Journal of Hydroscience and Hydraulic Engineering
Vol. 3, No. 1, April, 1985, pp.75-88.

APPROXIMATE SKEW OF THE SUM OF SKEWED AND CORRELATED VARIATES

By

Kiyoshi Hoshi

Civil Engineering Research Institute
Hokkaido Development Bureau
Sapporo 062, Japan

and

Quirino A. Romano

Division of Water Resources Engineering
Asian Institute of Technology
Bangkok 10501, Thailand

SYNOPSIS

A method is explored to compute the skewness coefficient of the sum of hydrologic variables, when the first three moments as well as covariances are known ; if quantitative probability statements of the sum variable are to be made, a three-parameter probability function can easily be fitted to it, given the first three moments. A primary task integral to this study is the determination of an appropriate matrix that transforms correlated variables to uncorrelated variables.

To determine the appropriate transformation matrix the proposed approaches are compared with the exact solution for identically distributed gamma variates with a stationary Markov process as well as with the results of Monte Carlo experiments where the three-parameter log normal (LN3) distribution is used to generate correlated variates with correlation schemes of AR(1) and ARMA(1, 1) processes. The results of the distribution-free approach reported here compare favorably with those of Monte Carlo experiments.

INTRODUCTION

The importance of multivariate analysis in the planning and management of water resource systems has been recognized in recent years. Of particular interest in multivariate problems are the statistical characteristics of the sum of hydrologic variables. The following are some of the examples in which the sum of skewed and correlated variates is frequently encountered in hydrologic applications:

(1) The annual flow is regarded as the sum of monthly flows. The seasonal AR(1) process, well known as a Thomas-Fiering model in operational hydrology does not necessarily guarantee preservation of the variance and lag-one serial correlation at the annual level. The inability of the assumed seasonal AR(1) model to replicate annual statistics has important consequences in reservoir design capacities (Hoshi et al., 1978) ; this example clearly demonstrates the importance of preserving all relevant correlations between monthly flows. In many modeling situations it is also necessary to model the form of the monthly flow marginal probability distributions; skewed marginal distributions widely used in hydrologic studies are the log normal and gamma families.

(2) Point rainfall amounts in a watershed are usually lumped to the mean areal amount in rainfall-runoff

analysis. The mean areal rainfall is regarded as the weighted sum of rainfall amounts at the sites of interest. There are many instances where point rainfall amounts are highly correlated. The degree of correlations depends on the distances among the sites as well as on the terrain of the watershed concerned. It has been found, however, that the correlation structure of point rainfalls does not follow fixed patterns; for example, space variations of rainfall vary with different time units. Moreover, it is unlikely that the frequency distributions of point rainfall amounts are approximated by identical marginal distributions. When the skewed marginal distributions of hydrologic variables vary spatially and temporally, it is unrealistic to use the same distributions to be fitted to them.

(3) A linear transfer function model (i. e., unit hydrograph theory) has long been used in watershed modeling approaches in which the discharge is given by the weighted sum of rainfall inputs. Major practical interest in this problem is how to estimate the frequency distribution of runoff outputs for any given distributions of inputs.

(4) The properties of partial sums of net inflows to a reservoir are relevant to the determination of the required design capacity. The dependence between sequences of inflows and the form of flow marginal distributions largely affect the magnitude of reservoir size.

It is extremely difficult to theoretically derive the probability density function for the sum of random variables which have particular marginal probability distributions and correlation structure. This is due to the fact that the reproductive property pertinent only to normal variates is not valid for log normally and gamma distributed variates; the sum of log normal or gamma variates which are dependent is not necessarily of the same type. Kotz and Neumann (1963) have developed the approximate distribution of the sum of identically distributed gamma variates whose correlation structure is defined by the first-order autoregressive, AR(1) process. The fundamental premise in their solution is that the sum itself is a gamma variate. The approximate distribution of sum variate is uniquely determined, because the first two moments define completely a two-parameter gamma distribution. Thom (1968) and Murota et al. (1974) have shown that the sum of n independently distributed gamma variates, each having different distribution parameters, can also be approximated by the gamma distribution. The accuracy of approximations was evaluated by the relative error of the third moment about the origin. Kotz and Adams (1964), and Nagao (1975) have derived the exact solution for the distribution of the sum of identically distributed gamma variates which are correlated according to a stationary AR(1) process.

There are some limited practical applications in the above approaches in which individual variates in the sum are described by two-parameter gamma distributions and their correlation links between variates are assumed to follow the stationary processes. In many practical stuations where asymmetrically distributed hydrologic data are encountered, it is often necessary to fit a three-parameter density function to them for preserving skewed marginal distributions. It is clear that the use of a two-parameter distribution fails to maintain the first three moments of the observed data. Quite frequently, the nonstationary process is required to accommodate the correlation structure of hydrologic variables in stochastic hydrologic modelings.

The present study is primarily motivated by the fact that the first two moments of the sum of correlated variates are distribution free and that no higher order moments more than the third are used to model skewed data in frequency analysis of hydrologic data. Once the skewness coefficient for the sum of any correlated variates is obtained by some appropriate method, it is quite straightforward to make quantitative probability statements (i. e., estimation of quantiles corresponding to probability levels) via use of a three-parameter density function. For example, Lettenmaier and Burges (1977) have shown that the theoretical Pearson type 3 (P3) and three-parameter log normal (LN3) distributions for a given skew coefficient γ and coefficent of variation β are almost identical for $\beta \leq 1$ and γ between *1* and *3*. It makes little difference, therefore, which distribution is used to represent skewed data ; the choice is one of operational convenience.

The objective of this paper is to develop the distribution-free approach by which the skewness

coefficient of the sum of variates can be computed given the first three moments of variates in the sum and any correlation structure. The proposed procedure was compared with the exact gamma solutions and Monte Carlo experiments.

EXACT SOLUTION OF GAMMA DISTRIBUTION

Consider a two-parameter gamma distribution with p.d.f. :

$$f(x) = (x/a)^{b-1} exp(-x/a) / [a\Gamma(b)] \qquad 0 \leq x \leq \infty \tag{1}$$

where a and b are scale and shape parameters, respectively. The parameters a and b are positive, and $\Gamma(b)$ denotes the gamma function. The population moments of the gamma variate x are given by

$$\begin{aligned} \mu &= ab \\ \sigma^2 &= a^2 b \\ \gamma &= 2/b^{1/2} \end{aligned} \tag{2}$$

where μ, σ^2, and γ are the mean, variance, and skewness coefficient, respectively. The relationship between coefficients of variation (β) and skewness (γ) for the gamma distribution is defined by

$$\gamma = 2\beta \tag{3}$$

Consider n random variables, x_1, x_2, ····· , x_n and their sum :

$$z = x_1 + x_2 + \cdots\cdots + x_n \tag{4}$$

The distribution-free mean μ_z and variance σ_z^2 of z are easily determined from

$$\mu_z = \sum_{i=1}^{n} \mu_i \tag{5}$$

$$\sigma_z^2 = \sum_{i=1}^{n} \sigma_i^2 + 2 \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \sigma_i \sigma_j \rho_{ij} \tag{6}$$

where μ_i and σ_i^2 are the mean and variance of variable x_i, respectively and ρ_{ij} is a correlation coefficient between variables x_i and x_j.

When assuming that the sum variable z is distributed as gamma defind by eq. 1, then the two parameters, a_z and b_z of the approximate gamma distribution are easily estimated by $a_z = \sigma_z^2/\mu_z$ and $b_z = \mu_z^2/\sigma_z^2$, using the relations of eqs. 2, 5, and 6.

Kotz and Adams (1964), and Nagao (1975) have shown the population statistics of variate z, defined by eq. 4 in which each variate x_i has the same marginal distribution of eq. 1 and the correlation structure is defined by a stationary AR(1) process. The r-th cumulant K_r $(r = 1, 2, 3, \cdots\cdots)$ of the sum variate, z is given by

$$K_r = (r-1)!\, b a^r \sum_{i=1}^{n} \lambda_i^r \quad (r = 1, 2, 3, \cdots\cdots) \tag{7}$$

where λ_i $(i = 1, 2, \cdots\cdots, n)$ are the eigenvalues of correlation matrix whose elements are given by $\rho_{ij} = \rho^{|i-j|/2}$ $(i, j = 1, 2, \cdots\cdots, n)$ and ρ is a lag-one serial correlation coefficient. Hence, the mean μ_z, variance σ_z^2, and skew coefficient γ_z of the sum variable z are given by

$$\begin{aligned} \mu_z &= ab \sum_{i=1}^{n} \lambda_i = nab \\ \sigma_z^2 &= a^2 b \sum_{i=1}^{n} \lambda_i^2 \end{aligned} \tag{8}$$

$$\gamma_z = (2a^3 b \sum_{i=1}^{n} \lambda_i^3) / \sigma_z^3$$

Some useful relations can be evolved from the statistical properties of eq. 8 as follows :
(1) Consider the case of $n=2$ for which $\lambda_1 = 1-\rho^{1/2}$ and $\lambda_2 = 1 + \rho^{1/2}$. Thus eq. 8 reduces to

$$\begin{aligned} \mu_z &= 2ab \\ \sigma_z^2 &= 2a^2 b(1+\rho) \\ \gamma_z &= 2(2b)^{-1/2}(1+3\rho)(1+\rho)^{-3/2} \end{aligned} \tag{9}$$

The solution given by eq. 9 is in agreement with that obtained by Nagao (1975).
(2) For the case of $\rho = 1$ (i.e., complete dependence), $\lambda_1 = n$ and $\lambda_i = 0$ $(i=2, 3, \cdots\cdots, n)$ and hence the first three moments of z are given by

$$\begin{aligned} \mu_z &= nab \\ \sigma_z^2 &= n^2 a^2 b \\ \gamma_z &= 2/b^{1/2} \end{aligned} \tag{10}$$

Equation 10 indicates that the skew of the sum of identically distributed gamma variates is identical with that of a univariate gamma distribution as shown in eq. 2.
(3) For the case in which x_i $(i=1, 2, \cdots\cdots, n)$ are independent and identical gamma variates (i.e., $\rho = 0$), the statistical properties of eq. 8 are found to be

$$\begin{aligned} \mu_z &= nab \\ \sigma_z^2 &= na^2 b \\ \gamma_z &= 2/(nb)^{1/2} \end{aligned} \tag{11}$$

because of $\lambda_i = 1$ $(i=1, 2, \cdots\cdots, n)$. The following observation can be made from eq. 11 that γ_z approaches zero as n becomes sufficiently large. This characteristic conforms to the central limit theorem.
For both cases of $\rho = 0$ and $\rho = 1$, the relation of $\gamma_z = 2\beta_z$ holds, where β_z is the coefficient of variation of z, defined by σ_z/μ_z, implying that the sum of gamma variates with a common distribution is also distributed as gamma.

DISTRIBUTION-FREE APPROACH

The preceding approach has limitations as to choices of marginal distributions and correlation schemes ; two-parameter identically distributed gamma variates in the sum are correlated according to the stationary lag-one Markov process. The method presented herein is based on a distribution-free approach in which there are no restrictions on choices of the skew coefficients and correlation links between variables. The main difficulty involved in multivariate problems is the correlation structure, which makes the pertinent issue for determining the skew of the sum of variables extremely difficult to resolve. One way of averting this complication is to first decompose the correlated variates into the uncorrelated ones. Second, the skewness coefficients of independent variates are expressed by skews of marginal distributions. Finally, a closed-form solution for the skewness of the sum of correlated variates is derived from skews of uncorrelated variates.

Transformation of Correlated Variates

Let x_i and g_i $(i=1, 2, \cdots\cdots, n)$ be the correlated variate with zero mean, and uncorrelated (independent) variate with zero mean and unit variance, respectively. When defining the transformation matrix B whose elements are denoted by b_{ij} $(i, j=1, 2, \cdots\cdots, n)$, the transformation relationship is given by

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} b_{11} & b_{12} & b_{13} & \cdots\cdots & b_{1n} \\ b_{21} & b_{22} & b_{23} & \cdots\cdots & b_{2n} \\ \cdots\cdots & b_{ij} & \cdots\cdots & & \\ b_{n1} & b_{n2} & b_{n3} & \cdots\cdots & b_{nn} \end{pmatrix} \begin{pmatrix} g_1 \\ g_2 \\ \vdots \\ g_n \end{pmatrix} \tag{12}$$

The general expression for x_i is obtained from eq. 12 as

$$x_i = b_{i1}g_1 + b_{i2}g_2 + b_{i3}g_3 + \cdots\cdots + b_{in}g_n \qquad (i = 1, 2, \cdots\cdots, n) \tag{13}$$

Equation 12 is also represented in matrix form as

$$X = BG \tag{14}$$

where

$$\begin{aligned} X^T &= [x_1 \; x_2 \; x_3 \; \cdots\cdots \; x_n] \\ G^T &= [g_1 \; g_2 \; g_3 \; \cdots\cdots \; g_n] \end{aligned} \tag{15}$$

B is an $(n \times n)$ transformation matrix, and the superscript T denotes the transpose of a matrix. Keeping in mind that x_i has zero mean and g_i is an independent standardized variate (i.e., zero mean and unit variance), the variance-covariance matrices of X and G are written as

$$\begin{aligned} V_x &= E(XX^T) \\ V_g &= E(GG^T) = I \end{aligned} \tag{16}$$

where V_x is an $(n \times n)$ symmetric matrix of variances and covariances of x_i ; I is an $(n \times n)$ identity matrix ; and $E(\cdot)$ denotes the expectation operator.
Postmultiplying eq. 14 by X^T and taking expectations gives

$$V_x = BB^T \tag{17}$$

The form of eq. 17 is frequently encountered in multivariate stochastic generation models. The solution for the transformation matrix B can be effected in several ways. The first method is well known as the Crout or Cholesky method (Young, 1968) in which a symmetric matrix is decomposed into the product of a lower triangular matrix, L and its transpose L^T. The second method uses the technique of principal components (Kendall, 1961; Fiering, 1964) to solve eq. 17 for B. The four methods proposed so far are summarized as follows :

$$\textit{Method 1 ;} \quad B = L \;\; (\textit{lower triangular matrix}) \tag{18}$$

$$\textit{Method 2 ;} \quad B = PD^{1/2}P^T \tag{19}$$

$$\textit{Method 3 ;} \quad B = PD^{1/2} \tag{20}$$

$$\textit{Method 4 ;} \quad B = PD^{1/2}P \tag{21}$$

where D is an $(n \times n)$ diagonal matrix of eigenvalues of V_x and P is an $(n \times n)$ matrix of corresponding normalized eigenvectors. The eigenvector matrix has an orthonormal property of $P^T = P^{-1}$ (inverse matrix of P) and hence, Methods 2, 3, and 4 satisfy the condition of $V_x = PDP^T$. Method 2 produces a symmetric matrix of B which Todini (1980) applied to generating skewed seasonal flows in the disaggregation scheme. All the methods preserve the first two moments of the sum of correlated variates ; therefore the question arises as to which scheme can be practically used to approximate the skew for the sum of skewed data. The remainder of this paper is directed toward the determination of an appropriate transformation matrix of B.

Skewness of Uncorrelated Variates

Because the elements of the column vector G in eq. 14 are mutually independent, the element g_i is univariate distributed with zero mean, unit variance, and skewness γ_{gi}. It is also proved that the following relations can hold between independent variates g_i $(i=1, 2, \cdots\cdots, n)$:

$$E(g_i g_j g_k) = \begin{cases} \gamma_{gi} & (i=j=k) \\ 0 & (i \neq j=k) \\ 0 & (i \neq j \neq k) \end{cases} \tag{22}$$

where γ_{gi} is the skewness coefficient of g_i.
Cubing both sides of eq. 13 and taking the expectations yields

$$\begin{aligned} E(x_i^3) = \gamma_i \sigma_i^3 = \sum_{j=1}^{n} b_{ij}^3 E(g_j^3) + 3 \sum_{\substack{i=1 \\ (j \neq k)}}^{n} \sum_{k=1}^{n} b_{ij} b_{ik}^2 E(g_j g_k^2) \\ + \sum_{\substack{j=1 \\ (j \neq k \neq m)}}^{n} \sum_{k=1}^{n} \sum_{m=1}^{n} b_{ij} b_{ik} b_{im} E(g_j g_k g_m) \quad (i=1, 2, \cdots\cdots, n) \end{aligned} \tag{23}$$

where σ_i and γ_i are the standard deviation and skewness coefficient of variate x_i, respectively. Substituting the condition of eq. 22 into eq. 23 gives

$$\gamma_i \sigma_i^3 = \sum_{j=1}^{n} b_{ij}^3 \gamma_{gj} \qquad (i = 1, 2, \cdots\cdots, n) \tag{24}$$

Equation 24 can be written in matrix form as

$$\Sigma_x^3 \Gamma_x = B_1 \Gamma_g \tag{25}$$

where

$$\Gamma_x^T = [\gamma_1 \ \gamma_2 \ \gamma_3 \ \cdots\cdots \ \gamma_n]$$

$$\Gamma_g^T = [\gamma_{g1} \ \gamma_{g2} \ \gamma_{g3} \ \cdots\cdots \ \gamma_{gn}]$$

$$\Sigma_x = \begin{pmatrix} \sigma_1 & & & & \\ & \sigma_2 & & 0 & \\ & & \sigma_3 & & \\ & 0 & & \ddots & \\ & & & & \sigma_n \end{pmatrix} \tag{26}$$

$$B_1 = \begin{pmatrix} b_{11}^3 & b_{12}^3 & b_{13}^3 \cdots\cdots & b_{1n}^3 \\ b_{21}^3 & b_{22}^3 & b_{23}^3 \cdots\cdots & b_{2n}^3 \\ \cdots\cdots & b_{ij}^3 & \cdots\cdots & \\ b_{n1}^3 & b_{n2}^3 & b_{n3}^3 \cdots\cdots & b_{nn}^3 \end{pmatrix}$$

Σ_x in an $(n \times n)$ diagonal matrix of standard deviations of x_i, and the elements of B_1 are computed by being cubed element by element in matrix, B.
A skewness vector of uncorrelated variates g_i can be obtained by solving eq. 25 for Γ_g as

$$\Gamma_g = B_1^{-1} \Sigma_x^3 \Gamma_x \tag{27}$$

where B_1^{-1} is the inverse matrix of B_1.
Equation 27 indicates that the skewness coefficients of original variables are explicitly incorporated to determine those of uncorrelated variables g_i.

Properties of the Sum of Correlated Variates

Premultiplying eq. 14 by a $(1\times n)$ unit vector gives the sum of variates x_i as

$$\begin{aligned} z &= x_1+x_2+\cdots\cdots+x_n \\ &= UX = UBG = CG \end{aligned} \tag{28}$$

where

$$\begin{aligned} U &= [\,1\ 1\ 1\cdots\cdots 1\,] \\ C &= UB \\ &= [\,1\ 1\ 1\cdots\cdots 1\,] \begin{pmatrix} b_{11} & b_{12} & \cdots\cdots & b_{1n} \\ b_{21} & b_{22} & \cdots\cdots & b_{2n} \\ \cdots\cdots & \cdots\cdots & \cdots\cdots & \cdots\cdots \\ b_{n1} & b_{n2} & \cdots\cdots & b_{nn} \end{pmatrix} \end{aligned} \tag{29}$$

C is a $(1\times n)$ known vector whose elements are given by the sum of the rows of matrix B as shown in eq. 29. The scalar form of eq. 28 is written as

$$z = c_1g_1 + c_2g_2 + \cdots\cdots + c_ng_n \tag{30}$$

where c_i is the element of vector C. By definition, X has zero means and hence the sum variable z has also zero mean. From eq. 30, the variance (σ_z^2) and skewness (γ_z) of z are obtained, using eqs. 16 and 22 as

$$\sigma_z^2 = E(z^2) = \sum_{i=1}^{n} c_i^2 \tag{31}$$

and

$$\gamma_z \sigma_z^3 = E(z^3) = \sum_{i=1}^{n} c_i^3 \gamma_{gi} \tag{32a}$$

or

$$\gamma_z = \left\{ \sum_{i=1}^{n} c_i^3 \gamma_{gi} \right\} \Big/ \sigma_z^3 \tag{32b}$$

where γ_{gi} is the skewness coefficient of g_i.
Equation 31 is another expression for the variance of z, which is statistically equivalent to eq. 6. It follows from eq. 32b that the skewness for the sum of correlated variates is expressible by functions of the skewness coefficients for uncorrelated variates.

PERFORMANCE COMPARISON

When the proposed procedure is used to estimate the skewness coefficient of the sum of correlated variates, an important question is : Which method would give a desired result for the skewness out of the four transformation matrices given in eq. 18 to eq. 21? In Chapter 2 the exact solutions were derived for the first three moments of the sum of identically distributed gamma variates whose correlation structure follows the stationary AR(1) process. As a preliminary screening tool to select the candidate methods of the proposed approach, a comparison was made between skew coefficients of the sum variable resulting from the exact gamma solution and each of the four transformation matrices. There are no presently existing studies on the statistical properties of the sum variable for the cases where variates in the sum have different skewed properties and take a nonstationary process. The resolution can best be made

through Monte Carlo experiments for such cases. As the second screening procedure for the selection of an appropriate transformation matrix, the results via distribution-free approaches were compared with those of Monte Carlo tests, given the first three moments of variables in the sum and correlation schemes.

Comparison between Exact Gamma and Distribution-Free Solutions

To determine the appropriate transformation matrix the results using the distribution-free approach are compared with those via exact gamma solutions for each of four transformation matrices given from eq. 18 to eq. 21. The methodology employed here is very simple. The population statistics of identical marginal distributions are given in eq. 2. The variance-covariance matrix is specified using the stationary AR(1) process, given the number of variates in the sum, n and a lag-one correlation coefficient ρ. To have an insight into the changes in magnitude of the skew of the sum variate, the values of $\rho = 0.1$ $(0.1)\,0.9$ and $n = 3$ and 6 are used in the experiments.

Tables 1 and 2 report the skewness coefficients γ_z of the sum of variates for $n = 3$ and $n = 6$, respectively, where the results of the exact gamma solution are compared with those using four transformation matrices. The values of γ_z corresponding to $\rho = 0$ and $\rho = 1$ are computed by using eqs. 11 and 10, respectively. As mentioned before, the four methods produce the same variance of the sum variate as the gamma solution does; this consistency was numerically checked by using eq. 8 for the gamma solution and eq. 31 for the distribution-free approach.

The results of Tables 1 and 2 show that the skewness coefficients estimated from Method 3 (eq. 20) and Method 4 (eq. 21) behave substantially differently from those of the exact solution; Methods 3 and 4 should be discarded to compute an approximate skew of the sum of skewed and correlated variates. Methods 1 and 2 (eqs. 18 and 19) produce skewness coefficients that are smaller than the exact ones for a gamma distribution, but the skew γ_z resulting from Method 1 is slightly larger than that from Method 2.

Comparison between Distribution-Free and Monte Carlo Results

It is found from the above experiments that Methods 3 and 4 are inappropriate for use in the computation of skewness coefficients of the sum variate through distribution-free approach. Insofar as the exact solution is only applicable to identical marginal distributions of gamma variates with stationary correlation structure, additional experiments are added to evaluating the performance of Methods 1 and 2 for any skewed and correlated variates.

The three-parameter log normal (LN3) distribution is used to model the marginal distributions of variates and covariance relationships between them. The convenience of the LN3 distribution lies in the fact that there exist theoretical relationships between covariances in the untransformed (real) and transformed (log) domains; the nonstationary correlations in real space can easily be transformed from stationary process in log space by changing real space variances. The LN3 distribution used here is parameterized in the Slade-type model of the form :

$$x_i = \tau_i + c_i \exp(k_i y_i) \qquad (i = 1, 2, 3, \cdots\cdots, n) \tag{33}$$

where x_i is a LN3 distributed variate, y_i has a standard normal distribution with zero mean and unit variance, and τ_i, c_i, and k_i are distribution parameters.
The mean μ_i, variance σ_i^2, and coefficient of skewness γ_i of a random variable x_i with a LN3 distribution are given by

$$\begin{aligned} \mu_i &= \tau_i + c_i \eta_i^{1/2} \\ \sigma_i^2 &= c_i^2 \eta_i (\eta_i - 1) \qquad (i = 1, 2, 3, \cdots\cdots, n) \\ \gamma_i &= (\eta_i + 2)(\eta_i - 1)^{1/2} \end{aligned} \tag{34}$$

Table 1 Comparison of Skews between Four Methods for the Sum of Identically Distributed Gamma Variates with AR(1) process ($a=2$, $b=4$, $\gamma=1.0$, $n=3$)

ρ^*	Gamma Solution	Method 1	Method 2	Method 3	Method 4
0.1	0.683	0.584	0.580	4.396	1.506
0.2	0.769	0.606	0.570	3.718	1.320
0.3	0.838	0.632	0.605	3.143	1.171
0.4	0.891	0.669	0.626	2.657	1.055
0.5	0.930	0.712	0.654	2.248	0.969
0.6	0.959	0.761	0.691	1.904	0.909
0.7	0.979	0.815	0.739	1.615	0.875
0.8	0.991	0.873	0.801	1.373	0.870
0.9	0.998	0.935	0.885	1.170	0.905

* lag-one serial correlation coefficient

Table 2 Comparison of Skews between Four Methods for the Sum of Identically Distributed Gamma Variates with AR(1) process ($a=2$, $b=4$, $\gamma=1.0$, $n=6$)

ρ^*	Gamma Solution	Method 1	Method 2	Method 3	Method 4
0.1	0.506	0.414	0.411	7.589	2.100
0.2	0.596	0.430	0.419	6.166	0.936
0.3	0.680	0.455	0.433	5.025	0.713
0.4	0.758	0.490	0.454	4.099	0.726
0.5	0.828	0.536	0.482	3.338	1.507
0.6	0.889	0.594	0.521	2.704	1.325
0.7	0.938	0.667	0.575	2.171	0.919
0.8	0.973	0.757	0.654	1.718	0.793
0.9	0.991	0.865	0.778	1.330	0.871

* lag-one serial correlation coefficient

where

$$\eta_i = exp(k_i^2) \tag{35}$$

The parameter η_i is determined from γ_i as

$$\eta_i = [s_i + (s_i^2-1)^{1/2}]^{1/3} + [s_i - (s_i^2-1)^{1/2}]^{1/3} - 1 \tag{36}$$

with $s_i = 1+\gamma_i^2/2$. The real space correlation coefficient $\rho_x(i, j)$ between x_i and x_j is given by

$$\rho_x(i, j) = \frac{exp[k_i k_j \rho_y(i, j)] - 1}{[exp(k_i^2) - 1]^{1/2}[exp(k_j^2) - 1]^{1/2}} \tag{37}$$

$$(i, j = 1, 2, 3, \cdots\cdots, n ; i \neq j)$$

where $\rho_y(i, j)$ is a correlation coefficient between standard normally distributed variates y_i and y_j.

Two different correlation schemes are chosen for $\rho_y(i, j)$ in this experiment. The first scheme is the short memory model represented by a stationary AR(1) process whose autocorrelation function is

defined by

$$\rho_y(i, j) = \rho^{|i-j|} \tag{38}$$

where ρ is a lag-one serial correlation coefficient.
The second is the long memory model using an autoregressive moving average (ARMA(1, 1)) process whose correlation structure is given by

$$\rho_y(i, j) = \rho\phi^{|i-j|-1} \qquad (|i-j| \geq 1) \tag{39}$$

The parameter ϕ was held fixed at $\phi = 0.95$ in this experiment. Three levels of $\rho = 0.3$, 0.5, and 0.8, reflecting the degree of dependence between variates, were used. Table 3 shows an example of serial correlation coefficients for AR(1) and ARMA(1, 1) processes in the log domain, conditioned on $\rho = 0.8$ and $\phi = 0.95$. It is observed in Table 3 that the autocorrelation function corresponding to an AR(1) process decays faster with lag values than that of ARMA(1, 1). The correlation matrix generated with stationary AR(1) and ARMA(1, 1) processes in the log domain is transformed to the real space covariance matrix using eq. 37. A special emphasis is placed on the fact that the covariance structure resulting from use of $\rho_x(i, j)$ is no longer AR(1) and ARMA(1, 1) sequences, but follows a nonstationary process, because the variances of x_i and x_j have different values. For each set of μ_i, σ_i, γ_i, and $\rho_x(i, j)$ combinations, the skewness coefficient of the sum of n random variables was computed using either Method 1 or Method 2 in the distribution-free approach.

The computational steps of Monte Carlo experiments where the sum of LN3 distributed variates was generated by using AR(1) and ARMA(1, 1) models, are summarized as follows :

(1) LN3 distribution parameters of τ_i, c_i, and k_i were determined corresponding to population statistics μ_i, β_i (coefficient of variation), and γ_i for the number of variates, n.
(2) As shown in eq. 17, the transformation matrix in log space was determined from a correlation coefficient matrix whose elements are given by $\rho_y(i, j)$ with AR(1) and ARMA(1, 1) processes ; Method 2 (eq. 19) was used to effect the solution for the transformation matrix (either of the four methods is applicable in the normal domain).
(3) Multivariate normally distributed deviates y_i $(i = 1, 2, 3\cdots\cdots, n)$ were generated from independent normal deviates using the transformation matrix computed in Step(2). The sum of LN3 distributed deviates x_i $(i = 1, 2, 3, \cdots\cdots, n)$ was computed through eq. 33.
(4) 50,000 independent sequences of the sum variate were generated, repeating the procedure of Step(3).
(5) The sample mean, variance, and skewness coefficient of the sum variate (taken over 50,000 realizations) were estimated for given values of n, μ_i, β_i, γ_i, and $\rho_y(i, j)$.

In all cases examined the number of variates in the sum was taken as $n = 6$. Two relationships were considered between population coefficients of variation (β_i) and skewness (γ_i) $(i = 1, 2, \cdots\cdots, n)$; the first scheme is the relationship of the two-parameter log normal distribution given by $\gamma_i = \beta_i^3 + 3\beta_i$ and the second is the case of $\gamma_i = 2\beta_i$, which is identical with the gamma relationship. For the latter case the LN3 distribution was force fitted to the gamma population.

Tables 4 and 5 compare the skew coefficients of the sum of 6 variates between Methods 1 (eq. 18) and 2 (eq. 19), and Monte Carlo results for AR and ARMA processes, using the relations of $\gamma_i = \beta_i^3 + 3\beta_i$ and $\gamma_i = 2\beta_i$, respectively. In these experiments all the mean values of variates were fixed at $\mu_i = 1$, while the coefficients of variation β_i ranged from 0.1 to 0.6 with increment 0.1. While not shown herein, the variances computed through Monte Carlo experiments were practically equal to those of theoretical relationship given by eqs. 6 or 31 ; the maximum relative error in percent was 2.6%. It should be noted that the results corresponding to use of Method 2 differ from those corresponding to Method 1. The comparisons between exact gamma and distribution-free solutions as shown in Tables 1 and 2 indicate that the results using Method 1 are always closer to the gamma solution than those of Method 2. The results

Table 3 Comparison of Autocorrelation Coefficients between AR (1) and ARMA(1,1) Processes for $\rho=0.8$ and $\phi=0.95$

Lag	1	2	3	4	5	6
AR(1)	0.800	0.640	0.512	0.410	0.328	0.262
ARMA(1,1)	0.800	0.760	0.722	0.686	0.652	0.619

shown in Tables 4 and 5, however, reveal that Method 2 gives rise to much closer results of Monte Carlo experiments than does Method 1. Of notable significance is the skew behavior produced in Method 1 for ARMA(1, 1) sequences ; for the three levels of lag-one correlation coefficients examined, Method 1 yielded the minimum skew of the sum variate at $\rho=0.5$. This inconsistency is not observed in the results from Method 2.

Further tests were conducted for the cases where the mean levels of variates in the sum were in a range between *1* and *6* with increment *1*, and other parameter relationships were the same as those in preceding experiments. Tables 6 and 7 report the results similar to those shown in Tables 4 and 5 for the different mean values of *6* variates and parameter relationships of $\gamma_i=\beta_i^3+3\beta_i$, and $\gamma_i=2\beta_i$, respectively. The behavior of skewness coefficient of the sum variate for different mean values bears out a similar conclusion as for identical mean levels of variates in the sum ; different values of mean and coefficient of variation did not affect the consistency of results via distribution-free approach using Method 2. A limited set of results from application of the methods reported here show convincingly that Method 2 might be a valuable tool in hydrologic data analysis where estimates of skews for the sum of skewed variates with nonstationary correlation structures are required. The principal appeal stems from the fact that the user does not have to assume a population from which the data were obtained.

CONCLUSIONS

There is no rationale to determine the theoretical distribution for the sum of correlated variates, except for the sum of gamma variates distributed identically and correlated according to the stationary lag-one Markov process. Different correlation schemes that hydrologic variables resemble aggravate this issue, wherever the exact gamma solution is no longer applicable. Consequently, the distribution-free approach is required to resolve this problem.

It is well known that the determination of the first two moments of the sum of correlated variates is distribution-free. Most of hydrologic data have nonsymmetrical distributions. Hence, full details of a method which appears to offer advantages for estimating the skew of sum variate directly from the first three moments and covariances have been given. The major advantage is that no underlying distributions and stationary processes have to be assumed for describing hydrologic sequences.

The most important ingredient of the present approach is the determination of an appropriate matrix which transforms the correlated variates to the uncorrelated ones. Of existing four transformation matrices, the matrix given by $PD^{1/2}P^T$, where D is a diagonal matrix having eigenvalues of variance-covariance matrix and P is a corresponding eigenvector matrix, became the most competitive to estimate the skew of the sum variate, based on the reported Monte Carlo results. The skewness coefficients using the above transformation matrix are slightly smaller than those of Monte Carlo experiments. The exact amount of deviation could not readily be determined, because it depends on correlation structures. It is seen, however, from Table 4 to Table 7 that the deviations tend to be the largest for lag-one correlation coefficient around *0.5*, and decrease with the values of correlation coefficient larger than or less than *0.5*.

Table 4 Comparison of Skews between Distribution-Free and Monte Carlo Results for the Sum of Correlated Variates
($\mu_i=1$, $\beta_i=0.1(0.1)0.6$, $\gamma_i=\beta_i^3+3\beta_i$, $n=6$)

	AR(1)			ARMA(1,1)		
ρ^*	Method 1	Method 2	Monte Carlo	Method 1	Method 2	Monte Carlo
0.3	0.762	0.837	0.891	0.543	0.848	0.943
0.5	0.762	0.876	1.021	0.508	0.919	1.028
0.8	0.841	1.081	1.179	0.605	1.146	1.215

Table 5 Comparison of Skews between Distribution-Free and Monte Carlo Results for the Sum of Correlated Variates
($\mu_i=1$, $\beta_i=0.1(0.1)0.6$, $\gamma_i=2\beta_i$, $n=6$)

	AR(1)			ARMA(1,1)		
ρ^*	Method 1	Method 2	Monte Carlo	Method 1	Method 2	Monte Carlo
0.3	0.471	0.513	0.555	0.337	0.520	0.582
0.5	0.478	0.541	0.628	0.319	0.567	0.698
0.8	0.534	0.671	0.738	0.382	0.711	0.760

* lag-one correlation coefficient in the log domain

Table 6 Comparison of Skews between Distribution-Free and Monte Carlo Results for the Sum of Correlated Variates
($\mu_i=1(1)6$, $\beta_i=0.1(0.1)0.6$, $\gamma_i=\beta_i^3+3\beta_i$, $n=6$)

	AR(1)			ARMA(1,1)		
ρ^*	Method 1	Method 2	Monte Carlo	Method 1	Method 2	Monte Carlo
0.3	0.996	1.106	1.113	0.750	1.114	1.119
0.5	0.982	1.136	1.187	0.675	1.171	1.291
0.8	1.071	1.328	1.366	0.774	1.379	1.374

Table 7 Comparison of Skews between Distribution-Free and Monte Carlo Results for the Sum of Correlated Variates
($\mu_i=1(1)6$, $\beta_i=0.1(0.1)0.6$, $\gamma_i=2\beta_i$, $n=6$)

	AR(1)			ARMA(1,1)		
ρ^*	Method 1	Method 2	Monte Carlo	Method 1	Method 2	Monte Carlo
0.3	0.607	0.669	0.695	0.455	0.675	0.718
0.5	0.607	0.691	0.764	0.415	0.713	0.784
0.8	0.670	0.813	0.843	0.480	0.844	0.867

* lag-one correlation coefficient in the log domain

ACKNOWLEDGEMENTS

The work reported in this paper was supported in part by the research grant from the Japan International Cooperation Agency while the senior author was at the Asian Institute of Technology (AIT), Bangkok, Thailand. Many cooperations and kind assistance provided by the AIT Regional Computer Center were highly appreciated.

REFERENCES

(1) Fiering, M.B.; Multivariate technique for synthetic hydrology, Journal of the Hydraulics Division, ASCE, Vol. 90, No.HY5, pp. 43-60, 1964.
(2) Hoshi, K., Burges, S.J. and Yamaoka, I.; Reservoir design capacities for various seasonal operational hydrology models, Proc. of the Japan Society of Civil Engineers, No. 273, pp. 121-134, 1978.
(3) Kendall, M.G.; A Course in Multivariate Analysis, 185 pp., Charles Griffin, London, 1961.
(4) Kotz, S. and Neumann, J.; On the distribution of precipitation amounts for periods of increasing length, Journal of Geophysical Research, Vol. 68, No.12, pp. 3635-3640, 1963.
(5) Kotz, S. and Adams, J.W.; Distribution of sum of identically distributed exponentially correlated gamma-variables, Annals of Mathematical Statistics, Vol. 35, No.1, pp. 277-283, 1964.
(6) Lettenmaier, D.P. and Burges, S.J.; An operational approach to preserving skew in hydrologic models of long-term persistence, Water Resources Research, Vol. 13, No.2, pp. 281-290, 1977.
(7) Murota, A., Etoh, T. and Tanaka, K.; Stochastic studies on sums of hydrologic variables, Proc. of the Japan Society of Civil Engineers, No.223, pp. 23-31, 1974 (in Japanese).
(8) Nagao, M.; Statistical estimation of theoretical curves between frequency and time distribution ratio of rainfall, Proc. of the Japan Society of Civil Engineers, No.243, pp. 33-46, 1975 (in Japanese).
(9) Thom, H.C.S.; Approximate convolution of the gamma and mixed gamma distributions, Monthly Weather Review, Vol. 96, No.1, pp. 883-886, 1968.
(10) Todini, E.; The preservation of skewness in linear disaggregation schemes, Journal of Hydrology, No.47, pp. 199-214, 1980.
(11) Young, G.K.; Discussion of "Mathematical assessment of synthetic hydrology" by N. C. Matalas, Water Resources Research, Vol. 4, No.3, pp. 681-683, 1968.

APPENDIX - NOTATION

The following symbols are used in this paper :

a = scale parameter of gamma distribution ;
b = shape parameter of gamma distribution ;
b_{ij} = element of transformation matrix B ;
B = $(n \times n)$ coefficient matrix ;
B_1 = $(n \times n)$ coefficient matrix computed by being cubed element by element in matrix B ;
B_1^{-1} = inverse matrix of B_1 ;
c_i = element of vector C, eq. 29 or parameter of three-parameter log normal (LN3) distribution, eq. 33 ;
C = $(1 \times n)$ coefficient vector ;
D = $(n \times n)$ diagonal eigenvalue matrix of V_x ;
$E(\cdot)$ = expectation operator ;

$f(\cdot)$ = probability density function ;
g_i = independent standardized variate ;
G = $(n\times 1)$ vector of uncorrelated variates g_i ;
I = identity matrix ;
k_i = parameter of LN3 distribution ;
K_r = r-th cumulant ;
L = $(n\times n)$ lower triangular matrix ;
n = number of variates ;
P = $(n\times n)$ normalized eigenvector matrix of V_x ;
s_i = parameter $(=1+\gamma_i^2/2)$;
T = transposition of matrix as superscript ;
U = $(1\times n)$ unit vector ;
V_x = $(n\times n)$ covariance matrix of X ;
x, x_i = random variable ;
X = $(n\times 1)$ vector of correlated variates x_i ;
y_i = standard normally distributed variate ;
z = sum of n variates ;
β, β_i = coefficient of variation ;
γ, γ_i = coefficient of skewness ;
ρ = lag-one serial correlation coefficient ;
ρ_{ij}, $\rho_x(i, j)$ = correlation coefficient between variates x_i and x_j ;
μ, μ_i = population mean ;
σ^2, σ_i^2 = population variance ;
ϕ = parameter of ARMA(1, 1) model ;
τ_i = location parameter of LN3 distribution ;
η_i = parameter $[= exp(k_i^2)]$;
λ_i = eigenvalue of correlation matrix ;
$\Gamma(\cdot)$ = gamma function ;
Γ_x = $(n\times 1)$ vector of skewness coefficient γ_i ; and
Σ_x = $(n\times n)$ diagonal matrix of standard deviation σ_i.