CHAPTER 6

Basic Concepts of Stochastic Hydrology

6. Introduction

Stochastic hydrology describes the physical processes involved in the movement of water onto, over, and through the soil surface. Quite often the hydrologic problems we face do not require a detailed discussion of the physical process, but only a time series representation of these processes. Stochastic models may be used to represent, in simplified form, these hydrologic time series. Some background in probability and statistics is necessary to fully understand this concept.

6.1 Time Series

The measurements or numerical values of any variable that changes with time constitute a time series. In many instances, the pattern of changes can be ascribed to an obvious cause and is readily understood and explained, but if there are several causes for variation in the time series values, it becomes difficult to identify the several individual effects. In Fig. 5.1, the top graph shows a series of observations changing with time along the abscissa; the ordinate axis represents the changing values of y with time, t. From visual inspection of the series, there are three discernible features in the pattern of the observations. Firstly, there is a regular gradual overall increase in the size of values; this trend, plotted as a separate component y1(t), indicates a linear increase in the average size of *y* with time. The second obvious regular pattern in the composite series is a cyclical variation, represented separately by y2(t), the periodic component. The third notable feature of the series may be considered the most outstanding, the single high peak half way along the series. This typically results from a rare catastrophic event which does not from part of a recognizable pattern. The definition of the function y3(t) needs very careful consideration and may not be possible. The remaining hidden feature of the series is the random stochastic component, y4(t), which represents an irregular but continuing variation within the measured values and may have some persistence. It may be due to instrumental of observational sampling errors or it may come from random unexplainable fluctuations in a natural physical process. A time series is said to be a random or stochastic process if it contains a stochastic component. Therefore,

most hydrologic time series may be thought of as stochastic processes since they contain both deterministic and stochastic components. If a time series contains only random/stochastic component is said to be a purely random or stochastic process.

The complete observed series, y(t), can therefore be expressed by:

$$y(t) = y1(t) + y2(t) + y3(t) + y4(t) \quad (6.1)$$

The first two terms are deterministic in form and can be identified and quantified fairly easily; the last two are stochastic with major random elements, and some minor persistence effects, less easily identified and quantified.

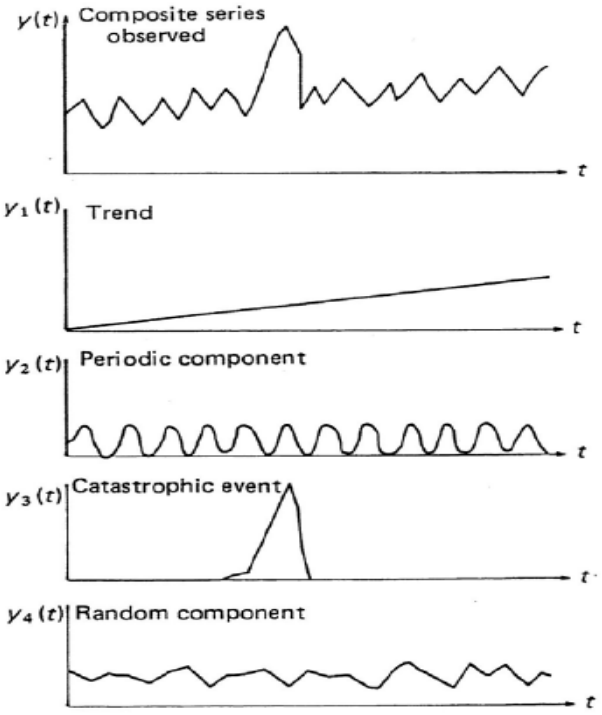


Figure 6 .1: The time series components

Properties of Time Series

The purpose of a stochastic model is to represent important statistical properties of one or more time series. Indeed, different types of stochastic models are often studied in terms of the statistical properties of time series they generate. Examples of these properties include: trend, serial correlation, covariance, cross-correlation, etc. Therefore, before reviewing the different types of stochastic models used in hydrology, some distribution properties of

stochastic processes need to be discussed. The following basic statistics are usually used for expressing the properties/characteristics of a time series.

$$Mean,\ \mu = E(X_1) = \bar{X} = \frac{1}{n}\sum_{t=1}^{n} X_t \qquad (6.2)$$

$$Variance,\ \sigma^2 = S^2 = \frac{1}{n-1}\sum_{t=1}^{n}(X_t - \bar{X})^2 \qquad (6.3)$$

$$Covariance,\ \lambda = Cov(X_t, X_{t+L}) = \frac{1}{n-L}\sum_{t=1}^{n-L}(X_t - \bar{X})(X_{t+L} - \bar{X})$$

Where L is the time lag.

Stationary time series:

If the statistics of the sample (mean, variance, covariance, etc.) as calculated by equations (6.2)-(6.4) are not functions of the *timing* or the *length* of the sample, then the time series is said to be stationary to the second order moment, weekly stationary, or stationary in the broad sense. Mathematically one can write as:

$$E(X_1) = \mu$$
$$Var(X_t) = \sigma^2$$
$$Cov(X_t, X_{t+L}) = \lambda_L$$

In hydrology, moments of the third and higher orders are rarely considered because of the unreliability of their estimates. Second order stationarity, also called covariance stationarity, is usually sufficient in hydrology. A process is strictly stationary when the distribution of Xt does not depend on time and when all simultaneous distributions of the random variables of the process are only dependent on their mutual time-lag. In another words, a process is said to be strictly stationary if its n-th (n for any integers) order moments do not depend on time and are dependent only on their time lag.

Non-stationary time series:

If the values of the statistics of the sample (mean, variance, covariance, etc.) as calculated by equations (6 .2)-(6.4) are dependent on the *timing* or the *length* of the sample, i.e. if a definite trend is discernible in the series, then it is a nonstationary series. Similarly, periodicity in a series means that it is non-stationary. Mathematically one can write as:

$$E(X_1) = \mu_t$$
$$Var(X_t) = \sigma_t^2$$
$$Cov(X_t, X_{t+L}) = \lambda_{L,t}$$

White noise time series:

For a stationary ties series, if the process is purely random and stochastically independent, the time series is called a white noise series. Mathematically one can write as:

$$E(X_1) = \mu$$
$$Var(X_t) = \sigma^2$$
$$Cov(X_t, X_{t+L}) = 0 \text{ for all } L \neq 0$$

Gaussian time series:

A Gaussian random process is a process (not necessarily stationary) of which all random variables are normally distributed, and of which all simultaneous distributions of random variables of the process are normal. When a Gaussian random process is weekly stationary, it is also strictly stationary, since the normal distribution is completely characterized by its first and second order moments.

Analysis of Hydrologic Time Series

Records of rainfall and river flow form suitable data sequences that can be studied by the methods of ***time series analysis***. The tools of this specialized topic in mathematical statistics provide valuable assistance to engineers in solving problems involving the frequency of occurrences of major hydrological events. In particular, when only a relatively short data record is available, the formulation of a time series model of those data can enable long sequences of comparable data to be generated to provide the basis for better estimates of hydrological behaviour. In addition, the time series analysis of rainfall, evaporation, runoff and other sequential records of hydrological variables can assist in the evaluation of any irregularities in those records. Cross-correlation of different hydrological time series may help in the understanding of hydrological processes.

Tasks of time series analysis include:

(1) Identification of the several components of a time series.

(2) Mathematical description (modelling) different components identified.

If a hydrological time series is represented by X1, X2, X3, ..., Xt, ..., then symbolically, one can represent the structure of the Xt by:

$X \Leftrightarrow [T_t, P_t, E_t]$

Where Tt is the trend component, Pt is the periodic component and Et is the stochastic component. The first two components are specific deterministic features and contain no element of randomness. The third, stochastic, component contains both random fluctuations and the self-correlated persistence within the data series. These three components form a basic model for time series analysis.

The aims of time series analysis include but not limited to:

(1) Description and understanding of the mechanism,

(2) Monte-Carlo simulation,

(3) Forecasting future evolution,

Basic to stochastic analysis is the assumption that the process is stationary. The modelling of a time series is much easier if it is stationary, so identification, quantification and removal of any non-stationary components in a data series is under-taken, leaving a stationary series to be modelled.

6.1.1 Trend component

This may be caused by long-term climatic change or, in river flow, by gradual changes in a catchment's response to rainfall owing to land use changes. Sometimes, the presence of a trend cannot be readily identified.

Methods of trend identification:

Different statistical methods, both *nonparametric tests and parametric tests*, for identifying trend in time-series are available in the literature. Two commonly used methods for identifying the trend are discussed briefly in this section.

(1) Mann-Kendall test

The test uses the raw (un-smoothed) hydrologic data to detect possible trends. The Kendall statistic was originally devised by Mann (1945) as a non-parametric test for trend. Later the exact distribution of the test statistic was derived by Kendall (1975). The Mann-Kendall test is based on the test statistic S defined as follows:

$$S = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \operatorname{sgn}(x_j - X_i) \qquad (6.5)$$

Where the Xj are the sequential data values, *n* is the length of the data set, and

$$\operatorname{sgn}(\theta) = \begin{cases} 1 if\ \theta > 0 \\ 0 if\ \theta = 0 \\ -1 if\ \theta < 0 \end{cases} \qquad (6.6)$$

Mann (1945) and Kendall (1975) have documented that when, the statistic S is approximately normally distributed with the mean and the variance as follows:

$$E(S) = 0 \qquad (6.7)$$

$$V(S) = \frac{n(n-1)(2n+5) - \sum_{p=1}^{q} t_p (t_p - 1)(2t_p + 5)}{18} \qquad (6.8)$$

Where n = number of data

tp = the number of ties for the pth value (number of data in the pth group)

q = the number of tied values (number of groups with equal values/ties)

The standardized Mann-Kendall test statistic ZMK is computed by

$$\operatorname{sgn}(\theta) = \begin{cases} \dfrac{S-1}{\sqrt{Var(s)}} & S > 0 \\ 0 & S = 0 \\ \dfrac{S+1}{\sqrt{Var(s)}} & S < 0 \end{cases} \qquad (6.9)$$

The standardized MK statistic Z follows the standard normal distribution with mean of zero and variance of one.

The hypothesis that there has not trend will be rejected if

$$|Z_{mk}| > Z_{1-\alpha/2} \qquad (7.10)$$

Where $Z_{1-\alpha/2}$ is the value read from a standard normal distribution table with α being the significance level of the test.

(2) Linear regression method

Linear regression method can be used to identify if there exists a linear trend in a hydrologic time series. The procedure consists of two steps, fitting a linear

regression equation with the time T as independent variable and the hydrologic data, Y as dependent variable, i.e.

$Y = \alpha + \beta.T$ (6.11)

and testing the statistical significance of the regression coefficient β.

Test of hypothesis concerning β can be made by noting that (β – β0)/Sβ has t distribution with n-2 degrees of freedom. Thus the hypothesis H0: β = β0 versus H0: β ≠ β0 is tested by computing

$$t = \frac{\beta - \beta_0}{S_\beta} \quad (6.12)$$

Where S_β is the standard deviation of the coefficient β with

$$S_\beta = \frac{S}{\sqrt{\sum_{i=1}^{n} (T_i - \bar{T})^2}} \quad (6.13)$$

And

$$S = \sqrt{\frac{1}{n-2} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2} \quad (6.14)$$

Where S is the standard error of the regression, *Yi and* $\hat{Y}$ are observed and estimated hydrologic variable from the regression equation, respectively.

The hypothesis, i.e. no trend, is rejected if $|t| > t_{1-\alpha/2,\, n-2}$

Models for trend:

The shape of the trend depends on the background of the phenomenon studied. Any smooth trend that is discernible may be quantified and then subtracted from the sample series. Common models for trend may take the following forms:

Tt = a + bt (a linear trend, as in Fig. 5.1) (6.15)

or

Tt = a + bt + ct2 + dt3 + ... (a non-linear trend)

(6.16)

The coefficients a, b, c, d ... are usually evaluated by least-squares fitting. The number of terms required in a polynomial trend being primarily imposed by the interpretation of the studied phenomenon. The number of terms is usually based on statistical analysis, which determines the terms contributing significantly to the description and the interpretation of the

time series. Restriction is made to the significant terms because of the principle of parsimony concerning the number of unknown parameters (constants) used in the model. One wishes to use as small a number of parameters as possible, because in most cases the addition of a complementary parameter decreases the accuracy of the other parameters. Also prediction- and control procedures are negatively influenced by an exaggerated number of parameters. This principle of parsimony is not only important with respect to the selection of the trend function but also with respect to other parts of the model.

6.1.2 Periodic component

In most annual series of data, there is no cyclical variation in the annual observations, but in the sequences of monthly data distinct periodic seasonal effects are at once apparent. The existence of periodic components may be investigated quantitatively by (1) Fourier analysis, (2) spectral analysis, and (3) autocorrelation analysis. Of which, the autocorrelation analysis method is widely used by hydrologists and will be discussed briefly in this section.

Identification of periodic component by autocorrelation analysis:

The procedure consists of two steps, calculating the autocorrelation coefficients and testing their statistical significance. For a series of data, Xt, the autocorrelation coefficient rL between Xt and Xt+L are calculated and plotted against values of L (known as the lag), for all pairs of data L time units apart in the series:

$$r_L = \frac{1}{n-L}\sum_{t=1}^{n-L}(X_t - \bar{X})(X_{t+L} - \bar{X}) \Big/ \frac{1}{n}\sum_{t=1}^{n}(X_t - \bar{X})^2 \qquad (6.17)$$

Where $\bar{X}$ is the mean of the sample of n values of Xt and L is usually taken for values from zero up to n/4. A plot of rL versus L forms the correlogram. The characteristics of a time series can be seen from the correlogram. Examples of correlograms are given in Fig 7.2. Calculation of equation (6.17) for different L gives the following cases:

• If L = 0, $rL = 1$. That is, the correlation of an observation with itself is one.

• If $rL \approx 0$ for all $L \neq 0$, the process is said to be a purely random process. This indicates that the observations are linearly independent of each other. The correlogram for such a complete random time series is shown in Fig 6 .2(a).

• If $rL \neq 0$ for some $L \neq 0$, but after $L > \tau$, then , the time series is still referred to as simply a random one (not purely random) since it has a 'memory' up to $L = \tau$. When τ r , the

process is said to have no memory for what occurred prior to time t-τ. The correlogram for such a non-independent stochastic process is shown in Fig 6.2(b). This is representative of an auto regressive process. Typically, such a correlogram could be produced from a series described by the Autoregressive model:

$$Xt = a1Xt\text{-}1 + a2Xt\text{-}2 + a3Xt\text{-}3 + \ldots + \varepsilon t \qquad (6.18)$$

where *ai* are related to the autocorrelation coefficients *ri* and εt is a random independent element.

• In the case of data containing a cyclic (deterministic) component, then $rL \neq 0$ for all $L \neq 0$, the correlogram would appear as in Fig. 7.2(c). Where T is the period of the cycle.

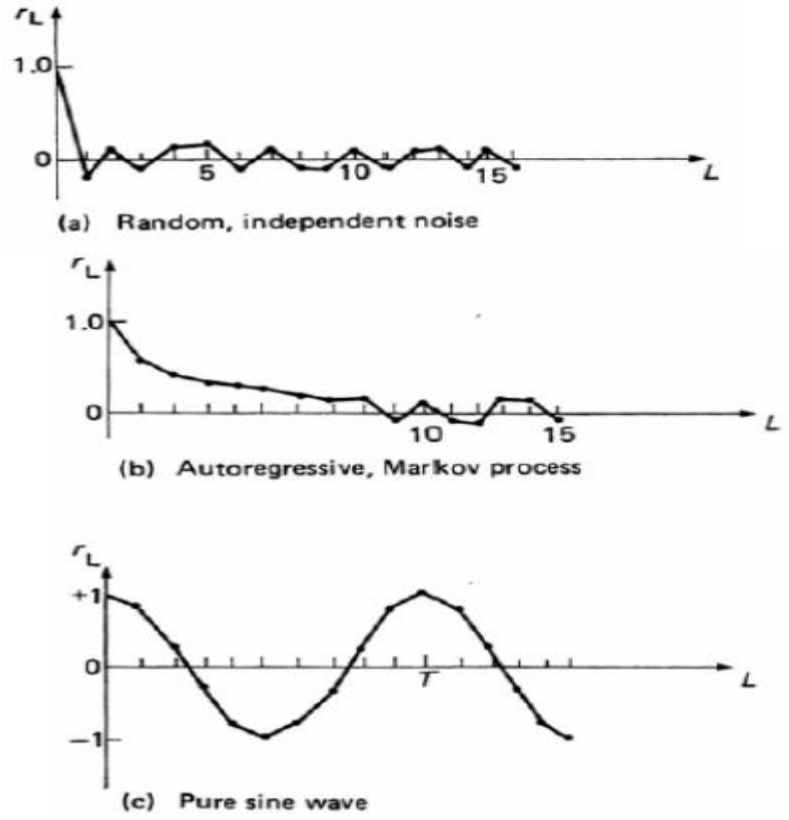


Figure 6 .2 Examples of Correlograms

Modelling of periodic component:

A periodic function Pt is a function such that

Pt+T = Pt for all t

The smallest value of T is called the period. The dimension of T is time, T thus being a number of time-units (years, months, days or hours, etc.) and we also have

Pt+nT = Pt for all t and for all integer n.

The frequency is defined as the number of periods per time-unit:

$$frequency = \frac{1}{period}$$

Trigonometric functions are simple periodic functions. For example, α sin (ωt + β) has a period of 2π/ω, because α sin [ω(t+2π/ω)+β] = α sin(ωt+2π+β) = α sin(ωt + β)

The pulsation or angular frequency is defined as

$$\omega = \frac{2\pi 1}{period} = 2\pi . frequency$$

the constant α is termed the amplitude and β the phase (with respect to the origin) of the sine-function.

A simple model for the periodic component may be defined as (for more discussions refer to the literature of Time Series Analysis):

$$Pt = m + C\sin(2\pi t/T) \qquad (6.19)$$

Where C is the amplitude of the sine wave about a level m and of wavelength T

The serial (auto) correlation coefficients for such a Pt are given by:

$$r_L = \cos(2\pi L/T) \qquad (6.20)$$

The cosine curve repeats every T time units throughout the correlogram with

r_L= 1 for L = 0, T, 2T, 3T,… Thus periodicities in a time series are exposed by regular cycles in the corresponding correlograms.

Once the significant periodicities, Pt, have been identified and quantified by μt (the means) and σt (the standard deviations) they can be removed from the original times series along with any trend, Tt, so that a new series of data, Et, is formed:

$$E_t = \frac{X_t - T_t - m_t}{S_t} \qquad (6.21)$$

Simple models for periodic component in hydrology can be seen in the literature. For example, in many regions, typical monthly potential evapotranspiration variation during the

year can be modelled more or less by a sinusoidal function, with a couple of parameters to tune the annual mean and the amplitude (Xu and Vandewiele, 1995).

This behavior leads to the idea to model by a truncated Fourier series:

ept = {a + bsin[(2π /12)(t-c)]}+

where again *t* is time in month. The plus sign at the end is necessary for avoiding negative values of *ep* which otherwise may occur in rare cases. Again parameters a, b and c are characteristics of the basin.

6.1.3 Stochastic component

ET represents the remaining stochastic component of the time series free from non-stationary trend and periodicity and usually taken to be sufficiently stationary for the next stage in simple time series analysis. This Et component is analysed to explain and quantify any persistence (serial (auto) correlation) in the data and any residual independent randomness. It is first standardized by:

$$Z_t = \frac{E_t - \bar{E}}{S_E} \qquad (6.22)$$

and *sE* are the mean and standard deviation of the Et series. The series, Zt, then has zero mean and unit standard deviation. The autocorrelation coefficients of Zt are calculated and the resultant correlograms is examined for evidence and recognition of a correlation and/or random structure.

For example, in Fig. 6.3a for a monthly flow, the correlogram of the Zt stationary series (with the periodicities removed) has distinctive features that can be recognized. Comparing it with Fig. 6.2, the Zt correlogram resembles that of an auto regressive (Markov) process. For a first order Markov model:

$$Z_t = r_1 \bar{Z}_{t-1} + e_t \qquad (6 .23)$$

Where r1 is the autocorrelation coefficient of lag 1 of the Zt series and et is a random independent residual. A series of the residuals et may then be formed from the Zt series and its known lag 1 autocorrelation coefficient, r1:

$$e_t = Z_t - r_1 Z_{t-1} \qquad (6.24)$$

The correlogram of residuals is finally computed and drawn (Fig.6.3b). For this data this resembles the correlogram of 'white noise', i.e. independently distributed random values. If there are still signs of autoregression in the et correlogram, a second-order Markov model is tried, and the order is increased until a random et correlogram is obtained. The frequency distribution diagram of the first order et values (Fig. 6.3c) demonstrates an approximate approach to the normal (Gussian) distribution.

At this stage, the final definition of the recognizable components of the time series has been accomplished including the distribution of the random residuals. As part of the analysis, the fitted models should be tested by the accepted statistical methods applied to times series. Once the models have been formulated and quantified to satisfactory confidence limits, the total mathematical representation of the time series can be used for solving hydrological problems by synthesizing non-historic data series having the same statistical properties as the original data series.

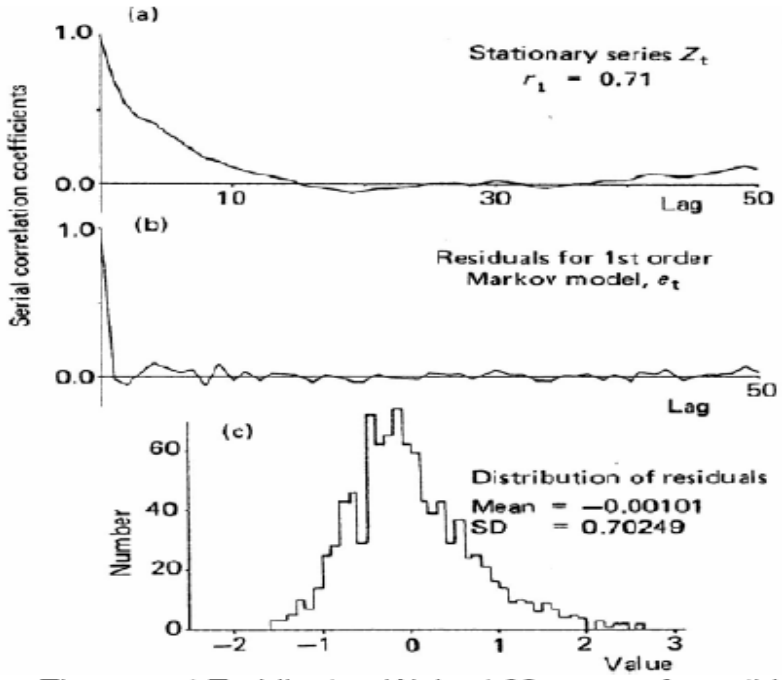


Figure 6.3 River Thames at Teddington Weirs (82 years of monthly flows, from Shaw, 1988)

6.2 Time Series Synthesis

The production of a synthetic data series simply reverses the procedure of the time series analysis. First, for as many data items as are required, a comparable sequence of random numbers, drawn from the et distribution, is generated using a standard computer package. Second, the corresponding synthetic Zt values are recursively calculated using equation 7.23 (starting the series with the last value of the historic Zt series as the Zt-1 value). Third, the Et series then derives from equation 7.22 in reverse:

$$E_t = Z_t S_E + \bar{E} \tag{6.25}$$

The periodic component Pt represented by mt and st for time period *t* is then added to the Et values to give:

$$X_t = T_t + E_t S_t + m_t \tag{6 .26}$$

The incorporation of the trend component Tt then produces a synthetic series of Xt having similar statistical properties to the historic data series.

6.3 Some Stochastic Models

Ultimately design decisions must be based on a stochastic model or a combination of stochastic and deterministic models. This is because any system must be designed to operate in the future. Deterministic models are not available for generating future watershed inputs in the form of precipitation, solar radiation, etc., nor is it likely that deterministic models for these inputs will be available in the near future. Stochastic models must be used for these inputs.

6.3.1 Purely random stochastic models

Possibly the simplest stochastic process to model is where the events can be assumed to occur at discrete times with the time between events constants, the events at any time are independent of the events at any other time, and the probability distribution of the event is known. Stochastic generation from a model of this type merely amounts to generating a sample of random observations from a univariate probability distribution. For example, random observations for any normal distribution can be generated from the relationship,

$$y = \sigma RN + \mu \qquad (6.27)$$

Where RN is a standard random normal deviate (i.e. a random observation from a standard normal distribution) and μ and σ are the parameters of the desired normal distribution of y. Computer routines are available for generating standard random normal distribution.

6.3.2 Autoregressive models

Where persistence is present, synthetic sequences cannot be constructed by taking a succession of sample values from a probability distribution, since this will not take account of the relation between each number of sequences and those preceding it. Consider a second order stationary time series, such as an annual time series, made up of a deterministic part and a random part. The deterministic part is selected so as to reflect the persistence effect, while it is assumed that the random part has a zero mean and a constant variance. One of the models to simulate such a series is the Autoregressive model. The general form of an autoregressive model is:

$$(y_t - \mu) = \beta_1 (y_{t-1} - \mu) + \beta_2 (y_{t-2} - \mu) + \ldots + \beta_k (y_{t-k} - \mu) + \varepsilon_t \qquad (6.28)$$

Where μ is mean value of the series, β is the regression coefficient, the {y1, y2, …, yt,…} is the observed sequence and the random variables εt are usually assumed to be normally and independently distributed with zero mean and variance . In order to determining the order k of auto regression required to describe the persistence adequately, it is necessary to estimate k+2 parameters: β1, β2, …βk, μ and the variance of residuals . Efficient methods for estimating these parameters have been described by Kendall and Stuart (1968), Jenkins and Watts (1968).

The first order auto regression:

$$y_t - \mu = \beta_1 (y_{t-1} - \mu) + \varepsilon_t \qquad (6.29)$$

has found particular application in hydrology. When equation (6.29) is used to model annual discharge series, the model states that the value of y in one time period is dependent only on the value of y in the preceding time period plus a random component. It is also assumed that εt is independent of yt. Equation (6.29) is the well-known first order Markov Model in the literature. It has three parameters to be estimated: μ, β1, and $2E$ σ .

For the moment method of parameter estimation, parameter μ can be computed from the time series as the arithmetic mean of the observed data. As for β1, the Yule-Walker equation (Delleur, 1991) shows that:

$$\rho_k = \sum_{j=1}^{P} \beta_j \rho_{k-j} \qquad k>0 \tag{6 .30}$$

the above equation, written for k = 1, 2, …, yields a set of equations. Where ρk is the autocorrelation coefficient for time lag k. As the autocorrelation coefficients ρ1, ρ2, …, can be estimated from the data using equation (4.17), these equations can be solved for the autoregressive parameters β1, β2, …, βp. This is the estimation of parameters by the method of moments. For example, for the first order autoregressive model, AR (1), the Yule-Walker equations yield

$$\rho_1 = \beta_1 . \rho_0 = \beta_1 \quad since \ \rho_0 = 1 \tag{6 .31}$$

Similarly we can derive the equations for computing β1 and β2 for the AR(2) model as

$$\beta_1 = \frac{\rho_1(1-\rho_2)}{1-\rho_1^2} \tag{6.32}$$

It can be shown that σ_E^2. is related to (the variance of the y_t series) b

$$\sigma_s^2 = \sigma_y^2(1-\beta_1^2) \tag{6 .33}$$

If the distribution of y is $N(\mu_y, \sigma_y^2)$ then distribution of ε is N(0, σ_E^2.). Random values y_t can now be generated by selecting ε_t randomly from a N(0, σ_E^2.) distribution. If z is N(0,1) then $Z\sigma_\phi \ or \ Z\sigma_y\sqrt{1-\beta_1^2}$ *is* $N(0,\sigma_s^2$. Thus, a model for generating Y's that are $N(\mu_y, \sigma_y^2)$ and follow the first order Markov model is

$$y_t = \mu_y + \beta_1(y_{t-1} - \mu_y) + Z_t\sigma_y\sqrt{1-\beta_1^2} \tag{6 .34}$$

The procedure for generating a value for y_t is:

(1) estimate μ_y, σ_y, and β_1 by $\bar{y}$, s_x and r_1

(2) select a z_t at random from a N(0, 1) distribution, and

(3) calculate y_t by Eq. (7.34) based on $\bar{y}$, s_x and β_1, and y_{t-1}

The first value of y_t, i.e. y_1, might be selected at random from a $N(\mu_y, \sigma_y^2)$. To eliminate the effect of y_1 on the generated sequence, the first 50 or 100 generated values might be discarded.

Eq. (6.34) has been widely used for generating annual runoff from watersheds

6.3.3 First order Markov process with periodicity: Thomas – Fiering model

The first order Markov model of the previous section assumes that the process is stationary in its first three moments. It is possible to generalise the model so that the periodicity in hydrologic data is accounted for to some extent. The main application of this generalisation has been in generating monthly streamflow where pronounced seasonality in the monthly flows exists. In its simplest form, the method consists of the use of twelve linear regression equations. If, say, twelve years of record are available, the twelve January flows and the twelve December flows are abstracted and January flow is regressed upon December flow; similarly, February flow is regressed upon January flow, and so on for each month of the year.

$$q_{jan} = \bar{q}_{jan} + b_{jan}(q_{dect} - \bar{q}_{dec}) + \varepsilon_{jant}$$

$$q_{feb} = \bar{q}_{feb} + b_{feb}(q_{jan} - \bar{q}_{jan}) + \varepsilon_{feb}$$

.....

Fig.6.4 shows a regression analysis of qj+1 on qj, pairs of successive monthly flows for the months (j+1) and j over the years of record where j = 1, 2, 3, ..., 12 (Jan, Feb, ... Dec) and when j = 12, j+1 = 1 = Jan (there would be 12 such regressions). If the regression coefficient of month j+1 on j is bj, then the regression line values of a monthly flow, $\hat{y}_{j+1}$, can be determined from the previous months flow qj, by the equation:

$$\hat{q}_{j+1} = \bar{q}_{j+1} + b_j(q_j - \bar{q}_j) \qquad (6.35)$$

To account for the variability in the plotted points about the regression line reflecting the variance of the measured data about the regression line, a further component is added:

$$Z.S_{j+1}\sqrt{(1-r_j^2)}$$

where is the standard deviation of the flows in month j+1, rj is the correlation coefficient between flows in months j+1 and j throughout the record, and Z =N(0, 1), a normally

distributed random deviate with zero mean and unit standard deviation. The general form may written as

$$\hat{q}_{j+1} = \bar{q}_{j+1} + b_j (q_j - \bar{q}_j) + Z_{j+1,i} . S_{j+1} \sqrt{(1 - r_j^2)} \qquad (6.36)$$

Where $b_j = r_j * S_j + 1 / S_j$, there are 36 parameters for the monthly model (q, for each month). The subscript j refers to month. For monthly synthesis j varies from 1 to 12 throughout the year. The subscript i is a serial designation from year 1 to year n. Other symbols are the same as mentioned earlier.

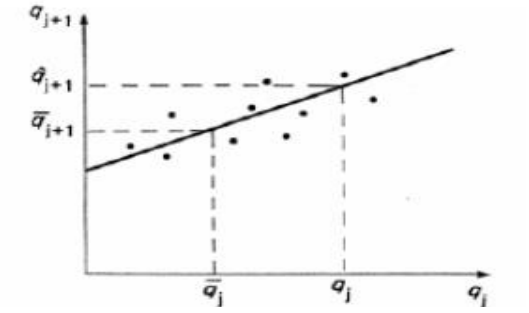


Fig.6.4: Thomas-Fiering model

The procedure for using the model is as follows:

(1) For each month, j = 1, 2, … 12, calculate

(a) The mean flow, $\bar{q}_{\beta} = \frac{1}{n} \sum q_{j,i}; \qquad (i = j, 12j, 24 + j, \ldots)$

(b) The standard deviation, $s_j = \sqrt{\frac{\sum (q_{j,i} - \bar{q}_j)^2}{n-1}}$

(c) The correlation coefficient with flow in the preceding month,

$$r_j = \frac{\sum (q_{j,i} - \bar{q}_j)(q_{j+1,i} - \bar{q}_{j+1})}{\sqrt{\sum_i (q_{j,i} - \bar{q}_j)^2 \sum_i (q_{j+1,i} - \bar{q}_{j+1})^2}}$$

(d) The slope of the regression equation relating the month's flow to flow in the preceding month:

$$b_j = r_j \frac{S_{j+1}}{S_j}$$

(2) The model is then set of twelve regression equations

$$\hat{q}_{j+1,i} = \bar{q}_{j+1} + b_j (q_{j,i-1} - \bar{q}_j) + Z_{j+1,i} \cdot s_{j+1} \sqrt{(1 - r_j^2)}$$

where Z is a random Normal deviate N(0, 1).

(3) To generate a synthetic flow sequence, calculate (generate) a random number sequence {Z_1, Z_2, … }, and substitute in the model.

6.4 The Uses of Stochastic Models

(1) To make predictions of frequencies of extreme events

Stochastic models have been used to make predictions about the frequency of occurrence of certain extreme events of interest to the hydrologist. Models such as that given by equation (6.29) are selected, and the residual is taken to be random variable with probability distribution whose parameters are specified. The parameters are estimated from data; so-called "synthetic" sequence {yt} can then be constructed, and the frequency with which the extreme event occurs in them can be taken as an estimate of the "true" frequency with which it would occur in the long run.

(2) For the investigation of system operating rules

A further use for synthetic sequences generated by stochastic models is in reservoir operation, such as the investigation of the suitability of proposed operating rules for the release of water from complex systems of interconnected reservoirs. By using the generated sequence as inputs to the reservoir system operated according to the proposed rules, the frequency with which demands fail to be met can be estimated. This may lead to revision of the proposed release rules; the modified rules may be tested by a similar procedure.

(3) To provide short-term forecasts

Stochastic models have been used to make forecasts. Given the values xt, xt-1, xt-2, ...; yt, yt-1, yt-2, ... assumed by the input and output variables up to time t, stochastic models have been constructed from this data for forecasting the output from the system at future times, t+1, t+2, ..., t+k, ... In statistical terminology, k is the lead-time of the forecast. Many stochastic models have a particular advantage for forecasting purposes in that they provide, as a byproduct of the procedure for estimating model parameters, confidence limits for forecasts (i.e. a pair of values, one less than the forecast and one greater, such that there is a given probability P that these values will bracket the observed value of the variable at time t+k). Confidence limits therefore express the uncertainty in forecasts; the wider apart the confidence limits, the less reliable the forecast. Furthermore, the greater the lead-time k for which forecasts is required, the greater will be the width of the confidence interval, since the distant future is more uncertain than the immediate.

(4) To "extend" records of short duration, by correlation

Stochastic models have been used to "extend" records of basin discharge where this record is short. For example, suppose that it is required to estimate the instantaneous peak discharge with a return period of T years (i.e. such that it would recur with frequency once in T years, in the long run). One approach to this problem is to examine the discharge record at the site for which the estimate is required, to abstract the maximum instantaneous discharge for each year of record , and to represent the distribution of annual maximum instantaneous discharge by a suitable probability density function. The abscissa, Yo, say, that is exceeded by a proportion 1/T of the distribution then estimates the T-year flood. It, however, frequently happens that the length of discharge record available is short, say ten years or fewer. On the other hand, a much longer record of discharge may be available for another gauging site, such that the peak discharges at the two sites are correlated. In certain circumstances, it is then permissible to represent the relation between the annual maximum discharges at the two sites by a regression equation and to use this fitted equation to estimate the annual maximum instantaneous discharges for the site with short record.

(5) To provide synthetic sequences of basin input

Suppose that the model has been developed for a system consisting of a basin with rainfall as input variable, streamflow as output variable. If a stochastic model were developed from which a synthetic sequence of rainfall could be generated having statistical properties resembling those of the historic rainfall sequence, the synthetic rainfall sequence could be used as input to the main model for transformation to the synthetic discharge sequence. The discharge so derived could then be examined for the frequency of extreme events. This approach to the study of the frequency of extreme discharge events is essentially an alternative to that described in paragraph (1) above. In the latter, a synthetic sequence is derived from a stochastic model of the discharge alone; in the former, a synthetic discharge sequence is derived by using a model to convert a synthetic sequence of rainfall into discharge.