainfall inferences. *Nat. Hazards* **22** (1), 31–51.

Kurothe, R. S., Goel, N. K. & Mathur, B. S. 1997 Derived flood frequency distribution of negatively correlated rainfall intensity and duration. *Water Resour. Res.* **33** (9), 2103–2107.

Li, S. T. & Hammond, J. L. 1975 Generation of pseudorandom numbers with specified univariate distribution and correlation coefficients. *IEEE Trans. Syst., Man Cybern.* 557–561.

Lurie, P. M. & Goldberg, M. S. 1998 An approximate method for sampling correlated random variables form partially-specified distributions. *Manage. Sci.* **44** (2), 203–218.

Morgan, M. G. & Henrion, M. 1990 *Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis*. Cambridge University Press, New York, pp. 172–217.

Muzik, I. 2002 A first-order analysis of the climate change effect on flood frequencies in a subalpine watershed by means of a hydrological rainfall–runoff model. *J. Hydrol.* **267** (1–2), 65–73.

Parrish, R. S. 1990 Generating random deviates from multivariate Pearson distributions. *Comput. Stat. Data Anal.* **9** (3), 283–295.

Rahman, A., Weinmann, P. E., Hoang, T. M. T. & Laurenson, E. M. 2002 Monte Carlo simulation of flood frequency curves from rainfall. *J. Hydrol.* **256** (3–4), 196–210.

Salvadori, G. & Michele, C. D. 2006 Statistical characterization of temporal structure of storms. *Adv. Water Resour.* **29** (6), 827–842.

Salvadori, G. & Michele, C. D. 2007 On the use of copulas in hydrology: theory and practice. *J. Hydrologic Eng.* **12** (4), 369–380.

Schweizer, B. & Wolff, E. F. 1981 On nonparametric measures of dependence for random variables. *Ann. Stat.* **9** (4), 870–885.

Segond, M.-L., Onof, C. & Wheater, H. S. 2006 Spatial–temporal disaggregation of daily rainfall from a generalized linear model. *J. Hydrol.* **331** (3–4), 674–689.

Serinaldi, F. & Grimaldi, S. 2007 Fully nested 3-copula: procedure and application on hydrological data. *J. Hydrol. Eng.* **12** (4), 420.

Smith, E. A., Ryan, P. B. & Evans, S. J. 1992 The effect of neglecting correlations when propagating uncertainty and estimating the population distribution of risk. *Risk Analysis* **12** (4), 467–474.

Thorndahl, S. & Willems, P. 2008 Probabilistic modeling of overflow, surcharge and flooding in urban drainage using the first-order reliability method and parameterization of local rain series. *Water Res.* **42**, 455–466.

Vořechovský, M. & Novák, D. 2009 Correlation control in small-sample Monte Carlo type simulations I: A simulated annealing approach. *Probabilist. Eng. Mech.* **24**, 452–462.

Yue, S. 2000 Joint probability distribution of annual maximum storm peaks and amounts as represented by daily rainfalls. *Hydrol. Sci. J.* **45** (2), 315–326.

Zhang, L. & Singh, P. V. 2007 Gumbel–Hougaard copula for trivariate rainfall frequency analysis. *J. Hydrol. Eng.* **12** (4), 409–419.

First received 4 February 2010; accepted in revised form 16 December 2011. Available online 31 July 2012