PROPERTIES OF NON-HOMOGENEOUS HYDROLOGIC SERIES

by

V. Yevjevich and R. I. Jeng

April 1969

HYDROLOGY PAPERS
COLORADO STATE UNIVERSITY
Fort Collins, Colorado

32

Several departments at Colorado State University have substantial research and graduate programs oriented to hydrology. These Hydrology Papers are intended to communicate in a fast way the current results of this research to the specialists interested in these activities. The papers will supply most of the background research data and results. Shorter versions will usually be published in the appropriate scientific and professional journals, or presented at national or international scientific and professional meetings and published in the proceedings of these meetings.

The investigations leading to this paper are part of a research project supported by the U. S. National Science Foundation, Grant No. GK-1661. The research was conducted at Colorado State University, Civil Engineering Department, Fort Collins, Colorado.

PROPERTIES OF NON-HOMOGENEOUS HYDROLOGIC TIME SERIES

by

V. Yevjevich

and

R. I. Jeng

HYDROLOGY PAPERS
COLORADO STATE UNIVERSITY
FORT COLLINS, COLORADO 80521

April 1969 No. 32

TABLE OF CONTENTS

TABLE OF CONTENTS (continued)

LIST OF FIGURES AND TABLES

LIST OF FIGURES AND TABLES (continued)

ABSTRACT

The effects of inconsistency (systematic errors) and non-homogeneity of data (created either by man-made or natural changes in the environment) on hydrologic variables and time series are investigated. It is assumed that both the inconsistency and the non-homogeneity are in the form of constant and linear jumps, of linear and polynomial trends, and of their subsequent combinations.

The independent sequences and the first order Markov linear dependent sequences are used in this study. The known jumps and trends are superposed on the stationary series. Changes in the probability density functions, including mean, variance, skewness, excess, and serial correlation are analytically determined for various cases of jumps and trends assumed in advance as the known non-homogeneity and/or inconsistency.

Inconsistency and non-homogeneity introduce the dependence into the independent series, and increase the dependence of the first order Markov linear models. Usually, the first serial correlation coefficient becomes either positive and increased or only increased, respectively. Some forms of inconsistency and non-homogeneity may transform the one-peak probability density functions into two-peak or multi-peak density functions. Consequently, the statistical parameters of a series, with inconsistent and non-homogeneous data, become significantly different from those of the original series. As the hydrologic time series are often subject to inconsistency and non-homogeneity, a portion of the positive dependence and the higher variance in such a series comes from these two factors, apart from other basic physical processes in nature.

PROPERTIES OF NON-HOMOGENEOUS HYDROLOGIC SERIES*

by

V. Yevjevich** and R. I. Jeng***

Chapter I

INTRODUCTION

1. Definition of inconsistency and non-homogeneity. The inconsistency in data is defined in this study as being systematic errors in measurements and compilations. Existing systematic fluctuations make a difference between the figures produced by observation or computations and those produced by true values. Inconsistency is introduced by systematic errors in the series which change from time to time or from place to place.

If discharge is observed only once or twice daily at a cross section of a river which has daily periodic fluctuations, and the values are then estimated, the systematic errors in daily flows may be inevitable. When the measuring technique is changed to a recording instrument, and the daily flows are more accurately determined, the two periods combined in a time series have an inconsistency in the amount of systematic errors for the previous period. If a rain gauge is installed in an area with a small amount of vegetation and with no surrounding buildings, and if the vegetation matures in time and buildings are constructed, the catch of the gauge may change slowly or suddenly due to a change of aerodynamic patterns during the storm periods. A systematic difference exists between the observed values and the true rainfall values in the form of a trend or a jump. Many similar examples may be given for various sources of inconsistencies in hydrologic data.

The non-homogeneity in data is defined in this text as being the changes in a hydrologic series which result from a substantial transformation in environment which are either man-made or are natural. Differences between the virgin values (values produced in observations if the causative factors remain unchanged with time) and the true values are called in this text the non-homogeneity to distinguish them from the classical concept of non-stationarity, although non-homogeneity remains a part. In a series of monthly values of precipitation or runoff, the annual periodic component is part of the non-stationarity concept. However, if trends or jumps are added as described above, they increase the non-stationarity. To separate these two types, the first one is called non-stationarity and the latter, non-homogeneity.

If one begins a successful cloud seeding operation, the precipitation series experiences a jump (slippage) in its mean and also in its variance and other parameters. Changes of river basin factors which affect runoff-rainfall relationships introduce the non-homogeneity in data, usually in the form of jumps and trends. The natural changes such as forest fires and later forest growth, landslides into the river valleys creating temporary lakes, erosion or rock dissolution processes, biological cycles or replacements in species in nature, and other similar changes produce the non-homogeneity of various types. Man-made changes are becoming more and more important factors of non-homogeneity. The man-produced release of heat, the discharge of various gases (which change the natural composition of the air) and fine particles into the air by industrial and other activities, are assumed to be slowly affecting the temperature, precipitation, and evaporation. These releases are expected to be the causative factors of present and future very slow climatic changes. However, in order to be meaningful, the changes must be proven significant by statistical tests of data. The increased water consumption through additional evaporation and evapotranspiration which is caused by man's measures and structures in river basins such as irrigation, diversions, pond construction, man-made lakes and reservoirs, trans-mountain diversions, intentional change of vegetation cover, etc. is the main source of non-homogeneity in a hydrologic series. The mean discharge of the Colorado River at the Lee Ferry gauging station between the Upper and Lower River was about 15 percent smaller by 1960 than it was around 1900, mainly because of various man-made water depletion measures and structures.

*A small version of this study is published under the name of "Effects of Inconsistency and Non-homogeneity on Hydrologic Time Series," Proceedings of Fort Collins International Hydrology Symposium, September 1967, Vol. I, pages 451-458, (Colorado State University, Engineering Research Center, Fort Collins, Colorado), [6].

**Professor of Civil Engineering and Professor-in-Charge of Hydrology Program, Colorado State University, Fort Collins, Colorado.

***Former Ph.D. Graduate of Colorado State University, Civil Engineering Department, Fort Collins, Colorado, now Assistant Professor, California State College, Department of Civil Engineering, Los Angeles, California.

The schematic definition of inconsistency and non-homogeneity is:

Type of variable values	Effect on series
1. Observed values with systematic errors in part of the series } →	Differences: Inconsistency
2. True values in nature with changes in causative factors in the part of the series } →	Differences: Non-homogeneity
3. Virgin (time invariant) values	

The first concept is a clearly man-made difference between nature and the data. The second is related to induced changes in nature. Because non-homogeneity is a more important problem in water resource activities, both concepts are often encompassed under the term "non-homogeneity" in this text.

In the present study, inconsistency and non-homogeneity in the form of jumps and trends are treated. If the differences between the historical true values and the virgin values are constant, it is called constant jump. If a constant discharge is diverted continuously from one river to another, the first has a constant but negative jump in its mean and the second has the opposite, or a constant but positive jump. This is a classical case of a slippage problem in statistics and stochastic processes. If $\{Y(t)\}$ is the stochastic process or a time series of historical values, then $\{Y(t)\} = \{X(t)\} + \delta$ where $\{X(t)\}$ is the time series of virgin values, and δ is the constant jump which can be either a positive or a negative value. This constant jump must be introduced in a series between 0 and N (for a discrete series) or between 0 and T (for a continuous series); 0, N, and T being excluded in order that a series becomes non-homogeneous.

Another kind of jump is the linear jump. If the difference between the historical and the virgin values shows a linear relation, the quantity of change is also linearly dependent on the virgin value. Then $\{Y(t)\} = \{(1+I)X(t)\}$ represents the linear jump where "I" is a constant value greater or smaller than minus one, or minus one excluded. The example of this non-homogeneity occurs in the model a diversion from one river basin to another happens to be proportional to the flow of the first river at a given gauging station. Another example is the case of weather modification when the artificial attainments in precipitation or runoff are proportional to the natural precipitation or runoff.

The other type of inconsistency and non-homogeneity is the trend. When the difference between the historical and virgin values continuously change with time, it is called the trend. If these differences are a linear function of time, then the non-homogeneity and inconsistency are in the form of a linear trend. It can be represented as $\{Y(t)\} = \{X(t)\} + at + b$, where a and b are constant values which define the linear trend. It is called a polynomial trend if the differences follow a polynomial function of time, or $\{Y(t)\} = \{X(t)\} + a_o + a_1 t + a_2 t^2 + \ldots + a_m t^m$, where coefficients $\{a_i\}$ are constant values for a finite time series.

2. Practical significance and examples of inconsistency and non-homogeneity. The planning, design, operation, and maintenance of water resource developments require statistical information in the form of various hydrologic series. A thorough understanding of the structure of hydrologic time series is a prerequisite for any reliable input data in the planning and operation of water resouce projects. Apart from the stochastic variation of hydrologic quantities with time, diverse sources of inconsistency and non-homogeneity superpose their changes to the stationary stochastic and non-stationary deterministic variations (cyclic movements). Therefore, a hydrologic series observed for a sufficiently long time cannot generally be considered the sample from only one population. Inconsistency and non-homogeneity are often encountered in hydrology. Many hydrologic observations have a higher or lower degree of non-homogeneity or inconsistency. Therefore, the study of the effects of non-homogeneity and inconsistency of data on the properties of a hydrologic time series is a very important subject for practical application.

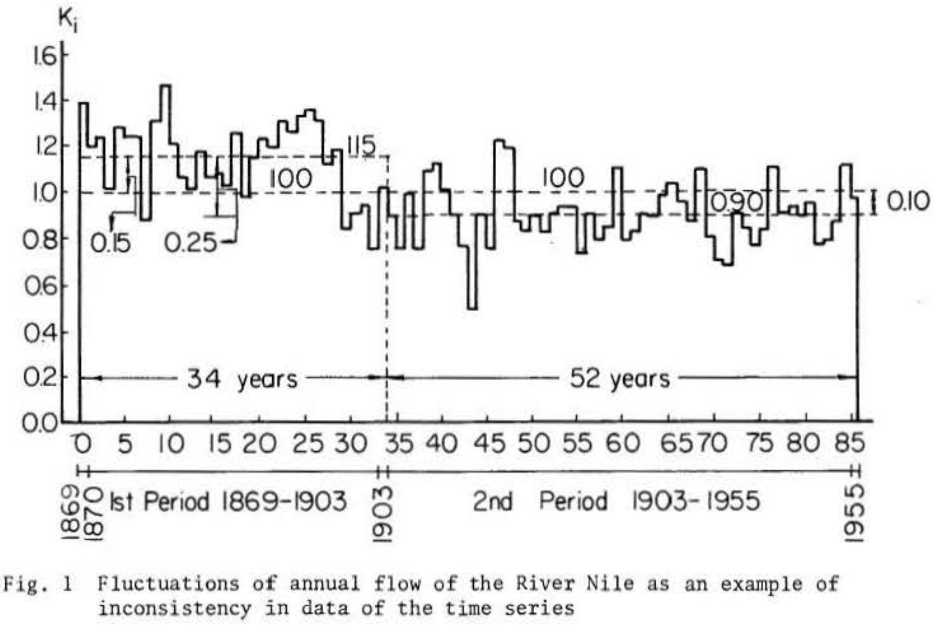

Fig. 1 Fluctuations of annual flow of the River Nile as an example of inconsistency in data of the time series

The essential practice in the water resource field is to use the statistical data of past observations, make an inference about the population of a hydrologic variable, and expect that the basic properties will hold true in future samples. However, if the past data show an inconsistency and non-homogeneity, the statistical inference about a unique population may not correspond to future samples. Future samples may not have non-homogeneity or they may experience another type. Two examples, the Nile River and the Colorado River, are discussed here for a better illustration of inconsistency and non-homogeneity in order to show both their importance and ways of treating them.

3. Two examples of inconsistency and non-homogeneity in data. The case of the River Nile at Aswan Dam, Fig. 1, is given here as an example of inconsistency [1]. Before the construction of Aswan Dam (1903), observations were made by using stage gauge downstream of the present dam-site. From 1903-1939, discharges were determined accurately enough by relating sluice measurements to the gauge-stages downstream. The subsequent rating curve was then applied in order to determine the discharges before 1903, from 1869 to 1902.

After the dam was put into operation, the downstream degradation through the removal of sediment islands and through bank erosion must have changed the rating curve which existed before the dam was built. The results [2] show that the mean discharge for 34 years before 1903 (1869-1902) was 3380 m^3/s, or 1.15 in modular coefficients, Fig. 1. The mean discharge for 52 years after 1903 (1903-1955) was 2650 m^3/s, or 0.90 in modular coefficients, Fig. 1. The four-year period prior to the operation of the reservoir from 1899-1902, shows a mean flow close to the mean of the second period, 1903-1955. One wonders if this was associated with the backwater regime above the dam due to its construction, and consequent deposition of coarse sediment upstream of the dam, or if the degradation of sediment, banks, and sediment islands downstream of the dam, was not really started in 1899 instead of 1903. Though the man-made reservoir increased the losses by evaporation and eventually by seepage, the difference in the mean flow of 730 m^3/s between the two periods cannot be chiefly explained in this manner. Four factors might have combined to show a 25% difference in means of the two periods:

(a) The natural stochastic variation may have been so that the 34-year period (1869-1902) was much wetter than the 52-year period (1903-1955). However, the difference of the two means of 0.25 in modular coefficients (25% of the mean for the 86-year period, 1869-1955) has only a 0.01% chance to be produced by a natural stochastic fluctuation. If the time-intervals are divided into 29-year (1869-1898) and 57-year periods (1899-1955), then this difference would be still greater, and the probability of its occurrence would be smaller than 0.01%.

(b) Inconsistency in data was brought about by the use of a rating curve which was produced after 1903 and applied to river gauging stages before 1903.

(c) Non-homogeneity in data was produced by the reservoir in the form of increased evaporation and percolation into the river banks, rocks and soils.

(d) Non-homogeneity in data was produced by an increase in upstream water consumption, by larger lake evaporation and by irrigations.

One might be tempted to assume a long-range persistence or periodicity in the annual flow of the Nile River. Due to the fact that a preponderant number of rivers in the world do not show long-range persistence or periodicity, the probability that it exists in the Nile River should be very small indeed. Therefore, one should postulate a hypothesis for this example. Specifically, the assumption is that the factors in the combination of stochastic variation, inconsistency in data, and non-homogeneity in annual river flows produce the graph of Fig. 1. The planners of the New Aswan Dam were, therefore, very wise in using the data of the period after 1903 for various water resource problems involving the reservoir and the Nile River water allocation, and particularly for their estimation of properties of inflows into the new reservoir.

Another instructive example is the non-homogeneity in the annual river flow of the Colorado River at Lee's Ferry Station, between the Upper and Lower river basins, as given in Fig. 2. Figure 2 shows the virgin and measured (historical) annual flows at this station. Until 1947, the information came from House Document No. 364, Washington, D. C., 1954 [3]: Colorado River Storage Project; and for years 1948-1959 the data were obtained from the U.S. Department of Interior, Bureau of Reclamation, Regional Office, Region 4, Salt Lake City, Utah. The following information and test of the example on non-homogeneity is mainly from a publication on the Colorado River Basin [4].

It might be that some inconsistency exists in data prior to 1914. In the reference [3], page 141, it is stated: "Although inaccuracies are risked with the extension of records prior to 1914, the Bureau of Reclamation made extensions to include the 1896-1947 period at Lee Ferry... ." In determining the depletion of the water yield, the same reference on page 143 states: "Stream depletions from upper basin development, therefore, have been estimated only at sites of use, and aggregate depletions so determined are considered representative of the depletion at Lee Ferry," and on the same page, "This includes depletions from all causes, such as irrigation and uses incident to irrigation, water exports to areas outside of the drainage basins, domestic and industrial uses, and evaporation from storage reservoirs. The estimate allows credits for water importations and channel salvage."

For the Lee's Ferry Station, Fig. 3 gives the relationships of three variables: (1) annual depletion, D, in 10^6 acre-feet; (2) annual virgin flow, V, in 10^6 acre-feet; and (3) time (as parameter). It is clearly shown that the depletion has fast increased from the turn of the century until the end of World War I, then stayed approximately constant for the period, 1920-1930, slowly increased from 1930-1950, then increased faster from 1954-1957. In this case, the historical annual flow at Lee's Ferry Station is an evolutive time series (and not a stationary time series). To make a series homogeneous (to compensate for depletions), the annual virgin flows which have been approximated give an insight as to what would be the flow if the hydrologic factors of Upper Colorado River Basin would remain unchanged by man's activities. Although the approximated depletions have errors (because they depend on many factors such as rough and approximate evaluations of net consumptive water uses and on net evaporation from the new water surfaces), and although the computed virgin flows are less accurate than in the case where they coincide with historical (measured) flows, they nevertheless show a measure of man-made non-honogeneity in the hydrologic records of the Colorado River Basin.

The following model is defined as the relation of annual depletion to annual virgin flow for a given interval of years, or for a single year. They are approximated by straight lines (Fig. 3) because a

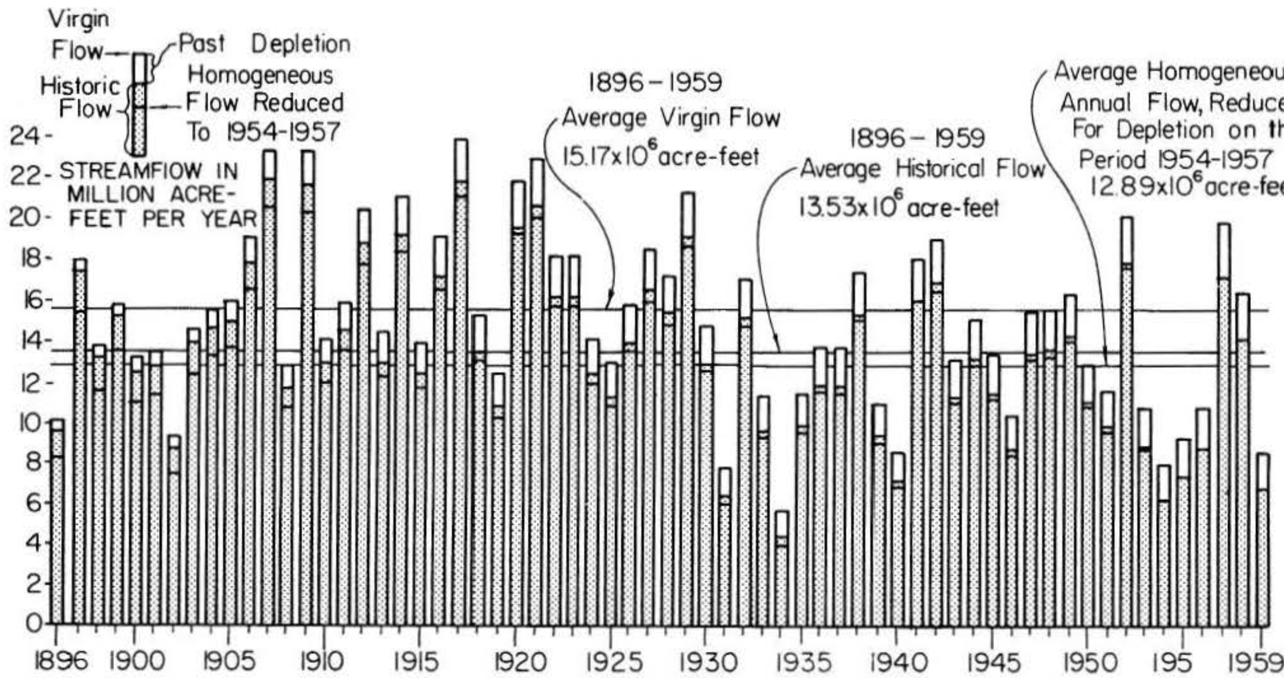

Fig. 2 The Colorado River annual flows at Lee Ferry, Arizona

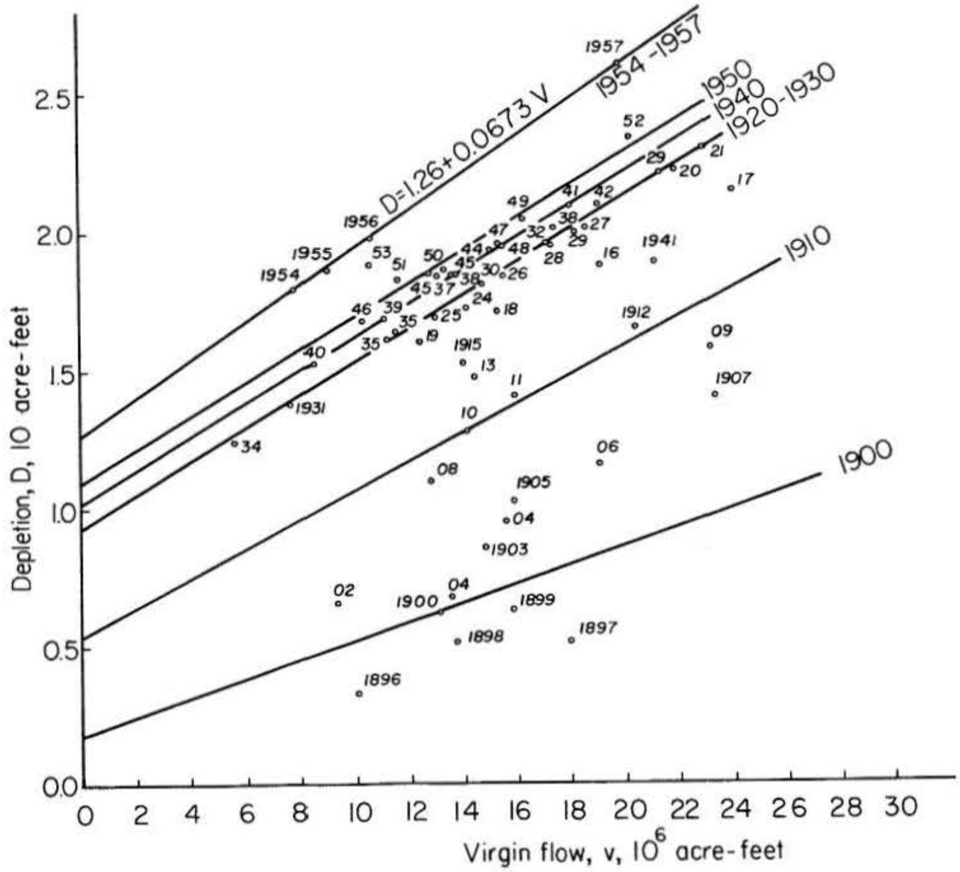

Fig. 3 The relation of the man-made annual flow depletion in the Upper Colorado River at Lee's Ferry Station, Arizona, to the annual virgin flow (historical flow plus depletion) for given time intervals or years, for the period 1896-1957.

complex model would not be justified in view of the errors which are inherent in the determination of depletions. For the period 1954-1957, the depletion model is

$$D = 1.26 + 0.0673\ V \tag{1}$$

Figure 3 shows that both coefficients A and B in equation $D = A + B\ V$ increase with time. This increase of A and B means a greater average depletion per year with time. An increase of B means that the depletion fluctuates more in function of the absolute value of virgin flow with time than in earlier depletions. The increase of B also means that the number of factors which affect the depletion, but are proportional to the virgin flow, increase with time (when more water is diverted to irrigation inside the basin, or when more water is diverted out of the basin during wet years, or when more evaporation occurs from reservoirs because of a greater mean free surface area in wet years than in dry years, and similar factors).

The historical annual flow of the Lee's Ferry Station is considered to be a non-homogeneous (non-stationary) time series. In statistical words, the information from 64 years of annual flows at Lee's Ferry Station is derived from a mixture of populations. Theoretically speaking, that mixture of populations and samples may be treated statistically or probabilistically if the law of change in time from one population to the next is known. Due to the fact that most of the changes are introduced by man's activity, and that the laws of change in time for runoff are complex and unpredictable, the approach of treating the mixture of populations in annual runoff is not feasible. This is the reason why the techniques of changing the non-homogeneous samples into homogeneous samples are introduced and practiced currently. It would be extremely difficult to project the depletion of annual runoff in amount and in time at the Lee's Ferry Station for the next three to five decades with sufficient accuracy. This means that all future storage reservoirs, diversion projects, and irrigation schemes would have to be predicted exactly in runoff amount and in time for the next 30-50 years. If that very approximate depletion projection would be acceptable for economic and engineering studies, it would be possible, at least theoretically, to treat future projected samples of runoff starting with the mixture of populations.

The computation of virgin flows for Lee's Ferry Station from 1896 to 1959 is also a procedure to determine a homogeneous sample. In other words, the computed virgin flow sample is drawn from the population that existed prior to any depletion, under the circumstances of a large number of natural causes which affect the runoff. Practical problems require however, that computations in engineering and economics be carried out with homogeneous samples, or one population is inferred from these samples (valid for the moment of computation, or for the time interval a project would normally serve). For the period of 1954-1957, applying the depletion model of eq. (1) to the virgin flows of the sample, 1896-1959, a new homogeneous sample valid only for the period 1954-1957 is obtained. Assuming that small changes in depletion have taken place from 1958-1960 in comparison with the period from 1954-1957, the new homogeneous sample can be considered the same as if drawn from the population of annual runoffs at Lee's Ferry Station, valid for the late fifties. The homogeneous sample reduced to the period, 1954-1957, is also given in Fig. 2.

By extrapolating the depletion model $D = A + B\ V$ (or any other more complex model) in the future by computing A and B as functions of time, it is possible to reduce the virgin flow sample, 1896-1959, to any future date. While planning the Upper Colorado River Basin development, it should be possible to project the depletion model, if not in function of time, than at least in function of future projects, and even in function of population growth. In this case, the new variate V_t, the annual flows at Lee's Ferry Station for a given date is

$$V_t = V - D_t = (1-B_t)\ V - A_t \tag{2}$$

where A_t and B_t are parameters of depletion model $D_t = A_t + B_t V$, at the date t. Assuming that $A_t(t)$ and $B_t(t)$ are given, then $V_t(V,t)$ would also be given. With the probability distribution of V, as well as the characteristics of sequence patterns of V given in analytical form, both probability distribution and sequence model for V_t can be derived as a function of t. This approach enables the computations of average hydrologic characteristics during a depreciation time for a water resource development project.

The above analysis leads to the conclusion that the computation of effects of man-made structures and measures in river basins has an important bearing on the reliability of hydrologic data used for further water resource developments and water project operations. The studies and calculations aimed to make hydrologic samples homogeneous (and also consistent by removing the eventual inconsistency in data) through computation and analysis of depletion models (or systematic errors), is a new and important task of hydrologic activities.

4. Subject. Changes introduced into a stationary time series by inconsistency and non-homogeneity of data in the form of jumps (changes suddenly introduced inside a series) and trends are the subject of this paper. The various models of jumps and trends may be superposed in a variety of ways. However, the usual case in many applied sciences is to search for changes which are introduced into the series as unknown jumps and trends. In this study, the approach is opposite. The second order stationary (time-invariant) series with known properties is subjected to known changes (jumps and trends) and the impact of these changes is investigated. The main attempt of this study is to determine the changes in the properties of various parameters and in the distributions which occur as a result of given types of jumps and trends so that a comparison with the observed series of unknown jumps and trends may be made. Specifically, the changes in density functions, in mean, variance, skewness, kurtosis, and serial correlation coefficients are studied. The objective is to find the statistical properties of the series when the known inconsistency and non-homogeneity are introduced, and to show how these factors affect the original homogeneous series.

5. Hypotheses for investigations. The general structure of a hydrologic time series usually has three main parts: trend and/or jump and periodic as well as stochastic components. For the purpose of this study, it is assumed that the periodic component is absent, or the cyclic component can be detected and removed from the original time series. The former case is well approximated by an annual runoff or annual precipitation series, or a series of similar variables of annual values. The procedure for detecting and

substracting periodic components from the time series are beyond the scope of the present study. Therefore, the general model of a hydrologic time series is given by

$$X_t = R_t + \varepsilon_t \tag{3}$$

with R_t the jump or trend component, and ε_t the stochastic (dependent or independent) component, and X_t is the resulting series. For the sake of simplicity, the ε_t-component in this study is assumed to be a sequence of mutually independent random variables of a second order stationarity. The trend and/or jump component is assumed to be a known mathematical function of time.

6. Research program. The independent stochastic time series is used for the investigation, and the inconsistency and non-homogeneity of various types are introduced into the series to produce the non-homogeneous series. In terms of stochastic processes, the virgin (stationary) values of an independent hydrologic series are assumed to be mutually independent and stationary random variables for the following investigations. In the same terms, these variables are identically distributed along a time series with the known probability density function.

Two independent stationary time series, one with the normal and the other with the lognormal probability density function are used. It is not considered very important to use the dependent stationary series for further investigation because it is expected that it will show results similar to the effect of inconsistency and non-homogeneity as in the case of an independent stationary series.

The types of non-homogeneity introduced into an independent stationary time series in this study are:

(a) Constant jump for a part of a series,

(b) Combination of constant jumps along a series,

(c) Linear jump for a part of a series,

(d) Combination of linear jumps along a series,

(e) Linear trend, and

(f) Polynomial trend.

The effects of non-homogeneity are studied on these properties of variables:

(a) Probability density functions,

(b) Expected values (means),

(c) Variances,

(d) Skewness coefficients,

(e) Kurtosis coefficients, and

(f) Serial correlation coefficients.

The first case of the constant jump in the series introduced at any place between its beginning and its end is given in detail in Chapter II in order to better illustrate the method of investigation and consequent results. The other cases of non-homogeneity in other chapters are presented as final results to strengthen the conclusions of Chapter II and to show the differences in effects of various types of non-homogeneity.

The majority of non-homogeneity and inconsistency types in hydrology are in the form of jumps and trends and their various combinations. The cases studied provide a sufficient general picture of the effects of non-homogeneity on the properties of a hydrologic time series.

EFFECT OF A CONSTANT JUMP

1. Definition of the change introduced by a constant jump. A constant change, δ, is introduced into a series of size N at the position m from the beginning of the series and n from the end of the series, with $N = m + n$, so that

$$x_t = \varepsilon_t, \text{ for } t \leq m$$
$$x_t = \varepsilon_t + \delta, \text{ for } t > m \tag{4}$$

where x_t is the historical value of a non-homogeneous hydrologic time series, ε_t is the virgin value of an independent stationary series, and δ is the constant jump being positive or negative throughout the last part, n, of the series. This is graphically represented in Fig. 4.

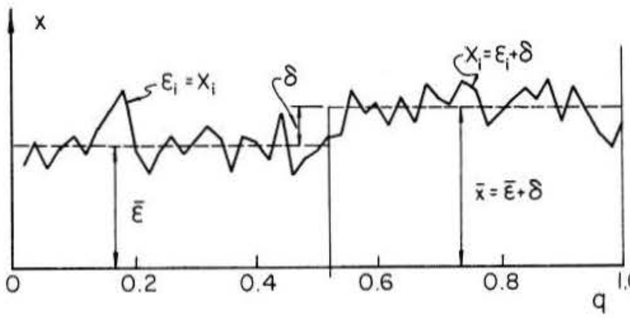

Fig. 4 A scheme of the constant jump introduced into an independent homogeneous series, ε_i

This case may simulate a constant annual diversion, δ, of water from one river basin to another, with a change δ or $-\delta$. It is clear that x_t values in the second part are biased in comparison with the first part or the converse. The position, m, corresponds to the time when the annual diversion begins. The ratios $p = m/N$ and $q = n/N$, with $p+q = 1$ are used in this text as the dimensionless measures of the relative position of the constant jump in the series. For the non-homogeneity to be present in a series, it must be shown that $0 < p < 1$ and $0 < q < 1$. In other words, the constant jump should not be at the beginning or the end of the series. In order to measure the importance of the constant jump on the series, the relative values $\delta/\bar{x}$ or δ/σ_x may be used where $\bar{x}$ and σ_x are the mean and the standard deviation of the series of a variable x.

2. Effect on a probability density function. It is assumed that ε_t is an independent and stationary random series which has $g(\varepsilon)$, a given probability density function. Then in the part of the series $x_t = \varepsilon_t$, and in the part of the series, $x_t = \varepsilon_t + \delta$, both have $g(\cdot)$ as their probability density functions. Furthermore, the density function of x_t is determined by giving the weights m and n to the densities of ε_t and $\varepsilon_t+\delta$, respectively, or

$$f(x) = p\ g(x) + q\ g(x-\delta)\ . \tag{5}$$

The probability density function, $g(\varepsilon)$, usually is assumed to be a known function. Normal and log-normal probability density functions are studied for this constant jump. For ε_t, an independent normal function with the mean of $\bar{\varepsilon}$ and the variance of σ_ε^2, the probability density function of non-homogeneous series x_t is

$$f(x) = \frac{p}{\sigma_\varepsilon\sqrt{2\pi}}\, e^{-(x-\bar{\varepsilon})^2/2\sigma_\varepsilon^2} + \frac{q}{\sigma_\varepsilon\sqrt{2\pi}}\, e^{-(x-\delta-\bar{\varepsilon})^2/2\sigma_\varepsilon^2} \tag{6}$$

because the constant jump does not change the variance σ_ε^2 in each part.

For ε_t, an independent log-normal function, a similar equation is obtained with some modification as

$$\left.\begin{aligned} f(x) &= \frac{p}{\sigma_n x\sqrt{2\pi}}\, e^{-(\ln x-\bar{\varepsilon}_n)^2/2\sigma_n^2} + \frac{q}{\sigma_n (x-\delta)\sqrt{2\pi}}\, e^{-[\ln(x-\delta)-\bar{\varepsilon}_n]^2/2\sigma_n^2}, \\ &\text{for } x > \delta, \text{ and} \\ f(x) &= \frac{p}{\sigma_n x\sqrt{2\pi}}\, e^{-(\ln x-\bar{\varepsilon}_n)^2/2\sigma_n^2}, \\ &\text{for } 0 \leq x \leq \delta \end{aligned}\right\} \tag{7}$$

where $\bar{\varepsilon}_n$ is the mean of $\ln \varepsilon$, and σ_n^2 is the variance of $\ln \varepsilon$. As the value x_t can be less than δ, but only in part m of the series, $f(x)$ must be separately represented for values above and below δ.

It is characteristic that eq. (6) gives either a two-peak or one-peak new probability density function. If the hydrologic variables are standardized, with $\bar{\varepsilon} = 0$, $\sigma_\varepsilon^2 = 1$, or $N(0,1)$, and a positive constant jump δ is introduced in the second part of the series, then eq. (6) becomes:

$$f(x) = \frac{p}{\sqrt{2\pi}}\, e^{-x^2/2} + \frac{q}{\sqrt{2\pi}}\, e^{-(x-\delta)^2/2}\ . \tag{8}$$

The first derivation of $f(x)$ with respect to x is

$$f'(x) = \frac{1}{\sqrt{2\pi}}\, e^{-x^2/2}\left[p(-x) + q\, e^{\delta(x-\delta/2)}(\delta-x)\right]. \tag{9}$$

Position x, for the maximum and the minimun of $f(x)$, are roots of $f'(x) = 0$, or

$$q\, e^{\delta(x-\delta/2)}(\delta-x) - p\, x = 0 \qquad (10)$$

because $\exp(-x^2/2)/\sqrt{2\pi}$ is a non-zero quantity. The solution of eq. (10) is found on the digital computer by using the trial-and-error method. The results show that the number of roots in eq. (10) is either one or three, depending on the values q and δ. Whenever there is only one solution in eq. (10), the probability density function has one maximum (peak). When there are three roots, there are two maxima (peaks) and one minimum in the probability density function.

It is clear that the probability density function of a non-homogeneous series f(x) is composed of two normal density functions with two different means, zero and δ, respectively, but each with the same variance of unity. Intuitively, one can expect that there are two peaks in the probability density function of a non-homogeneous series f(x), one around zero and the other around δ, if δ is large enough and q is close to 0.5. The exact positions of peaks for a given δ, depend entirely on the value of q. Obviously, the value of q indicates which of the two normal density functions has more weight. Small values of q mean that the standard normal density function has more weight, whereas large values of q mean that the normal density function with the mean of δ has a greater weight. The closer the value q is to 0.5, and the greater value of δ, the more distinguished are the two peaks in the density function f(x). In the opposite case, the new probability density function is a one-peak distribution.

3. Criterion for one-peak or two-peak distributions. The criterion for having one or two peaks in the probability density function of non-homogeneous series f(x) is obtained from the results of computations on the digital computer. Equation (10) is solved for various values of q and δ. The results tell at what value δ the function f(x) begins to be two-peak distribution for a given value of q. For values of q far away from 0.5, δ must be very large to sufficiently separate the two normal probability density functions in order to obtain the two peaks. When q is close to 0.5, the two normal probability density functions with means zero and δ, respectively, are almost equally weighted. In that case, even a comparatively small value of δ will be sufficient to create the two-peak probability density curves.

Figure 5 shows the regions (1) and (2) necessary for distributions to have one and two peaks in terms of critical values q_o and δ_o, respectively. The line separating the two regions is obtained by connecting the 12 computed points of (q_o, δ_o) of eq. (10). This division will be discussed in details further in the text. Figure 5 will be replotted on different coordinate scales and an empirical equation of $q_o = \psi(\delta_o)$ will be developed. For $q_o = 0$, or $q_o = 1$, a two-peak probability density function will never occur for any value of δ_o. For $q_o = 0.5$, the critical values δ_o can be obtained analytically.

For q = 0.5, eq. (10) becomes

$$e^{\delta(x-\delta/2)}(\delta-x) - x = 0 \; . \qquad (11)$$

By inspection, $x = \delta/2$ is one of the roots of eq. (11). Therefore, f(x) will be either the maximum or the minimum at $x = \delta/2$. In order to detect whether

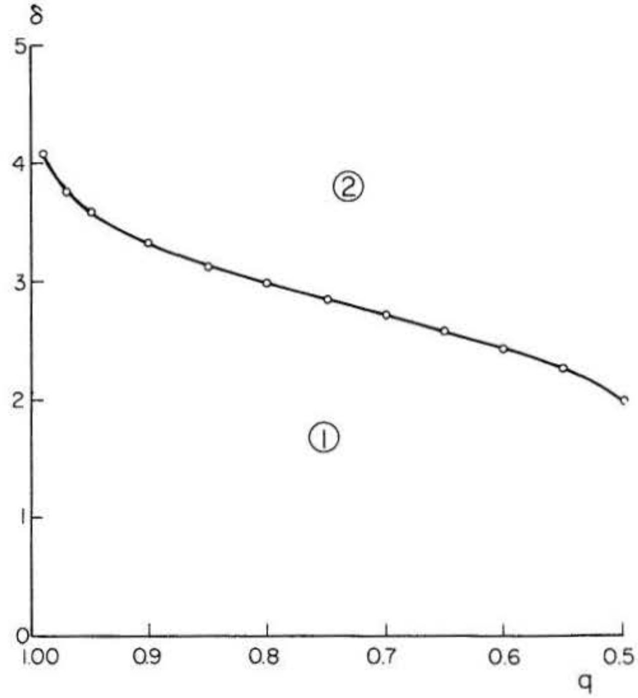

Fig. 5 The two regions (1) and (2) indicating the values of q = n/N and δ (constant jump) and designating if the probability density function of a non-homogeneous series, f(x), has one or two peaks: (1) One-peak region, (2) two-peak region

it is the maximum or minimum of f(x), the second deviation of f(x) is

$$\frac{d^2}{dx^2} f(x) = \frac{1}{2\sqrt{2\pi}} e^{-x^2/2}\left[-1-e^{\delta(x-\delta/2)} - (x-\delta)\delta e^{\delta(x-\delta/2)} + x^2 + x(x-\delta)e^{\delta(x-\delta/2)}\right], \qquad (12)$$

and

$$\left.\frac{d^2}{dx^2} f(x)\right|_{x=\delta/2} = \frac{1}{2\sqrt{2\pi}} e^{-\delta^2/8}\left(\frac{\delta^2}{2} - 2\right) . \qquad (13)$$

It is obvious that for $\delta = 2$, the second derivative of f(x) is equal to zero at $x = \delta/2$. For $|\delta| > 2$, the second derivatives are positive. For $|\delta| < 2$, the second derivatives are negative. This analysis tells that the probability density function of a non-homogeneous series for q = 0.5 has both its maximum and its points of inflection at x = 1 for $\delta = 2$. Therefore, it has only one peak which is very flat. For $|\delta| > 2$, the probability density function has its minimum at $\delta/2$, which implies that it has two peaks. For $|\delta| < 2$, the probability density function of a non-homogeneous series has its maximum at $\delta/2$ which implies that it has only one peak. Therefore, the critical value of δ necessary for the probability density function to have one or two peaks is $\delta_o = 2$ for the case $q_o = 0.5$.

Both the probability density at x for the maximum and for $|\delta| \leq 2$, and the probability density at x for the minimum and for $|\delta| > 2$, in the case q = 0.5, are equal to

$$\frac{1}{\sqrt{2\pi}} e^{-\delta^2/8} . \qquad (14)$$

It is obvious for q = 0.5 that the normal curve of eq. (14) is the locus of the above maxima or minima of the probability density of a non-homogeneous series for various values of δ. For q = 0.5 and the four values of δ ($\delta = 0.0$, $\delta = 1.00$, $\delta = 2.0$ and $\delta = 3.0$), the four probability density functions are given in Fig. 6.

For the values of q different from 0.5, the solutions of eq. (10) for x cannot be obtained in the explicit form. As was discussed earlier in this text, the trial-and-error method is used to find the solutions. The critical value of δ_o for a given q_o is the maximum value of δ giving only one solution in eq. (10). Eleven values of δ_o and the corresponding symmetrical values of q_o are listed in Table 1. They are plotted in Fig. 7 in logarithmic-probability scales in such a way that $q_o = 0.5$ corresponds to percentage zero, $q_o = 0$ corresponds to percentage 100, and δ_o is plotted on the logarithmic scale. Then $q_o = \psi(\delta_o)$ as a straight line in this paper. The mathematical representation of q_o in terms of δ_o is then

$$q_o = \frac{1}{2}\left\{1 - \int_o^{\ln|\delta_o|} \frac{1}{0.174\sqrt{2\pi}} e^{-(\ln|\delta_o|-1.047)^2/2(0.174)^2} d\delta_o\right\} \qquad (15)$$

Constants 1.047 and 0.174 are estimated from Fig. 7. If $q_o > q > 1 - q_o$, the distribution has only one peak, otherwise it has two peaks. It should be noted that the new density function has only one peak for any value of q if $|\delta| \leq 2$.

TABLE 1 Critical values of δ_o for various values of q, so that two peaks occur in a density function if $\delta > \delta_o$ for given q_o

q_o	δ_o
0.01 0.99	4.09
0.03 0.97	3.76
0.05 0.95	3.59
0.10 0.90	3.32
0.15 0.85	3.14
0.20 0.80	2.99
0.25 0.75	2.85
0.30 0.70	2.72
0.35 0.65	2.59
0.40 0.60	2.44
0.45 0.55	2.28
0.50	2.00

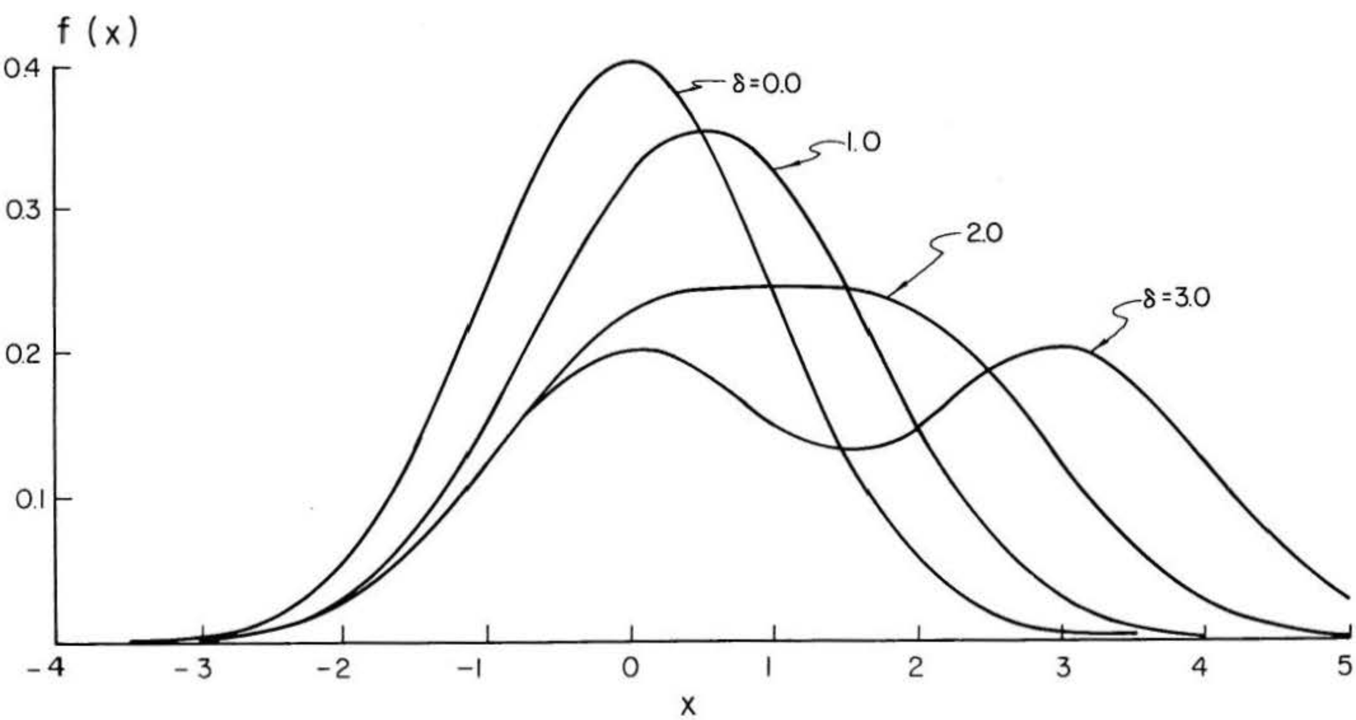

Fig. 6 Probability density functions for various values of δ with $\bar{\varepsilon} = 0$, $\sigma_\varepsilon = 1.0$ and q = 0.5

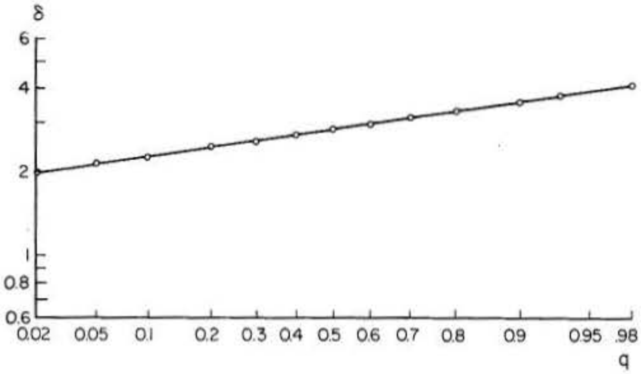

Fig. 7 The critical value, q_o, as related to, δ_o (constant jump), for determining whether the new probability density function will have one or two peaks

The value of δ indicates the distance between two different weighted probability density functions. Therefore, the positive and negative signs of δ indicate the relative position of these two probability density functions. The two-peak density function depends on the value of δ regardless if it is positive or negative. Based on this explanation, the absolute value of δ is used in eq. (15) for the general case. To find the critical value δ_o for a given value q_o from eq. (15), a trial-and-error procedure is necessary. If the resulting frequency curve has two peaks and a minimum of $f(x)$ between them, the x-values are designated by m_1(first peak), m_2 (minimum), and m_3(second peak). For a one-peak density function, the mode is designated by m.

It is obvious that the position m of the peak in the one-peak probability density function is closer to either $\bar{\varepsilon}$ or $\bar{\varepsilon} + \delta$, which depends only on values q and δ. From eq. (6) it follows that the smaller the q, the closer m is to $\bar{\varepsilon}$ and the converse, and the larger the q, the closer m is to $\bar{\varepsilon} + \delta$. For the two-peak density function, m_1 and m_3 are around $\bar{\varepsilon}$ and $\bar{\varepsilon} + \delta$, respectively. The positions of m_1 and m_3 obtained on digital computer are plotted in Fig. 8 versus δ by using q as parameter. The distances between the positions of one peak and the mean of zero and the second peak and $+\delta$, are equal in a symmetrical probability density function. For $q = 0.5$ and $\delta < 2$, the position m_1 has a linear relationship with δ, or $m_1 = \delta/2$. The position m_1(or m_3) linearly increases with δ up to $\delta = 2$. For the other values of q and for $q = 0.5$ but in the region of $\delta > 2$, the positions of m_1 or m_3 are non-linear relationships to δ. Therefore, the values m_1 and m_3 depend on δ, as shown in Fig. 8. They will increase as δ increases until they reach a maximum, then they decrease as δ further increases. It is evident from Fig. 8 that for the standardized variable $\varepsilon(\bar{\varepsilon} = 0,\ \sigma_\varepsilon = 1)$ and for $\delta > 4$, the positions m_1 and m_3 are approximately at 0 and δ, respectively. For non-standardized variables and $\delta > 4\,\sigma_\varepsilon$, these positions are approximately $\bar{\varepsilon}$ and $\bar{\varepsilon} + \delta$. For one-peak density curves and $\delta > 4$, the position m is approximately either at 0 (for small q) and at δ (for large q). From practical view point and $\delta > 4$ (or $\delta > 4\sigma_\varepsilon$) the values q and δ barely affect the positions m_1, m_3, and m.

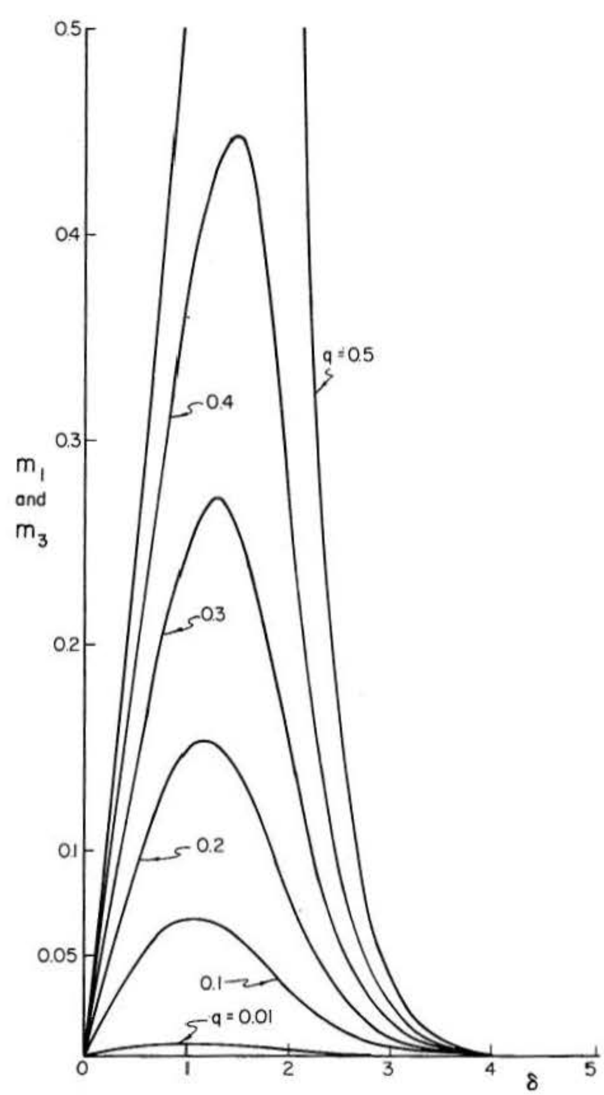

Fig. 8 The positions m_1 and m_3 of peaks in the probability density function for various values of q and δ, with $m_3 = \delta - m_1$

The position of the minimum, m_2, in two-peak distributions increases as δ increases. The value m_2, is equal to $\delta/2$ for $q = 0.5$. For $q \neq 0.5$, m_2 is calculated on a digital computer and is shown in Fig. 9. The family of curves of m_2 is symmetrical about the $q = 0.5$ for $q > 0.5$. For small δ, m_2 deviates significantly from the value $\delta/2$ when the minimum exists. For large δ the deviation of m_2 from $\delta/2$ as $(m_2 - \delta/2)$ is a constant. This constant increases as $|q - 0.50|$ increases.

The absolute difference between the probability densities of peaks, in two-peak distributions, is

$$D = \frac{1}{\sqrt{2\pi}\ \sigma} \left[e^{-\sigma^2/2}\ (2q-1) + (1-2q)\right]. \qquad (16)$$

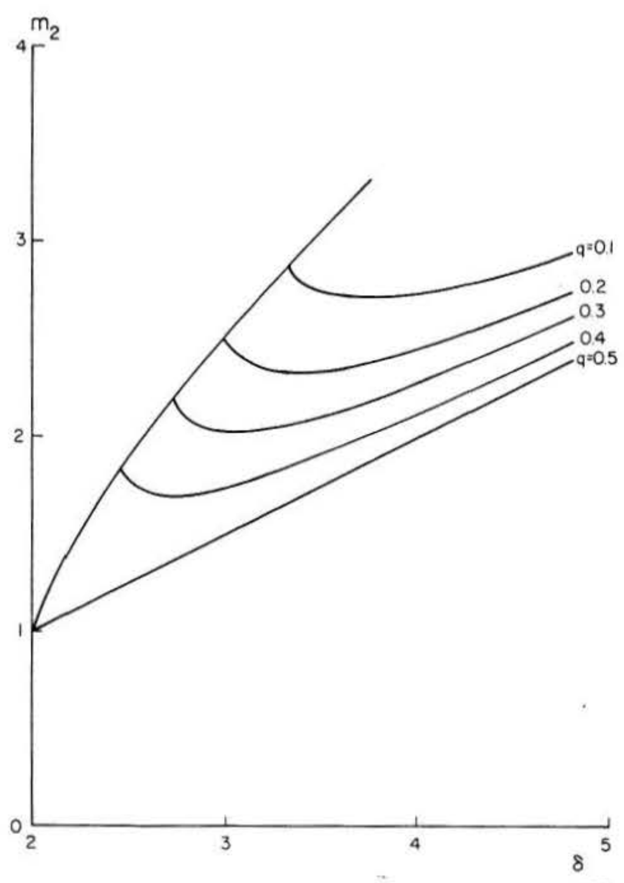

Fig. 9 The position, m_2, of the minimum of f(x) between the two peaks in the probability density function of a non-homogeneous series for various values of q and δ.

For large δ, $e^{-\delta^2/2}$ will converge fast to unity, and then the difference, D, depends only on q. This is shown in Fig. 10, on which the family of D curves versus δ are plotted with q as parameter. Practically, curves become horizontal lines for $\delta > 4$ for all values of q.

4. Relevance of above analysis to hydrology. Some hydrologic variables (daily flows, monthly flows, hourly, daily or monthly precipitation, and so forth) exhibit two-peak probability density curves under particular conditions. In most cases, the time series of these variables are composed of a periodic (of day or year) and a stochastic component. The periodic component is mainly present in the means and standard deviations of these variables. The coefficient of variation, the skewness coefficient and the covariances of the remaining stochastic component (obtained by removing the periodic movement in the mean and in the standard deviation) show relatively small or no periodic movement. If μ_τ and σ_τ are the mean and standard deviation at any hour of the day, or any day, or any month of the year, then $\varepsilon_i = (x_i - \mu_\tau)/\sigma_\tau$ is considered as a stochastic component, with x_i the original variable, and μ_τ and σ_τ the periodic functions, with periods of either a day or a year.

If the periodic component of μ_τ and σ_τ are transformed to a duration curve, and the first derivative is determined, the shape of this new curve will be similar to a U-probability density function.

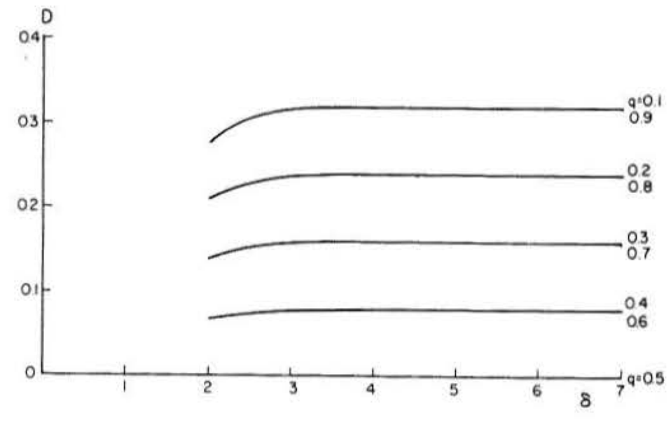

Fig. 10 The absolute difference, D, between the probability densities of peaks in a two-peak distribution for various values of q and δ.

The U-curve may have the same effect on x-density function as a random variable with a U-shape density curve. The stochastic components are mainly bell-shaped or J-shaped distributions. The combination of a U-shaped and a bell-shaped distribution often produces a two-peak density function. A combination of a periodic component (with U-type the first derivative of its duration curve) and a random component (with a bell-shaped probability density function) produces, under certain particular conditions, a two-peak density function of the variable x. Therefore, whenever hydrologic variables are composed of clear within-the-year or within-the-day periodic components and a stochastic component and treated as a univariate (instead of multivariates), then it is expected that some variables might exhibit two-peak probability density functions. Indeed, this fact has been observed quite frequently. Thus, a constant jump, as an equivalent of the periodic component, is shown in Fig. 11, which may be interpreted as a periodic component. Therefore, it is expected to produce two-peak density functions for particular values of q and δ.

Various explanations for two-peak density curves in hydrology may be encountered if one scans the literature. The first explanation is that two distinct climatic regimes exist in river basins, when either precipitation or runoff phenomena and their daily or

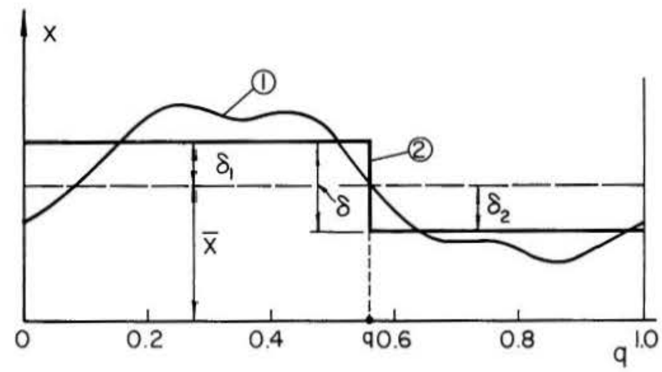

Fig. 11 A periodic movement (1) may be approximated by a constant jump scheme (2), with $\delta_1 + \delta_2 = \delta$

monthly variable values are considered, and the total effect of these two regimes produces two-peak density curves (or four inflection point distribution curves). The second explanation is that the river flows at a station below the confluence of two tributaries with different flow regimes may have two-peak density curves (two-peak histograms) if these two regimes are sufficiently distinct. The third explanation is that the two peaks and the minimum between them are simply a sampling product, and if observations are continued the frequency density curves will slowly converge to one-peak curves.

The first alternative is right only if there are two very distinct regimes of precipitation, one regime producing small amounts part of the year and nearly nothing the remaining time, and the other regime producing large amounts during the period that the first regime does not produce, and then producing nothing or very little in other periods. In other words, a large periodicity inside the year is created.

It is easy to contest the second explanation. Assume that the two tributaries, with river flows x and y, respectively, have each a bell-shaped frequency density curve. Then the sum x+y also shows a similar curve. If the skewness coefficients of x and y are approximately equal, the variable x+y is less skewed but is also bell-shaped distributed. Only when the conditions of the first alternative are met by the two distinct and different tributary regimes in time, the two-peak frequency density curves may occur as the result of periodicity within the year.

The third case is one of sampling errors, with two or more peaks in histograms occurring by the pure chance. It can be shown that in the cases of a composite series (periodic and stochastic components) the differences between peaks, between frequency densities of peaks, and between the frequency density of the minimum and frequencies of the two peaks, are statistically significant, even for medium size samples.

5. Effect of constant jump non-homogeneity on the mean. The mean of the distribution of a constant jump non-homogeneous series is

$$E(x) = \bar{x} = \int_{-\infty}^{\infty} xf(x)dx = \int_{-\infty}^{\infty} x(1-q)g(x)dx + \int_{-\infty}^{\infty} x\ q\ g(x-\delta)dx$$

$$= (1-q)\ \bar{\varepsilon} + q\bar{\varepsilon} + q\delta = \bar{\varepsilon} + q\delta, \qquad (17)$$

because $\int_{-\infty}^{\infty} xg(x)dx = \bar{\varepsilon}$ and $\int_{-\infty}^{\infty} g(x-\delta)dx = 1$.

It is obvious from a simple analysis that the change in the mean produced by this type of non-homogeneity is $q\delta$. The absolute change of mean increases with q and $|\delta|$.

6. Effect on variance. The variance of a constant jump non-homogeneous series is

$$\sigma_x^2 = \int_{-\infty}^{+\infty} (x-\bar{x})^2 f(x)dx = \sigma_\varepsilon^2 + (1-q)\ q\delta^2 . \qquad (18)$$

Non-homogeneity will introduce the change in the variance with quantity of $(1-q)\ q\delta^2$. Since $0 < q < 1$, $(1-q)\ q\delta^2$ is always a positive quantity regardless of δ positive or negative. Equation (18) proves, therefore, that the variance of a non-homogeneous series is always greater on the average than the variance of an original homogeneous series. The variance of a series is always increased by introducing the inconsistency or non-homogeneity into data. This increase has a maximum value for $q = 0.50$, for a given value of σ_ε^2 and δ, or for a given dimensionless parameter $\alpha = \delta/\sigma_\varepsilon$ as shown in Fig. 12.

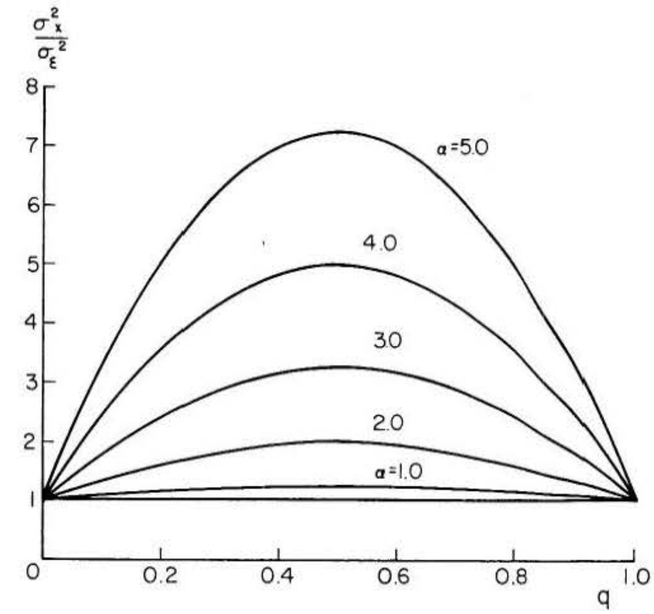

Fig. 12 The ratio of variances of a non-homogeneous and a homogeneous series, created by a constant jump, versus q for various values of α, with $\alpha = \delta/\sigma^2$

7. Effect on skewness. The skewness coefficient of a constant jump non-homogeneous series is

$$\beta = \frac{1}{\sigma_x^3} \int_{-\infty}^{+\infty} (x-\bar{x})^3\ f(x)\ dx = \left(\frac{\sigma_\varepsilon}{\sigma_x}\right)^3 \beta_\varepsilon + (2q-1)(q-1)q\alpha^3 \quad (19)$$

where β_ε is the skewness of the original homogeneous series, and $\alpha = \delta/\sigma_\varepsilon$. When $q < 0.50$ but δ are negative, or $q > 0.50$ but δ are positive, the quantity $(2q-1)(q-1)q\alpha^3$ is negative. In these cases $\sigma_x^3\beta < \sigma_\varepsilon^3\beta_\varepsilon$. As $\sigma_x > \sigma_\varepsilon$, in these cases β is definitely smaller than β_ε.

A general method to determine which of the two, β or β_ε, is larger, can be developed. If the difference between the skewness of a non-homogeneous series and an original homogeneous series is denoted by $\Delta\beta$, then

$$\Delta\beta = \beta - \beta_\varepsilon = \frac{1}{\sigma_x^3}\left\{\beta_\varepsilon(\sigma_\varepsilon^3 - \sigma_x^3) + (2q-1)(q-1)q\delta^3\right\} .$$

Since σ_x^3 is always a positive quantity, $\Delta\beta$ is positive only when

$$\beta_\varepsilon(\sigma_\varepsilon^3 - \sigma_x^3) + (2q-1)(q-1)\ q\delta^3 > 0 . \tag{20}$$

From eq. (20) it follows that the critical value β_c of β_ε is

$$\beta_c = \frac{(2q-1)(q-1)\ q\ \delta^3}{\sigma_x^3\ - \sigma_\varepsilon^3} . \tag{21}$$

The skewness of a non-homogeneous series will then equal that of the original series if $\beta_\varepsilon = \beta_c$. Then $\beta > \beta_\varepsilon$ if $\beta_\varepsilon < \beta_c$, and $\beta < \beta_\varepsilon$ if $\beta_\varepsilon > \beta_c$. This holds true particularly for the symmetrical probability distribution such as the normal distribution $\beta_\varepsilon = 0$, and when the skewness is not changed by non-homogeneity only if $q = 0.5$. If $q > 0.50$ and δ is negative, or $q < 0.50$ and δ is positive, the skewness coefficient β is positive. If $q > 0.50$ and δ is positive, or $q < 0.50$ and δ negative, the skewness coefficient β is negative. These properties are shown in Fig. 13.

8. Effect on kurtosis. The kurtosis, γ, of a constant jump non-homogeneous series is

$$\gamma = \frac{1}{\sigma_x^4}\int_{-\infty}^{+\infty} (x-\bar{x})\ f(x)\ dx =$$

$$= \frac{1}{\sigma_x^4}\left\{\sigma_\varepsilon^4\ \gamma_\varepsilon + 6(1-q)q\delta^2\sigma_\varepsilon^2 - q\delta^4\ (q-1)(3q^2-3q+1)\right\} \tag{22}$$

where γ_ε is the kurtosis of the original homogeneous series, and all other symbols are defined as stated. The difference of the kurtosis of the two series (after and before the non-homogeneity is introduced) denoted by $\Delta\gamma$, is

$$\Delta\gamma = \gamma - \gamma_\varepsilon = \frac{1}{\sigma_x^4}\left\{\gamma_\varepsilon(\sigma_\varepsilon^4-\sigma_x^4) + (1-q)q\delta^2[6\sigma_\varepsilon^2+(3q^2-3q+1)\delta^2]\right\}.$$

Because σ_x^4 is a positive quantity, $\Delta\gamma$ is a positive value only when

$$\gamma_\varepsilon^3(\sigma_\varepsilon^4-\sigma_x^4) + (1-q)q\delta^2[6\sigma_\varepsilon^2+ (3q^2-3q+1)\delta^2] > 0 . \tag{23}$$

From eq. (23) it follows that the critical value of γ_ε is

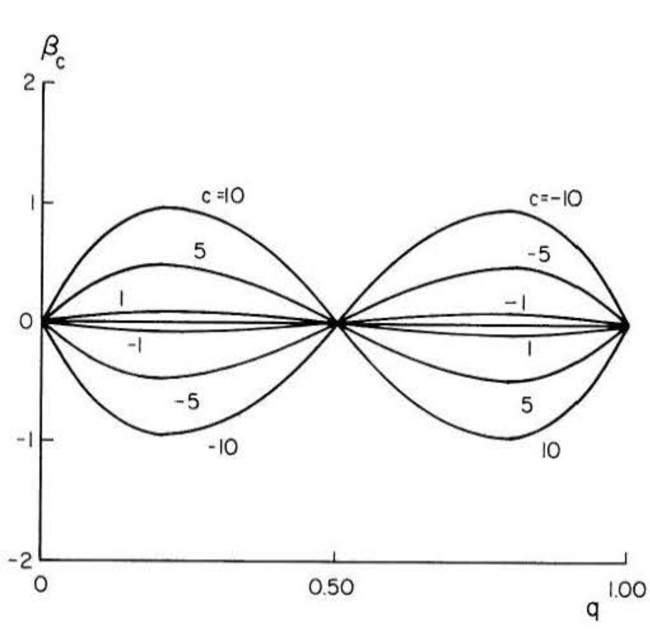

Fig. 13 The criterion, β_c, versus q for various values of c with $c = \alpha^3/(\frac{\sigma_x^3}{\sigma_\varepsilon^3} - 1)$

$$\gamma_c^3 = (1-q)q\delta^2\ \frac{[6\sigma_\varepsilon^2 + (3q^2 - 3q+1)\delta^2]}{\sigma_x^4 - \sigma_\varepsilon^4} =$$

$$= \frac{[6\sigma_\varepsilon^2 + (3q^2 - 3q + 1)\delta^2]}{\sigma_x^2 + \sigma_\varepsilon^2} . \tag{24}$$

When $\gamma_\varepsilon = \gamma_c$, then $\Delta\gamma$ is zero, or $\gamma = \gamma_\varepsilon$. If $\gamma_\varepsilon < \gamma_c$, then $\gamma > \gamma_\varepsilon$; and if $\gamma_\varepsilon > \gamma_c$, then $\gamma < \gamma_\varepsilon$. For the normal probability density function $\gamma_\varepsilon = 3$, and the kurtosis of a non-homogeneous series γ_ε is 3 if $q = (1\pm\sqrt{1/3})/2$, for all values of δ. For any value of δ, if q is between $\frac{1}{2} - \frac{\sqrt{3}}{6}$ and 0, or $\frac{1}{2} + \frac{\sqrt{3}}{6}$ and 1, the effect of non-homogeneity increases the kurtosis. If q is between $\frac{1}{2} - \frac{\sqrt{3}}{6}$ and $\frac{1}{2} + \frac{\sqrt{3}}{6}$, the non-homogeneity in data decreases the kurtosis.

9. Effect on serial correlation coefficients. Expected values of the sample serial correlation coefficients of a homogeneous independent random series are zero (except of $r_o = 1$), because the series is assumed to be sequentially independent. However, in the parts p and q of non-homogeneous series, the values are still mutually independent, but for the total series, the expected sample serial correlation coefficients are not zero because the computed mean is not zero. The variance is not σ_ε^2 and the expected covariances are not zeros.

The computational mathematical expression for the sample serial correlation coefficients used in this study is

$$r_k = \frac{\frac{1}{N-k}\sum_{i=1}^{N-k} (x_i-\bar{x})(x_{i+k}-\bar{x})}{\sigma_x^2} \tag{25}$$

where $\sigma_x^2 = \sigma_\varepsilon^2 + (1-q)q\sigma^2$ is assumed to be a constant parameter. The expected value of $\Sigma(x_i-\bar{x})(x_{i+k}-\bar{x})$ can be found either according to k or to the relative value k/N. The region of k/N is assumed to be less than the smaller value of q and 1-q, so that

$$\sum_{i=1}^{N-k} (x_i-\bar{x})(x_{i+k}-\bar{x}) = (N-k)q^2\delta^2 + (n-k)\delta^2 - q\delta^2\ (2n-k)$$

because $\Sigma(\varepsilon_i - \bar{\varepsilon})(\varepsilon_{i+k}-\bar{\varepsilon}) = 0$ for $k \neq 0$, and $\Sigma(x_i - \bar{\varepsilon}) = \Sigma(x_{i+k} - \bar{\varepsilon}) = 0$, and n = the part of N between the position of constant jump δ and the end of the series. In the region $m \leq k \leq n$, where m is the part of N between the beginning of the series and the position of constant jump,

$$\Sigma(x_i-\bar{x})(x_{i+k}-\bar{x}) = (N-k)\ q^2\delta^2 - (2n+m-2k)\ q\delta^2 + (n-k)\ \delta^2.$$

If $n < k < m$, then $\Sigma(x_i-\bar{x})(x_{i+k}-\bar{x}) = (N-k)q^2\delta^2 - nq\delta^2$. If the value of k is greater than the larger value of m and n, but of course is less than the value of N, then $\Sigma(x_i-\bar{x})(x_{i+k}-\bar{x}) = -q\delta^2(N-k)(1-q)$. From these results, the expected values of the sample serial correlation coefficients are various function of the lag k, according to the value of k and the position of the jump, m. In summary,

$$\begin{aligned}
r_k &= \alpha^2[q^2 + \frac{q(1-2q)-p\lambda}{1-\lambda}], && \text{for } \frac{1}{N} \leq \lambda \leq \text{Min }(p,q) \\
r_k &= p\alpha^2\ (\frac{q-\lambda}{1-\lambda} - q)\ , && \text{for } p \leq \lambda < q \\
r_k &= \alpha^2 q^2\ (\frac{\lambda}{\lambda-1})\ , && \text{for } q \leq \lambda < p \\
r_k &= -pq\alpha^2\ , && \text{for } \max(p,q) \leq \lambda \leq 1
\end{aligned} \tag{26}$$

where $\alpha = \delta/\sigma_x$, $p = n/N$, $q = n/N$ and $\lambda = k/N$ are dimensionless parameters.

The correlograms of eq. (26) are shown in Fig. 14 for various values of q. It should be noted that the correlograms are identical for the values q and 1-q. Whenever k is greater than m and n (whichever is larger), the expected sample serial correlation coefficients are constants and are independent of k but

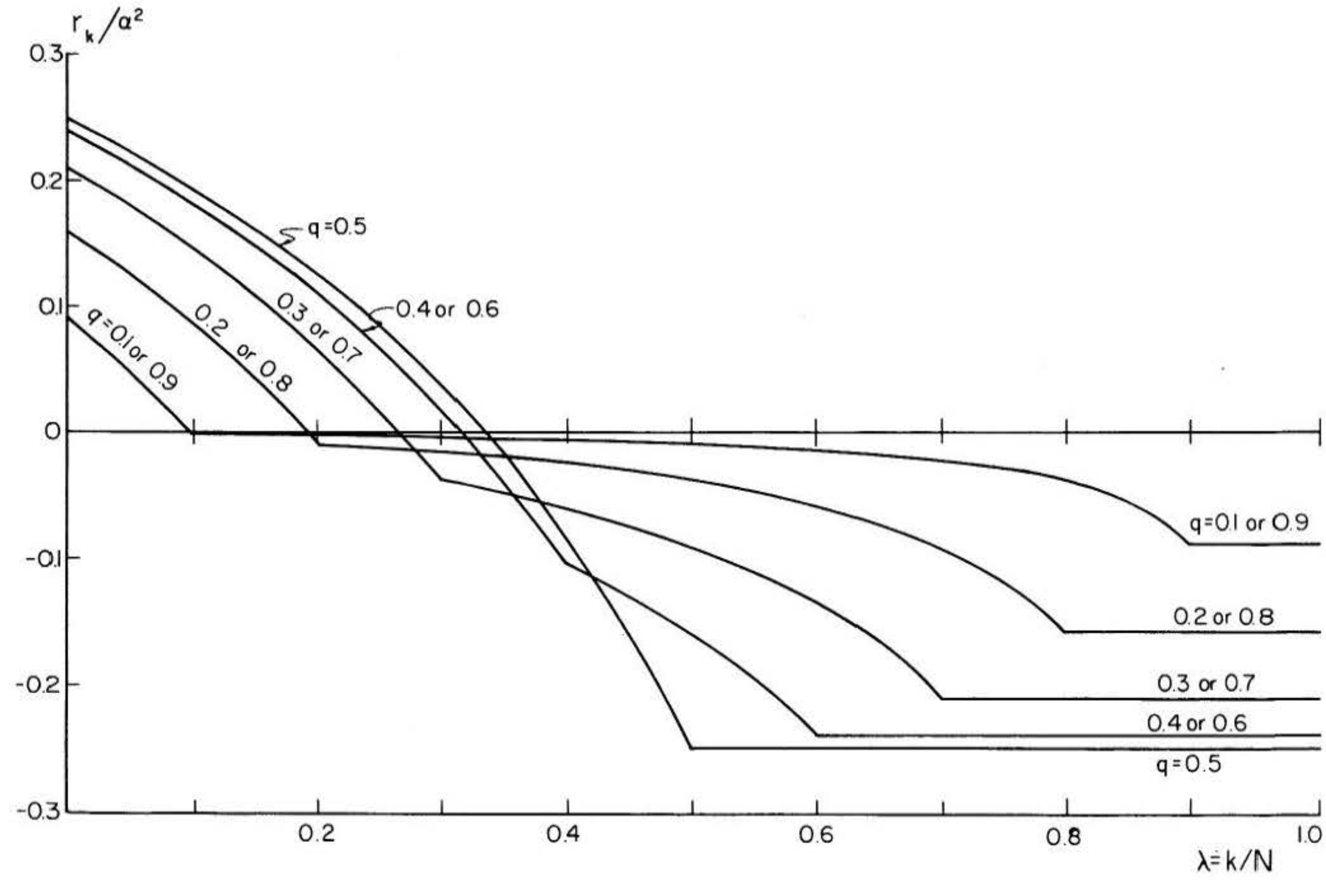

Fig. 14 The relationship of r_k/α^2, with $\alpha = \delta/s_x$, versus $\lambda = k/N$, for various values of $q = n/N$, and for the case of a constant jump, δ, introduced into a homogeneous series at the position $p = m/N$ or $q = n/N$.

depend on q, δ, and α. As shown by eq. (26), the expected serial correlation coefficients of the non-homogeneous series always have values different from the expedted coefficients of a homogeneous series. As the hydrologic series is often subject to inconsistency and non-homogeneity, a portion of the positive dependence at the initial parts of the correlograms (see Fig. 14) also comes from these two factors, apart from the effect of other basic physical processes of nature.

If r_k/α^2 is differentiated with respect to q and set-up equal to zero, one solution can be found in both regions, $\frac{1}{N} < \lambda < \mathrm{Min}(p,q)$ and $\mathrm{Max}(p,q) \leq \lambda \leq 1$. Therefore, r_k/α^2 will have the maximum value when $q = 1/2$ for a given value α. The maximum value of r_k/α^2 is $\frac{1}{4}(\frac{1-3\lambda}{1-\lambda})$ in the region $\frac{1}{N} < \lambda < \mathrm{Min}(p,q)$, and $(-\frac{1}{4})$ in the region $\mathrm{Max}(p,q) \leq \lambda \leq 1$.

The correlogram of a non-homogeneous series has the property that the serial correlation coefficients are constant negative values for the lag greater than a certain value. This property may provide a simple way for detecting the non-homogeneity in a time series. If the analysis of a hydrologic time series shows that the serial correlation coefficients follow approximately a negative constant value after the lag, k, is larger than a certain value, the hypothesis that there may be non-homogeneity present in the series is attractive for further investigation.

Chapter III

EFFECTS OF A COMBINATION OF CONSTANT JUMPS

1. Definition of change by constant jumps. The current hydrologic practice shows that changes occur in a sequence which may be composed of constant jumps at several positions in a series. In this case, several positive or negative constant changes, δ_i, occur along the series at various positions $q_i = n_i/N$, where n_i = length of the series after the jump δ_i and before the jump δ_i+1, so that $\sum_{i=1}^{h} q_i = 1$, or $\sum_{i=1}^{h} n_i = N$, with h = total number of subseries, the first length included. This case would correspond to many constant diversions of water in or out of a river basin with varying values of δ_i and q_i.

2. Effect on probability densities. The probability density function of a non-homogeneous series x_t of constant-jumps is determined by giving the weights q_i to probability densities of $(\varepsilon_t - \delta_i)$ respectively, or

$$f(x) = \sum_{i=1}^{h} q_i\, g(x-\delta_i) \tag{27}$$

where $g(\cdot)$ is the basic probability density function of ε. If $g(\cdot)$ is the normal probability density function, then

$$f(x) = \sum_{i=1}^{h} \frac{q_i}{\sigma_\varepsilon\sqrt{2\pi}}\, e^{-\frac{1}{2\sigma_\varepsilon^2}(x-\bar{\varepsilon}-\delta_i)^2} \quad . \tag{28}$$

When $g(\cdot)$ the lognormal probability density function,

$$f(x) = \sum_{i=1}^{h} \frac{q_i}{\sigma_n(x-\delta_i)\sqrt{2\pi}}\, e^{-\frac{1}{2\sigma_n^2}[\ln(x-\delta_i) - \mu_n]^2} ,$$

for $x > \max(\delta_1, \delta_2, \ldots, \delta_h)$, and

$$f(x) = \sum_{i<v} \frac{q_i}{\sigma_n(x-\delta_i)\sqrt{2\pi}}\, e^{-\frac{1}{2\sigma_n^2}[\ln(x-\delta_i)-\mu_n]^2} \tag{29}$$

where v is the subset of i for which δ_i is less than x. The probability density function f(x) may be multimodal (several peaks) which depends on values q_i and δ_i.

3. Effect on the mean. The mean of the new series is

$$E(x) = \bar{x} = \int_{-\infty}^{\infty} x f(x)\, dx = \bar{\varepsilon} + \sum_{i=1}^{h} q_i\delta_i \; . \tag{30}$$

In this case, the change in the mean because of non-homogeneity in the form of several constant jumps is $\sum_{i=1}^{h} q_i\delta_i$. As δ_i may be positive or negative, it can happen that the mean is unchanged if $\sum_{i=1}^{h} q_i\delta_i = 0$, or it can also indicate there is no change in the total balance of water in a river basin for the sample size N. It is often encountered in practice by the concept of water replacement or interchange, without changing the water balance of adjacent river basins. However, the other properties of river regimes may be highly affected even in this case of unchanged mean.

4. Effect on the variance. The variance of a non-homogeneous series is

$$\sigma_x^2 = \int_{-\infty}^{+\infty} (x-\bar{x})^2 f(x)dx = \sigma_\varepsilon^2 + \sum_{i=1}^{h} q_i\delta_i^{\,2} - (\sum_{i=1}^{h} q_i\delta_i)^2 . \tag{31}$$

It is obvious that the second term in eq. (31) is always greater than the third term. Hence the variance of series is increased by the non-homogeneity of constant jumps δ_i. If the mean is unchanged, as shown in the above discussion, the variance can still be significantly increased.

5. Effect on the skewness. Similarly, as in the previous text, the skewness coefficient β of a multiple constant-jump non-homogeneity is

$$\beta = \frac{1}{\sigma_x^3}\left\{\beta_\varepsilon\sigma_\varepsilon^3 + [\Sigma\, q_i\delta_i][2(\Sigma\, q_i\delta_i)^2 - 3\ \Sigma q_i\delta_i^2] + \Sigma q_i\delta_i^{\,3}\right\} \tag{32}$$

where sums in eq. (32) are all from i = 1 to i = h. If ε is a normal variable with $\beta_\varepsilon = 0$, the skewness of the non-homogeneous series will be either positive ($\beta > 0$) or negative ($\beta < 0$) depending on the values q_i and δ_i. If $\beta_\varepsilon \neq 0$, the skewness coefficient, β, may be greater or smaller than β_ε. In other words, the combination of constant jumps and their positions do not work only in one direction, such as is the case with the variance (the second central moment).

6. Effect on kurtosis. The kurtosis coefficient is also well affected by the non-homogeneity in the form of multiple constant jumps, so that

$$\gamma = \frac{1}{\sigma_x^4}\left\{\gamma_\varepsilon\sigma_\varepsilon^4 + 6\sigma_\varepsilon^2[\Sigma q_i\delta_i^{\,2} - (\Sigma q_i\delta_i)^2] - 3(\Sigma q_i\delta_i)^4 \right.$$
$$\left. + (\Sigma q_i\delta_i)[6(\Sigma q_i\delta_i^2)\ (\Sigma q_i\delta_i) - 4\ \Sigma q_i\delta_i^3] + \Sigma q_i\delta_i^4\right\} \tag{33}$$

with all sums in eq. (33) from i = 1 to i = h. The non-homogeneity may cause the kurtosis to increase or decrease which will depend on values q_i and δ_i. However, an anlysis of eq. (33) shows that on the average, the value, γ, would be mainly greater than γ, or the distribution of x should tend to have a flatter central part than the variable, ε.

7. Effects on serial correlation coefficients. The expected values of sample serial correlation coefficients are given here only for the parameter $\lambda = k/N$ which is less than the minimum value of $q_i (k < q_{min} N$, which is $k < n_{min}$,) or

$$r_k = \frac{1}{N-k}\left\{ k\left[\sum_{j+1}^{h-1}(\delta_j - \sum_{i=1}^{h} q_i\delta_i)(\delta_{j+1} - \sum_{i=1}^{h} q_i\delta_i)\right] + \sum_{j=1}^{h}(n_j - k)(\delta_j - \sum_{i=1}^{h} q_i\delta_i)^2\right\} \tag{34}$$

where $k \leq \min(n_1, n_2, \ldots, n_h)$.

Many hydrologic processes may be exposed to shifts (slippages) in the form of changing systematic errors of constant jumps. A similar case occurs with water withdrawals to or from a river, especially in complex water resource developments and operation schemes. Therefore, an independent series (say approximate the series of annual precipitation) becomes a time dependent variable by a combination of various constant jumps along the series. Or, if a series is already dependent, the non-homogeneity of this type increases the dependence, on the average. This statement "on the average" needs further explanation.

A sample from an independent normal variable may show a time dependence (though not significant). By adding a constant jump to this sample, it may occur by pure chance that the dependence of the non-homogeneous sample becomes smaller than in the homogeneous sample. However, if the same procedure is applied to many samples, the average result will show that the dependence is produced by constant jumps. If dependence already exists in the homogeneous series, it will increase in the large majority of samples.

An independent variable ε will become a dependent variable x by a sequence of the constant-jump source of non-homogeneity.

EFFECT OF LINEAR JUMPS

1. Definition of the change introduced by a linear jump. A non-homogeneity is introduced at positions t between m and N of the series in the form of

$$x_t = (1+I)\varepsilon_t, \quad \text{for } m \leq t \leq N$$
$$x_t = \varepsilon_t, \qquad \text{for } 0 \leq t \leq m \ , \tag{35}$$

or along the part n, with I = constant. In other words, x_t is proportional to ε_t by a constant different from unity along the second part of the series, where x_t is the historical value of a non-homogeneous series, while ε is the virgin value of an independent stationary series.

This case may simulate a diversion of river flows which are proportional to discharges at a river gauging station. In other words, if a partition of river discharge is made by a simple rule of proportionality, the change in the river from which water is diverted is equal to a negative linear jump. If the evaporation from a reservoir is approximately proportional to the annual inflow (higher reservoir levels and the corresponding larger evaporation surfaces may be linearly related to annual inflows), then the loss of water from the initiation of reservoir operation may be closely described by a linear jump. If systematic errors are proportional to the values of a hydrologic variable, the linear jump of eq. (35) may well describe this inconsistency in data from beginning to end.

2. Effect on probability density function. The variable $x/(1+I)$ will follow $g(\cdot)$ along the part n. The probability density function of the non-homogeneous series is then given as

$$f(x) = p\ g(x) + q\ g(\frac{x}{1+I}) \tag{36}$$

with $g(\cdot)$ the probability density function of the homogeneous series.

If it is assumed that the homogeneous series, ε_t, follows the normal distribution with mean $\overline{\varepsilon}$ and variance σ_ε^2, then

$$f(x) = \frac{p}{\sigma_\varepsilon\sqrt{2\pi}}\, e^{-(x-\overline{\varepsilon})^2/2\sigma_\varepsilon^2} + \frac{q}{(1+I)\sigma_\varepsilon\sqrt{2\pi}}\, e^{-[x-(1+I)\overline{\varepsilon}]^2/2(1+I)^2\sigma_\varepsilon^2} . \tag{37}$$

If the basic probability density function $g(\cdot)$ is lognormal with the mean $\overline{\varepsilon}_n$ of $\ln\varepsilon$ and the variance σ_n^2 of $\ln\varepsilon$ then

$$f(x) = \frac{p}{\sigma_n\, x\,\sqrt{2\pi}}\, e^{-\frac{1}{2\sigma_\varepsilon^2}[\ln x-\overline{\varepsilon}_n]^2} + \frac{q}{\sigma_n\, x\,\sqrt{2\pi}}\, e^{-\frac{1}{2\sigma_n^2}[\ln x-\overline{\varepsilon}_n-\ln(1+I)]^2} \tag{38}$$

for $0 < x < \infty$ and $I > -1$.

When $g(\cdot)$ is the normal probability density function of zero mean and variance σ_ε^2, f(x) is composed of two normal density functions both with the mean zero but different variances, σ_ε^2 and $(1+I)^2\,\sigma_\varepsilon^2$, respectively. In this case, regardless of values q and "I," the probability density function of a non-homogeneous series is always one peak function located at x = 0 and symmetrical about it. For $I < 0$, and particularly for q large, the variance of a non-homogeneous series is less than σ_ε^2 of a homogeneous series. If $I < 0$, the variance of $(1+I)\varepsilon_t$ will be smaller than the variance of ε_t, and for a large value of q, the probability density function with a small variance is more weighted than the unchanged part of series, so that it results in the variance of non-homogeneous series being smaller, or $\sigma_x^2 < \sigma_\varepsilon^2$.

3. Effect of a linear jump of a non-homogeneous series on the mean. The expected value of the new series of x_t is

$$E(x) = \overline{x} = \int_{-\infty}^{\infty} x f(x) dx = (1+qI)\overline{\varepsilon} . \tag{39}$$

The change of mean because of non-homogeneity in the form of a linear jump is $qI\overline{\varepsilon}$. The change depends not only on the properties of the non-homogeneity of q and I, but also on the mean, $\overline{\varepsilon}$. If $\overline{\varepsilon} = 0$, the mean $\overline{x} = 0$ also.

4. Effect on the variance. The variance of the non-homogeneous series is

$$\sigma_x^2 = \int_{-\infty}^{\infty}(x-\overline{x})^2 f(x)\,dx = (1+2qI+qI^2)(\sigma_\varepsilon^2+\overline{\varepsilon}^2)-(1+qI)^2\overline{\varepsilon}^2 =$$
$$= (1+2qI + qI^2)\ \sigma_\varepsilon^2 + (1-q)\ qI^2\overline{\varepsilon}^2 \ . \tag{40}$$

This relationship of eq. (40) is shown in Fig. 15. The negative value of "I" means that the second part values (or the biased part of the series) have a smaller mean and a smaller variance than the first part. Therefore, the variation created by non-homogeneity is cancelled by the smaller variance in the

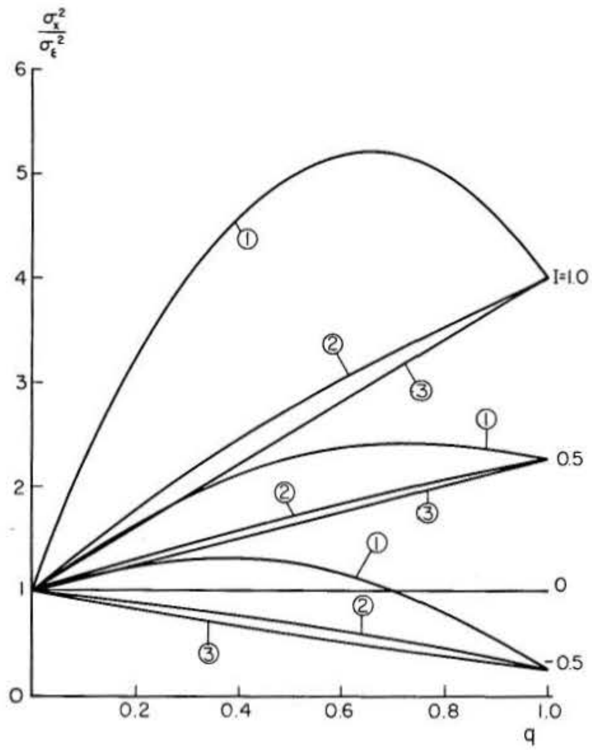

Fig. 15 The ratio of variances between a non-homogeneous and homogeneous series in the form of a linear jump for various values of I and $(\frac{\varepsilon}{\sigma_\varepsilon})^2$: (1) $(\bar{\varepsilon}/\sigma_\varepsilon)^2 = 10$ (2) $(\bar{\varepsilon}/\sigma_\varepsilon)^2 = 1$ (3) $(\bar{\varepsilon}/\sigma_\varepsilon)^2 = 0$

biased part, rather than in the original homogeneous series. In this case, the variance of the non-homogeneous series is not necessary always greater than for homogeneous series. This property of a non-homogeneous series formed by a linear jump is different from a non-homogeneous series formed by a constant jump.

If the change in variance produced by the non-homogeneity of a linear jump is denoted by $\Delta\sigma_x^2$, then

$$\Delta\sigma_x^2 = \sigma_x^2 - \sigma_\varepsilon^2 = qI(2+I)\sigma_\varepsilon^2 + qI^2(1-q)\bar{\varepsilon}^2 \quad . \tag{41}$$

When the variances of homogeneous and non-homogeneous series are equal, and $\Delta\sigma_x^2 = 0$, then eq. (41) gives

$$I = \frac{-2}{1 + (1-q)\bar{\varepsilon}^2/\sigma_\varepsilon^2} \quad . \tag{42}$$

Since "I" cannot be less than minus one, therefore, $(1-q)\bar{\varepsilon}^2$ must be greater than σ_ε^2, then

$$1 - q > (\frac{\sigma_\varepsilon}{\bar{\varepsilon}}) = \eta^2 \quad . \tag{43}$$

The inequality (43) gives

$$q < 1 - \eta^2 \quad , \tag{44}$$

where $\eta = \sigma_\varepsilon/\bar{\varepsilon}$ is the coefficient of variation of the homogeneous series ε_i. Because q is a quantity which is always positive and less than the unity, then the inequality (44) can only occur when the coefficient of variation is smaller than the unity, or when the standard deviation is smaller than the mean.

A conclusion can be determined from the above analysis. When $\eta < 1$ for the homogeneous series, $q < 1 - \eta^2$ and $I = -2/[1+(1-q)\varepsilon^2/\sigma_\varepsilon^2]$, then the variances of the homogeneous and non-homogeneous series are equal. The non-homogeneity in this particular case does not affect the variance. If $I < -2/[1+(1-q)/\eta^2]$, the variance of the non-homogeneous series is smaller than for the homogeneous series. Only for $I > -2/[1+(1-q)/\eta^2]$ does the non-homogeneity increase the variance.

5. Effect on the skewness. The skewness of the non-homogeneous series of a linear jump is

$$\beta = \frac{1}{\sigma_x^3}\int_{-\infty}^{+\infty}(x-\bar{x})^3 f(x)dx = \frac{1}{\sigma_x^3}\left\{\beta_\varepsilon (1+3qI+3qI^2+qI^3)\sigma_\varepsilon^3 + 3pqI^2\bar{\varepsilon}\sigma_\varepsilon^2(2+I) + p^3I^3\bar{\varepsilon}^3\right\} \quad , \tag{45}$$

where $p = 1-q$.

The skewness, β, of the probability density function of a non-homogeneous series will now be different from β_ε of the homogeneous series. In general, $\bar{\varepsilon}$ in this case will play an important role whether or not the effect will be a positive or a negative skewness of a non-homogeneous series. The degree of change mostly depends on q.

6. Effect on kurtosis. The kurtosis of a non-homogeneous series is

$$\begin{aligned}\gamma = \frac{1}{\sigma_x^4}\Big\{&(1+4qI+6qI^2+4qI^3+qI^4)(\gamma_\varepsilon\sigma_\varepsilon^4 + 4\beta_\varepsilon\bar{\varepsilon}\sigma_\varepsilon^3) - \\ &- (1+3qI+3qI^2+qI^3)(4\beta_\varepsilon\bar{\varepsilon}\sigma_\varepsilon^3 - 6\bar{\varepsilon}\sigma_\varepsilon^2) \\ &+ 6qI(1+I)^3\bar{\varepsilon}^2\sigma_\varepsilon^2 + (\bar{\varepsilon}-x)^4 + qI\bar{\varepsilon}[(4+6I+4I^2+I^3)\bar{\varepsilon}^3 \\ &- 4(3+3I+I^2)\bar{\varepsilon}^2\bar{x} + 6(2+I)\bar{\varepsilon}\bar{x}^2 - 4\bar{x}^3]\Big\} \quad .\end{aligned} \tag{46}$$

Kurtosis of a non-homogeneous series is usually less than for a homogeneous series. In other words, the probability density function of a non-homogeneous series has a relatively flat-topped peak when a linear jump is introduced into a homogeneous series.

7. Effects on serial correlation coefficients. Based on the same argument as in the case of a constant jump, the expected values of the sample serial correlation coefficients of a non-homogeneous series in the form of linear jump are obtained by a similar procedure which gives:

$$\left.\begin{aligned}
r_k &= \alpha^2[q^2 + \frac{q(1-2q) - p\lambda}{1-\lambda}], && \text{for } \frac{1}{N} < \lambda \leq \text{Min } [p,q] \\
r_k &= p\alpha^2 \ (\frac{q-\lambda}{1-\lambda} - q) && \text{for } p \leq \lambda < q \\
r_k &= \alpha^2 q^2 \ (\frac{\lambda}{\lambda-1}) && \text{for } q \leq \lambda < p \\
r_k &= -pq\alpha^2 && \text{Max } [p,q] < \lambda \leq 1
\end{aligned}\right\} \quad (47)$$

where $\alpha = \bar{\varepsilon}I/\sigma_x$, $p = m/N$, $q = n/N$, and $\lambda = k/N$. The expressions for the serial correlation coefficients of a non-homogeneous series of the linear jump given in eq. (47), and of the constant jump given in eq. (26), are identical except for the differences in definition of α . It should be noted that the expected value of the mean and the expected values of the sample serial correlation coefficients are not affected by the linear jump when theoriginal homogeneous series has a mean equal to zero. In other words, if the original series has zero mean and the linear jump is introduced into the series, the non-homogeneous series also has the expected mean of zero and the expected serial correlation coefficients of zero. This last statement means that the linear change of positive and negative values ε_i around $\bar{\varepsilon} = 0$ by a multiplier $(1 + I)$, even with $-1 < I < 0$ and for a part of series $(0 < q < 1)$, will not affect the serial correlation coefficients when passing from ε-series to x-series. In other words, the covariances are changed by the same proportion as the variances. However, as most hydrologic series are positively-values variables with $\bar{\varepsilon} > 0$, the linear jump will affect both the mean and serial correlation coefficients, so that $\bar{x} \neq \bar{\varepsilon}$ and $r_k(x) \neq r_k(\varepsilon)$.

8. Definition of a combination of linear jumps. The combination of linear jumps is defined as

$$\left.\begin{aligned}
x_t &= \varepsilon_t(1+I_1) && \text{for } 0 \leq t \leq n_1 \\
x_t &= \varepsilon_t(1+I_2) && \text{for } n_1 < t \leq n_1 + n_2 \\
&\ldots\ldots\ldots\ldots \\
x_t &= \varepsilon_t(1+I_i) && \text{for } \sum_{j=1}^{i-1} n_j < t \leq \sum_{j=1}^{i} n_j \\
&\ldots\ldots\ldots\ldots \\
x_t &= \varepsilon_t(1+I_\ell) && \text{for } \sum_{j=1}^{\ell-1} n_j < t \leq \sum_{j=1}^{\ell} n_j \ .
\end{aligned}\right\} \quad (48)$$

This is equivalent to many discrete changes along the series during the observations in the samples of size N. It is then assumed that

$$\sum_{j=1}^{\ell} n_j = N \qquad (49)$$

with ℓ = number of linear jumps.

9. Effects of the combination of linear jumps on a probability density curve of this non-homogeneous series, and on its parameters. These effects are given in condensed form for various properties of the resulting non-homogeneous series:

(a) Effects on probability density function.

$$f(x) = \sum_{i=1}^{\ell} q_i g \ (\frac{x}{1+I_i}) \ . \qquad (50)$$

(b) Effect on the mean.

$$E(x) = \bar{x} = (1 + \sum_{i=1}^{\ell} q_i I_i)\bar{\varepsilon} \ . \qquad (51)$$

(c) Effect on the variance.

$$\sigma_x^2 = [1+2\Sigma q_i I_i + \Sigma q_i I_i^2]\sigma_\varepsilon^2 + [\Sigma q_i I_i^2 - (\Sigma q_i I_i)^2]\bar{\varepsilon}^2 \qquad (52)$$

with all sums of eq. (52) and following eqs. (53) and (54) being between $i = 1$ and $i = \ell$.

(d) Effect on the skewness.

$$\beta = \frac{1}{\sigma_x^3}\Big\{\beta_\varepsilon\sigma_\varepsilon^3 \ [1+3\Sigma q_i I_i + 3\Sigma q_i I_i{}^2 + \Sigma q_i I_i{}^3] + 3(\bar{\varepsilon}-\bar{x})\sigma_\varepsilon^2 \ [1+2\Sigma q_i I_i + \Sigma q_i I_i{}^2] + 3\bar{\varepsilon}\sigma_\varepsilon^2[\Sigma q_i I_i(1+I_i)^2] + (\bar{\varepsilon}-\bar{x})^3+3(\bar{\varepsilon}-\bar{x})^2\bar{\varepsilon}(\Sigma q_i I_i)+3(\bar{\varepsilon}-\bar{x})\bar{\varepsilon}[\Sigma q_i I_i{}^2]+\bar{\varepsilon}^3\Sigma q_i I_i^3\Big\} . \qquad (53)$$

(e) Effect on the kurtosis.

$$\gamma = \frac{1}{\sigma_x^4}\Big\{\gamma_\varepsilon\sigma_\varepsilon^4 \ [\Sigma(1+I_i)^4 q_i]+4(\bar{\varepsilon}-\bar{x}) \ \beta_\varepsilon\sigma_\varepsilon^3[\Sigma(1+I_i)^3 q_i] + 4\bar{\varepsilon}\sigma_\varepsilon^3 \ \beta_\varepsilon[\Sigma q_i I_i(1+I_i)^3] + 6(\bar{\varepsilon}-\bar{x})^2\sigma_\varepsilon^2[\Sigma q_i(1+I_i)^2] + 12 \ \bar{\varepsilon} \ (\bar{\varepsilon}-\bar{x}) \ \sigma_\varepsilon^2 \ [\Sigma q_i I_i(1+I_i)^2] + 6\bar{\varepsilon}^2\sigma_\varepsilon^2[\Sigma q_i I_i{}^2(1+I_i)^2] + 4(\bar{\varepsilon}-\bar{x})^3\varepsilon(\Sigma q_i I_i) + 6 \ (\bar{\varepsilon}-\bar{x})^2\bar{\varepsilon}[\Sigma q_i I_i{}^2] + 4(\bar{\varepsilon}-\bar{x})\bar{\varepsilon}^3[\Sigma q_i I_i{}^3] + \bar{\varepsilon}^4(\Sigma q_i I_i{}^4) + (\bar{\varepsilon}-\bar{x})^4\Big\}. \qquad (54)$$

(f) Effects on the serial correlation coefficients.

$$r_k = \frac{\bar{\varepsilon}}{\sigma_\varepsilon^2}\Big\{(\bar{\varepsilon}-\bar{x})^2 + \frac{1}{N-k} \ [\sum_{j=2}^{\ell} n_j I_{j-1}^2 -k \ (\sum_{i=1}^{\ell-1} I_i{}^2) + (\bar{\varepsilon}-\bar{x})(2\sum_{i=1}^{\ell} n_i I_{i-1} - kI_{\ell-1}) + k \ (\sum_{i=1}^{\ell-2} I_i I_{i+1})]\Big\}, \text{ for } 1 < k \leq \text{Min } (n_j). \qquad (55)$$

Chapter V

EFFECTS OF LINEAR AND NON-LINEAR TRENDS

1. Definition of linear and non-linear trends. The linear trend is defined as

$$x_t = \varepsilon_t + a + bt \tag{56}$$

which is a random variable ε_t superposed by a linear trend $a + bt$.

A non-linear trend in this study is of a polynomial type where

$$x_t = \varepsilon_t + a_o + a_1 t + a_2 t^2 + \dots + a_m t^m \tag{57}$$

which is a random variable, ε_t, superposed by a non-linear function of an unknown equation but developed in the power series form. In both cases when $0 \leq t \leq N$, a, b, and a_o, a_1, ... , a_m are coefficients of the trend, respectively, for linear and non-linear type.

The trends are very common features of hydrologic non-homogeneity and inconsistency. An irrigation project does not enter into full development when the irrigation starts, so the depletion of water by evapotranspiration occurs slowly over a period of time until the full implementation of the project. The same procedure happens with the return flow. The systematic errors (say the slow change of environment around a precipitation gauge) may be shown to be a trend. The change of water quality is usually of a slow trend type. Examples include the effects of river basin conservation or anti-conservation activity on sediment transport, or of the effects of temperature pollution when there is a slow increase of the use of water for cooling purposes, etc.

The fact that many techniques are developed in a time series analysis for the detection and statistical inference of various trends, testifies of the importance of this subject to many disciplines including hydrology.

2. Effects on the probability density function. The distributions of a series with linear and non-linear trends are, respectively,

$$f(x) = \frac{1}{N}\int_o^N g(x-a-bt)dt \tag{58}$$

and

$$f(x) = \frac{1}{N}\int_o^N g(x-a_o-a_1 t-\dots-a_m t^m)dt \tag{59}$$

where N is the length of the series, and $g(\cdot)$ is the original probability density function of a homogeneous ε_t-series.

If ε_t is normal, $N(\bar{\varepsilon},\sigma_\varepsilon^2)$, and the time independent variable, then the linear trend a+bt produces a new probability density function

$$f(x) = \frac{1}{Nb}\left\{\Phi\left(\frac{Nb+\bar{\varepsilon}+a-x}{\sigma_\varepsilon}\right) - \Phi\left(\frac{a+\bar{\varepsilon}-x}{\sigma_\varepsilon}\right)\right\}, \tag{60}$$

where $\Phi(y)$ is the standard normal distribution function with y given either as $y = (Nb+\bar{\varepsilon}+a-x)/\sigma_\varepsilon$, or $y = (\bar{\varepsilon}+a-x)/\sigma_\varepsilon$.

Figure 16 gives four probability density functions of eq. (60) with $\theta = Nb$, $\bar{\varepsilon} = 0$, $\sigma_\varepsilon = 1$, independent with $E(r_k) = 0$, $k \neq 0$, and with the linear trend symmetrical for the series (the trend passes through the mean $\bar{\varepsilon} = 0$, at the center of series at N/2).

If the distribution of ε_t is lognormal, and the variable is independent, $\lambda(\bar{\varepsilon}_n, \sigma_n^2, 0)$, or the mean of $\ln \varepsilon$ is $\bar{\varepsilon}_n$, and the variance of $\ln \varepsilon$ is σ_n^2, with $E(r_k) = 0$, except for $k = 0$ for ε_t variable, the trend a+bt then produces the probability density function

$$f(x) = \frac{1}{Nb}\left\{\Phi\left[\frac{\ln(x-a)-\bar{\varepsilon}_n}{\sigma_n}\right]\right\}, \text{ for } a \leq x \leq (a + bN), \text{ and}$$

$$f(x) = \frac{1}{Nb}\left\{\Phi\left[\frac{\ln(x-a)-\bar{\varepsilon}_n}{\sigma_n}\right] - \Phi\left[\frac{\ln(x-a-bN)-\bar{\varepsilon}_n}{\sigma_n}\right]\right\},$$

$$\text{for } x > a + bN . \tag{61}$$

If b is negative,

$$f(x) = -\frac{1}{Nb}\left\{\Phi\left[\frac{\ln(x-a-bN)-\bar{\varepsilon}_n}{\sigma_n}\right]\right\}, \text{ for } a + bN < x < a$$

and (62)

$$f(x) = -\frac{1}{Nb}\left\{\Phi\left[\frac{\ln(x-a-bN)-\bar{\varepsilon}_n}{\sigma_n}\right] - \Phi\left[\frac{\ln(x-a)-\bar{\varepsilon}_n}{\sigma_n}\right]\right\}, \text{for } x > a.$$

Figure 17 gives four probability density functions of eqs. (61) - (62) with $\theta = Nb$, $\bar{\varepsilon} = 0$, var $\varepsilon_t = 1$, independent ε with $E(r_k) = 0$ except for $k = 0$, and with the linear trend a+bt symmetrical for the series of size N.

3. Effects on the mean. The linear trend gives the mean

$$E(x) = \bar{x} = a + \frac{1}{2}bN + \bar{\varepsilon} . \tag{63}$$

In the case of a symmetrical linear trend, with

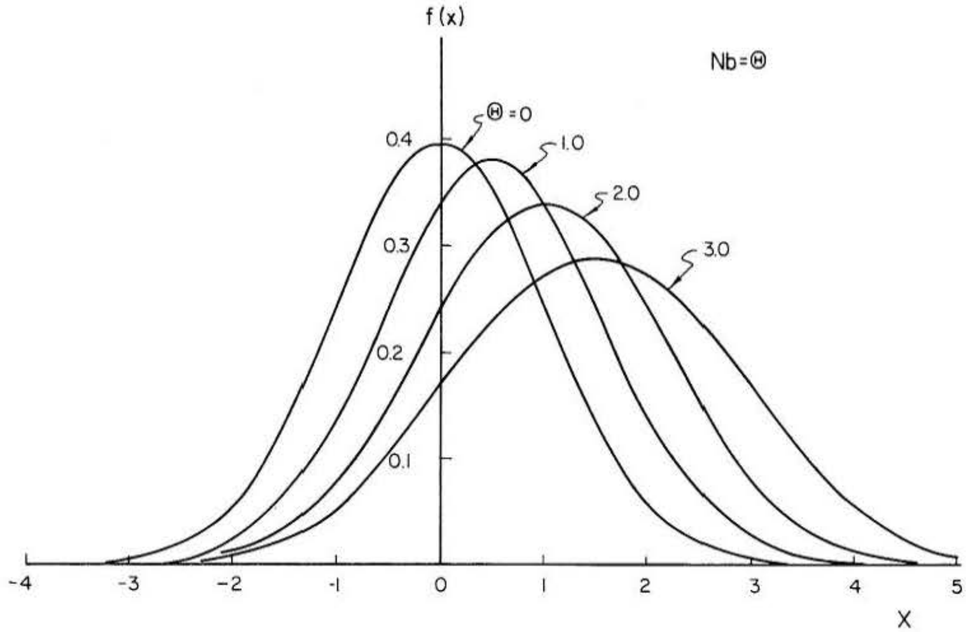

Fig. 16 Probability density functions of the linear trend, $x_t = a + bt + \varepsilon_t$, with ε_t the independent standard normal function N(0, 1, 0), for four values of $\theta = Nb$

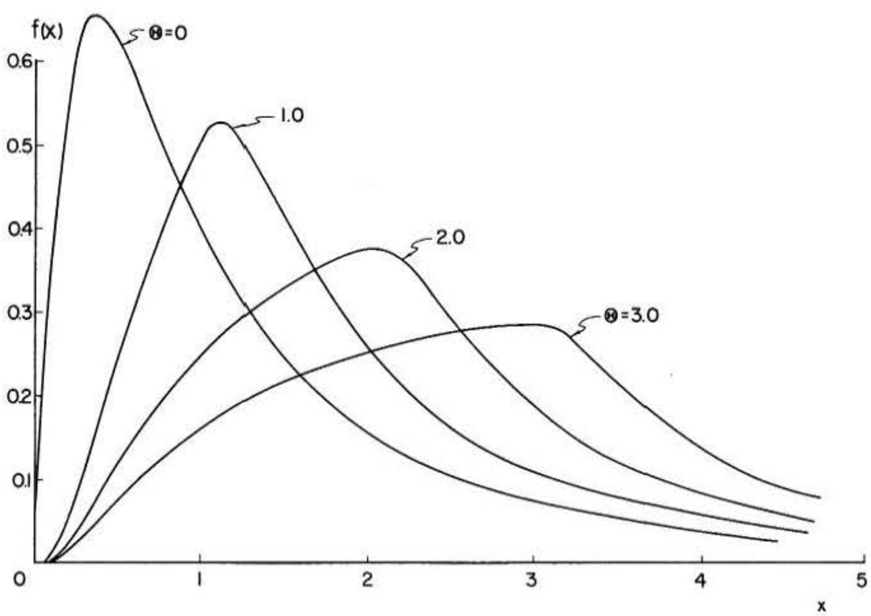

Fig. 17 Probability density functions of the linear trend, $x_t = a + bt + \varepsilon_t$, with ε_t the independent lognormal variable, for four values of $\theta = Nb$

$a = - bN/2$, eq. (63) becomes $\bar{x} = \bar{\varepsilon}$, so that the trend does not change the mean. However, if the trend starts at $N = 0$, then $a = 0$, eq. (63) becomes $\bar{x} = \bar{\varepsilon} + bN/2$.

The polynomial trend has the mean

$$E(x) = \bar{x} = \bar{\varepsilon} + a_o + \frac{1}{2} a_1 N + \frac{1}{3} a_2 N^2 + \dots + \frac{1}{m} a_m N^m =$$

$$= \bar{\varepsilon} + \sum_{i=1}^{m} \frac{1}{i+1} a_i N^i \ . \qquad (64)$$

For the start of the trend at $a_o = \bar{\varepsilon}$, then a_o in eq. (64) should be replaced by $\bar{\varepsilon}$.

4. Effects on the variance. The linear trend always increases the variance, so that

$$\sigma_x^2 = \sigma_\varepsilon^2 + \frac{b^2 N^2}{12} \qquad (65)$$

and for the polynomial trend, the variance is

$$\sigma_x^2 = \sigma_\varepsilon^2 + \sum_{i=1}^{m} \sum_{j=1}^{m} a_i a_j \left[\frac{N^{i+j}}{(i+j+1)} - \frac{N^{i+j}}{(i+1)(j+1)} \right] =$$

$$= \sigma_\varepsilon^2 + \sum_{i=1}^{m} \sum_{j=1}^{m} a_i a_j N^{i+j} \left[\frac{ij}{(i+1)(j+1)(i+j+1)} \right] \quad (66)$$

It can be shown from eq. (66) that $\sigma_x^2 > \sigma_\varepsilon^2$ regardless of the type of polynomial and its various coefficients a_i, a_j. In other words, any trend of the type of eqs. (56) and (57) increases the variance of a non-homogeneous series x_t in comparison with ε_t series.

5. Effect on the skewness. The linear trend changes the skewness so that

$$\beta = \frac{\sigma_\varepsilon^2}{\sigma_x^2} \beta_\varepsilon \quad (67)$$

or for $\beta_\varepsilon = 0$, also $\beta = 0$. The linear trend superposed on a series of independent normal variable does not introduce the skewness. However, as $\sigma_x^2 > \sigma_\varepsilon^2$, then the linear trend decreases the skewness of the dependent skewed variable ε_t, because $|\beta| < |\beta_\varepsilon|$ in that case, regardless whether or not β_ε is positive or negative.

For the polynomial trend, the skewness of x_t-series is

$$\beta = \frac{1}{\sigma_x^3} \left\{ \beta_\varepsilon \sigma_\varepsilon^3 + \sum_{i=1}^{m} \sum_{j=1}^{m} \sum_{k=1}^{m} a_i a_j a_k N^{i+j+k} \left[\frac{1}{i+j+k+1} - \right.\right.$$

$$- \frac{1}{(i+1)(j+k+1)} - \frac{1}{(j+1)(i+k+1)} - \frac{1}{(k+1)(i+j+1)} +$$

$$\left.\left. + \frac{2}{(i+1)(j+1)(k+1)} \right] \right\}. \quad (68)$$

Whether or not $\beta > \beta_\varepsilon$ or $\beta < \beta_\varepsilon$, depends then on the type of polynomial trend. If $\beta_\varepsilon = 0$, eq. (68) shows that β can be either positive or negative, which de-depends on the coefficients of eq. (57).

6. Effects on kurtosis. For the linear trend, the kurtosis of the x_t-series is

$$\gamma = \frac{1}{\sigma_x^4} \left\{ \gamma_\varepsilon \sigma_\varepsilon^4 + \frac{1}{2} (bN)^2 \sigma_\varepsilon^2 + \frac{1}{80} (bN)^4 \right\} \quad (69)$$

while the kurtosis of a series with a polynomial trend is

$$\gamma = \frac{1}{\sigma_x^4} \left\{ \gamma_\varepsilon \sigma_\varepsilon^4 + 6\sigma_\varepsilon^2 \left[\sum_{j=1}^{m} \sum_{j=1}^{m} a_i a_j N^{i+j} \frac{ij}{(i+1)(j+1)(i+j+1)} \right] + \right.$$

$$+ \sum_{i=1}^{m} \sum_{j=1}^{m} \sum_{k=1}^{m} \sum_{\ell=1}^{m} a_i a_j a_k a_\ell N^{i+j+k+\ell} \left[\frac{1}{i+j+k+\ell+1} \right.$$

$$- \frac{1}{(i+1)(j+k+\ell+1)} - \frac{1}{(j+1)(i+k+\ell+1)} + \frac{1}{(i+1)(j+1)(k+\ell+1)}$$

$$- \frac{1}{(k+1)(i+j+\ell+1)} + \frac{1}{(i+1)(k+1)(j+\ell+1)} + \frac{1}{(j+1)(k+1)(i+\ell+1)}$$

$$- \frac{1}{(\ell+1)(i+j+k+1)} + \frac{1}{(i+1)(\ell+1)(k+j+1)}$$

$$+ \frac{1}{(j+1)(\ell+1)(i+k+1)} + \frac{1}{(\ell+1)(k+1)(i+j+1)}$$

$$\left.\left. - \frac{3}{(i+1)(j+1)(k+1)(\ell+1)} \right] \right\}. \quad (70)$$

7. Effects on serial correlation coefficients. The linear trend produces a time dependence in x_t-series even if ε_t-series is independent, so that in this case $r_k(x)$ is

$$E(r_k) = \frac{b^2}{4\sigma_x^2} \left\{ 3N(N-2k) + 4N + 2k(k-2) + 1 \right\}, \text{ for } k \geq 1 \text{ but } k \ll N. \quad (71)$$

Figure 18 gives the expected correlogram of a series composed of an independent normal variable ε_t and a linear trend $a + bN$, for a sample size $N = 100$, and for four values of the parameter $\psi = (\sigma_\varepsilon/b)^2$ being 1, 100, 500, and 1000. The correlogram shows a highly dependent non-homogeneous series if b is sufficiently large in comparison with σ_ε. The value $\psi = 1$ corresponds to $\sigma_\varepsilon = 1$ (standardized ε_t variable) and a trend of 45°. Even a trend of the slope 1:1000, for $\sigma_\varepsilon = 1$, gives the first serial correlation coefficients of the order $r_k = 0.4$.

It is likely that the time dependence of many series of annual precipitation is partly or fully produced by a trend in data, which is due to an inconsistency in data.

The expected serial correlation coefficients of a variable x_t produced by an independent variable ε_t (0, 1, 0), and a polynomial trend is

$$E(r_k) = \frac{1}{\sigma_x^2} \left\{ \left[\sum_{i=1}^{m} \frac{a_i}{i+1} N^i \right]^2 - \frac{1}{N-k} \left[\left(\sum_{i=1}^{m} \frac{a_i}{i+1} N^i \right) \left(\sum_{i=1}^{m} \frac{a_i}{i+1} (N-k)^{i+1} \right) \right. \right.$$

$$+ \left(\sum_{i=1}^{m} \frac{1}{i+1} a_i N^i \right) \left[\sum_{i=1}^{m} \frac{a_i}{i+1} (N^{i+1} - k^{i+1}) \right]$$

$$\left.\left. - \sum_{j=1}^{m} \sum_{k=1}^{m} \frac{(N-k)^{j+1} a_i a_j}{\frac{(i+j+1)!}{(i-1)!}} \left(\sum_{R=0}^{j} (i+j-R)! k^R N^{j-R} \frac{j!}{(j-R)!} \right) \right] \right\}. \quad (72)$$

Equation (72) shows that $E(r_k)$ of x_t are not zeros for $k > 0$, as is the case for $E(r_k)$ for ε_t.

The linear and non-linear trends are, therefore, producers of time dependence if a homogeneous series is independent. They increase the dependence, on the average, if a series to which trends are applied are already time dependent.

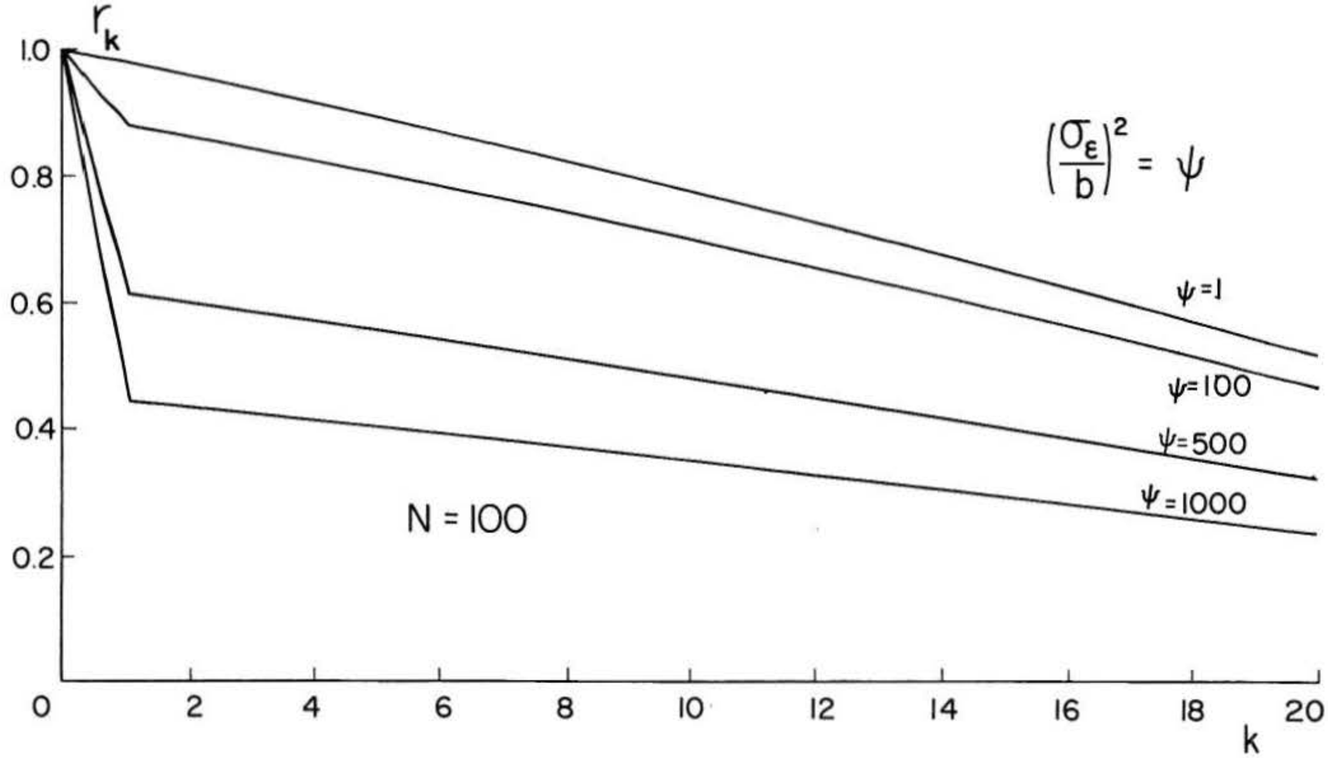

Fig. 18 Correlograms of the linear trend series, $x_t = a + bt + \varepsilon_t$, with ε_t the independent variable for the sample size, $N = 100$, and for four values of the parameter, $\psi = (\sigma_\varepsilon/b)^2$

CHAPTER VI

EXAMPLES OF EFFECTS OF NON-HOMOGENEITY

1. Types of examples. The hypothetical cases were selected in order to investigate the effects of non-homogeneity that are introduced into an independent stationary stochastic series. In two cases, this variable is normal and independent, with the mean zero and the variance unity, $N(0,1,0)$. In the other two cases, the variable is normal and dependent, $N(0,1,\rho)$, of the first order Markov linear model. In each case, five samples of 200 random independent normal numbers are generated. Practically, 1000 numbers, ε_i, are generated, and each divided into five samples.

For each sample of N = 200, the non-homogeneity is introduced in five parts or subsamples (0-40, 41-80, 81-120, 121-160, and 161-200). The first and third examples have the following non-homogeneities:

(1) For i = 1-40, a small linear trend is introduced in the form, $x_i = \varepsilon_i - 0.10 + 0.005i$;

(2) For i = 41-80, no change is made;

(3) For i = 81-120, a slippage (jump) of $\delta = 0.25$ is added, so that $x_i = \varepsilon_i + \delta$;

(4) For i = 121-160, another linear trend is introduced in the form, $x_j = \varepsilon_j + 0.12 - 0.006j$, with j = 1 for i = 121, and j = 40 for i = 160; and

(5) For i = 161-200, a linear jump is introduced in the form, $x_i = 0.85\ \varepsilon_i$.

The second and fourth examples made all of these changes stronger, and under (2) another slippage (constant jump) is added, so that the second and fourth examples have:

(1) For i = 1-40, a linear trend, $x_i = \varepsilon_i - 0.30 + 0.020i$;

(2) For i = 41-80, a slippage of $\delta = 0.50$ is introduced, so that, at any position, i = 81-120, the values $x_i = \varepsilon_i + 0.50$;

(3) For i = 81-120, another slippage of $\delta = -0.40$ is added, so that $x_i = \varepsilon_i - 0.40$;

(4) For i = 121-160, the linear trend is introduced, so that $x_j = \varepsilon_j + 0.45 - 0.30j$, where j = 1 for i = 121, and j = 40 for i = 160; and

(5) For i = 161-200, a linear jump is produced by $x_i = 1.25\ \varepsilon_i$.

The third and fourth examples of non-homogeneity are applied to the dependent stationary variable, $\eta_i = \rho\ \eta_{i-1} + \varepsilon_i$, with $\rho = 0.20$. The non-homogeneous variables thus obtained are y_i.

Altogether, there are two times two examples: the independent variable, ε_i, with two types of non-homogeneity, weak and strong, which produce x_i-variables; and the dependent variable, η_i, with these two same types of weak and strong non-homogeneity, to produce y_i-variables.

Each of the 200 samples of all four examples, is changed by the corresponding five non-homogeneities in the 40 member subsamples, so that the non-homogeneous samples of 200 values of x_i or y_i are produced. In this way, five samples of x_i and five samples of y_i are obtained for each of the four examples: (a) a weak non-homogeneity approach of the first type, as x_i- and y_i-variables; and (b) a strong non-homogeneity of the second type, also as x_i- and y_i-variables.

2. Computations. For all four types of variables, ε_i, x_i, η_i and y_i, and for each of the samples of N = 200, the following parameters are computed: the mean, the standard deviation, the variance, the skewness coefficient, the excess coefficient, and the first fifty serial correlation coefficients. The average values of five 200-value long samples are then computed for these parameters.

For the same variables, ε_i, x_i, η_i and y_i, and for the samples of N = 1000 that are composed as the sequence of five samples of N = 200, the same parameters as previously shown are computed. Frequency distributions are also determined.

3. Results of the first example. This example is the series of ε_i, independent stationary normal variable, with the weak type of non-homogeneity introduced in each of the five subsamples of 40 values, for every sample of N = 200 that produces the non-homogeneous variable, x_i.

Figure 19, dashed line (1), gives the frequency distribution of 1000 values of ε_i in the Probability-Cartesian scales. It is well fitted by a straight line (or the generated numbers follow the normal distribution). The same figure, solid line (2), shows the frequency distribution of 1000 values of x_i of the first example. It departs somewhat from the distribution of ε_i, but still it is approximately normally distributed.

Table 2 shows, for the first example, the parameters of ε_i- and x_i-variables for five samples of N = 200, their averages, and the sample of N = 1000 (composed of five successive samples of N = 200). Only the first three serial correlation coefficients are given. By comparing the average parameters of five samples of N = 200 between ε_i- and x_i-variables, or the parameters of samples of N = 1000 for these variables, it is clear that some changes have occurred in nearly all parameters, though in a small way.

Figure 20 shows, for the first example, the correlograms of ε_i and x_i, as $r_k(\varepsilon_i)$ and $r_k(x_i)$, graph lines (1) and (2), respectively, for the sample of N = 1000. The line (3) gives their differences, $\Delta r_k = r_k(x_i) - r_k(\varepsilon_i)$. It is clear that these differences, though

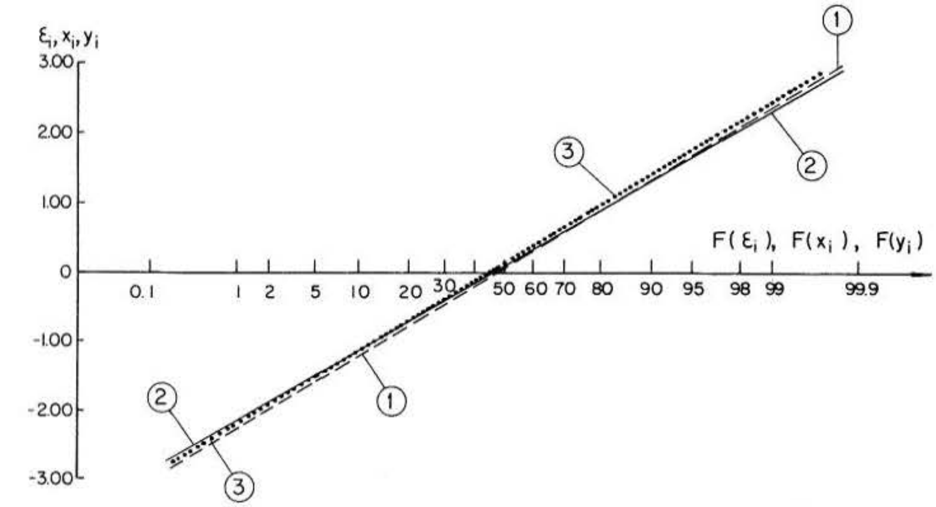

Figure 19 Frequency distributions of three variables in the Probability-Cartesian scales for the first and third examples, derived from a sample of N = 1000: (1) distribution of ε_i, standard normal independent variable; (2) distribution of the non-homogeneous variable, x_i, obtained by a superposition of weak non-homogeneities to the variable, ε_i; and (3) distribution of the non-homogeneous variable, y_i, obtained by a superposition of weak non-homogeneities, to the variable $\eta_i = 0.20\ \eta_{i-1} + \varepsilon_i$

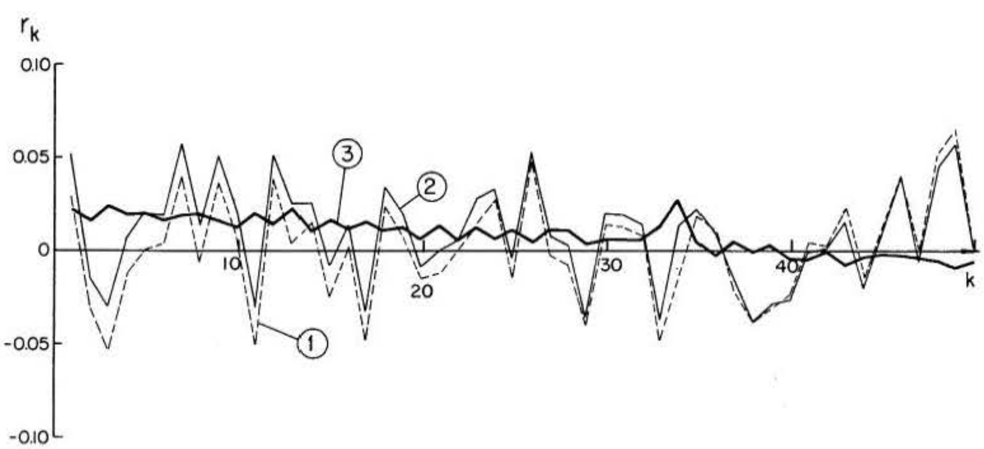

Figure 20 Correlograms of the two series of the first example: (1) the series, ε_i, for the sample of N = 1000; (2) the non-homogeneous series, x_i, of the weak non-homogeneities introduced to ε_i, for N = 1000; and (3) the difference, $\Delta r_k = r_k(x_i) - r_k(\varepsilon_i)$

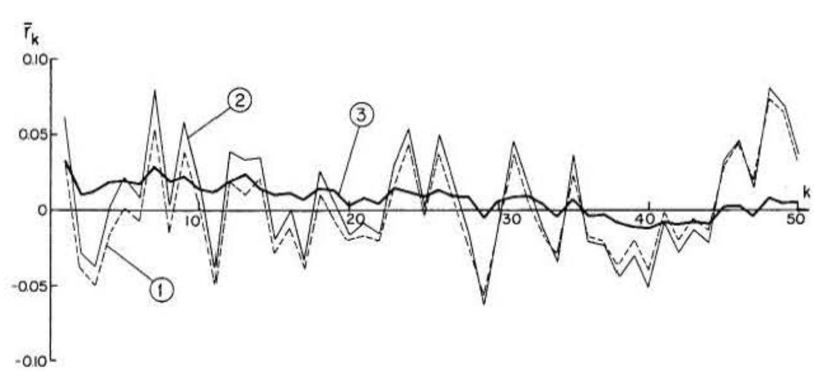

Figure 21 Correlograms of the two series in the first example: (1) the average values of $r_k(\varepsilon_i)$ for five samples of N = 200 of the series, ε_i; (2) the average values of $r_k(x_i)$ for five samples of N = 200 of the non-homogeneous series, x_i, and the weak non-homogeneities introduced to ε_i; (3) the differences, $\Delta\bar{r}_k = \bar{r}_k(x_i) - \bar{r}_k(\varepsilon_i)$

TABLE 2

PARAMETERS OF THE FIRST EXAMPLE OF NON-HOMOGENEITY EFFECTS

PARAMETERS		ε - VARIABLE							x_i - VARIABLE						
	Symbols	1st Sample N = 200	2nd Sample N = 200	3rd Sample N = 200	4th Sample N = 200	5th Sample N = 200	Average of 5 Samples	Sample of N=1000	1st Sample N = 200	2nd Sample N = 200	3rd Sample N = 200	4th Sample N = 200	5th Sample N = 200	Average of 5 Samples	Sample of N=1000
Mean	μ	0.01521	0.04903	0.05537	0.00078	0.09615	0.04331	0.04300	0.06963	0.09519	0.10692	0.04860	0.14169	0.09261	0.09241
Standard Deviation	σ	0.97843	1.01045	1.00894	1.04976	0.94196	0.99791	0.99713	0.96926	1.00115	0.98771	1.01899	0.92571	0.98056	0.97964
Skewness	β	0.00993	-0.08289	-0.10741	0.00365	0.03778	-0.02779	-0.03542	0.04527	-0.05367	0.02727	0.03864	0.02856	0.01721	0.01133
Excess	γ	-0.11946	-0.13518	0.35190	-0.50856	-0.42035	-0.16633	-0.17332	-0.17014	0.00382	0.70525	-0.39298	-0.34803	-0.04042	-0.04553
1-Serial Correlation Coefficient	ρ_1	0.02625	-0.04969	0.13300	-0.02098	0.06916	0.03155	0.02900	0.04927	-0.01372	0.15516	-0.01761	0.09988	0.06091	0.05191
2-Serial Correlation Coefficient	ρ_2	-0.05116	0.00677	-0.09097	-0.01286	-0.03928	-0.03750	-0.03176	-0.02677	0.03659	-0.09490	-0.00329	-0.01462	-0.02810	-0.01546
3-Serial Correlation Coefficient	ρ_3	-0.01017	0.03291	0.02451	-0.14829	-0.14682	-0.04957	-0.05248	0.00818	0.06693	0.03954	-0.15003	-0.09978	-0.03695	-0.02961

TABLE 3

PARAMETERS OF THE SECOND EXAMPLE OF NON-HOMOGENEITY EFFECTS

PARAMETERS		ε - VARIABLE							x_i - VARIABLE						
	Symbols	1st Sample N = 200	2nd Sample N = 200	3rd Sample N = 200	4th Sample N = 200	5th Sample N = 200	Average of 5 Samples	Sample of N=1000	1st Sample N = 200	2nd Sample N = 200	3rd Sample N = 200	4th Sample N = 200	5th Sample N = 200	Average of 5 Samples	Sample of N=1000
Mean	μ	0.01521	0.04903	0.05537	0.00078	0.09615	0.04331	0.04300	0.01667	0.06425	0.06161	0.00909	0.11242	0.05282	0.05281
Standard Deviation	σ	0.97843	1.01045	1.00894	1.04976	0.94196	0.99791	0.99713	1.08614	1.15429	1.10232	1.18249	1.07575	1.12020	1.11934
Skewness	β	0.00993	-0.08289	-0.10741	0.00365	0.03778	-0.02779	-0.03542	-0.08705	-0.09990	0.09407	-0.03639	0.11715	-0.00242	-0.01130
Excess	γ	-1.19459	-0.13518	0.35190	-0.50856	-0.42035	-0.16633	-0.17332	-0.18548	-0.11599	0.84655	-0.31934	-0.19447	0.00625	-0.01252
1-Serial Correlation Coefficient	ρ_1	0.02625	-0.04969	0.13300	-0.02098	0.06916	0.03155	0.02900	0.10253	0.08933	0.18431	0.07463	0.18141	0.13275	0.12341
2-Serial Correlation Coefficient	ρ_2	-0.05116	0.00677	-0.09097	-0.01286	-0.03928	-0.03750	-0.03176	0.03529	0.13870	0.02754	0.06545	0.09368	0.06463	0.07728
3-Serial Correlation Coefficient	ρ_3	-0.01017	0.03291	0.02451	-0.14829	-0.14682	-0.04957	-0.05248	0.07185	0.14232	0.09825	-0.05309	-0.03619	0.03471	0.03964

they are relatively small, are positive up to about $k = 40$ of the order $\Delta r_k = 0.02 - 0.03$.

Figure 21 shows the correlograms of ε_i and x_i of the first example, similar as in Fig. 20, but in this case, $\overline{r}_k(\varepsilon_i)$ and $\overline{r}_k(x_i)$ are the averages of five values, each obtained for one sample of N = 200. The patterns are similar as for Fig. 20, and the differences, $\Delta\overline{r}_k$, given as line (3), are positive up to k = 40.

This first example of each non-homogeneity shows that the effects are small but are not negligible when it comes to the crucial differences of a dependent or independent series.

4. Results of the second example. This example is the series of ε_i, independent stationary normal variable, with the strong type of non-homogeneity introduced in the five subsamples of 40 values, for every sample of N = 200 that produces the non-homogeneous variable, x_i.

Figure 22, dashed line (1), gives the frequency distribution of 1000 values of ε_i in the Probability-Cartesian scales. The solid line (2) gives the frequency distribution of 1000 values of x_i, a strongly non-homogeneous variable. Though these graphs do not permit the investigation of detailed differences between ε_i- and x_i-distributions, x_i-distribution is approximately symmetrical with a greater slope (larger standard deviation) than ε_i, as a result of strong non-homogeneity.

Table 3 gives parameters of ε_i- and x_i-variables for five samples of N = 200, their averages, and the sample of N = 1000. This table is analogous to Table 2. Only the first three serial correlation coefficients are given. The table shows differences in parameters, especially in the standard deviation and in the serial correlation coefficients.

Figure 23 shows, for the second example, the correlograms of ε_i and x_i, as $r_k(\varepsilon_i)$ and $r_k(x_i)$, graph lines (1) and (2), respectively, for the sample of N = 1000. The line (3) gives their differences $\Delta r_k = r_k(x_i) - r_k(\varepsilon_i)$. It is clear that these differences are significant and positive up to k = 35, and

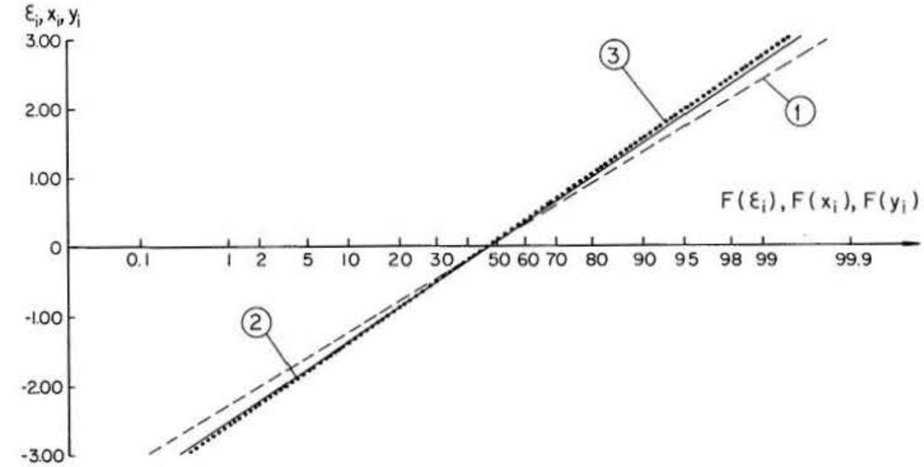

Figure 22 Frequency distributions of three variables in the Probability-Cartesian scales for the second and fourth examples, derived from a sample of N = 1000: (1) distribution of ε_i, standard normal independent variable; (2) distribution of the non-homogeneous variable, x_i, obtained by a superposition of strong non-homogeneities to the variable, ε_i; and (3) distribution of the non-homogeneous variable, y_i, obtained by a superposition of strong non-homogeneities to the variable, $\eta_i = 0.20\ \eta_{i-1} + \varepsilon_i$

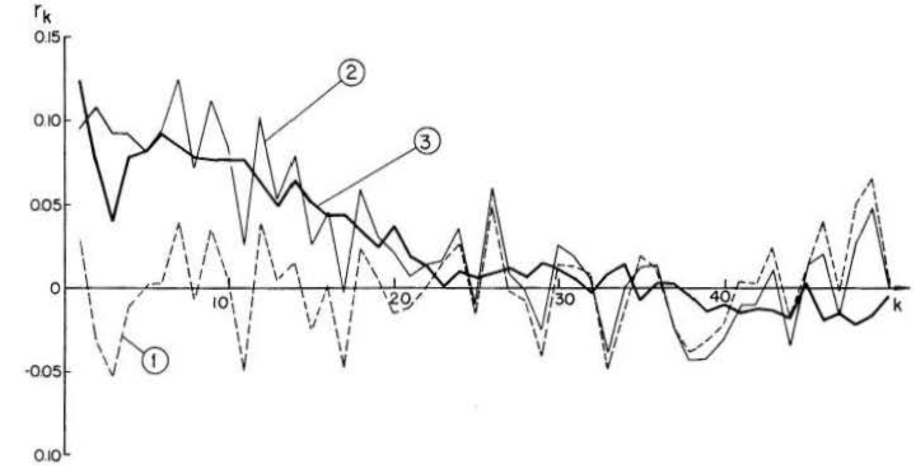

Figure 23 Correlograms of the two series of the second example: (1) for the series, ε_i, of the sample of N = 1000; (2) the non-homogeneous series, x_i, of the strong non-homogeneities introduced to ε_i, for N = 1000; and (3) the differences, $\Delta r_k = r_k(x_i) - r_k(\varepsilon_i)$

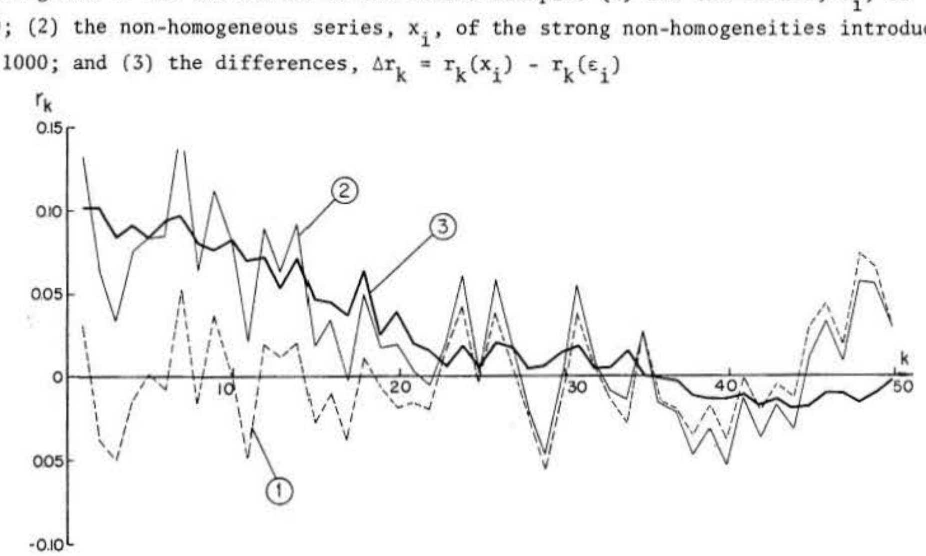

Figure 24 Correlograms of the two series of the second example: (1) the average values, $\bar{r}_k(\varepsilon_i)$, for five samples of N = 200 of the series, ε_i; (2) the average values of $\bar{r}_k(x_i)$, for five samples of N = 200 of the non-homogeneous series, x_i, and the strong non-homogeneities introduced to ε_i; and (3) the differences, $\Delta\bar{r}_k = \bar{r}_k(x_i) - \bar{r}_k(\varepsilon_i)$

then negative beyond. The differences decrease from $r_1(x_i) = 0.10$, on.

Figure 24 shows the correlograms of ε_i and x_i of the second example, similar as for Fig. 23, but in this case, $\bar{r}_k(\varepsilon_i)$ and $\bar{r}_k(x_i)$ are the averages of five values, each obtained for one sample of N = 200. The patterns are nearly identical as for Fig. 23 and the differences, $\Delta\bar{r}_k$, are positive and decrease from about $\Delta\bar{r}_1 = 0.10$ for k = 1 to zero, approximately k = 35.

This second example of strong non-homogeneities shows that the affects of non-homogeneity are significant and cannot be neglected in the analysis of properties of a hydrologic time series.

5. Results of the third example. This example is the series of $\eta_i = 0.2\ \eta_{i-1} + \varepsilon_i$, a dependent stationary normal variable of the first order Markov linear model, with the weak non-homogeneities introduced in the five subsamples of 40 values each, for every sample of N = 200 that produces the non-homogeneous variable, y_i.

Figure 19, line (3) shows the distribution of variable, y_i. The variance of η_i is then var η_i = = var $\varepsilon_i/(1-\rho^2)$ = var $\varepsilon_i/0.96$ = 1.042 var ε_1 = 1.042, for var $\varepsilon_i = 1$, so that the variance of distribution of y_i has both the affect of var $\eta_i > 1.00$ and the affects of non-homogeneity.

Table 4 shows the parameters of the y_i-variable of the third example, both for the averages of five samples of N = 200 and for the unique sample of N = 1000.

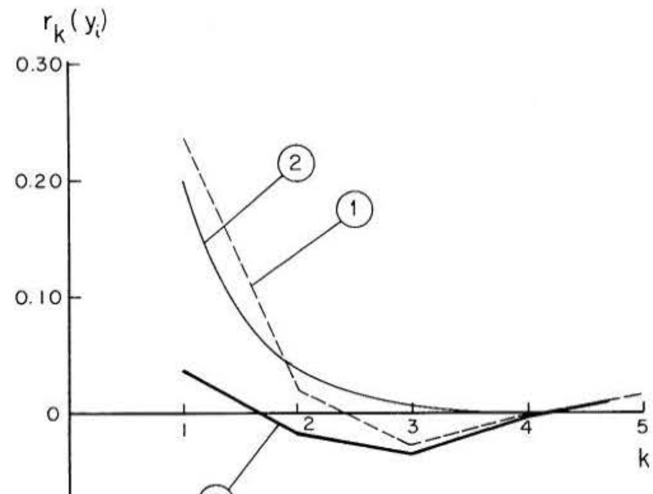

Figure 25 Correlograms for the third example: (1) for the variable, y_i, as $r_k(y_i)$ of N = 1000, in the case the weak non-homogeneities are introduced to the dependent stationary and normal variable, $\eta_i = 0.20\ \eta_{i-1} + \varepsilon_i$; (2) the expected correlogram of η_i; $\rho_k = \rho^k = 0.20^k$; and (3) differences, $\Delta r_k = r_k(y_i) - 0.20^k$

Figure 25 shows the correlograms of the third example of weak non-homogeneity, for the sample of N = 1000 of the y_i-variable. Line (1) shows $r_k(y_i)$, line (2) the expected correlogram $\rho_k = \rho^k = 0.20^k$, and line (3) their differences, $\Delta r_k = r_k(y_i) - 0.20^k$. Only for k = 1 is there a small positive difference, Δr_1 = = 0.038.

Figure 26 shows the correlograms of the third example, similar to Fig. 25, but in this case, $\bar{r}_k(y_i)$ are the averages of the five values of r_k, each obtained for one sample of N = 200. The patterns are similar to those of Fig. 25.

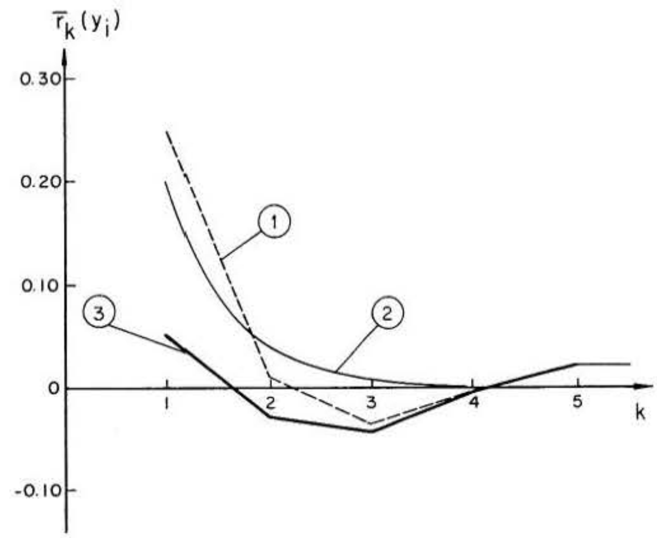

Figure 26 Correlograms for the third example: (1) for the variable, y_i, as the average value, $\bar{r}_k(y_i)$, for five samples of N = 200, for the weak non-homogeneities superposed to the variable, η_i = = $0.20\ \eta_{i-1} + \varepsilon_i$; (2) the expected correlogram of η_i; $\rho_k = \rho^k = 0.20^k$; and (3) differences, $\Delta r_k = \bar{r}_k(y_i) - 0.20^k$

The third example of weak non-homogeneities superposed to a dependent normal variable, with $\rho = 0.20$ of the first order Markov linear model, shows a relatively small change in comparison with the original dependent variable, η_i.

6. Results of the fourth example. This example is the series of $\eta_i = 0.20\ \eta_{i-1} + \varepsilon_i$, a dependent stationary normal variable of the first order Markov linear model, with the strong non-homogeneities introduced in the five subsamples of 40 values each, for every sample of N = 200 that produces the non-homogeneous variable, y_i.

Figure 22, line (3), gives the distribution of the y_i-variable. Part of the increased standard deviation is due to the larger variance (1.042) of η_i in comparison to the variance (1.000) of ε_i, while the other part is due to a strong non-homogeneity effect.

TABLE 4

Parameters of the third example of non-homogeneity effects

PARAMETERS	Symbols	1st Sample N = 200	2nd Sample N = 200	3rd Sample N = 200	4th Sample N = 200	5th Sample N = 200	Average of 5 Samples	Sample of N = 1000
Mean	μ	0.14630	0.19507	0.22843	0.09085	0.31099	0.19433	0.10192
Standard Deviation	σ	1.38389	1.41894	1.41994	1.44768	1.30907	1.39590	1.00387
Skewness	β	-0.10501	-0.20179	-0.23315	-0.09598	-0.34152	-0.19349	0.03029
Excess	γ	-1.55378	-1.50082	-1.10470	-1.71004	-1.45899	-1.46567	-0.00807
1-Serial Correlation Coefficient	ρ_1	0.23280	0.19129	0.32119	0.17196	0.26919	0.24947	0.23792
2-Serial Correlation Coefficient	ρ_2	0.01574	0.07832	-0.01894	-0.01016	0.02155	0.01168	0.02227
3-Serial Correlation Coefficient	ρ_3	0.01110	0.06900	0.03675	-0.16036	-0.09556	-0.03520	-0.02576

TABLE 5

Parameters of the fourth example of non-homogeneity effects

PARAMETERS	Symbols	1st Sample N = 200	2nd Sample N = 200	3rd Sample N = 200	4th Sample N = 200	5th Sample N = 200	Average of 5 Samples	Sample of N = 1000
Mean	μ	0.03753	0.13493	0.13751	0.01219	0.25446	0.11532	0.06252
Standard Deviation	σ	1.55001	1.64179	1.58459	1.68233	1.52871	1.59749	1.14284
Skewness	β	-0.07528	-0.14923	-0.07463	-0.04949	-0.16553	-0.10283	0.01523
Excess	γ	-1.57910	-1.60560	-1.09944	-1.70492	-1.49496	-1.49680	-0.01086
1-Serial Correlation Coefficient	ρ_1	0.26498	0.27287	0.34320	0.24343	0.33289	0.31802	0.29097
2-Serial Correlation Coefficient	ρ_2	0.06424	0.17542	0.07981	0.06200	0.11245	0.11171	0.10393
3-Serial Correlation Coefficient	ρ_3	0.06261	0.14935	0.10137	-0.06675	-0.04364	0.04753	0.04043

Table 5 shows the parameters of the y_i-variable in the fourth example, both for the averages of five samples of N = 200, and for the unique sample of N = 1000.

Figure 27 shows, for the fourth example, the correlograms of strong non-homogeneity, for the sample of N = 1000 of y_i-variable. Line (1) shows $r_k(y_i)$, line (2) the expected correlogram $\rho_k = \rho^k = 0.20^k$, and line (3) their differences. It shows a significant increase in the time dependence by non-homogeneity.

Figure 28 gives the correlograms for the fourth example, similar to Fig. 27, but in this case, $\bar{r}_k(y_i)$ are the averages of the five values of r_k, each obtained for one sample of N = 200. The patterns are similar to these of Fig. 27.

The fourth example of strong non-homogeneities superposed to a dependent normal variable, with ρ = 0.20 of the first order Markov linear model, shows a significant change in comparison with the original dependent variable, η_i.

7. Conclusions. All four hypothetical examples show that the non-homogeneity affects the properties of the stationary time series. The degrees of these changes are functions of types and degrees of non-homogeneity.

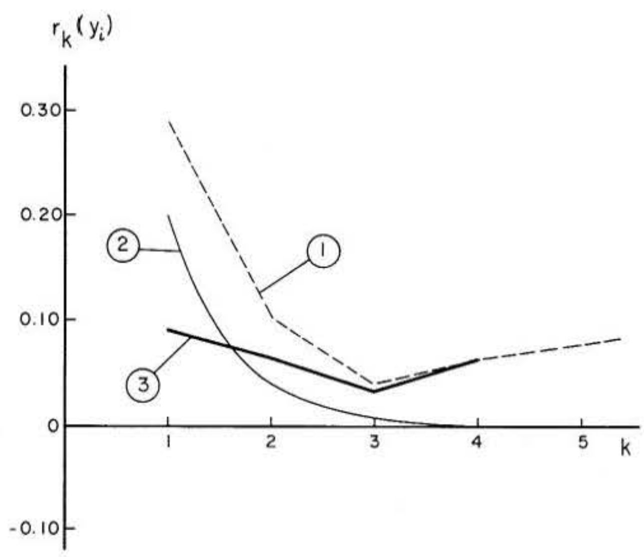

Figure 27 Correlograms for the fourth example: (1) for the variable, y_i, as $r_k(y_i)$ of N = 1000, in the case the strong non-homogeneities are introduced to the dependent and normal variable, $\eta_i = 0.20\ \eta_{i-1} + \varepsilon_i$; (2) the expected correlogram of η_i; $\rho_k = \rho^k = 0.20^k$; and (3) their differences, $\Delta r_k = r_k(y_i) - 0.20^k$

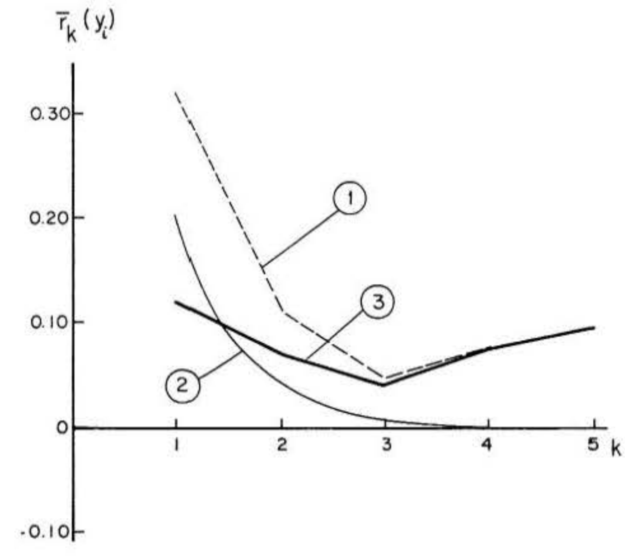

Figure 28 Correlograms for the fourth example: (1) for the variable, y_i, as the averages, $\bar{r}_k(y_i)$, for five samples of N = 200, for the strong non-homogeneity superposed to the variable, $\eta_i = 0.20\ \eta_{i-1} + \varepsilon_i$; (2) the expected correlogram of η_i; $\rho_k = \rho^k = 0.20^k$ and (3) their differences, $\Delta r_k = \bar{r}_k(y_i) - 0.20^k$

Chapter VII

DISCUSSION OF RESULTS AND CONCLUSIONS

1. Main implication of results in previous chapters on hydrologic information. The types of basic homogeneous series were simple in the analysis of previous chapters, such as independent normal and log-normal random variables. Simplified and generalized cases of non-homogeneity and inconsistency were treated. Regardless of these two factors, the relatively small changes introduced into a time series have often shown significant changes in variable properties. The implication is that the data of many hydrologic random variables contain unknown amounts of non-homogeneity and inconsistency. This in turn makes a difference in variable properties in comparison with the underlying homogeneous series.

A typical example is the analysis of the first serial correlation coefficient of the annual precipitation series for a large number of gauging stations in Western North America [5]. The series of annual precipitation are divided in two groups: (a) homogenous or consistent with no significant change in station position or environment, a total of 1141 stations, and (b) non-homogeneous or inconsistent with a significant change of the vertical or horizontal position of stations during the period of observation (or with other changes which occurred during that time), a total of 473 stations. For the period of observation of 30 years, 1931-1960, the first group of the series produced the average first serial correlation coefficient of $\overline{r}_1 = 0.028$, while the second group of the non-homogeneous series produced $\overline{r}_1 = 0.053$. The total records available gives an average length of 54 years for homogeneous series, and of 57 years for non-homogeneous series. The average first serial correlation coefficients are $\overline{r}_1 = 0.055$ and $\overline{r}_1 = 0.071$, respectively. The non-homogeneous series always gave on the average a higher value of $\overline{r}_1$ than did the homogeneous series.

As one could assume, the many series included here in the group of homogeneous series may contain the unidentified non-homogeneity or inconsistency. This is an attractive assertion to make, namely that the values $\overline{r}_1 = 0.028$ and $\overline{r}_1 = 0.055$ for the homogeneous series and for the record of 30 and 54 years, respectively, may have been in part determined by inconsistency in the data. One should take into account that, through the averaging used by a large number of station series (1141), any non-homogeneity increases the first serial correlation coefficient. Therefore, even the first group of station series may be considered as quasi-stationary (or quasi-homogeneous) because of difficulties in detecting the relatively small amounts of non-homogeneity and inconsistency. If it could be possible to remove all sources of inconsistency in the data, the above values $\overline{r}_1 = 0.028$ and $\overline{r}_1 = 0.055$ would be further reduced. This analysis leads to the statement that annual precipitation is very close to being an independent hydrologic random variable.

The hydrologic information is therefore very often biased by the presence of either a neglected or unidentifiable non-homogeneity and inconsistency in a time series. In some cases, this bias may even show a significant effect on water resource decision making. The bias in the mean results in an incorrect prediction of water resources available. The bias in the standard deviation and the first serial correlation coefficient, to the upper side, also means a greater storage requirement, all other factors being the same.

The detection and removal of inconsistency and non-homogeneity in hydrologic data is an important part of processing the data and of extracting the maximum information from a given amount of data.

2. Conclusions. The previous six chapters show that many types of sources of non-homogeneity and inconsistency have impacts on properties of hydrologic random variables, leading to the following conclusions:

(1) On the average, a constant jump or a combination of constant jumps change all properties of a sample.

(2) On the average, a linear jump or a combination of linear jumps produce in a sample of positively valued variables a change in all properties of a time series.

(3) A linear or polynomial trend also affects all properties of a sample.

(4) In the majority of cases, the non-homogeneity and inconsistency in a time series alters the variance of a variable.

(5) In nearly all cases, any type of non-homogeneity and inconsistency in a time series produces the correlograms which show an increase in autocorrelation coefficients to positive values, at least in its initial part.

(6) The study of the effect of various types, amounts, and sources of inconsistency and non-homogeneity, the investigation of methods for identification by statistical inference and physical analyses, the procedures for their removal and prediction, and similar works are important subjects of modern hydrology.

REFERENCES

1. V. Yevjevich, Fluctuations of Wet and Dry Years, Part I, Research Data Assembly and Mathematical Models. Colorado State University Hydrology Paper No. 1, July 1963, Fort Collins, Colorado.
2. Hurst, Black, Simaika, The Nile Basin, Vol. 7, The Future Conservation of the Nile: S.O.P. Press, Cairo, Egypt, 1946.
3. 83rd U.S. Congress, 2nd Session, House Document No. 364, Colorado River Storage Project, Chapter IV, p. 141-155. U.S. Government Printing Office, Washington, D.C. 1954.
4. V. Yevjevich, Some General Aspects of Fluctuations of Annual Runoff in the Upper Colorado River Basin. Colorado State University Publication, October 1961, Fort Collins, Colorado.
5. V. Yevjevich, Fluctuations of Wet and Dry Years, Part II, Analysis by Serial Correlation. Colorado State University, Hydrology Paper No. 4, June 1964, Fort Collins, Colorado.
6. V. Yevjevich and R.I. Jeng, Effects of Inconsistency and Non-Homogeneity on Hydrologic Time Series. Proceedings of the Fort Collins International Hydrology Symposium, Vol. I, pages 451-458.

Key words: hydrology, time series, non-homogeneity, trends, jumps in series

Abstract: The affect of inconsistency (systematic errors) and non-homogeneity in data (created either by man-made or natural changes in the environment) on hydrologic variables and time series are investigated. It is assumed that inconsistency and non-homogeneity are in the form of constant and linear jumps, linear and polynomial trends, and their subsequent combinations. The independent variables are used with superposed jumps and trends. Changes in the probability density function, including mean, variance, skewness, excess, and serial correlation coefficients are determined for various cases of jumps and trends. Inconsistency and non-homogeneity introduce the dependence into the independent time series, with the first serial correlation coefficient becoming mainly positive. As the hydrologic time series are often subject to inconsistency and non-homogeneity, a portion of the positive dependence in such a series also comes from these two factors, apart from other basic physical processes in nature.

Reference: V. Yevjevich and R.I. Jeng, Colorado State University, Hydrology Paper No. 32 (April 1969), "Properties of Non-Homogeneous Hydrologic Time Series."

Key words: hydrology, time series, non-homogeneity, trends, jumps in series

Abstract: The affect of inconsistency (systematic errors) and non-homogeneity in data (created either by man-made or natural changes in the environment) on hydrologic variables and time series are investigated. It is assumed that inconsistency and non-homogeneity are in the form of constant and linear jumps, linear and polynomial trends, and their subsequent combinations. The independent variables are used with superposed jumps and trends. Changes in the probability density function, including mean, variance, skewness, excess, and serial correlation coefficients are determined for various cases of jumps and trends. Inconsistency and non-homogeneity introduce the dependence into the independent time series, with the first serial correlation coefficient becoming mainly positive. As the hydrologic time series are often subject to inconsistency and non-homogeneity, a portion of the positive dependence in such a series also comes from these two factors, apart from other basic physical processes in nature.

Reference: V. Yevjevich and R.I. Jeng, Colorado State University, Hydrology Paper No. 32 (April 1969), "Properties of Non-Homogeneous Hydrologic Time Series."

Key words: hydrology, time series, non-homogeneity, trends, jumps in series

Abstract: The affect of inconsistency (systematic errors) and non-homogeneity in data (created either by man-made or natural changes in the environment) on hydrologic variables and time series are investigated. It is assumed that inconsistency and non-homogeneity are in the form of constant and linear jumps, linear and polynomial trends, and their subsequent combinations. The independent variables are used with superposed jumps and trends. Changes in the probability density function, including mean, variance, skewness, excess, and serial correlation coefficients are determined for various cases of jumps and trends. Inconsistency and non-homogeneity introduce the dependence into the independent time series, with the first serial correlation coefficient becoming mainly positive. As the hydrologic time series are often subject to inconsistency and non-homogeneity, a portion of the positive dependence in such a series also comes from these two factors, apart from other basic physical processes in nature.

Reference: V. Yevjevich and R.I. Jeng, Colorado State University, Hydrology Paper No. 32 (April 1969), "Properties of Non-Homogeneous Hydrologic Time Series."

Key words: hydrology, time series, non-homogeneity, trends, jumps in series

Abstract: The affect of inconsistency (systematic errors) and non-homogeneity in data (created either by man-made or natural changes in the environment) on hydrologic variables and time series are investigated. It is assumed that inconsistency and non-homogeneity are in the form of constant and linear jumps, linear and polynomial trends, and their subsequent combinations. The independent variables are used with superposed jumps and trends. Changes in the probability density function, including mean, variance, skewness, excess, and serial correlation coefficients are determined for various cases of jumps and trends. Inconsistency and non-homogeneity introduce the dependence into the independent time series, with the first serial correlation coefficient becoming mainly positive. As the hydrologic time series are often subject to inconsistency and non-homogeneity, a portion of the positive dependence in such a series also comes from these two factors, apart from other basic physical processes in nature.

Reference: V. Yevjevich and R.I. Jeng, Colorado State University, Hydrology Paper No. 32 (April 1969), "Properties of Non-Homogeneous Hydrologic Time Series."

PREVIOUSLY PUBLISHED PAPERS

Colorado State University Hydrology Papers

No. 23 "An Objective Approach to Definitions and Investigations of Continental Hydrologic Droughts," by Vujica Yevjevich, August 1967.

No. 24 "Application of Cross-Spectral Analysis to Hydrologic Time Series," by Ignacio Rodriguez-Iturbe, September 1967.

No. 25 "An Experimental Rainfall-Runoff Facility," by W. T. Dickinson, M. E. Holland and G. L. Smith, September 1967.

No. 26 "The Investigation of Relationship Between Hydrologic Time Series and Sun Spot Numbers," by Ignacio Rodriguez-Iturbe and Vujica Yevjevich, April 1968.

No. 27 "Diffusion of Entrapped Gas From Porous Media," by Kenneth M. Adam and Arthur T. Corey, April 1968.

No. 28 "Sampling Bacteria in a Mountain Stream," by Samuel H. Kunkle and James R. Meiman, March 1968.

No. 29 "Estimating Design Floods from Extreme Rainfall," by Frederick C. Bell, July 1968.

No. 30 "Conservation of Ground Water by Gravel Mulches," by A. T. Corey and W. D. Kemper, May 1968.

No. 31 "Effects of Truncation on Dependence in Hydrologic Time Series," by Rezaul Karim Bhuiya and Vujica Yevjevich, November 1968.

Colorado State University Fluid Mechanics Papers

No. 4 "Experiment on Wind Generated Waves on the Water Surface of a Laboratory Channel," by E. J. Plate and C. S. Yang, February 1966.

No. 5 "Investigations of the Thermally Stratified Boundary Layer," by E. J. Plate and C. W. Lin, February 1966.

No. 6 "Atmospheric Diffusion in the Earth's Boundary Layer--Diffusion in the Vertical Direction and Effects of the Thermal Stratification," by Shozo Ito, February 1966.

Colorado State University Hydraulics Papers

No. 1 "Design of Conveyance Channels in Alluvial Materials," by D. B. Simons, March 1966.

No. 2 "Diffusion of Slot Jets with Finite Orifice Length-Width Ratios," by V. Yevjevich, March 1966.

No. 3 "Dispersion of Mass in Open-Channel Flow," by William W. Sayre, February 1968.