PROBABILISTIC METHODS AND STOCHASTIC HYDROLOGY

G. G. S. Pegram
Civil Engineering Department, University of Natal, Durban, South Africa

Keywords : stochastic hydrology, statistical methods, model selection, stochastic models, single sites, multiple sites, random fields, space-time, simulation, forecasting, data infilling, data repairs, disaggregation.

Contents

Summary

Stochastic hydrology is the statistical branch of hydrology that deals with the probabilistic modeling of those hydrological processes which have random components associated with them. A stochastic hydrologist will suggest appropriate models and means of estimating the parameters of those models, and will go on to suggest techniques for simulating the processes and perhaps performing forecasts using those models. Testing the validity of the models is an important final step before the model is applied in practice. Phenomena whose models are described in this article are rainfall and stream flow, seen as both as point and spatial processes. The content is introductory in nature but attempts to offer as wide a spread of applications as may be useful, keeping the mathematical development at an expository level.

1. Introduction

Hydrology is the science that attempts to catalogue, understand and model the processes of precipitation of water and its passage through, over and under the earth’s land surfaces. Hydrology has strong links with other study areas, such as ecology, geology, meteorology, economics, politics, social science, and law. This article addresses the modeling, estimation, simulation and forecasting of two of the more variable processes in hydrology—precipitation and stream flow. These two processes are selected from a list of all hydrological processes which extends to include evaporation, infiltration and groundwater flow, and require the measurement, cataloguing and modeling of temperature, humidity, wind velocity, radiation, geology, soil and vegetation cover, geomorphology, etc. The restriction of the attention in this article to rainfall and stream flow is justified because these constitute the primary variables of interest to water resources engineers (see *Fluids at Rest and in Motion*; *Measurement of Free Surface Flow*).

The problem of the water resources engineer is that water is either too abundant, too lacking, or too dirty. Storage of water in reservoirs behind dams on rivers is a way of saving excess water from a time of plentiful supply, to be made available during periods of low flow, which can occur seasonally or over some years.

Estimates of reliability of the provision of water have to be made, together with the cost of supply in relation to its benefit. The dams and associated structures involved in the storage of water have to be protected from natural hazards, such as floods and earthquakes, and the chance that these might occur has to be assessed. These are tasks that face the water resources engineer, who has to understand how to extract the relevant information from the available rainfall and river flow data (see *Hydrological Data Acquisition Systems*).

The understanding and modeling of the processes, *ipso facto*, depends on what has been observed in the past, remembering that the future is not what the past used to be. The possibility of change in climate must be anticipated, modeled, and accounted for. Short-term precipitation forecasts, in hours for flash flooding, days for weather bulletins, and months for agricultural scientists, is a field of endeavor with possibly rich rewards, and finds its place in hydrology (see *Hydraulic Methods and Modeling*).

Stochastic models are used to describe the physical processes that are observed, and about which, data are recorded. Modeling is a process undertaken to understand and to find associations between processes, so that predictions of behavioral response can be conjectured and tested. Simulating the natural processes to produce possible future scenarios offers a means of testing the reliability of structures or schemes, and adds value to the information inherent in the measured data.

These data sets are often woefully short, inaccurately recorded, or at worst, are completely lacking, in which case the stochastic hydrologist is powerless to invent information. However, by careful inference, interpolation and cautious extrapolation, it is sometimes possible to transfer information to ungauged locations; this is an ongoing endeavor (see *Applied Hydraulics, Flow Measurement and Control*).

In this article, physical processes as such are not described, but rather the modeling of the data that present themselves in various forms in measurement of rainfall and stream flow. The branch of hydrology which deals with these tools is called Stochastic Hydrology. The word “stochastic” is a statistical term describing time series when they are not purely random, but exhibit dependence in time. Physical data collection is described elsewhere (see *Hydrological Data Acquisition Systems*).

There are two main sections following this introduction: Section 2 deals with Statistics, Probability and Model Selection; and Section 3 addresses Stochastic Models.

2. Statistics, Probability and Model Selection

This section presents some statistics and statistical tools, probability distributions, and techniques of model selection, which are commonly used in stochastic hydrology.

Statistics are numbers derived from data in such a way as to summarize neatly the major features of the behavior of the data; examples are location, spread, shape, bounds, etc. These descriptors include the mean, median, variance, skewness, and range of data, which can be calculated independently of any assumptions about the underlying distribution. They may suggest appropriate probability models (as distribution functions) which might describe or explain the way the data present themselves. Probability distribution functions are chosen to represent data derived from observations on processes in an economical way using a few parameters to define the functions. The exercise of choosing a suitable function is dealt with under model selection.

2.1 Statistics and Statistical Methods

Well-known statistics such as the mean and variance of a sample of data are the simplest of the quantities to derive. The mean is the average, (i.e., the sum of the data divided by the number of data) while the variance is the average of the squares of the differences of the data from their sample mean. The second statistic is an example of a product moment. A statistic which is more difficult to calculate is the median. The median is an example of an order statistic, computed by ranking the data in size from small to large or vice versa; the median is the middle of the ranked data, i.e., the value exceeded by half of the data in the sample.

Thus, given a sample of n data values x_i: i = 1, 2, …, n, the values are given by Equations (1) to (3):

$$\text{Mean:}\, m = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad (1)$$

$$\text{Variance:} \quad s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - m)^2 \qquad (2)$$

$$\text{Median:} \quad med = x^{[n/2]} \text{ or } \frac{\left(x^{[(n-1)/2]} + x^{[(n+1)/2]}\right)}{2} \qquad (3)$$

if n is even or odd, where $x^{[j]}$ is the j^{th} ordered sample value.

Other order statistics are the “quantiles” such as quartiles, being the values exceeded by ¼ or ¾ of the data. The median is an example of a robust statistical measure of location. The word robust in this context refers to the insensitivity of a statistic to outliers. Outliers are those data which do not seem to “belong” to the sample; they may be values which are erroneous, or may have been sampled from a different population than the rest of the data. It is often difficult to determine which is the right deduction and more needs to be found out about the collection and recording of the data, and especially of the outliers, to ascertain whether to accept the latter or exclude them from further calculations.

There are some useful statistical tests for outliers. One of these is to determine whether suspect data lie outside lower or upper cutoffs C_L and C_U calculated from the order statistics as in Equation (4):

$$C_L = F_L - 1.5 d_F \quad \text{and} \quad C_U = F_U + 1.5 d_F \qquad (4)$$

where F_L and F_U are the lower and upper quartiles and $d_F = F_U - F_L$, and is called the inter-quartile range. It may be that the data are very skewed, in which case they may require a transformation to make the data symmetrical before applying the outlier test. Note that if the data were truly normally distributed (see section 2.2) then d_F = 1.35s (where s is the standard deviation) and the cutoffs are at ± 2.7s, which would, on the average, identify seven values out of a thousand as outliers.

For finding relationships between data sets, a commonly used statistical tool is linear regression. Here, an underlying assumption is that there is some causal (physical) relationship between one variable and another. A hydrological example is the peak flow of a flood and its volume—if one is large (or small) the other tends to follow suit. A measure of the strength of this relation is a statistic called the correlation coefficient. This is calculated from a set of n pairs of data (x_j, y_j): j = 1, 2,…, n as given in Equation (5):

$$r_{x,y} = corr(x,y) \;\; \frac{1}{n}\sum_{i=1}^{n}(x_i - m_x)(y_i - m_y)/s_x s_y \qquad (5)$$

where the subscripted *m* and *s* are the sample means and standard deviations of the respective variables *x* and *y*. Note that *r* is bounded: $|r| < 1$. A high value of *r* (> 0.6, say) is indicative of a strong association between the random variables and vice versa.

Linear regression exploits these ideas to suggest models for transferring information from one variable to another. A hydrological example of the application of regression is the use of long rainfall and stream-flow sequences to extend short ones, perhaps after a transformation to normality.

2.2 Probability Distribution Functions

The probability that an event *A* happens is written as P[A], and is interpreted as a number from 0 to 1. The event A can be, for example, anything like the following:

- River flow rate at a given dam site exceeds 1000 m^3/s at least once a year.
- The rainfall collected in a gauge on January 7 is between 5 and 10 mm.
- There were exactly ten dry days in March.

In general, *A* is written $[X \le x]$, which is the event that a random variable *X* is less than a given value *x*, in which case, $P[X \le x]$ is written as $F_X(x)$ and is called the cumulative distribution function (*cdf*) of the random variable *X*. The probability that *X* exceeds a given value *x* is written $G_X(x) = P[X > x]$; note that $G_X(x) = 1 - F_X(x)$ and is sometimes called the survivor function.

If the variable *X* is continuous, in the sense that it can take the value of any real number over a range, then a probability density function (*pdf*), which is the derivative of $F_X(x)$, can be defined, labeled $f_X(x)$.

If the variable concerned is discrete, having come from a counting process, then a probability mass function (*pmf*) can be defined, usually on integer values, so that the equivalent notations are the following: $p_j = p_X(j) = P[X = j]$, and can be interpreted as $p_j = F_X(x) - F_X(x - \varepsilon)$, where ε is a small number. Thus in describing the distribution of a discrete random variable, the *cdf* is a series of steps, whereas with a continuous random variable, it is a smooth curve.

2.2.1 Continuous Distributions

Among commonly occurring distribution functions used in stochastic hydrology are the normal (Gaussian), log-normal, exponential, gamma and generalized extreme value distribution functions. Only an outline of their properties is presented here, as they are described in detail in readily available standard statistical texts (see *Statistical Methods*).

The normal distribution has some very attractive properties, which tempt hydrologists to transform the data mathematically, so that they appear normally distributed. Taking logarithms of the data, which of course must be nonnegative, is a commonly applied transformation of variables such as rain rates measured by a radar or derived from annual maximum flood peaks. If the logarithms (logs) of the data are judged to be normally distributed, then the data are assumed to obey the log-normal distribution. These two distributions, normal and log-normal, are compared in Figure 1.

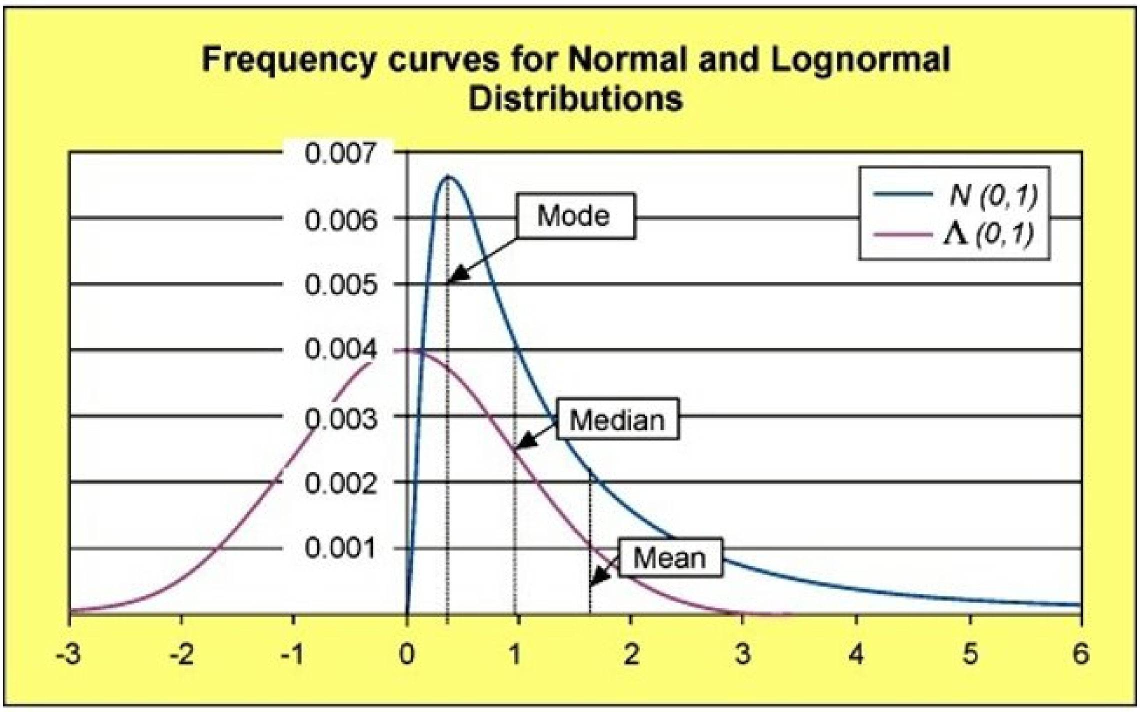

Figure 1. Comparison of the distributions of the normal, **N** and the log-normal, Λ probability density functions

The attractive properties of the normal distribution include the following:

- its two parameters are best estimated by the sample mean and variance, and in addition, in multidimensional cases, the sample covariance;
- the sums of variates from a normal distribution are normally distributed; and
- the sums of a large number of variates from arbitrary distributions are also normally distributed.

Unfortunately, there is no closed-form mathematical formula to express the normal *cdf*. Its *pdf* is given by Equation (6), wherein μ and σ^2 are the population mean and variance:

$$\phi_X(x) = (2\pi\sigma^2)^{-1} \exp[-\{(x-\mu)/\sigma\}^2/2] \qquad (6)$$

The normal distribution has been used to describe annual rainfall and evaporation totals, among other hydrological variables. Most other variable data sets are too skewed to be satisfactorily modeled directly by the normal distribution function.

The exponential distribution (sometimes called the negative exponential) is described by one parameter, and is possibly the simplest continuous distribution to work with. It is commonly used to describe the waiting times between events, and is an exact model of such a point process, if the events are randomly distributed in time. It is also an approximate model for the amount of rain falling into a rain gauge in a short interval, such as an hour, and is used to describe daily rainfall totals.

The exponential distribution has a *cdf* is given by Equation (7):

$$F_X(x) = 1 - \exp(-t/k) \qquad (7)$$

and its *pdf* is given by Equation (8):

$$f_X(x) = 1/k\ \exp(-t/k) \qquad (8)$$

Here, *k* is the mean of the distribution. A close relative of the exponential is the gamma distribution, which in some parametric representations is very similar to the log-normal distribution. The gamma distribution can be derived as a convolution of the exponential distribution, and for example, can be used to describe the waiting times between one or more occurrences of an independent point process. In the case of the probability of waiting time to the next arrival in such a point process, it becomes a specialized distribution that reverts to the exponential distribution.

The gamma distribution also is a special case of the Pearson Type 3 distribution, the related distribution of which, the log Pearson Type 3, has been recommended for the description of annual flood peaks observed in Australia and the US.

The Generalized Extreme Value (GEV) distribution is a function which has, as special cases, three distributions which have been suggested for the modeling of extremes, such as floods and droughts. Of these three, the best known is the Gumbel distribution, which Gumbel called the Type I distribution.

The Gumbel distribution has a *cdf* as given in Equation (9) below, which is easy to manipulate:

$$F_X(x) = \exp[-\exp\{-(x-\xi)/\alpha\}] \qquad (9)$$

where ξ and α are location and scale parameters, whose estimates are simply related to the sample mean and variance. Although routinely used by many practitioners to describe such variables as annual maximum stream flows, the Gumbel distribution suffers from the restriction that its coefficient of skewness is constant at the value 1.14. The GEV is a 3-parameter function that is not so limited, and is now preferred as a candidate for modeling extreme events. Its *cdf* is as given in Equation (10):

$$F_X(x) = \exp[-\{1 - \kappa(x-\xi)/\alpha\}^{1/\kappa}] \qquad (10)$$

which reduces to the Gumbel distribution of Equation (9), when κ = 0.

2.2.2 Discrete Distributions

The Bernoulli distribution describes binary processes like the occurrence of wet and dry days. It is the fundamental discrete distribution. Events, such as the occurrence of wet/dry days or a tossed coin turning up "heads" or "tails," are usually assigned one of two numbers: 0 or 1. In such cases, the probability mass function (*pmf*) is simply as given by Equations (11) and (12):

$$p_0 = P[X = 0] = q = 1 - p \qquad (11)$$

where

$$p_1 = P[X = 1] = p \qquad (12)$$

and is described by a single parameter, *p*, which is estimated by the sample mean.

The Binominal distribution is a generalization of the Bernoulli distribution and describes processes modeled by the outcomes of several (*n*) Bernoulli trials at each stage. A hydrological example of such a random variable is the number of wet days in a week. The *pmf* is given by Equation (13):

$$p_j = P[X = j] = \left(\frac{(n-j)!\,j!}{n!} \right) p^j q^{n-j} \text{ for } j = 0,1,\ldots,n \qquad (13)$$

where *q* = 1–*p* as in the Bernoulli distribution. The mean of the Binominal distribution is *np*.

The Poisson distribution describes the probability that a chosen number of occurrences of an independent point process will occur within a given interval. It is thus related to the exponential distribution which described above. If the mean interval between arrivals is *k* time units, where *k* is the parameter of the exponential, then the mean rate of arrival per unit time interval is λ = 1/*k*. The number of arrivals in an interval of *t* time units is then described by the Poisson distribution, whose *pmf* is given by Equation (14):

$$p_n = P[X = n] = \frac{(\lambda t)^n}{n!} \exp(-\lambda t) \text{ for } n = 0,1,2,\ldots \qquad (14)$$

A hydrological example of a random variable described by a Poisson distribution, is the number of storms in a given month of the year, in which case λ would change from month to month being larger in wet than in dry months.

The Poisson distribution is the fundamental building block in a class of models called Generalized Poisson processes, and another set called Rectangular Pulse models used in modeling gauged rainfall on a continuous basis.

Bibliography

Benjamin J. R. and Cornell C. A. (1970). *Probability, Statistics, and Decision for Civil Engineers.* New York: McGraw-Hill. 684 pp. [A fundamental subject treatise aimed at sound statistics-based decision making.]

Box G. E. P. and Jenkins G. M. (1970). *Time Series Analysis: Forecasting and Control.* Holden Day, San Francisco. 617 pp. [A good reference for hydrologists and hydraulic engineers involved with hydrological time series data analysis.]

Cressie N. (1991). *Statistics for Spatial Data.*, New York: Wiley. 900 pp. [A fundamental reference on the subject.]

Dooge J. C. I. (1973). *Linear Theory of Hydrologic Systems.* Technical Bulletin No 1468, Washington, DC: ARS, US Department of Agriculture. 327 pp. [An authoritative treatise by an expert on the subject.]

Kemeny J. G. and Snell J. L. (1960). *Finite Markov Chains.* New York: Van Nostrand. 210 pp. [A classic theoretical exposition of the subject, essential for understanding hydrologic time-series modeling.]

Linhart H. and Zucchini W. (1986). *Model Selection.* New York: Wiley. 310 pp. [A good review of the topic of Statistical Modelling.]

Maidment D. R., ed. (1993). *Handbook of Hydrology*. New York: McGraw-Hill. 1329 pp. [A basic text-book for the student and practicing hydraulic engineer.]

Salas J. D., Delleur J. W., Yevjevich V., and Lane W. L. (1980). *Applied Modeling of Hydrologic Time Series.* Littleton Colorado: Water Resources Publications. 484 pp. [An advanced level treatise for practitioners in hydrologic modeling.]

Vanmarke E. (1983). *Random Fields; Analysis and Synthesis.* Cambridge, Massachusetts: The MIT Press. 382 pp. [A good review of the various data-field modeling techniques.]

Wichman B. A. and Hill I. D. (1985). An efficient and portable pseudo-random number generator. In P. Griffiths and I. Hill, eds. *Applied Statistics Algorithms*. Chichester (UK): Ellis Horwood, for The Royal Statistical Society, London. 307 pp. [Deals with the difficulties of, and solutions to, finding suitable non-cyclic or repeating random number generators.]

Biographical Sketch

Geoffrey Pegram is Professor of Hydraulic Engineering in the Department of Civil Engineering at the University of Natal in Durban, South Africa. His Bachelors and Masters degrees in Engineering were obtained at the University of Natal and his doctorate was awarded by the University of Lancaster Mathematics Department for work on Probability Theory as applied to Storage. His expertise lies in hydraulic and hydrological modeling, stochastic hydrology and radar rainfall modeling. Apart from rain fields and rainfall modeling, his research interests include river flood hydraulics, flood protection and forecasting, as well as large reservoir system reliability. He has published in *Stochastic Hydrology, Water Resources and Hydraulics*, and has a current interest in the space-time modeling of rain fields measured by weather radar. He is the representative of the International Association of Hydrological Sciences

(IAHS) on the International Commission on Remote Sensing and Data Transmission (ICRSDT). He is a member of the South African National Committee of the IAHS (SANCIAHS) for 2000 to 2003.