

TRAINING COURSE
ON
SOFTWARE FOR GROUNDWATER
DATA MANAGEMENT

UNDER
WORLD BANK FUNDED HYDROLOGY PROJECT

LECTURE NOTES
ON

TIME SERIES ANALYSIS

BY

A K KESHARI
&
VIJAY KUMAR

ORGANISED BY

NATIONAL INSTITUTE OF HYDROLOGY
ROORKEE - 247 667
INDIA

TIME SERIES ANALYSIS

1. INTRODUCTION

A time series is a sequence of values arrayed in order of their occurrence which can be characterized by statistical properties. The set of observations in a time series is usually taken at specified times, usually at equal intervals. Mathematically the sequence of values can be represented by:

$$x(t) = x(t_1), x(t_2), x(t_3), \dots; \quad t_1 < t_2 < t_3 \dots \quad (1)$$

where x is any hydrologic variable, t is time, and $x(t_1), x(t_2), x(t_3), \dots$ are values of the variable x at the specified times, t_1, t_2, t_3, \dots . The specified interval may be hourly, daily, monthly, yearly, etc. The daily hydrograph is a graphical representation of time series of daily discharges. Other examples of hydrologic time series are the sequences of groundwater levels, hydraulic heads, pumping from and recharge to the aquifers; sequences of values of different water quality parameters or concentration of chemical constituents present in wells, rivers at various times; daily sequences rainfall, temperature and other climatic parameters; and annual sequences of floods, low flows and mean discharges.

A time series may be a function of time explicitly or a function of any single variable which takes the place of time. Examples of sequences ordered by distance rather than time are the width and roughness of a stream channel as a function of distance. Most of the hydrologic variables fall under the category of time series data. Data formats for such kind of data have a special structure. Time series data may be either continuous or discrete. The discretized data may be regular or irregular in time.

Much statistical theory is concerned with random samples of independent observations. The special feature of time series analysis is the fact that successive observations are usually not independent and that the analysis must take into account the time order of the observations. When successive observations are dependent, future values may be predicted from past observations. If a time series can be predicted exactly, it is said to be deterministic. But most time series are stochastic in nature, which indicates that the future is only partly determined by past values. For stochastic series, exact predictions are impossible and must be replaced by the idea that future values have a probability distribution which is conditioned by a knowledge of past values.

In reality, all hydrological processes are more or less stochastic. Mathematically speaking, a stochastic process is a family of random variables $X(t)$ which is a function of time (or other parameters) and whose variate x_t is running along in time t within a range T . If the chance of occurrence of the variables is taken into consideration, and the concept of

probability is introduced in formulating the model, the process and its models are described as stochastic or probabilistic. Fig. 1 illustrates the basic classification of hydrologic processes.

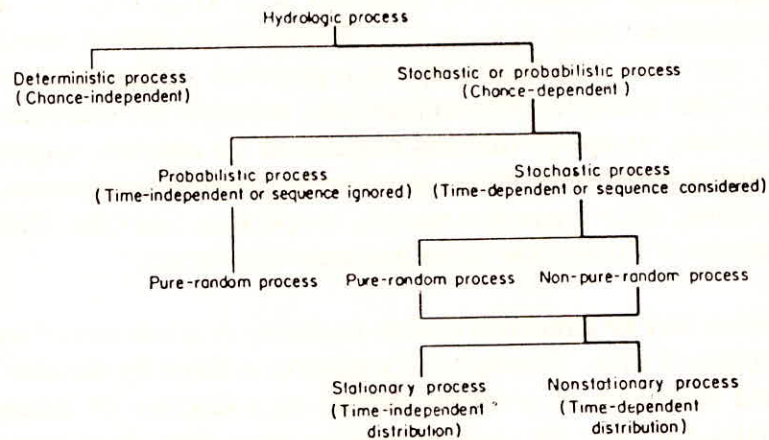


Fig.1 Classification of Hydrological Processes

2. CLASSIFICATION OF TIME SERIES

Time series data are classified in two categories as per data formats or measurement policy. They are: (1) continuous time series, and (2) discrete time series. A time series is said to be continuous when observations are made continuously in time. The term continuous is used for series of this type even when the measured variable can only take a discrete set of values. A time series is said to be discrete when observations are taken only at specific times, usually equally spaced. The term discrete is used for series of this type even when the measured variable is a continuous variable.

Discrete time series can arise in several ways. Given a continuous time series, we could read off the values at equal intervals of time to give a discrete series, such series may be called a sampled series. Another type of discrete series occurs when a variable does not have an instantaneous value but we can aggregate (or accumulate) the values over equal intervals of time. Examples of this type are rainfall measured daily, annual groundwater draft. Some time series may be inherently discrete.

Depending upon the method of analysis or statistical properties, a time series can be classified as either stationary or nonstationary. Broadly speaking, a time series is said to be stationary if there is no systematic change in mean (no trend), if there is no systematic change in variance, and if strictly periodic variations have been removed. Thus, a time series is said to be stationary when the general structure and statistical parameters representing the series, such as the mean, standard deviation, etc. do not change from one segment of the series to another; whereas a time series is said to be nonstationary if these statistical properties are changing from one segment of the series to another.

To visualize the concept of stationary and nonstationary time series more clearly, let us assume that a time series is divided into several segments and that a statistical parameter such as the mean is used to characterize the data within each section. If the expected value of the statistical parameter is the same for each section, the time series is said to be stationary. If the expected values are not the same, the time series is nonstationary. In stationary time series, absolute time is not important, and the series may be assumed to have started somewhere in the infinite past. However, in nonstationary time series, it is necessary to consider absolute time since the series cannot be assumed to have begun prior to the time of the initial observation.

Most of the probability theory of time series is concerned with stationary time series, and for this reason, time series analysis often requires one to turn a nonstationary series into a stationary one so as to use this theory. For example, one may remove the trend and seasonal variation from a set of data and then try to model the variation in the residuals by means of a stationary stochastic process.

Hydrologic time series can be divided in two basic groups: (1) single time series at a specified point, and (2) multiple time series at several points or multiple series of different kinds at one point. Single time series are also called univariate series while multiple time series are called multisite, multipoint or multivariate time series. In any case, they constitute sets of mutually related time series of individual points along a line, over an area, across a space, or as sets of time series of mutually related variables of various kinds.

3. COMPONENTS OF A TIME SERIES

A time series may be assumed to consist of four components, namely; trend, cyclic, seasonal and random. The trend and periodicity components are regarded as deterministic

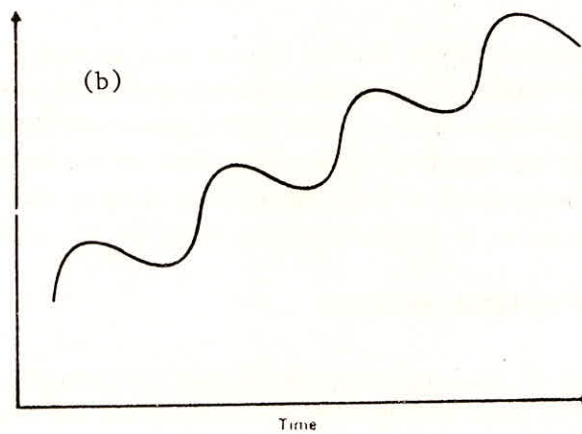
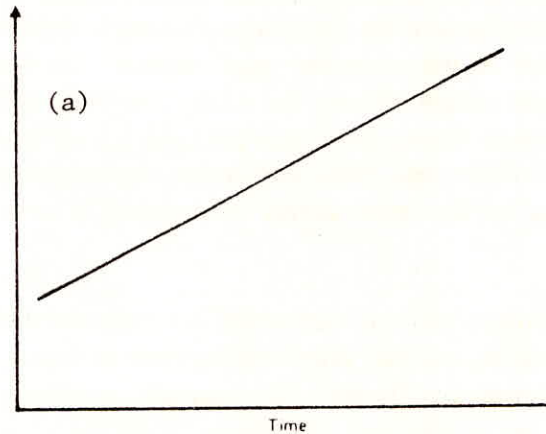
component, whereas the persistence and random components are regarded as stochastic component. Deterministic component is thought to be made of trend or long term movement, oscillations or fluctuations and periodic component or a component due to seasonality.

Mathematically, a time series may be represented by:

$$Y = T + C + S + I \tag{2}$$

where Y is an any time series, T is the trend, C is the cyclical movement, S is the seasonal movement and I is the random or irregular movement.

Fig. 2 (a) shows a long-term trend line, whilst (b) shows a cyclical variation imposed upon it. Irregular movements have been superimposed in Fig. 2 (c), and the resulting time series appears like a time series occurring in real life situation for most hydrologic variables.



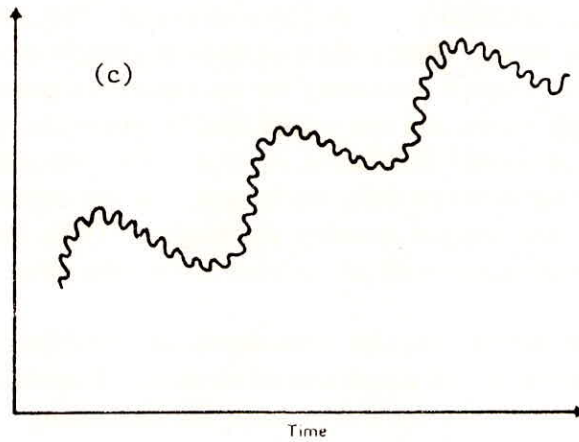


Fig. 2 (a) A long-term trend line, (b) a cyclic variation imposed on a trend line, (c) irregular movements imposed on a trend line

Fig. 3 shows a time series containing stochastic and several types of deterministic components.

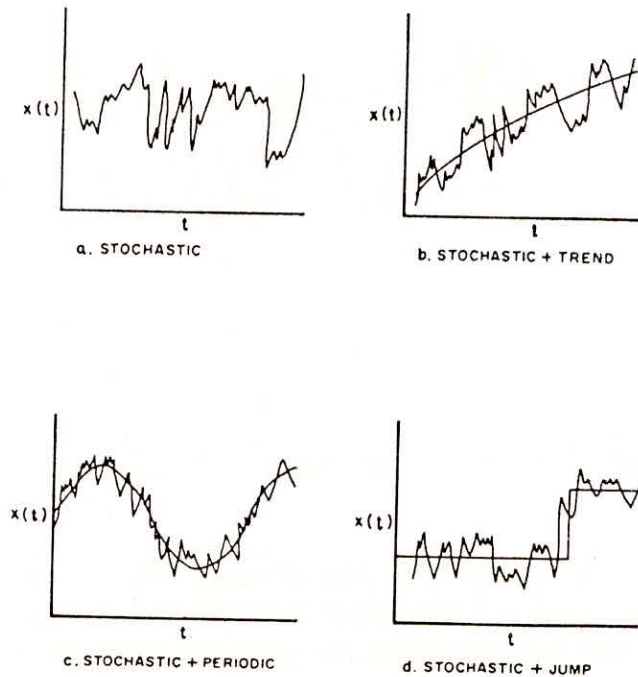


Fig. 3 A time series containing stochastic and several types of deterministic components.

4. CHARACTERISTICS OF TIME SERIES

Most of the statistical methods used in hydrologic studies are based on the assumption that the observations are independent of all previous events. This assumption is not always valid for hydrologic time series. Observations of daily discharges do not change appreciably from one day to the next. There is a tendency for the values to cluster, in the sense that high values tend to follow high values and low values tend to follow low values. Thus, the daily discharges are not independently distributed in time. The dependence between monthly discharges is less than that between daily discharges, and the dependence between annual discharges is less than that between monthly discharges. Thus, the dependence between hydrologic observations decreases with an increase in the time base.

Hydrologic time series may be considered as composed of the sum of two components: a random element and a nonrandom element. A nonrandom element is said to exist when observations separated by k time units are dependent. If the values of x_i are linearly dependent upon the values of x_{i+k} , then the correlation between x_i and x_{i+k} may be taken as the measure of dependence. This correlation is referred to as the k th-order serial correlation.

The serial correlation coefficient is analogous to the product-moment correlation coefficient for two sets of data. If x_i and x_{i+k} are considered as two sets of data then the k th order serial correlation coefficient is defined as:

$$r_k = \frac{\frac{1}{N-k} \sum_{i=1}^{N-k} x_i x_{i+k} - \frac{1}{(N-k)^2} \left(\sum_{i=1}^{N-k} x_i \right) \left(\sum_{i=1}^{N-k} x_{i+k} \right)}{\left[\frac{1}{N-k} \sum_{i=1}^{N-k} x_i^2 - \frac{1}{(N-k)^2} \left(\sum_{i=1}^{N-k} x_i \right)^2 \right]^{1/2} \left[\frac{1}{N-k} \sum_{i=1}^{N-k} x_{i+k}^2 - \frac{1}{(N-k)^2} \left(\sum_{i=1}^{N-k} x_{i+k} \right)^2 \right]^{1/2}} \quad (3)$$

where N is the length of the time series. For $k = 0$, it follows that $r_0 = 1$, and for $k \geq 1$, $-1 \leq r_k$.

If a time series is random, $r_k = 0$ for all values of $k \geq 1$. However, for a sample of finite size, computed values of r_k may differ from zero because of sampling errors. Since N is small for most hydrologic sequences, the sampling errors are very large, so that it is necessary to test the values of r_k to determine if they are significantly different from zero.

4.1 Time Series Representation

A time series is represented pictorially by drawing a graph, which is sometimes called a historigram. The time series representation may show up important features such as trend, seasonality, discontinuities and outliers. Plotting the data is not as easy as it sounds. The choice of scales, the size of the intercept, and the way the points are plotted (e.g. as a continuous line or as separate dots) may substantially affect the way the plot looks, and so the analyst must exercise care and judgment. Plotting the data may indicate if it is desirable to transform the values of the observed variable.

A time series is subject to certain variations or movements. The analysis of these movements assists in forecasting future movements. The characteristics movements of a time series may be classified as follows:

(a) Long-term or secular movements

These refer to the trend of the graph of a series over a long period of time. Long-term movement may be represented by a trend curve (Fig. 4a) or a trend line (Fig. 4b).

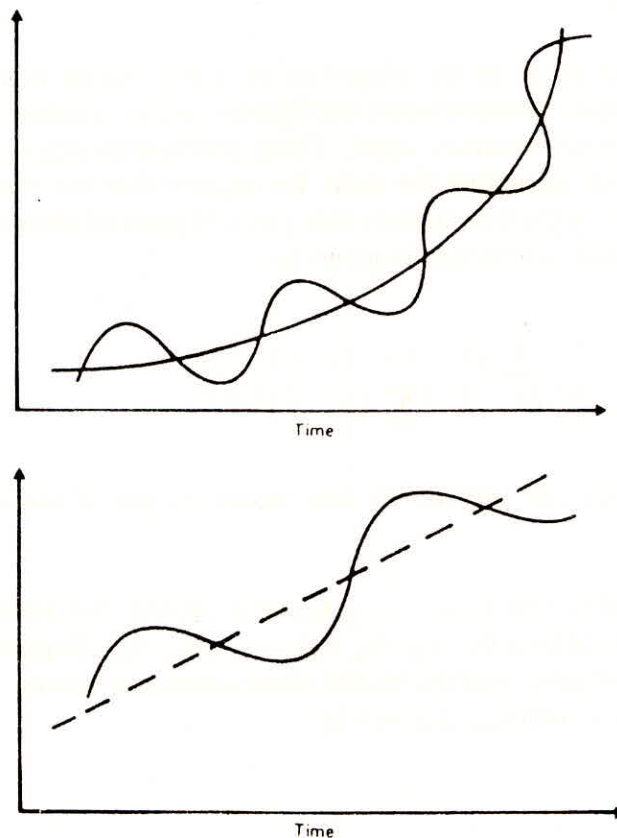


Fig. 4 (a) Trend curve with cyclic movement, (b) Long term trend and cyclic movement

(b) Cyclic movements

It refers to swings about the trend line or curve. These cycles are often periodic, that is they follow similar patterns after equal periods of time. In Figs. 4a and 4b, the cyclic movements are clearly shown.

(c) Seasonal movements

These occur because of events which take place each year. Seasonal movements refer to all periods of time, i.e. hourly, daily, weekly, monthly, annually, etc. depending upon the information available.

(d) Irregular movements

It occurs because of chance events. Basically, it is a random event and will be governed by the probability theories. Usually the events producing irregular events last only a short time, but sometimes they can produce new cyclic variations.

4.2 Autocorrelation

An important guide to the properties of a time series is provided by a series of quantities called sample autocorrelation coefficients, which measure the correlation between observations at different distances apart. These coefficients often provide insight into the probability model which generated the data. We assume that the reader is familiar with the ordinary correlation coefficient, namely that given N pairs of observations on two variables x and y , the correlation coefficient is given by

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{[\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2]^{1/2}} \quad (4)$$

A similar idea can be applied to time series to see if successive observations are correlated.

Given N observations x_1, x_2, \dots, x_N , on a discrete time series we can perform $(N-1)$ pairs of observations, namely $(x_1, x_2), (x_2, x_3), \dots, (x_{N-1}, x_N)$. Regarding the first observation in each pair as one variable, and the second observation as a second variable, the correlation coefficients between x_t and x_{t+1} is given by

$$r_1 = \frac{\sum_{t=1}^{N-1} (x_t - \bar{x}_{(1)}) (x_{t+1} - \bar{x}_{(2)})}{\left[\sum_{t=1}^{N-1} (x_t - \bar{x}_{(1)})^2 \sum_{t=1}^{N-1} (x_{t+1} - \bar{x}_{(2)})^2 \right]^{1/2}} \quad (5)$$

Where, $\bar{x}_{(1)}$ and $\bar{x}_{(2)}$ are the mean of the first (N-1) and last (N-1) observations respectively. But as $\bar{x}_{(1)} \approx \bar{x}_{(2)} \approx \bar{x}$, the above equation can be approximated as

$$r_1 = \frac{\sum_{t=1}^{N-1} (x_t - \bar{x}) (x_{t+1} - \bar{x})}{(N-1) \sum_{t=1}^N (x_t - \bar{x})^2 / N} \quad (6)$$

Some authors also drop the factor $N/(N-1)$ which is close to one for large N to give the even simpler formula

$$r_1 = \frac{\sum_{t=1}^{N-1} (x_t - \bar{x}) (x_{t+1} - \bar{x})}{\sum_{t=1}^N (x_t - \bar{x})^2} \quad (7)$$

As the above equation measures correlation between successive observation, it is called an autocorrelation coefficient. In a similar way we can find the correlation between observations a distance k apart, which is called the autocorrelation coefficient at lag k and is given as

$$r_k = \frac{\sum_{t=1}^{N-k} (x_t - \bar{x}) (x_{t+k} - \bar{x})}{\sum_{t=1}^N (x_t - \bar{x})^2} \quad (8)$$

4.3 Correlogram

A useful aid in interpreting a set of autocorrelation coefficients is a graph called a correlogram in which r_k is plotted against the lag k . Examples are given in Figs. 5-7. A visual inspection of the correlogram is often very helpful.

(a) A random series

If a time series is completely random, then for large N , $r_k \approx 0$ for all non-zero values of k .

(b) Short-term correlation

Stationary series often exhibit short-term correlation characterized by a fairly large value of r_1 followed by 2 or 3 more coefficients which, while significantly greater than zero, tend to get successively smaller. Values of r_k for longer lags tend to be approximately zero. An example of such a correlogram is shown in Fig. 5. A time series which gives rise to such a correlogram, is one for which an observation above the mean tends to be followed by one or more further observations above the mean, and similarly for observations below the mean. A model called an autoregressive model, may be appropriate for series of this type.

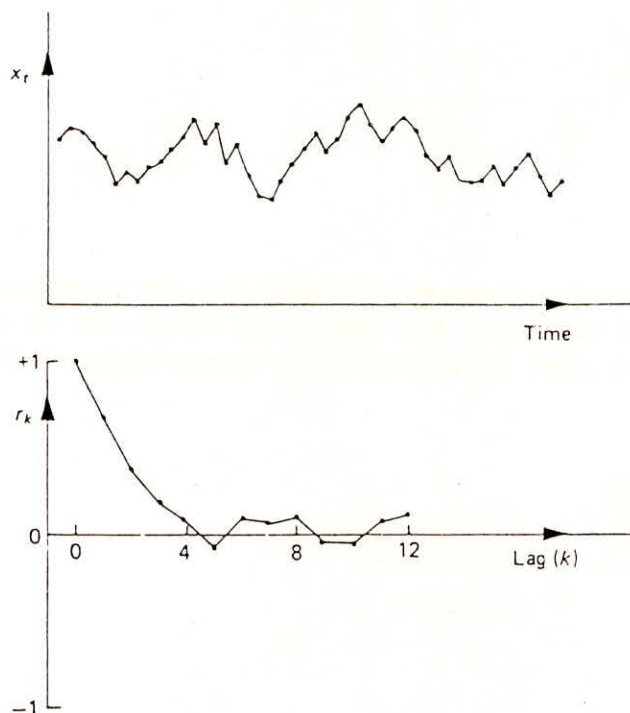


Fig. 5 A time series showing short term correlation together with its correlogram

(c) Alternating series

If a time series has a tendency to alternate, with successive observations on different sides of the overall mean, then the correlogram also tends to alternate. The value of r_1 will be negative. However, the value of r_2 will be positive as observations at lag 2 will tend to be on the same side of the mean. A typical alternating time series together with its correlogram is shown in Fig. 6.

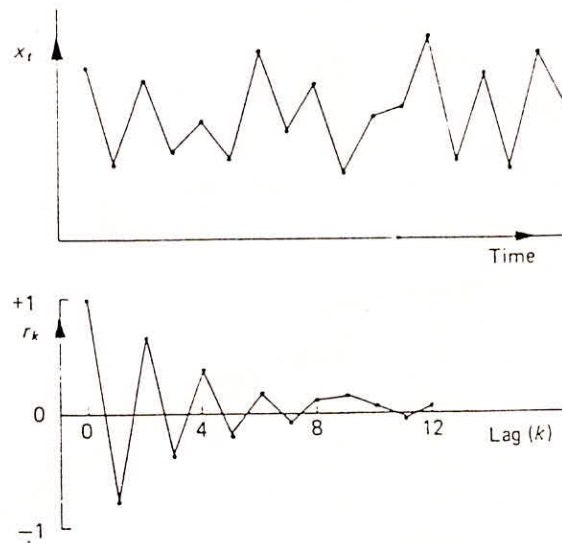


Fig. 6 An alternating time series together with its correlogram

(d) Non-stationary series

If a time series contains a trend, then the value of r_k will not come down to zero except for very large values of the lag. This is because an observation on one side of the overall mean tends to be followed by a large number of further observations on the same side of the mean because of the trend. A typical non-stationary time series together with its correlogram is shown in Fig. 7. Little can be inferred from a correlogram of this type as the trend dominates all other features.

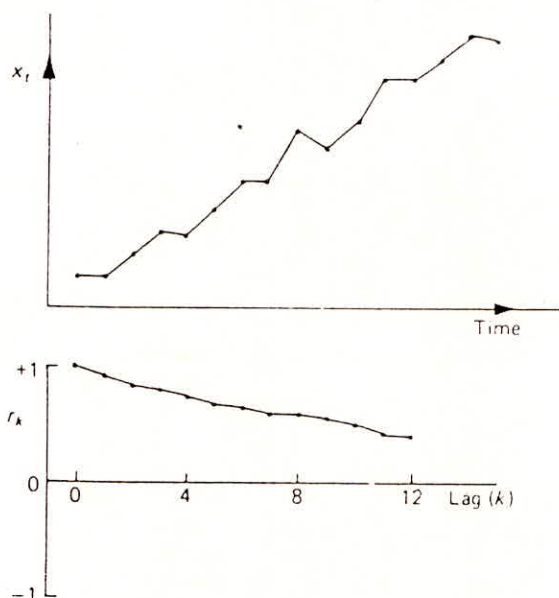


Fig. 7 A nonstationary time series together with its correlogram

(e) Seasonal fluctuations

If a time series contains a seasonal fluctuation, then the correlogram will also exhibit an oscillation at the same frequency. Fig. 8 shows the correlogram of the monthly air temperature data. The sinusoidal pattern of the correlogram is clearly evident, but for seasonal data of this type the correlogram provides little extra information as the seasonal pattern is clearly evident in the time plot of the data. If the seasonal variation is removed from seasonal data, then the correlogram may provide useful information.

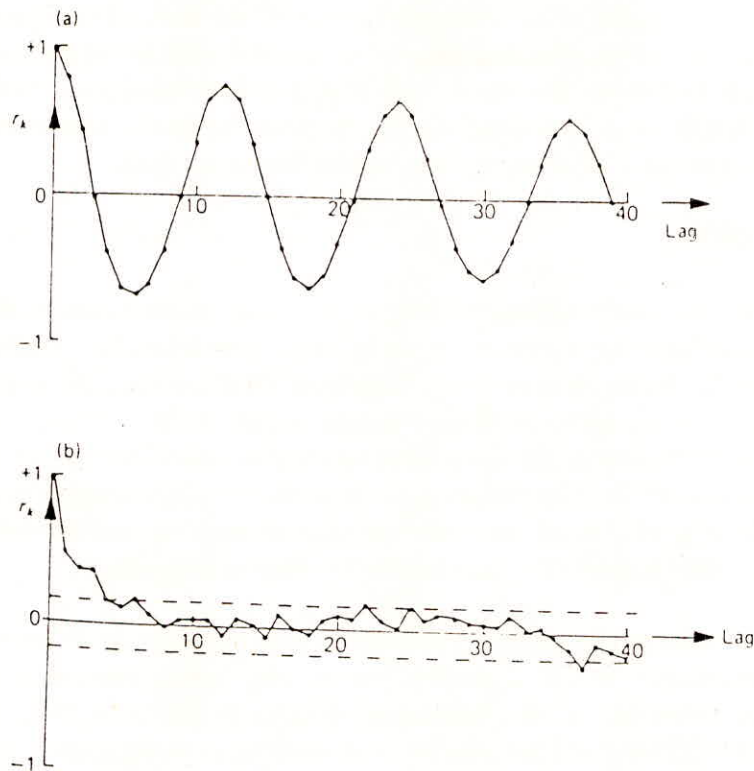


Fig. 8 The correlogram of monthly observation on air temperature for the
(a) raw data, (b) seasonality adjusted data

(f) Outliers

If a time series contains one or more outliers, the correlogram may be seriously affected and it is usually advisable to adjust outliers.

(g) General remarks

Clearly considerable experience is required in interpreting autocorrelation coefficients. In addition we need to study the probability theory of stationary series and discuss the classes of model which may be appropriate. We must also discuss the sampling properties of r_k .

4.4 Properties of the Nonrandom Element

The nonrandom element may be composed of both a trend, or a long-term movement, and an oscillation about the trend. Both of these parts need not be present in a particular time series. The first step in analyzing a time series is to separate the nonrandom element from the random element.

Trend is usually thought of as a smooth motion of the series over a long period of time. For any given time series, the sequence of values will follow an oscillatory pattern. If this pattern indicates a more or less steady rise or fall, it is defined as a trend. However, no matter what the length of a time series is, it can never be stated with certainty that an apparent trend is not part of a slow oscillation, unless the series ends.

5. TREND ANALYSIS

The trend may be loosely defined as long term change in the mean. A difficulty with this definition is deciding what is meant by long term. For example, climatic variables sometimes exhibit cyclic variation over a very long time-period such as 50 years. If one just had 20 years data, this long term oscillation would appear to be a trend, but if several hundred years data were available, the long-term oscillation would be visible. Nevertheless in the short term, it may still be more meaningful to think of such a long-term oscillation as a trend. Thus, in speaking of a trend, we must take into account the number of observations available, and should make a subjective assessment of what is long term.

Groundwater levels measured at a particular measuring point at different times form a time series of groundwater levels. The objective of time series analysis is to develop a comprehensive characterization of the underlying system in the form of a mathematical model. Time series modelling have been shown to provide systematic empirical methods for simulating and forecasting the behaviour of uncertain hydrological systems and for quantifying expected accuracy of the forecasts.

The most important and useful characteristic of each time series is the dependence or correlation between observations. Expressing the dependence by some mathematical model enables prediction of the future values of the system from the past values.

The following approach can be adopted for the analysis of a given set of observations:

- (i) Identify, quantify and remove the deterministic trend and periodic components.
- (ii) Choose an appropriate stationary stochastic process to describe the generating mechanism of the stochastic components.

The deterministic trend can be identified and can be removed by the regression analysis. Figs. 9-11 show graphically the trends in the groundwater levels. In Fig. 9, there is no trend in the series. A nonlinear decreasing trend is shown in Fig. 10, and a linear increasing trend is shown in Fig. 11.

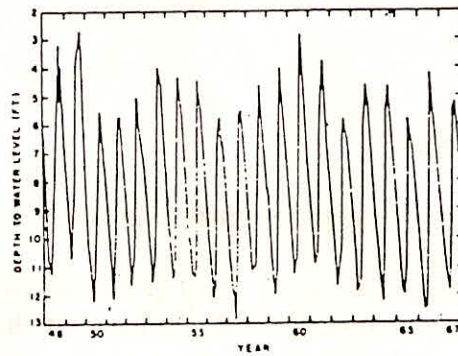


Fig. 9 Trendless Behaviour of Groundwater Depth

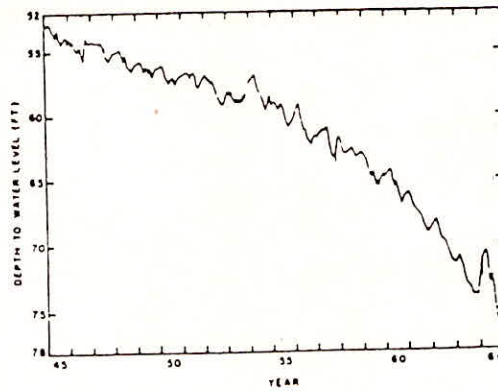


Fig.10 Nonlinear Decreasing Trend in Groundwater Depth

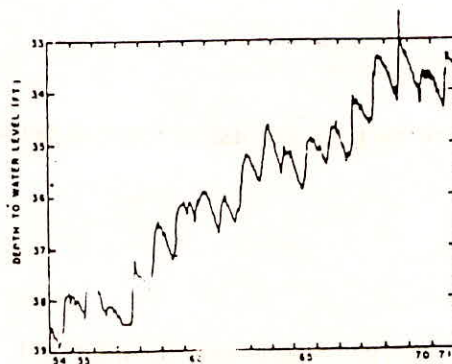


Fig.11 Linear Increasing Trend in Groundwater Depth

The trends present in the data may be in general approximated by the polynomial equation of the form

$$Z = a + b*t + c*t^2 + \dots \tag{9}$$

where

- Z = observed data
- t = time
- a, b, c = coefficients of the polynomial regression

In many cases, the linear trend i.e. $Z = a + b*t$ is sufficient; however, higher order terms may be necessary when the trend is far from linear. It should be noted that fitting a straight line or a curve to a set of observations does not imply that the data actually follow a straight line relationship. It indicates a general rise or fall of the water table levels with time.

The regression coefficients in the above equation are determined by the least square procedure. The polynomial equation representing the groundwater level at any particular well, can be represented by:

$$Z_i = a + b*t_i + c*t_i^2 + \dots \tag{10}$$

It can be written as:

$$\begin{aligned} Z_1 &= a + b*t_1 + c*t_1^2 + \dots \\ Z_2 &= a + b*t_2 + c*t_2^2 + \dots \\ Z_3 &= a + b*t_3 + c*t_3^2 + \dots \\ &\vdots \\ &\vdots \\ &\vdots \\ Z_n &= a + b*t_n + c*t_n^2 + \dots \end{aligned}$$

The above equation can be written in the matrix form as below:

$$[Z] = [A] [B] \tag{11}$$

where

[Z] = column matrix of observed water table of order n*1

[B] = column matrix of coefficients to be calculated $m \times 1$

[A] = square matrix of order $n \times m$

n = no. of observation points

m = no. of coefficients to be calculated

To find the value of coefficients a, b, c, \dots , the least square technique uses the criterion of minimization of square of error between the approximated and observed value i.e.

$$\text{Min. } e^2 = \sum (Z_i^* - Z_i)^2 \quad (12)$$

The minimisation of e^2 can be obtained by equating to zero the partial derivatives of the above equation with respect to the coefficients. The general solution of the above matrix is given by:

$$[B] = [A^T A]^{-1} [A^T Z] \quad (13)$$

In case of linear trend a and b are given as:

$$b = \frac{\sum Z_t - \frac{(\sum Z \sum t)}{n}}{\sum t^2 - \frac{(\sum t)^2}{n}} \quad (14)$$

$$a = \bar{Z} - b \bar{t} \quad (15)$$

The degree of polynomial, which best fits the data, can be decided on the basis of criterion of minimum standard error given by:

$$s = \sqrt{\frac{\sum (Z_i^* - Z_i)^2}{n - m}} \quad (16)$$

The significance of a regression can be tested by performing an analysis of variance. It gives the goodness of fit i.e. how well does the regression equation accounts for variations in the dependent variable. Numerically it is given by coefficient of determination (R^2).

$$R^2 = \text{RSS/TSS} \quad (17)$$

where

$$\begin{aligned} \text{RSS} &= \text{Regression (explained) sum of squared deviation} \\ &= \sum (Z_i^* - \bar{Z})^2 \end{aligned} \quad (18)$$

$$\begin{aligned} \text{TSS} &= \text{Total sum of squared deviation} \\ &= \sum (Z_i - \bar{Z})^2 \end{aligned} \quad (19)$$

R^2 indicates the explanatory power of the regression model. The possible values of the measure range from '+1' to '0'. When R^2 is near a value of +1, it shows a good fit of data.

Assuming each aquifer follows its present trend, the calculated regression lines can be used to predict future mean water table levels. Parametric and nonparametric methods are available for separating the periodic component present in the time series.

As we know, groundwater levels are variant in space and time. A long term annual time series records of groundwater levels helps in assessing the declining or rising trend of groundwater table and whether groundwater is being over exploited or not. A short term analysis of groundwater levels provides the seasonal variation of groundwater levels in a particular regime. By superimposing time series records of rainfall and groundwater levels, one can estimate how much recharge is taking place.

6. TIME SERIES MODELLING

Traditional methods of time-series analysis are mainly concerned with decomposing a series into a trend, a seasonal variation, and other irregular fluctuations. After trend and cyclic variations have been removed from a set of data, we are left with a series of residuals, which may or may not be random. We shall examine various techniques for analysing series of this type to see if some of the apparently irregular variation may be explained in terms of probability models, such as moving average or autoregressive models. Alternatively, we can see if any cyclic variation is still left in the residuals.

A mathematical model representing a stochastic process is called a stochastic model or a time series model. The techniques and procedures for finding a mathematical model which could represent a sample time series is called time series modelling. Time series modelling is a process which can be simple or complex, depending on the characteristics of the available sample series, the type of the model to use and the selected techniques of

modelling. For instance, series with statistical characteristics that do not vary with time usually lead to models and modelling techniques which are simpler than those of series with time varying characteristics. There are several types of model, varying from simple to sophisticated and complex, which can be used to represent a time series. The sophisticated analysis includes spectral and frequency domain analyses. Much of the simplicity or complexity of the modelling process ultimately depends on the modeler, such as modeler's theoretical knowledge and practical experience. There are several possible objectives in analysing a time series. These objectives may be classified as description, explanation, prediction and control, and will be considered in turn. In general, time series modelling has mainly two uses in hydrology and water resources:

- Generation of synthetic hydrologic time series
- Forecasting future hydrologic series

A systematic approach to hydrologic time series modelling may be composed of following six main phases (Salas and Smith, 1980):

- Identification of model composition
- Selection of model type
- Identification of model form
- Estimation of model parameters
- Testing goodness of fit of the model
- Evaluation of uncertainties

Fig. 12 shows the interaction between the above six modelling phases. In general, in any modeling of hydrologic time series, one has to decide whether the model will be a univariate or multivariate model, or a combination of a univariate and a disaggregation models, or a combination of a multivariate and a disaggregation models, etc. This decision is referred herein as the identification of the model composition. Such identification generally depends on the characteristics of the hydrologic time series and the modeler's input.

Several stochastic models are utilized for modelling hydrologic time series. They include AutoRegressive (AR), AutoRegressive Moving Average (ARMA), AutoRegressive Integrated Moving Average (ARIMA), Fractional Gaussian Noise (FGN), Broken Line (BL), Shifting Level (SL), ARMA-Markov, Disaggregation, etc. models. It has been observed that the judicious use of AR, ARMA, ARIMA and Disaggregation models will generally produce satisfactory results for most practical cases of operational hydrology.

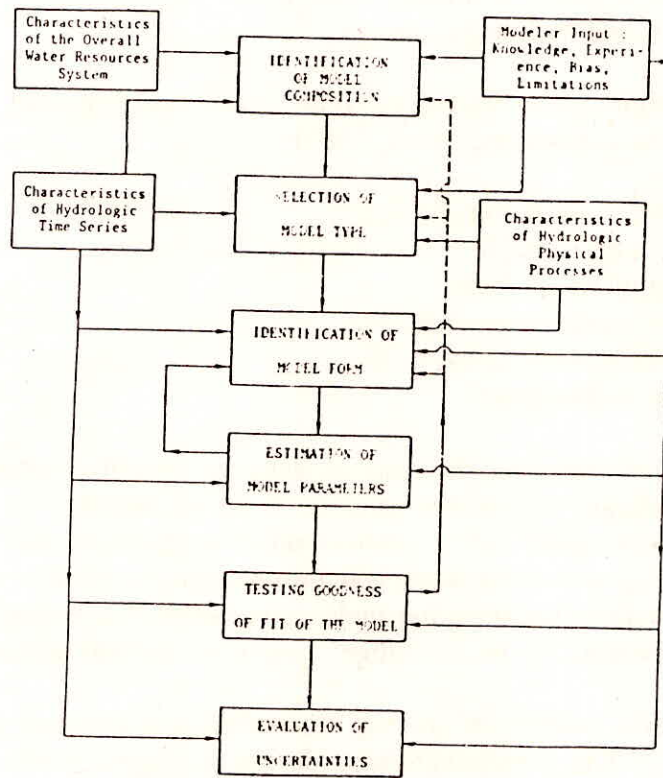


Fig. 12 Systematic approach of hydrologic time series modelling

BIBLIOGRAPHY

Chow, V.T., 1964. Handbook of Applied Hydrology. McGraw-Hill Book Company, New York.

Haan, C.T., 1994. Statistical Methods in Hydrology, The Iowa State Univ. Press, Ames.

Law, A.G., 1974. Stochastic Analysis of Groundwater Level Time Series in the Western United States. Hydrological Paper No. 68, Colorado State Univ., Colorado.

Salas, J.D., J.W. Delleur, V. Yevjevich and W.L. Lane, 1980. Applied Modeling of Hydrologic Time Series. Water Resources Publications, BookCrafters, Inc., Michigan, U.S.A.

Salas, J.D. and R.A. Smith, 1980. Uncertainties in hydrologic time series analysis. Paper presented at the vASCE Spring Meeting, Portland, Oregon, Preprint 80-158.
