APPLICATIONS OF STOCHASTIC METHODS IN HYDRAULICS

By

I. V. NAGY

Department of Water Management, Institute of Water Management and Hydraulic Engineering, Technical University, Budapest

(Received: February 15, 1977)

1. Mathematical modelling principles

During the early years of hydraulic design of structures, the approach has predominantly been a *deterministic* one, but over the past few years practicing hydrologists and hydraulic engineers began to search for better methods of hydraulic design. Recognition of the fact of not knowing all initial, boundary and geometric conditions propels some to a recognition that exact laws governing most hydraulic and hydrologic phenomena are very complex and in most cases they can only be approximated by *stochastic* methods. All processes in the natural environment are known to have a physical basis and randomness is but an accumulation of numerous unknown causal events into complex units. On the present level of knowledge *the stochastic approach provides a basis for greatly extending our model building and problem solving capability.*

The stochastic model is practically a mathematical representation of a natural process wherein the known behaviour of the set of random variables is distributed in a manner controlled by probabilistic laws. Of course, modelling is always an approximation of the prototype for the purpose of evaluating the performance of the prototype, but we know by experience that the stochastic relations give not only the deterministic solution as a special case but also the variance and covariance structure of predicted state variables. At the same time it must be remembered that stochastics is a very effective aid but not a total problem-solving technique, and the question is not "deterministic or probabilistic approach", but under what circumstances should either, or a combination of both, be used because, when using the averages of random variables, several hydraulic phenomena may appear as deterministic.

According to YEVYEVICH [1], when passing from a microscale to a macroscale in space and time, the stochastic process may become deterministic. We know by experience that the functions of random variables are also random variables, although they sometimes can be related by a deterministic function. The *deterministic solution is usually a particular case of the general probabilistic*

solution, but there exists a real danger herein, namely a considerable part of the available information may thus go lost and incorrect solutions may be obtained. In our understanding, the distinction between deterministic and stochastic processes is an artificial one, depending basically on the chosen scale of time and space. In the mathematical description of the prototype the differentiation between deterministic and indeterministic models can be assisted by relating them to the concepts of *certainty and uncertainty*, because in the latter case the risk aspects of uncertainties can be predicted with an element of probability. In fact, this does not imply that all uncertainties can be treated in terms of probability.

The stochastic simulation approach does not imply that the deterministic approach is less powerful or must not be used. The stochastic method is an effective way for studying complicated processes in hydrology we cannot understand otherwise by means of already known physical laws.

The so-called distinction between *stochastic hydraulics* and *stochastic hydrology* seems arbitrary because only the space-time patterns of water movement make the difference. *Hydrology* has been concerned with relatively infrequent large-scale phenomena and deals with problems of more environmental uncertainties than hydraulic problems do. *Hydraulics* often exhibit rapidly changing phenomena, but in principle it is not impossible to have extreme events in hydraulic phenomena either.

In reality the hydraulic magnitudes and parameters are random variables. The discussion of the implied issues may suggest practical solutions common to both disciplines.

The stochastic approach to hydraulic problems involving random elements has developed only in the last decade or so but there is now a widespread acceptance of stochastic models of complex physical processes that were treated earlier with limited success as deterministic.

The real difficulty lies in the fact that the majority of hydraulic problems accessible to theoretical treatment are restricted to one- or two-dimensional phenomena of water movement under steady and non-viscous conditions. As a result, the properties of water in a state of motion are described by equations of hydrodynamics due to Euler, Lagrange and Stokes, and consequently the practical results sometimes offered little help in understanding the *three-dimensional flow of real fluids*. Inherent to these dynamic equations is a limited capability for describing processes in space-time because of many uncertainties in model parameters and initial conditions and because of the stochastic character of governing functions in space and time. New problems involving *probabilistic characteristics in the basic equation of motion* are assuming importance.

It is well known that the hydraulic systems may be represented by differential equations derived by deterministic and/or stochastic methods, but

in the case of turbulent motion, there is a lot of uncertainty as to the validity of the classical basic equations of motion.

Consequently, in extending the model for real cases many of its assumptions must be critically examined and modified in some respects. One major difficulty is due to the fact that inputs are often uncertain, random, and the averages of system inputs seldom give the averages of observed outputs. A random input to a deterministic system gives a random output. Sometimes we may have an impression that the variance of the output is very small and then the stochastic process can approximately be replaced by the deterministic process. At the same time the possibility of misleading results must be remembered. The problem arises usually when the variances of random variables do not converge rapidly enough to zero with an increase of chosen scale of time and space.

A major difficulty originates from the fact that means of dealing with purely random variations alone have been developed to a relatively high degree in the hydraulics, but at the same time the investigation of systems responding to random inputs where a *time-dependent effect* exists has only been partly pursued in few cases.

According to CHOW [2] a simple modelling concept can be adopted by assuming hydraulic events to be purely random variables. In this case, measured data can be analyzed by many mathematical models of *probability distribution*. In reality, however, the response of a given system to any particular input may depend on the state of the system and when the state is affected by that input, it is usually difficult to evaluate the response of a system on a probability basis alone, given the probabilities of various inputs.

The occurrence of a hydraulic event may be affected by its *antecedents*. This means that hydraulic events may not occur in random sequences. A probability distribution is the distribution of a random variable whose given value cannot be predicted exactly except in terms of chances. If the distribution function is formulated for a given case, it is independent of when or where it occurs except under either a given or an average condition. In real situations, however, a random variable may have a *different probability distribution* for each point on the time scale and in space co-ordinates. These families of random variables constitute the *stochastic process*. Consequently, the deterministic process and the purely probabilistic process are but *two special cases* of the general stochastic process. When the probability or certainty of the random variable equals one then the stochastic process is a *deterministic* one, and when this probability is independent of any parameter index (time or space) and the family of random variables belongs to the same population, the stochastic process becomes purely *probabilistic*, including no deterministic component [2]. On a scale of probability from 0 to 1, the purely probabilistic and the deterministic processes will occupy respectively the two extremities,

while the stochastic process may occur anywhere between them, depending on the character of the investigated phenomena.

In our days, stochastic modelling is the highest level of modelling in hydrology and hydraulics, although it has not been well developed in view of many practical difficulties yet to overcome. The recorded random variables represent usually highly dependent processes and they are functions of very large numbers of known or unknown independent space, sequence, or time variables. It means that the statistical parameters should be related to the *hydrodynamics* of the flow and the *geometry* of the channel. Experiments are necessary to permit evaluating the influence of these parameters on the distribution functions of the observed events. Here a real danger exists, namely that sometimes the established relations between random variables do not show a real physical character because they are produced only by spurious correlation. Consequently, the interpretation to understand and explain the physical meaning of analytical results must be emphasized.

2. A general descriptor for the measure of stochastic relationships between random variables

Taking the considerable importance of knowing the *effect of antecedents* for the occurrence of an event in case of a given phenomenon into account, the use of the information theory is suggested in order to construct a new characteristic coefficient for the determination of the measure of stochastic relationship between random variables.

The *coefficient of correlation* is known to give first of all an information about the *linearity* of the relationship between variables only.

The proposed new method is, however, more general and may be useful for

(1) determining the measure of stochastic relationship between two or more variables,

(2) evaluating the Markovian character of given time series,

(3) testing the independence between variables.

Let us denote by X a random variable with specific values $x_1, x_2, \ldots, x_n$ and corresponding probabilities $p_1, p_2, \ldots, p_n$. The general expression of this distribution is

$$X: \begin{pmatrix} x_1, & x_2, & \ldots, & x_n \\ p_1, & p_2 & \ldots, & p_n \end{pmatrix}.$$

The entropy of the distribution of the variable X is described by the function:

$$H(X) = -\sum_{k=1}^{n} P_k \cdot \overset{2}{\log} p_k .$$

It was an original idea of V. NAGY and REIMANN [5] to use this entropy as a characteristic number of *uncertainty*. The uncertainty refers to the fact that X may assume different possible values in the course of the following observation. The uncertainty becomes maximum if X assumes any one of its possible values with uniform probability *(purely probabilistic process)*. Then the probability distribution of variable X is:

$$X: \begin{pmatrix} x_1, & x_2, & \ldots, & x_n \\ \frac{1}{n}, & \frac{1}{n}, & \ldots, & \frac{1}{n} \end{pmatrix}.$$

This statement is easy to prove by using the Jensen inequality. If $f(x)$ is an arbitrarily chosen convex function, then taking arbitrary values $x_1, x_2, \ldots, x_n$, the following inequality exists:

$$f\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) \leq \frac{1}{n}\sum_{i=1}^{n} f(x_i)\,. \tag{1}$$

By using the substitutions:

$$f(x) = x \log x, \qquad x_i = p_i$$

and taking into account that by definition $\sum_{i=1}^{n} p_i = 1$

then

$$\frac{\sum p_i}{n} \log \frac{\sum p_i}{n} = \frac{1}{n} \log \frac{1}{n} \leq -\frac{1}{n}\sum_{i=1}^{n} p_i \log p_i\,;$$

$$-\log n \geq \sum p_i \log p_i\,;$$

$$H(X) = -\sum p_i \log p_i \leq \log n = -\sum_{1}^{n} \frac{1}{n} \log \frac{1}{n}\,.$$

It is seen that if one of probabilities p_i equals 1 and all others equal 0, then $H(X) = 0$. In this case,

$$H(X) = 1 \log 1 + 0 \cdot \log 0 + 0 \cdot \log 0 + \ldots = 0$$

$$(x \log x = 0, \quad \text{for } x = 1).$$

Here the random variable X will assume the value x_i with a probability or certainty $p_i = 1$ and the *process is a deterministic one.*

Let us have now two discrete random variables X and Y with the corresponding probability distributions:

$$X: \begin{pmatrix} x_1, & x_2, \ldots, & x_n \\ p_1, & p_2, \ldots, & p_n \end{pmatrix}, \quad Y: \begin{pmatrix} y_1, & y_2, \ldots, & y_n \\ q_1, & q_2, \ldots, & q_n \end{pmatrix}$$

and

$$P(X = x_i) = p_i,\ p(Y = y_i) = q_j,\ P(X = x_i, Y = y_i) = r_{ij}$$
$$i, j = 1, 2, \ldots, n.$$

The entropy of joint distribution of random variables (X, Y) is defined a

$$H(X, Y) = -\sum_i \sum_j r_{ij} \log r_{ij}. \tag{2}$$

In the case X and Y are independent random variables, we have

$$r_{ij} = p_i q_j$$

and

$$H^*(X, Y) = -\sum_i \sum_j p_i q_j \log (p_i q_j) = \sum_i \sum_j p_i q_j \log p_i -$$

$$-\sum \sum p_i q_j \log q_j = -\sum_j \left(q_j \sum_i p_i \log p_i \right) - \sum_i \sum_j (q_j \log q_j) =$$

$$= H(X) + H(Y).$$

It is seen that

$$H(X, Y) \leq H^*(X, Y)$$

because from the definition of conditional probability it follows:

$$r_{ij} = P(X = x_i, Y = y_j) = P(X = x_i / Y = y_j)\, P(Y = y_j) =$$
$$= P(Y = y_j / X = x_i)\, P(X = x_i).$$

Consequently,

$$H(X, Y) = -\sum_i \sum_j \frac{r_{ij}}{q_j} q_j \log \frac{r_{ij}}{q_j} q_j = \sum_i \sum_j P(X = x_i / Y = y_j) \cdot$$

$$\cdot P(Y = y_j) \left[\log P(X = x_i / Y = y_j) + \log P(Y = y_j) \right] =$$

$$= \sum_i \sum_j \left[P(X = x_i / Y = y_j) \log P(X = x_i / Y = y_j) \right] P(Y = y_j) -$$

$$-\sum_i \sum_j P(X = x_i / Y = y_j)\, P(Y = y_j) \log P(Y = y_j) =$$

$$= \sum_j P(Y = y_i)\, H(X / Y = y_j) - \sum \left[P(Y = y_j) \cdot \right.$$

$$\left. \cdot \log P(Y = y_i) \right] \sum_i (X = x_i / Y = y_j).$$

Let us introduce the equality:

$$\sum_j P(Y = y_j)\, H(X/Y = y_j) = H(X/Y)$$

the so-called *conditional entropy* of random variable X with respect to Y. Then

$$H(X, Y) = H(X/Y) + H(Y) \tag{3}$$

and consequently

$$H^*(X, Y) - H(X, Y) = H(X) - H(X/Y) \geq 0. \tag{3'}$$

On the basis of the Jensen inequality it is seen that

$$H(H/Y) \leq H(X).$$

The relationship (3') essentially expresses the degree by which the uncertainty of the expected value or the variable X decreases if the value of Y is known. *The more the uncertainty decreases the more information is gained from Y with respect to X.*

Let us denote this quantity of information by $I(X, Y)$. Then,

$$I(X, Y) = H^*(X, Y) = H(X) - H(X/Y) \geq 0.$$

The value $I(X, Y)$ may be referred to as a *coefficient of stochastic relationship (CSR)* between random variables. For practical calculation purposes, however, use of the following relationship seems more suitable, namely

$$CSR(X, Y) = \frac{I(X, Y)}{H(X)} = \frac{H(X) - H(X/Y)}{H(X)} = 1 - \frac{H(X/Y)}{H(X)}, \tag{4}$$

percentage decrease of the uncertainty with respect to X. This relationship may be rewritten into a symmetrical form, by considering

$$I(X, Y) = H(X) - H(X/Y) = H(Y) - H(Y/X).$$

Then,

$$CSR(X, Y) = \frac{2I(X, Y)}{H(X) + H(Y)} = 1 - \frac{H(X/Y) + H(Y/X)}{H(X) + H(Y)}. \tag{5}$$

If there is a functional *(deterministic)* relationship between X and Y, i.e. $X = \varphi(Y)$, then in the expression

$$H(X/Y) = -\sum_i y_i \sum_k P(X = x_k/Y = y_i) \log P(X = x_k/Y = y_i), \tag{6}$$

For $Y = y_i$, $\varphi(y_i) = X_j$, and

$$P(X = x_j/Y = y_i) = 1$$

and

$$P(X = x_k/Y = y_i) = 0, \quad \text{for} \quad k \neq j.$$

Hence:

$$H(X/Y = y_i) = 0$$

and consequently

$$H(X/Y) = 0\,.$$

In the second case, for a *purely probabilistic process* i.e., X and Y are independent, in the expression (6)

$$P(X = x_k/Y = y_i) = P(X = x_k) = p_k$$

and then

$$H(X/Y) = -\sum_i q_i \sum_k p_k \log p_k = -\sum_k p_k \log p_k \,(\Sigma q_i) = H(X)$$

and

$$CSR(X, Y) = 0\,.$$

Consequently, properties of the proposed coefficient are

1. $CSR(X, Y) = CSR(Y, X)$,
2. $0 \leq CSR(X, Y) \leq 1$,
3. $CSR(X, Y) = 0$, if and only if X and Y are *independent*,
4. $CSR(X, Y) = 1$, if and only if there is a *functional relationship* between X and Y.

The above method may be generalized for *three or more variables*. For example, in the case of three variables X, Y and Z:

$$CSR(Z, X, Y) = 1 - \frac{H(Z/X, Y)}{H(Z)}$$

or

$$CSR(X, Y, Z) = 1 - \frac{H(X/Y, Z) + H(Y/X, Z) + H(Z/X, Y)}{H(X) + H(Y) + H(Z)} \tag{7}$$

It is an interesting conclusion that between the traditional *coefficient of correlation* and the new *characteristic coefficient* proposed by us there is a very simple connection if the joint distribution of X and Y follows a *normal* distribution [5].

Summary

The paper deals with several fundamental questions of describing hydraulic phenomena on the basis of deterministic or stochastic approaches. It is stressed that exact laws governing most hydraulic and hydrologic phenomena are very complex in nature and in many cases they can only be approximated by stochastic methods. There is no essential difference between deterministic and stochastic processes. The distinction depends basically on the chosen scale of time and space.

In the mathematical interpretation of the measured data, differentiation between deterministic and indeterministic models can be assisted by relating them to the concepts of *uncertainty*. The recorded random variables usually represent highly dependent processes and are functions of very many known or unknown independent space, sequence, or time variables. It means that there is a practical and theoretical need for better describing the stochastic relationships between random variables.

On the basis of the information theory a new method is suggested which may be useful for

(1) determining the measure of the stochastic relationships between two or more variables,

(2) evaluating the Markovian character of given time series,

(3) testing the independence between variables.

The method is developed to the point of practical application but the possible theoretical implications must not be underestimated.

References

1. YEVYEVICH, V. M.: Stochastic Processes in Hydrology. Water Resources Publications. Fort Collins, Colorado, USA, 1972.
2. CHOW, V. T.: Hydrologic Modelling. The Seventh John R. Preeman Memorial Lecture. Boston Society of Civil Engineers, 1972.
3. AMOROCHO, J., ORLOB, G. T.: Non-Linear Analysis of Hydrologic Systems. Univ. Calif. Publ. 1961.
4. CHIU, C. L.—LEE, T. S.: Stochastic Simulation in Study of Transport Processes in Irregular Natural Streams. Proc. of the First Internat. Symp. on Stochastic Hydraulics, Pittsburgh, USA, 1971.
5. NAGY, I. V.—REIMANN, J.: Forecasting of the Water Level Stages of Lake Balaton on the Basis of Information Theory. Proc. of the Helsinki Symposium, 1973.
6. VÁGÁS, I.: The Passage Theory.* ATIVIZING, Report, Szeged, 1974.
7. STAROSOLSZKY, Ö.: Considerations on the Future of Stochastic Hydraulics in the Computer Era. Proc. of the First Internat. Symp. on Stoch. Hydraulics, Pittsburgh, 1971.
8. KARTVELISVILI, N. A.: Stokhasticheskaya gidrologia (Stochastic hydrology). Gidrometeoizdat, Leningrad, 1975.

Prof. Dr. IMRE V. NAGY, H-1521 Budapest

* In Hungarian.