Proc. Indian Acad. Sci., Vol. C 2, Pt. 1, March 1979, pp 133–140. © Printed in India.

TECHNICAL NOTE

New super-kurtic probability density function for use in computer simulation

P KUMARASWAMY
Institute of Hydraulics and Hydrology, Poondi 602 023

MS received 8 November 1977; revised 14 December 1977

Abstract. The philosophy of computer simulation and its application to hydrological processes is described in this paper. The structure of natural hydrologic time processes is indicated and the techniques to filter out white noise is explained. The limitations of the well-known probability density functions (PDF) such as the Gaussian, Pearson's and Johnson's etc. in hydrologic applications are set forth. A new super-kurtic PDF developed by the author specially for hydrological processes is introduced and a numerical example is given.

Keywords. Computer simulation; probability density functions; hydrological processes; sine power probability density function.

1. Introduction

Whenever an experimental solution to a problem is either impossible or too expensive or the problem too complicated for analytical treatment, Monte Carlo analysis techniques could be useful. This analysis implies the solution of a mathematical model, which may be either probabilistic or non-probabilistic, through simulation of a stochastic process, whose probability distribution satisfies the mathematical relations of the model. With the advent of fast electronic digital computers, generation of large samples within short time intervals and their subsequent analysis through computer simulation, has become practical for hydrologists. Computer simulation can be defined as a technique of reproducing real-time phenomena in a time-scaled mathematical model, utilising a digital or analogue computer. This paper however deals only with digital computer applications.

2. Computer simulation

2.1. *Features of simulation models*

A mathematical model must necessarily embody elements of two conflicting attributes—realism and simplicity. Hydrologic models consist of four elements: components, variables, parameters and functional relationships. Examples in a reservoir system model are given below.

Components:	Reservoir, canal, ayacut.
Variables:	Surface water inflow into reservoir. Subsoil flow into reservoir. Outflow from the reservoir. Reservoir water level.
Parameters:	Rainfall intensity and distribution. Soil moisture conditions. Watershed characteristics such as slope, ruggedness, shape, extent.
Functional relationships:	Probability distribution of rainfall, rainfall-runoff relationships, depth-storage equation of the reservoir.

Variables may be exogenous or endogenous. Exogenous variables are the independent or input variables of the model, which act upon the system and are not acted upon by the system, e.g. rainfall over a watershed. Exogenous variables can be grouped into either controllable or non-controllable variables. Infiltration of water to the aquifers is a controllable variable, whereas rainfall is a non-controllable variable.

Endogenous variables are the dependent or output variables of the system and are generated from the interaction of the system's exogenous and status variables. Status variables describe the state of a system or one of its components either at the beginning of a time-period or at the end of a time-period. Soil moisture condition of a watershed is a status variable and surface runoff is an endogenous variable.

2.2. *Classification of simulation models*

Simulation models can be classified as deterministic, stochastic, static and dynamic. In *deterministic models*, neither the exogenous variables nor the endogenous variables are permitted to be random variables, and the operating characteristics are assumed to be exact relationships rather than probability density functions. Those models in which at least one of the operating characteristics is given by a probability function are said to be *stochastic* models. *Static* models are those models which do not explicitly take the variable time into account. Mathematical models that deal with time-varying interactions are said to be *dynamic* models. Hydrologic models are of the stochastic and dynamic variety, with significant aspects of determinism and statics built in.

3. Structure of natural hydrological time processes

Natural hydrologic time processes are those time series of various hydrologic variables, viz., inputs, states of the system and outputs. Daily rainfall, groundwater levels, and weekly runoff into a reservoir are a few examples. All these processes are periodic-stochastic processes, with periodicities caused by astronomical cycles, and stochasticity introduced by random processes of the earth's environment including the atmosphere. For a better understanding and mathematical modelling of hydrologic stochastic processes, and for development of computer simulation techniques for generating new samples of these processes, structural analysis of the time series is to be taken up first.

Long-range trends (of a century, for example) in natural hydrological processes are basically results of systematic errors and man-made changes. Trend and cyclicity in samples of data may be produced by chance combinations of low and high values in a series. Powerful discrimination techniques of mathematical statistics are utilised to remove such pseudo-trends and pseudo-cyclicities.

Annual series of natural hydrological processes are treated as temporary stationary stochastic processes, with no periodicities. However daily, weekly and monthly series show periodicities with harmonics of the annual period of revolution of the earth around the sun, the period of self-rotation of the earth and the period of revolution of the moon around the earth.

Annual hydrologic series do not indicate any 'persistence' or Markov effect. But daily, weekly and monthly series show a great deal of persistence due to the capacitance effects of the hydrologic system. The atmosphere has a memory of about one to two weeks; and the watersheds and aquifers possess much longer memories, sometimes even for six to ten months (Markovic 1965).

When the periodic components in the mean, standard deviation, autocorrelation coefficients and skewness coefficients are identified and removed from a hydrologic time-process, a *serially-dependent* stationary stochastic component is produced. This component can be modelled as a linear Markovian type, with the residue as the *serially-independent* stochastic component, otherwise called 'white noise'.

The basic technique in computer simulation of stochastic hydrological processes is to fit a suitable probability distribution function to this white noise, and use it for generating samples of the hydrological process.

Multiregression methods are employed to correlate the *white noises* of the related variables, and use the estimated parameters for prediction of future variables. This paper deals with the development of a new PDF suitable for hydrological processes. It is considered that this PDF may play a successful role in many non-hydrological processes also.

4. Probability density functions used in hydrology

The usual probability density functions used in hydrology are Gaussian, gamma, beta, Pearson's, Johnson's lognormal, double exponential and Poisson (Quimpo 1967). It has been observed that many hydrological processes, especially short-period processes, such as *daily* rainfall, do not follow any of the PDFs mentioned above. The processes are highly skewed and super-kurtic (Todorovic 1969). Given the four important statistics of a hydrological white noise; namely mean $(\bar{z})$, standard deviation $(s(z))$, skewness coefficient $(g_1(z))$ and kurtosis coefficient $(g_2(z))$, it is not possible to fit the data satisfactorily to any of the well-known PDFs so that all the four parameters are in agreement. The numerical example given in § 6 will bring out this point clearly. Therefore hydrologists were in need of a PDF whose parameters could be reliably estimated *from only* the four statistics given above. It should also be possible to compute the random variable, given its cumulative probability, through a closed mathematical equation, without recourse to any tables. This last feature is very important for reducing the cost of computer time for simulation experiments since, in practice, von Neumann's rejection technique is found to consume too much of computer time, in the inverse process of getting the random variable, given the CDF. Noteworthy is the fact that only PDFs have been defined for the well-known types, and *not* CDFs which invariably are complicated integral expressions, and are not directly useful in simulation work, because these values can be obtained only through tabular interpolation and not through straightforward equations.

5. Sine power probability density function

An important feature of a hydrological process, such as daily rainfall, is that its lower bound is zero and no negative values are physically possible. The CDF of its lower bound need *not* be zero. It can have any value between 0 and 1. For example, in the month of April, the number of rainless days may be 25 at a given locality, which means that F_0 (CDF of lower bound) is equal to 0·83. The CDF of daily rainfall at this place can vary only between 0·83 and 1·00.

This process has also an upper bound. The maximum probable daily rain at any place *cannot* be infinite, but should be finite and larger than the so-far-observed maximum. The PDF may have a mode occurring between the two bounds, or its mode may lie at the lower or upper bound.

To satisfy all the criteria given above, a new PDF suitable for hydrological purposes, entitled 'sinepower probability density function' was developed by the author (Kumaraswamy 1976). If z_{min}, z_{m}, z_{max} represent the lower bound, mode and upper bound of a random variable $\{z\}$, the CDF and PDF equations are derived as given below. Let

$$x = (z - z_{\text{min}})/(z_{\text{max}} - z_{\text{min}}), \tag{1}$$

$$x_{\text{m}} = (z_{\text{m}} - z_{\text{min}})/(z_{\text{max}} - z_{\text{min}}), \tag{2}$$

$$n = -\ln(\pi/2)/\ln x_{\text{m}}, \tag{3}$$

$$m = s^2\, 1 - (1 - 1/n) \tan 1, \tag{4}$$

$$y = x/x_{\text{m}}, \tag{5}$$

$$F(z) = F_0 + (1 - F_0) \sin^m y^n, \tag{6}$$

where $F(z)$ is the CDF of z; and

$$f(z) = \frac{1 - F_0}{z_{\text{m}} - z_{\text{min}}}\, m n y^{n-1} \cos(y^n)\, [\sin(y^n)]^{(m-1)}, \tag{7}$$

where $f(z)$ is the PDF of z.

Given the CDF, $F(z)$, it is possible to compute z through the following equations.

$$\phi = (F - F_0)/(1 - F_0), \tag{8}$$

$$z = z_{min} + (z_m - z_{min}) [\sin^{-1} (\phi^{1/m})]^{1/n}. \quad (9)$$

Given $\bar{z}$, $s(z)$, $g_1(z)$, $g_2(z)$ of a sample, it is possible to estimate the parameters z_{min}, z_m, z_{max} and F_0 (Kumaraswamy 1977).

6. Numerical example

The daily rainfall at a specified locality for a period of 130 days is given below.

No rain on 104 days; Rain in mm on the remaining 26 days:

7·9, 24·0, 3·6, 6·4, 17·7, 0·5, 2·1, 16·0, 61·3, 16·5, 55·8, 12·5, 36·4, 10·9, 7·0, 5·6, 26·2, 11·0, 24·9, 38·8, 60·9, 35·4, 53·8, 77·3, 32·0, 0·1.

The observed CDF has been plotted in figure 1. The values of the four statistics are computed to be

$\bar{z}$ = 4·96 mm (mean)

$s(z)$ = 13·83 mm (unbiassed standard deviation)

g_1 = 3·28 (unbiassed skewness coefficient)

g_2 = 14·29 (unbiassed kurtosis coefficient)

The problem of fitting various PDFs utilising *all* the above four statistics is now discussed.

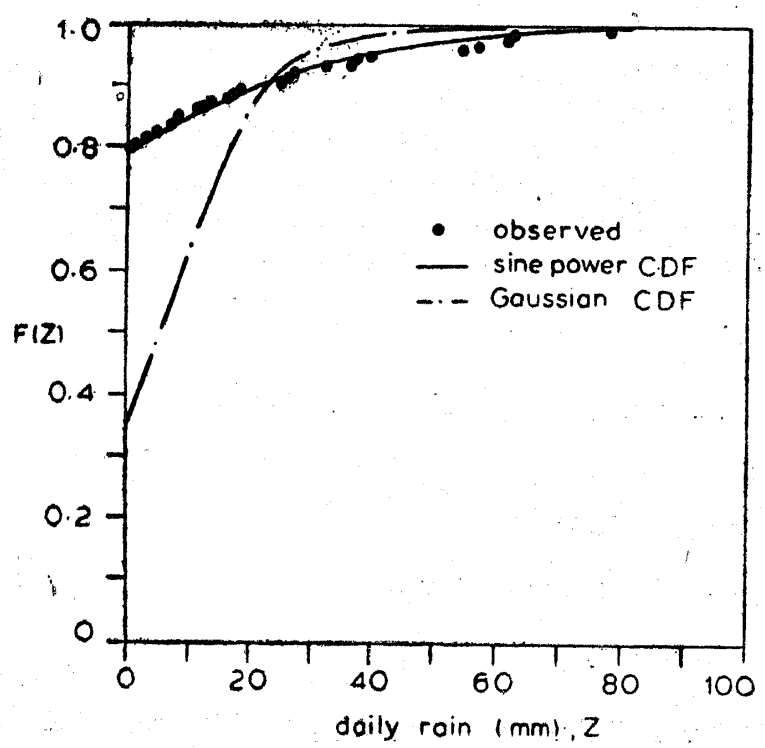

Figure 1. Cumulative distribution functions

Gaussian PDF

$g_1=0$ and $g_2=3$ for this distribution. Therefore, for the observed values of g_1 and g_2 equal to 3·28 and 14·29 respectively, this PDF is not suitable. However for the purpose of illustration, the CDF of this distribution has been calculated using the values of $\bar{x}$ and $s(z)$ and standard tables and plotted in figure 1. It is clear that the Gaussian PDF does not fit daily rainfall data.

Lognormal PDF

2-parameter lognormal PDF is computed. Using standard formulae for deriving μ_n and σ_n from z and $s(z)$, we get $\mu_n=0{\cdot}63$ and $\sigma_n=1{\cdot}43$. Therefore g_1 works out to 25·06 and $g_2=4578$ which are vastly different from the observed values of g_1 and g_2. This PDF is therefore not suitable.

Pearson's PDF

Using standard formulae given in Ven te Chow (1964) it can be seen that none of the seven PDFs fits to give the four statistics $\bar{z}$, $s(z)$, g_1 and g_2.

Johnson's PDF

To fit Johnson's PDF, it is necessary to know the values of the fractiles $z_{0.05}$, $z_{0.50}$ and $z_{0.95}$ in addition to $\bar{z}$, $s(z)$, g_1 and g_2. Since in this case the lower bound $\epsilon=0$ and $z_{0.05}=0$, it becomes impossible to fit Johnson's PDF (Hahn & Shapiro 1967).

Beta distribution

Since the PDF for daily rainfall is double-bounded, this is a distribution which has been applied extensively in hydrological work. Computing from the values of $\bar{z}$ and $s(z)$ we get $\alpha=-0{\cdot}9105$ and $\beta=0{\cdot}6251$, which indicate infinite density at the lower bound, which is not valid in hydrology. Moreover for the values of α and β, g_1 and g_2 are computed to be 3·57 and 16·81. This PDF has been plotted in figure 2 and it can be seen that the fit is not satisfactory.

Sine power PDF

Based on the parameters g_1 and g_2, F_0 is computed to be 0·80 and the other parameters as

$$z_{\mathrm{min}} = 0; \quad z_{\mathrm{m}} = 5{\cdot}0;$$

$$z_{\mathrm{max}} = 100{\cdot}0; \quad m = 12{\cdot}20;$$

$$n = 0{\cdot}1507.$$

Both the CDF and PDF are computed and plotted in figures 1 and 2. It can be seen that the PDF fits the observed data very well.

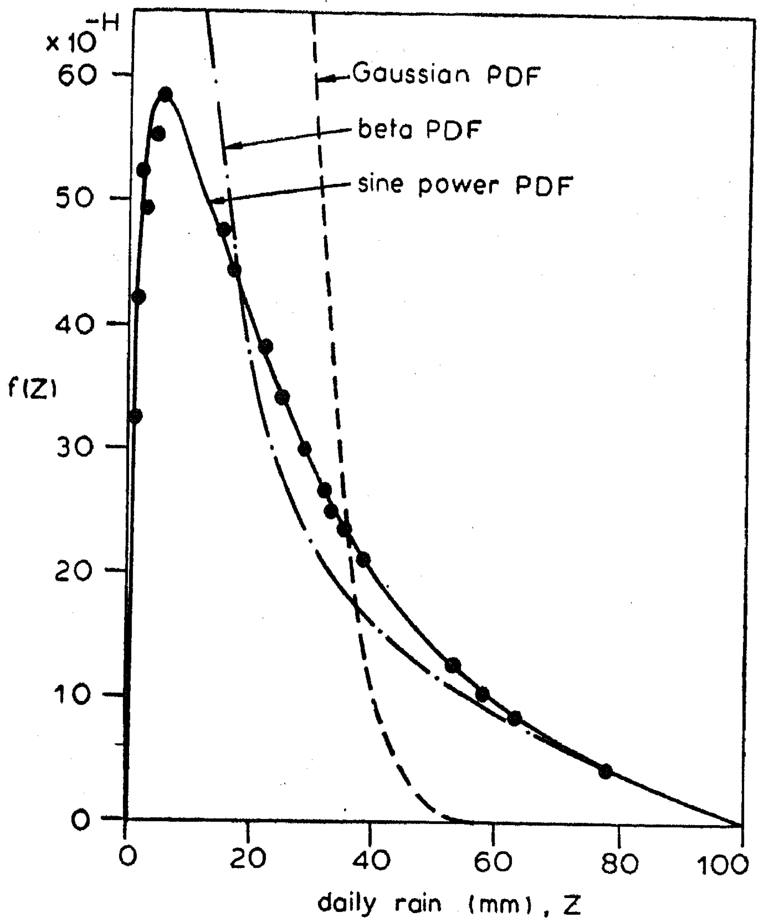

Figure 2. Probability density functions

The additional advantage with the sine power PDF is that, given $F(z)$, z can be computed through a straight equation which saves computer time enormously in large scale simulation work, unlike the Rejection method of von Neumann. This statement is made after testing about 500 samples of hydrological variables.

7. Conclusions

In computer simulation of hydrological variables, the sine power probability density function can be used to transform uniformly distributed $U(0, 1)$ random numbers into white noise. After adding the effects of persistence, periodicity and trend, the original hydrological process can be simulated and experiments conducted on the generated samples. By using this technique, computer simulation has been extensively adopted at Poondi in working out optimum operating schedules of water resources such as reservoirs and ground water aquifers.

References

Hahn G J & Shapiro S S 1967 *Statistical models in engineering* (New York: John Wiley)
Kendall M G & Stuart A 1958 *The advanced theory of statistics* Vol. 1.
Kumaraswamy P 1976 *J. Hydrol.* **31** 181
Pearson E S & Hartley H O 1972 *Biometrika table for statisticians* Vol. 2.
Markovic R D 1965 Probability functions of best fit to distributions of annual precipitation and run-off, Hyd. paper No. 8, Colorado State University, Colorado, USA
Quimpo R G 1967 Stochastic model of daily river flow sequences, Hyd. paper No. 18, Colorado State University, Colarado, USA
Todorovic P & Yevjevich V 1969 Stochastic process of precipitation, Hyd. paper No. 35, Colorado State University, Colorado, USA
Ven Te Chow 1964 *Handbook of applied Hydrology* (New York: McGraw Hill) **8** 13