THE EFFECTS OF HISTORICAL RECORD LENGTHS ON GENERATING SYNTHETIC DATA USING A STOCHASTIC MODEL OF THE MARKOV CHAIN

By

KENNETH L. PERRY

Bachelor of Science

Oklahoma State University

Stillwater, Oklahoma

1961

Submitted to the
faculty of the Graduate College of the
Oklahoma State University in partial
fulfillment of the requirements
for the Degree of
MASTER OF SCIENCE
May, 1969

THE EFFECTS OF HISTORICAL RECORD LENGTHS
ON GENERATING SYNTHETIC DATA USING A
STOCHASTIC MODEL OF THE MARKOV CHAIN

Thesis Approved:

m Namdy Bechir
Thesis Adviser

Raymond A. Mill.

Don F Kincannon

N. D. Durham
Dean of the Graduate College

ACKNOWLEDGEMENTS

The author wishes to express his sincere appreciation to his principal adviser, Dr. M. H. Bechir, for his valuable guidance and constant encouragement, without which the completion of this thesis would not have been possible.

The author is appreciative to the other members of his thesis committee, Dr. D. F. Kincannon and Dr. R. A. Mill, for their careful reading of the thesis and the valuable suggestions offered.

The author wishes to thank Mrs. Grayce Wynd for her careful and accurate typing of the manuscript.

Lastly, the author wishes to express his most sincere appreciation to his dear wife, Karen, and his son, Lyndon, who sacrificed so much, yet always endured with extended love and patience.

The author wishes to acknowledge financial support from the United States Army Corps of Engineers, and expresses his sincere gratitude for the opportunity to participate in its "Program of Advance Study for Professional Employees."

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER I

INTRODUCTION

A. General

Recent developments have made modern civilization far more dependent on water than any civilization of the past. This increased importance of water to meet the needs of an expanding population has increased the importance of water resources engineering to where there is no prospect of a decline of activity in this field in the forseeable future. Present indications are for a continued expansion in water resources engineering for many years. In fact, the increasing pressure for water is forcing the development of marginal projects which would not have been considered a few years ago. The planning for these marginal projects must be done with more care and thought, using more efficient and accurate design methods than were required for the more obvious projects of the past. This entails more accurate hydrologic methods employed in estimating available water.

The design of water resources projects is based primarily on hydrologic and economic data. A major type of hydrologic data is streamflow records. Surface streamflow data have two major uses; the first is to provide general regional information. It represents "natural" conditions and may be used in combination with similar data at other sites to gain a regional description of the streamflow of an area. The second major use is for project operation and design purposes.

The reliability of design results from streamflow data depends on errors existing in the data used. Errors in the data result from both measurement and sampling errors. Measurement errors can be controlled by expending more resources in terms of equipment and manpower so as to obtain more accurate or more frequent measurements of flow. Sampling error is a more important factor in ultimate water resources project design than error in streamflow measurement, and is usually overriding. Observed streamflow records collected over a period of time at a site provide an estimate of future occurrences at that site. The deviation of this estimate of the future from what will actually occur during a period of interest such as the economic life of a project, is primarily the result of sampling error.

B. Objectives

The primary objective of this study is to investigate the effects of length of historical streamflow records on the statistical parameters derived from it and used in a stochastic model of the Markov Chain developed by Thomas and Fiering to generate synthetic flow for a 500-year record (17). Since the statistical parameters of the population of the generated data are necessarily the same as those estimated from historical data, the new information is limited by errors of measurement and sampling that are inherent in the observed record. In short, the quality of the synthetic record is no better than the historical record from which the synthetic record was generated. This study attempts to determine the minimum length of historical record that is reliable enough to generate synthetic records which can be used to produce as many combinations of hydrologic sequences as desired for use in hydrologic analysis of reservoir operations and in the design of complex water resources systems.

C. Justification of This Research

The effects of the length of historical streamflow records on Thomas and Fiering's stochastic model of the Markov Chain is important as most streamflow records are less than 50 years in length. In order to produce synthetic streamflow data that represents the population streamflow of a given stream, a short record must be evaluated to determine if it is a representative sample.

This research sets forth a method to evaluate streamflow records, and determine their suitability for generating purposes on the basis of their length and statistical characteristics. The results of this study will provide the Arkansas-Oklahoma Compact Committee with a generating technique to (1) analyze reservoir operations, and (2) help determine the risk and uncertainty associated with the proposed investment.

D. Organization of the Research Report

In conducting this investigation, three important steps in the hydrologic analysis were performed: (a) selecting the streamflow stations under study and computing the correlation and regression coefficient for use in the generating process, (b) generating 500 years of record for each of the usable stations as the length of historical record is varied from 39 years to 15 years in length, (c) plotting five 100-year hydrographs and analyzing the results to determine the effect of varying the length of record on the generated data. The succeeding chapters of this report present these aspects.

CHAPTER II

LITERATURE SURVEY

A. Generating Techniques

When properly done, sequential generation yields hydrologic information in a form of great practical use in water resource systems for analyzing reservoir operations and design. At this point, however, the present techniques themselves must be further refined; suitable stochastic models, better than those presently proposed, must be developed; generated information must be rigorously tested statistically for precision and validity. The practical value of sequential generation is without question, but this field of study will require further research and investigation. Potentially, it has a promising future as a design tool to the hydrologist and water resources planner.

Sequential generation of streamflow data is a statistical process that applies the Monte Carlo method to generating sequentially synthetic hydrologic records. The Monte Carlo method is a process by which data can be produced synthetically by some form of random number generator.

The concept of the Monte Carlo method as applied to sequential generation of hydrologic data is not new. As early as 1914, Allen Hazen combined the annual mean flow for fourteen streams to generate a runoff sequence of 300 years (10). In 1927, C. E. Sudler used a generating technique by sampling of cards, the simplest method of generating hydrologic data, to obtain an artificial runoff record of 1000 years. He accomplished this by dealing twenty times a deck of fifty cards on each of which was printed a representative annual streamflow (16). F. B. Barnes in 1955 used a similar method, except he labeled the cards in accordance with a normal probability distribution approximating annual flows of a stream in order to provide a realistic distribution of hydrologic data. He synthesized a 1000-year sequence of streamflows using a table of random variables and assigning the synthetic flows as normal variates with the same mean and standard deviation as the historical record he used (2).

Another approach which simplifies the sampling of cards procedure is using random number tables. These tables have been subjected to standard statistical tests for randomness and are considered acceptable for general sampling use. Most of these published tables of random numbers have a rectangular distribution; however, it is possible to develop random variables of any given distribution from these tables. The extensive use of a random numbers table can be a laborious task in complicated problems. To eliminate this, mathematical programs for generating pseudorandom numbers have been developed and recorded on tapes and computer cards, which can be used as input to high speed computers. M. R. Brittan simulated streamflows in the Colorado River by selecting from a table of random numbers 100 random samples of five, each corresponding to a 5-year runoff sequence. The samples were chosen subject to the following constraints: (1) the annual runoff should have a range between the upper and lower limits set by the historical record, and (2) the 5-year sequences of runoff should be distributed according to the distribution of the mean and the ratio of the range to the mean of the historical data. Then 100 samples of thirty inflows (six samples of 5-year sequence) were chosen at random from the 100 samples. These simulated flows will exhibit statistical characteristics similar to the historical flows, as required by the constraints (4).

The assumptions necessary to use the sampling of cards method and table of random numbers are that the magnitudes of the synthetic data will be the same as those of the historical data, and that the hydrologic data is purely random. These assumptions are not realistic, and therefore have given way to better methods that are now being used.

With the theoretical work done in the field of mathematical statistics and probability in recent years, a new emphasis is placed on the subject of generating techniques. M. A. Benson used random numbers of extreme-value distribution to develop 1000 synthetic flood peaks which correspond to 1000 numbers in random order, representing annual peak flows that fit an extreme-value distribution (3). M. R. Brittan developed synthetic hydrologic records at

Lees Ferry, Arizona, on the Colorado River by two probability approaches: one by determination of the probability distribution of mean flows in relation to the range, and the other by use of a Markov Chain model (4). P. R. Julian generated synthetic hydrologic data of yearly flows on the Colorado River at Lees Ferry by means of a simple autoregressive model, or a first-order Markov Chain (11). The Markov Process or Markov Chain, as it is usually called, is based on the assumption introduced by a Russian mathematician, A. A. Markov, that the outcome of any trial such as mean monthly streamflow data depends on the outcome of the directly preceding trial such as the mean monthly streamflow data of the preceding month. This assumption led to the formulation of the classical concept of a stochastic process known as the Markov Chain. In the Markov Chain, the probability at any time of a system being in a given state depends only on the knowledge of the state of the system at the immediately preceding time.

In Julian's autoregression model, a first-order Markov process, he introduced the following equation:

$$x_t = rx_{t-1} + e(y)_t \tag{1}$$

where x_t = annual runoff at year t

x_{t-1} = annual runoff at the preceding, or the (t-1)st year

r = Markov Chain coefficient, a first order serial correlation coefficent for the runoff

$e(y)_t$ = a random uncorrelated component due to annual rainfall

He used this equation to generate annual runoff from runoff of the preceding year, and the random component due to rainfall. This is the same as a power spectrum of x_t. The power spectrum is the distribution of the variance of x_t on a frequency scale. He applied a chi-square test of fit between the actual spectrum of the Lees Ferry runoff and the generated Markov spectrum. It did not show any significant difference on the five per cent level. This test was not considered conclusive, however, because the length of record he used was not sufficient (11).

Brittan used the historical record of runoff at Lees Ferry to generate twenty sequences of 50 years each by means of the following Markov Chain model of a type similar to Equation 1.

$$x_t = rx_{t-1} + (1-r)\bar{x} + s_x(1-r^2)e \tag{2}$$

where x_t = annual runoff at year t

x_{t-1} = annual runoff at the preceding, or the (t-1)st year

$\bar{x}$ = mean annual flow computed from the historical record

s_x = standard deviation of the historical runoff

r = Markov coefficient

e = random variant assumed normally distributed with mean = 0 and standard deviation = 1

Brittan found that the generated flow contains negative values, and thus explained that this may be due to the incorrect assumption of a normal distribution for random component e.

B. Markov Chain Model Applied to This Research

Harold A. Thomas, Jr. and Myron B. Fiering used essentially the same Markov Chain model as presented in Equation (2) to synthesize 500 years of streamflow sequence or streamflow hydrograph from 32 years of observed record monthly runoffs and 6-hour flood flows of the Clearwater River and its tributaries in Idaho. In applying this model to generating monthly flows by serial correlation of monthly flows, the following equation was used:

$$Q_{i+1} = \bar{Q}_{j+1} + b_j(\bar{Q}_i - \bar{Q}_j) + s_{j+1}\,(1-r_j^2)^{\frac{1}{2}} e_i \qquad (3)$$

where Q_i and Q_{i+1} = discharges during the ith and (i+1)st month, respectively

$\bar{Q}_j$ and $\bar{Q}_{j+1}$ = mean monthly discharges during the jth and (j+1)st month, respectively

b_j = regression coefficient for estimating flow in the (j+1)st from the jth month

s_{j+1} = standard deviation of flows in the (j+1)st month

r_j = correlation coefficient between flows of the j th and (j+1)st month

e_i = random normal deviate with a zero mean and unit variance

This method of synthesizing streamflow as compared to those earlier methods used by Hazen, Sudler and Barnes, has the advantages that make possible its use for weekly, monthly, seasonal, and annual flows. It incorporates serial correlation between successive flows so as to accord with observed streamflow. It does not require that the flow data be normally distributed, and may be used with skewed distributions, as well.

C. Small Sampling Theory

Historical streamflow data is statistically considered a sample from an infinite population of streamflow. Streamflow data can be considered as a sample drawn from an infinite population of streamflow because, for practical purposes, sampling from a finite population which is very large can be considered as sampling from an infinite population. Hydrologic data are obtained by observations and by further appraisal of observed values; the hydrologic series are subject to human errors (random and systematic) and are often nonhomogenous. Random errors are always present because of the inaccuracy in measurements and observations. Systematic errors, or errors of inconsistency, refer to errors occurring in one direction, such as trends or jumps in the series. Nonhomogeneity of the data results from changes due either to natural catastrophies, such as fires, removal of forests and vegetation, landslides, etc., or to man-made developments. Any changes that would affect the basin characteristics will cause nonhomogeneity in hydrologic data such as streamflow recording. With this in mind, it is advisable to appraise the data in terms of their probable errors and nonhomogeneity before using them for statistical analysis and to consider the validity of the data in drawing conclusions concerning the reliability of the statistical parameters and relationships determined from them. This is especially true and important when the available data series represents a small sample, that is, where the sample size is smaller than about thirty to fifty observations.

It is noted that if the sample size is large enough, the sampling distributions are normal or nearly normal, and large sampling methods can be used to analyze the sample. When the samples are small, $N \leq 30$, the theory of small samples, or exact sampling theory as it is sometimes called, must be used to analyze the sample. This is because the smaller the sample size, the worse the approximation of a sample fitting a normal distribution, so that appropriate modification must be made. The results obtained from the small sampling theory holds for large samples as well as for small samples.

One important distribution that satisfies the small sampling theory is the "Student's" distribution. It was introduced W. S. Gosset (1876-1937), a student of Karl Pearson and a scientist of the Guinness firm of brewers. Karl Pearson (1857-1936) initially a mathematical physicist, spent nearly a half century in serious research on statistics. He founded the journal, Biometrika, and a school of statistics. While Pearson was concerned with large samples, large sample theory was proving somewhat inadequate for experimenters with necessarily small samples. Gosset was particularly concerned with the task of finding exact distributions of the sample standard deviation, of the ratio of the sample mean to the sample standard deviation, and of the correlation coefficient. His mathematics appeared to be insufficient for the task. Consequently he resorted to drawing shuffled cards, computing, and compiling empirical frequency distributions. Papers on the results appeared in Biometrika in 1908 under the name "Student," Gosset's pseudonym while with Guinness. Today "Student's" t is a basic tool of statiticians and experimenters, and its use is widespread. The equations for computing t to test whether or not the linear regression and linear correlation differ significantly from zero are presented later in this report.

CHAPTER III

DESCRIPTION OF THE DRAINAGE BASINS

A. Arkansas River Basin

1. Station 1645

Three gaging stations were selected for use in this study. Two are on the Arkansas River at Tulsa and Muskogee, and one is on the Illinois River near Tahlequah, Oklahoma.

Gaging Station 1645 (Figure 1) is located on the Arkansas River at Tulsa, Oklahoma, near left bank on downstream side of pier of the Eleventh Street bridge, 15.1 miles downstream from Keystone Reservoir, and at river mile 523.7. The drainage area is 74,615 square miles, and the period of record is from October, 1925, to September, 1964. Records are available through September, 1966; however, Keystone Reservoir started regulating flows in September, 1964. This would cause a nonhomogeneity in the streamflow data of the last two years. Prior regulation by John Martin Reservoir in Colorado, and Great Salt Plains Reservoir in Oklahoma are considered minor and to have little or no effect on the homogeneity of the streamflow data. This station was selected for this study because of its large drainage area and unregulated flow.

The climatological factors vary greatly, and none will

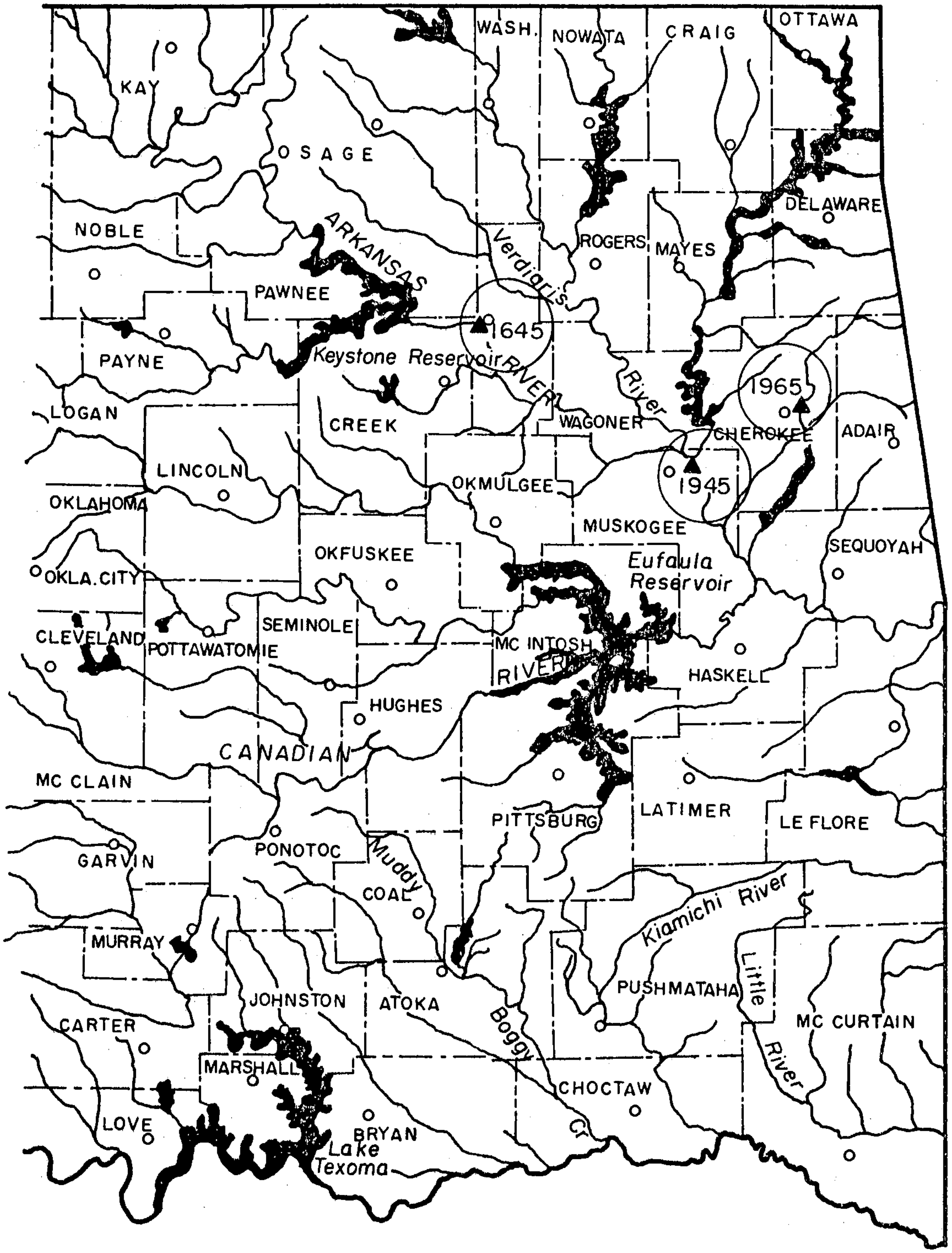

Figure 1 - Map of the area under study.

be presented in this study for Stations 1645 and 1945 because of the large drainage area and the length of the Arkansas River. The runoff for each of the two gaging stations on the Arkansas River varies from quick runoff due to the steep slopes of the Colorado mountains, to moderate runoff in Kansas and Oklahoma.

2. Station 1945

Station 1945 (Figure 1) is located on the downstream side of left pier of bridge on U. S. Highway 62, 1.7 miles downstream from Neosho River, 3.5 miles northeast of Muskogee and at river mile 457.8, 65.9 miles downstream of Station 1645. This station is influenced by an additional 22,059 square miles of drainage area (23 per cent of the total drainage area above this station) that is made up of two subbasins, one drained by the Verdigris River and the other by the Grand Neosho River. These two subbasins have highly regulated flows as the result of the multiple-purpose reservoirs in operation on each subbasin. This station was selected for this study because of its large drainage area, and it is the first gaging station upstream of the Oklahoma-Arkansas Compact Water Resources study area. It is hoped that results from this investigation can be utilized by the Oklahoma-Arkansas Compact Committee.

B. Illinois River Basin

The third station used in this study is Station 1965 (Figure 1) on the Illinois River, 2.2 miles northeast of Tahlequah, Oklahoma, 6.5 miles upstream from Barren Fork, and at river mile 55.8. The drainage area is 959 square miles, 58 per cent of the total subbasin, and the period of historical record is from October, 1935, to September, 1966. This station was selected on the basis of its size, and the fact that it has a good record of historical streamflow data with no regulated flow. This subbasin originates in northwest Arkansas as Osage Creek, and flows westward until it meets with Muddy Fork, which in turn drains Clear and Goose Creeks. The Muddy Fork system drains the southern portion of the tributary area of the Illinois River in the State of Arkansas, while Osage Creek and the upper reaches of Flint Creek drain the northern portion of the tributary area. The Illinois River then crosses the Oklahoma-Arkansas state line, and continues running westward. It drains tributaries such as Wedington Creek and Ballard Creek. After Flint Creek joins the Illinois River, the river flows in a southerly direction into the Tenkiller Ferry Reservoir located downstream of Station 1965. The major tributaries joining the river in this reach are Barren Fork and Caney Creek. After leaving the Tenkiller Ferry Reservoir, the Illinois River flows southward for a distance of approximately seven miles and drains into the Arkansas River just upstream of the proposed location of the Robert S. Kerr Lock and Dam.

The Illinois River subbasin is largely mountainous or hilly terrain with numerous tributary streams. This mountainous terrain is typical of most of the subbasin. Rocky, impervious soils and steep slopes of the tributary drainage areas indicate quick runoff.

Rainstorms over the basin are normally of long duration and high intensity; storms occur frequently in the spring, late fall, and winter months. The normal annual precipitation over the basin averages about forty-four inches.

TABLE IV

METHODS AND PROCEDURES

A. Method of Generating Synthetic Monthly Flows

1. Assumptions

Rarely does one have as many as fifty years of recorded observations of streamflow data for which the hydrologic system has been stable, and even with a record of that length it is quite probable that it lacks a critical sequence of years of low and high runoff. If the more severe droughts and floods on record are not representative of the statistical population, it becomes obvious that any design based on this data would be distorted.

A short historical record may not identify the true frequency of years or seasons of unusually low or high flows, but will usually provide a fairly precise estimate of mean annual and mean seasonal flows and their variances. These statistical parameters make it possible to construct a stochastic model that will generate synthetic flow sequences for as long a period of time as desired. Since most large water resource systems such as multi-purpose reservoirs are designed with a 100-year economic life, 500 years of mean monthly synthetic streamflow was generated to provide five replicate streamflow sequences of 100 years each. Synthetic hydrographs from alternate 100-year periods would enable project planners to test a given system design more exhaustively and with less chance of a distorted design than would be possible with the observed flows alone.

It is the function of the stochastic model to create synthetic patterns of low and high runoff that are probably not included in brief records of historical streamflow but that, based on statistical considerations, would be expected to be a part of an actual record of sufficient length. The synthetic flows should have peaks and lows that are higher and lower than those in the historical record. If the stochastic model of flow sequences is suitable in all respects, it should be impossible to distinguish real and synthesized hydrographs by the usual statistical test of significance.

2. Correlation Analysis

The stochastic model of the Markov Chain used by Thomas and Fiering proves to be a satisfactory model to generate synthetic flows, as it satisfies the assumptions mentioned above. To use the Markov Chain model by Thomas and Fiering, an introduction to the application of serial correlation to streamflow data is necessary. The term "serial correlation" connotes a month-to-month relationship associated with seasonal fluctuations of discharge. This relationship, in turn, induces a small amount of year-to-year serial correlation in the synthesized flows, which accords with that found in the observed flows.

To introduce the mathematical correlation, let X_i and Y_i represent numerical values corresponding to successive pairs of mean monthly flows, and let there be N such pairs. For each value of X_i, an estimate of the most probable value of the associated Y_i value is needed. Let $\bar{X}$ and $\bar{Y}$ be the observed arithmetic means of the respective concurrent measurements. If a linear-regression relationship is assumed and the method of least squares gives the best estimate, $\hat{Y}$, of the Y_i value corresponding to a given X_i value, then the following equation gives this relationship.

$$\hat{Y} = \bar{Y} + b(X_i + \bar{X}) \qquad (4)$$

This equation represents a straight line that best fits a plot of several pairs of X_iY_i values. This line must satisfy the least square criterion that requires the sum of the squares of the deviations, $\sum(Y_i - \bar{Y})^2$, of the observed points from a straight line moving average be a minimum. For such a "fitted" line, b is called the regression coefficient.

To determine the regression coefficient, b, requires the sum of cross products of the deviations of the observations from their corresponding means, and the sum of squares of X_i. The computing formula for the regression coefficient is given as follows:

$$b = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{X^2 - \frac{(\sum X)^2}{N}} \qquad (5)$$

This equation represents the slope of a line represented by Equation (4).

If the assumption that N values of X_i and Y_i observations is a sample from bivariate normal distribution, then the following relationship can be presented (17):

$$r = b \frac{S_x}{S_y} \tag{6}$$

where r = correlation coefficient

b = regression coefficient

S_x & S_y = standard deviations for X_i and Y_i, respectively

The correlation coefficient, in this case more properly called the "product-moment" correlation, is a means of presenting in numerical terms the measure of the degree to which variables vary, or a measure of the intensity of association. Equation (6) can be presented in a more useful form for computation purposes.

$$r = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sqrt{\left(\sum X^2 - \frac{(\sum X)^2}{N}\right)\left(\sum Y^2 - \frac{(\sum Y)^2}{N}\right)}} \tag{7}$$

The correlation coefficient, r, will have a value between -1 and +1. When $r = \pm 1$, there is perfect linear correlation between the variables X_i and Y_i, and when $r = 0$, there is no linear correlation between X_i and Y_i.

If the observed variances of the Y_i and X_i values are denoted by $S_y^2 = (1/N) \sum (Y_i - \bar{Y})^2$, and $S_x^2 = (1/N) \sum (X_i - \bar{X})^2$, respectively, the standard error of estimate $S_{y \cdot x}$, of the Y_i

values is defined by $S_y(1 - r^2)^{\frac{1}{2}}$, which is a measure of the random or unexplained variation of the Y_i (17). The proportion of the total variance of Y_i that can be attributed to variation of the X_i values is equal to r^2. Thus, for $r = 0$, $S_y(1-r^2)^{\frac{1}{2}} = S_y$, there is no explained variance; and for $r = \pm 1$, $S_y(1 - r^2)^{\frac{1}{2}} = 0$, and there is no unexplained variance. For all intermediate values of r, the observed values of Y_i are distributed about $\hat{Y}$, with a closeness of grouping about the regression line related to r^2. The quantity denoted by r^2 is called the coefficient of "determination."

3. Introduction of a Random Component

If we further assume that the Y_i values for a given X_i value have a probability distribution about $\hat{Y}$, Equation (4) may be rewritten to include a random component as follows:

$$\hat{Y}_R = \bar{Y} + b(X_i - \bar{X}) + S_y(1 - r^2)^{\frac{1}{2}}e_i \qquad (8)$$

$\hat{Y}_R$ denotes $\hat{Y}$ with a random component added, and e_i is a standardized random, normal, and independently distributed variate with zero mean and unit variance. This random component has the effect of adding to $\hat{Y}$ in Equation (4) a positive or negative component that exceeds in magnitude the band width of one standard error, $S_{y \cdot x}$, 31.7% of the time and the band width of two standard errors 4.5% of the time (12). It should be pointed out that the distribution of the e_i values is such that parameters computed from a sample of many estimates of the type in Equation (8) will not differ significantly from those in Equation (4).

4. Stochastic Model

Equation (3) represents a bivariate stochastic model of Equation (8) for unit time intervals of months. For serial correlation of monthly flows at each of the three stations investigated, the assumption is that the statistical population of streamflows at each station was normally distributed and that the twelve population correlation between successive pairs of months at each station were the same as those calculated from the sample of historical flows for each station. The variation in sign and magnitude of the random additive component makes for a continuous, unbounded, and serially correlated sequence of data for use in simulation studies.

B. Method of Testing Statistical Parameters

Two statistics used in the stochastic model that need to be tested for their statistical significance are the regression coefficient and correlation coefficient. Since streamflow records are rarely more than fifty years in length, it will be necessary to test for significance using the "small sample theory."

To test a theoretical population coefficient of correlation denoted by P, which is estimated by the sample correlation coefficient r, to see if P equals zero, a statistic involving Student's distribution can be used. This statistic is defined as follows:

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}} \qquad (9)$$

and has N - 2 degrees of freedom, where N is the number of observations. This value is compared with a "Student's t" table at a chosen level of significance and if the calculated t is a larger value, then the hypothesis that P equals zero can be rejected. In other words, it can be said that the sample of X_iY_i pairs was selected from a correlated population of X_iY_i pairs with a certain probability.

The regression coefficient, b, is computed by using Equation (5). To test the hypothesis, a sample drawn from a population that has a correlation coefficient, B, equal to zero, use the following statistic:

$$t = \frac{b - B}{\sqrt{\frac{S^2_{y.x}}{\sum(X_i - \bar{X})^2}}} \qquad (10)$$

This statistic has Student's distribution with N - 2 degrees of freedom. This value is also compared with the value found in the "Student's t" table for a chosen level of significance. If the computed t is larger than the value in the "Student's t" table for a given level of significance, such as 0.10 level using a 2-tail test, then the hypothesis that b equals B where B equals zero is rejected.

C. Procedure for Computing Synthetic Data

Equation (3) is a stochastic model in which the discharge in the (i+1)st month is composed of a component linearly related to that in the ith month and a random additive component. With a table of normal random deviates provided in most statistics books or on computer tapes and calculated statistics of monthly flows ($\bar{Q}_j$, $\bar{Q}_{j+1}$, S_j, S_{j+1}, b_j, and r_j), the computation of Q_{i+1} is a straightforward matter of arithmetic. It is only since the advent of computers that such stochastic models can be used to analyze hydrologic data due to the long tedious task involved in the computations.

A separate computer program using Fortran IV scientific language on an IBM computer was used to compute the required statistics of monthly streamflow and the t statistic mentioned above. The t statistic computed in this program is represented by Equation (10) and is used to determine if the coefficient of regression is at the selected 0.10 level of significance. This was done for three gaging stations, two of which were on the Arkansas River at Tulsa and near Muskogee, and one on the Illinois River near Tahlequah. The stations on the Arkansas River had consecutive thirty-nine years of historical streamflow data that were usable, and the station on the Illinois River had thirty years of consecutive streamflow data. This procedure was repeated for each station, varying the length of historical streamflow data for the two stations on the Arkansas River to provide statistical values for each station for 39 years, 35 years, 25 years, and 15 years of historical streamflow records. The historical streamflow record for the station on the Illinois River near Tahlequah was varied from 30 years to 25 years and 15 years of length, and statistical values were computed to apply the stochastic model for the different record lengths.

This program did not provide the t statistic presented by Equation (9) that was used to test the level of significance of the correlation coefficients. A separate program was written for this purpose and is presented in Appendix A. Station 1965 on the Illinois River was dropped from further study after an evaluation was made on the correlation coefficients. A further discussion of this is presented in Chapter VI, which discusses the results of this investigation. The remaining two stations on the Arkansas River had significant correlations at the 10 per cent level for the 35th and 39th year record lengths, and Equation (3) was applied to generate 510 years of synthetic monthly streamflow data. The 25-year record for Station 1645 on the Arkansas River also had significant correlation at the 10 per cent level. The extra ten years were generated to assure a random start at the beginning of the first 100 years' sequence. The generation of the synthetic data was accomplished using the program presented in Appendix B.

Some negative flows were experienced using this stochastic model. The flows were subsequently set to zero, and hydrographs of mean annual flows for five 100-year periods for each recorded period were plotted for each station using the computer program presented in Appendix C. A further discussion on negative flows is presented in the results section.

CHAPTER V

RESULTS

A. Gaging Station 1965

This gaging station, located on the Illinois River near Tahlequah, Oklahoma, was selected to evaluate the effect of the length of available historical record on the correlation coefficients at the ten per cent significance level because it typified the small drainage subbasins in the study area, and has 30 years of historical streamflow record that provides homogenous flow characteristics needed for using the Markov Chain model. This station gages the discharges from 959 square miles of a subbasin that has a mountainous terrain with steep slopes that indicate quick runoff. The average discharge at this station over the 30 years of record is 847 cfs. The maximum discharge of 150,000 cfs occurred on May 10, 1950, and a minimum daily discharge of 0.1 cfs occurred October 10-14, 1956.

This station has a period of record covering the water years 1937 up to the present time (30 water years). The hydrograph of the historical mean annual flows is presented in Figure 2. Twelve correlation coefficients for consecutive months were computed using a computer program for the first 15 years of the record (1937-1951), then the next

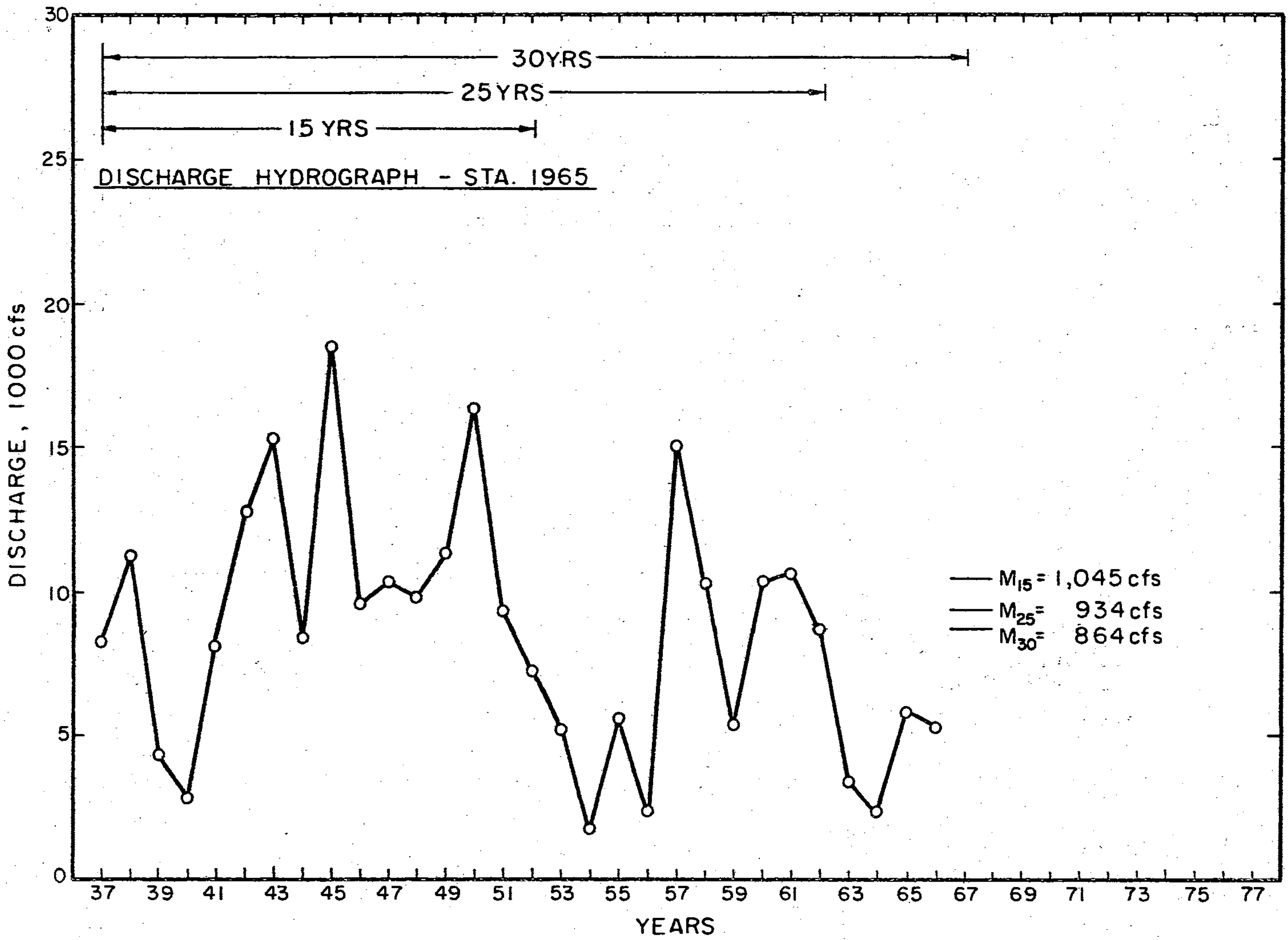

Figure 2 - Station 1965, discharge hydrograph.

10 years of record was added and another set of twelve correlation coefficients for 25 years of record (1937-1961) was computed. The last set of twelve correlation coefficients was computed for the entire length of record of 30 years (1937-1966). Correlation coefficients and their level of significance at the ten per cent level for 15, 25, and 30 years are shown in Table I. Due to the poor correlation coefficients of the consecutive months, regardless of the length of record utilized, it was decided not to synthesize flows nor plot mean annual hydrographs for this station, since these results would have been meaningless. This was also considered as proof of the unsuitability of the drainage area to be treated by the Markov Chain Model for monthly intervals. These points are discussed in greater detail in the discussion section.

B. Gaging Stations 1645 and 1945

Stations 1645 and 1945, located on the Arkansas River at Tulsa and near Muskogee, respectively, were selected to evaluate the effect of the length of available historical record on the correlation coefficients at the ten per cent level. The results of the investigation of these two stations are presented together as they both have a period of record covering the same span of time for water years 1926 up to the present time (41 water years) and both have large drainage areas. Station 1645 has a drainage area of 74,615 square miles, and Station 1945 has 96,674 square miles.

Station 1645 has a 41-year average discharge of

TABLE I

CORRELATION COEFFICIENTS FOR VARIOUS LENGTHS OF RECORD AT STATION 1965

Period Years	15	25	30
Correlation Coefficient at 10% Level of Significance	0.4409	0.3355	0.3060
Oct-Nov	0.56752	0.61912	0.61668
Nov-Dec	0.76219	0.74456	0.75117
Dec-Jan	-0.00364	0.12530	0.16005
Jan-Feb	0.01992	0.21030	0.23432
Feb-Mar	0.26205	0.32683	0.35072
Mar-Apr	0.67429	0.58270	0.57444
Apr-May	-0.12087	0.09357	0.10965
May-June	0.16945	0.42948	0.49094
June-July	0.41809	0.09830	0.19848
July-Aug	0.30303	0.37792	0.39033
Aug-Sept	0.20831	0.34706	0.35254
Sept-Oct	0.04313	0.11671	0.15825

6,504 cfs with a maximum discharge of 246,000 cfs occurring October 12 and 13, 1956. Station 1945 has a 41-year average discharge of 19,570 cfs with a maximum discharge of 700,000 cfs on May 21, 1943, and a minimum discharge of 66 cfs on October 9, 1956. The flood of May 21, 1943, is the greatest known since June, 1833, when a similar stage was probably reached. The record for this station is considered good, however, natural flow of 23 per cent of the total drainage area is affected by storage reservoirs and power development.

Only 39 years of record were utilized for the purposes of this study since Keystone Reservoir started regulation of flow on the Arkansas River in the latter part of September, 1964. Twelve correlation coefficients for consecutive months were computed for the first 15 years of the record (1926-1940). The next 10 years of record were added and a set of correlation coefficients was computed for 25 years of record (1926-1950). The next 10 years of record were added and a set of correlation coefficients was computed for 35 years of record (1926-1960). The last set of correlation coefficients was computed for the entire length of usable record of 39 years (1926-1964). The correlation coefficient and the desired ten per cent level of significance are presented in Table II.

To determine acceptability of the correlation coefficients for use of the Markov Chain model, the level of significance for each correlation coefficient was determined by using the calculated t values from Equation (9). The results were plotted on semi-log paper for Stations 1645 and 1945 for the 15, 25, 35, and 39 years of

TABLE II

CORRELATION COEFFICIENTS FOR VARIOUS LENGTHS OF RECORD AT STATIONS 1645 and 1945

Station	1645				1945			
Period, Years	15	25	35	39	15	25	35	39
10% Level of Significance	0.4409	0.3379	0.2831	0.2676	0.4409	0.3379	0.2831	0.2676
Oct-Nov	-.03509	.34322	.33552	.32479	.10865	.61787	.56560	.54206
Nov-Dec	.90163	.47767	.55181	.60224	.70077	.40368	.50864	.62462
Dec-Jan	.91844	.37445	.49788	.52731	45450	.29301	.43090	.44623
Jan-Feb	.69351	.83082	.83961	.84668	.45421	.55068	.60773	.62505
Feb-Mar	.63616	.56781	.55380	.54870	.43011	.29562	.33840	.35271
Mar-Apr	.74500	.58092	.47195	.47315	.60955	.59527	.53164	.53548
Apr-May	.23904	.28382	.23414	.24796	.34709	.13751	.17923	.19569
May-June	.68014	.56050	.72074	.77489	.52696	.30207	.49927	.47601
June-July	.71603	.29830	.56343	.56744	.62464	.25329	.30479	.31489
July-Aug	.03388	.48636	.34436	.34982	.19846	.54880	.38061	.39340
Aug-Sept	.29851	.37110	.37507	.34494	.14072	.28740	.34179	.33761
Sept-Oct	.57666	.51073	.45655	.43947	.87531	.69074	.51484	.47727

historical record, and are presented in Figures 3 and 4 for each station, respectively. The values below ten per cent significance level for the paired months were acceptable, and those above did not meet the specified criteria. A comparison of the mean monthly flows of the first 15 years, 25 years, 35 years, and 39 years of historical record and that of the synthesized flows based upon 15 years, 25 years, 35 years, and 39 years of record for both Stations 1645 and 1945 are shown in Tables III and IV, respectively.

The computed values for the other statistics needed for the Markov Chain model (regression coefficients, standard deviations, and standard error of estimates) for stations 1645 and 1945 for 15, 25, 35, and 39 years of historical record are presented in Tables V, VI, and VII, respectively.

Hydrographs of the historical mean annual flows at Stations 1645 and 1945 are presented in Figures 5 and 6, respectively. The mean for each historical record length of 15, 25, 35, and 39 years are presented to show that the shorter record lengths are not long enough to provide an adequate cycle of low and high mean annual flows needed to stabilize the true mean annual flows.

Synthetic hydrographs of 500 years for Station 1645 for 25, 35, and 39 years of record are presented in Figures 7 through 9. Synthetic hydrographs for Station 1945 for 35 and 39 years of record are presented in Figures 10 and 11. They show a favorable comparison between the historical and

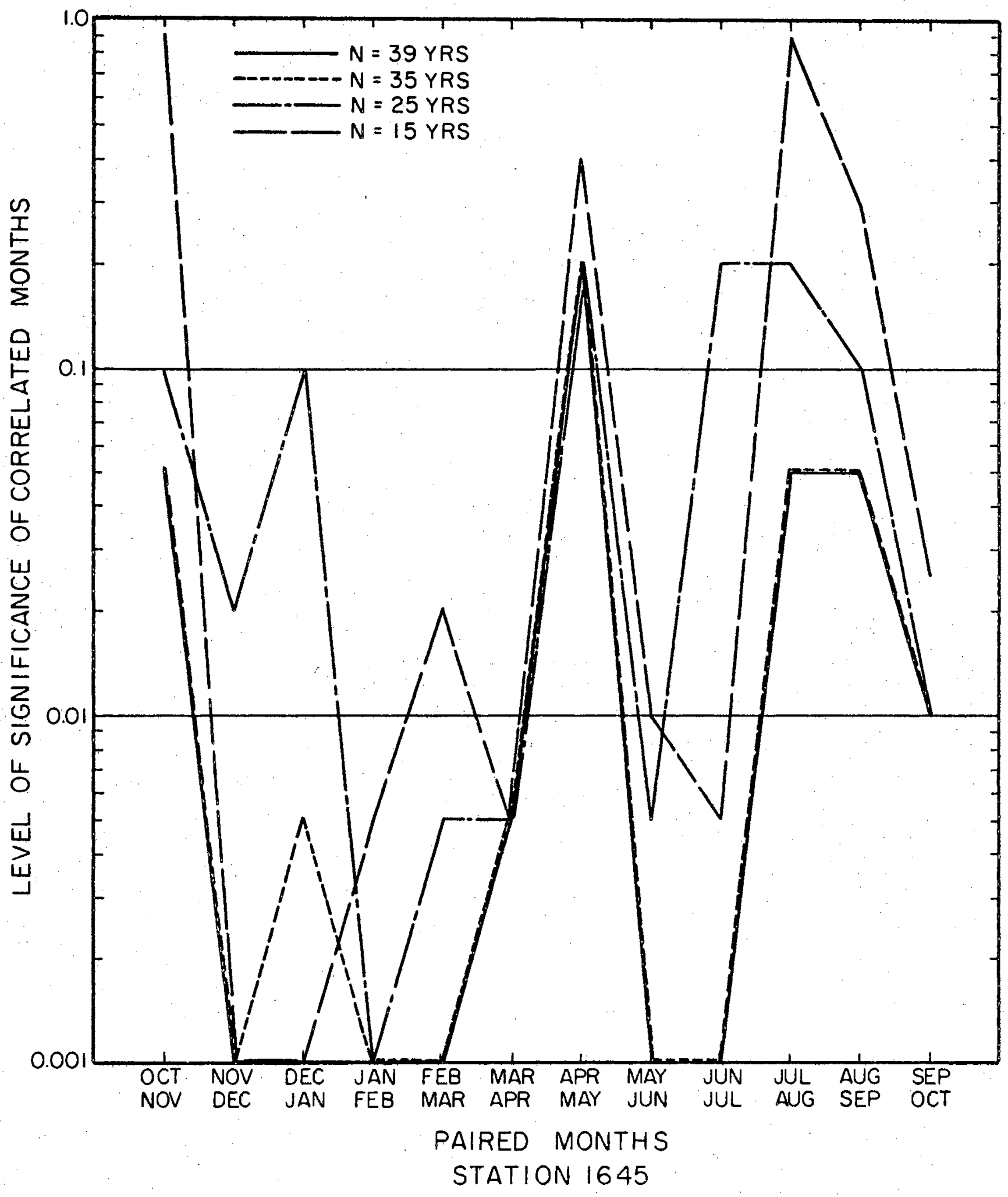

Figure 3 - Station 1645, correlation coefficients and level of significance for various lengths of record.

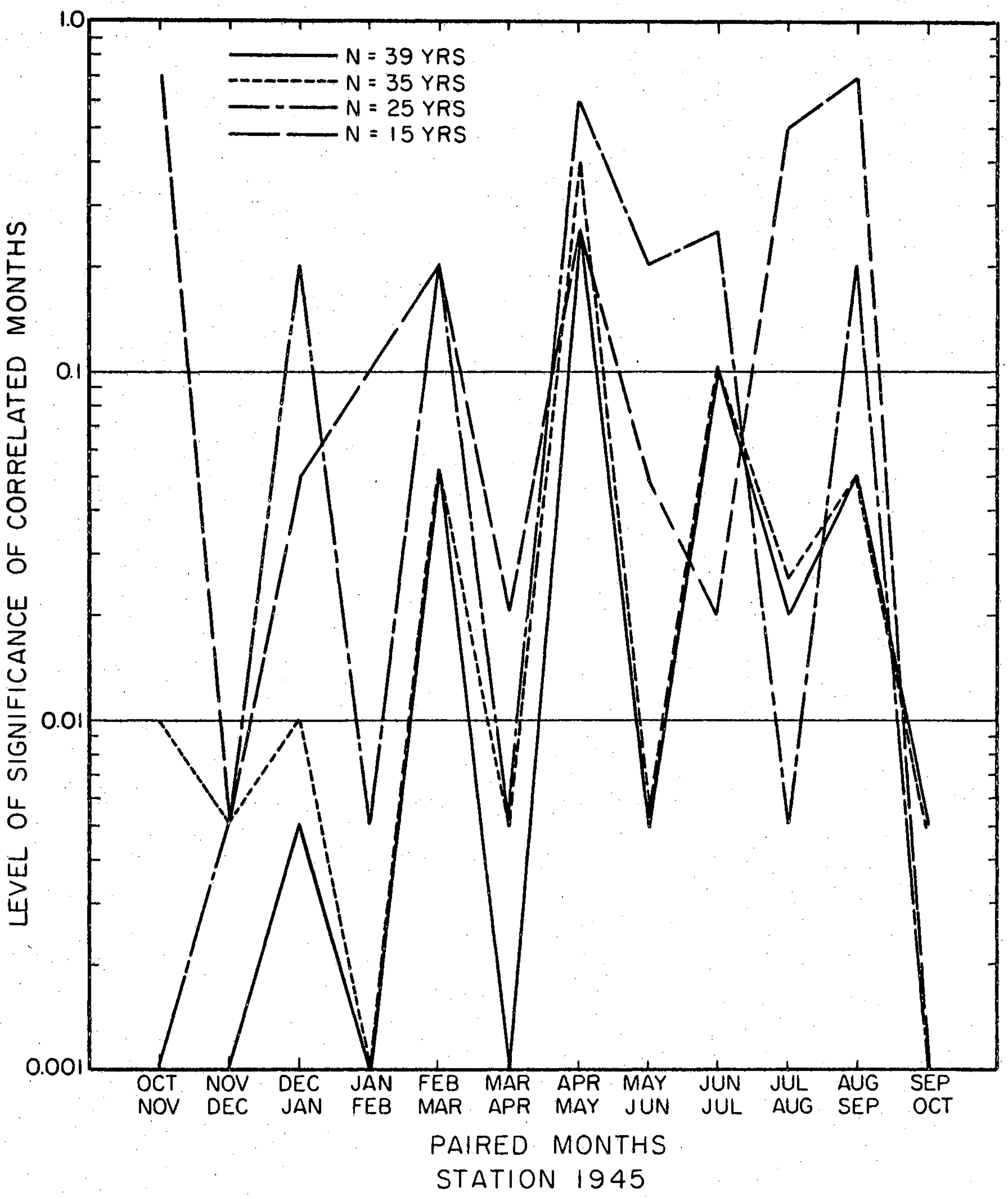

Figure 4 - Station 1945, correlation coefficients and level of significance for various lengths of record.

TABLE III

MEAN MONTHLY FLOW, cfs, FOR STATION 1645

Month	39 Years		35 Years		25 Years		15 Years	
	Historical	Synthetic (500 yr)	Historical	Synthetic (500 yr)	Historical	Synthetic (500 yr)	Historical	Synthetic (500 yr)
Oct	6674	6584	6572	6611	5574	5609	3806	4123
Nov	3706	3921	3118	3075	3342	3629	2975	3064
Dec	2782	2884	2615	2594	2849	3114	2000	2278
Jan	2561	3545	2526	3410	2730	4092	1934	1995
Feb	3275	3526	3141	3432	3306	3750	1753	1947
Mar	4222	7327	4278	7379	4035	8138	2344	5192
Apr	9270	11746	9808	12108	11281	13017	6720	8092
May	13423	14674	13774	14514	12743	12895	9416	10215
June	12754	13846	13107	13491	12242	11997	11514	11559
July	8991	9273	9169	7086	7750	4638	5656	
Aug	5407	5850	5439	5674	5911	6171	4782	4707
Sept	5490	7678	4852	7896	4679	6500	3830	4764

TABLE IV

MEAN MONTHLY FLOW, cfs, FOR STATION 1945

	39 Years		35 Years		25 Years		15 Years	
Month	Historical	Synthetic (500 yr)	Historical	Synthetic (500 yr)	Historical	Synthetic (500 yr)	Historical	Synthetic (500 yr)
Oct	19365	20251	19401	19478	19570	20759	14289	14025
Nov	12775	12009	11686	11594	13058	12950	10880	10813
Dec	9154	9807	8690	9281	9885	10647	8974	9908
Jan	9917	10916	10136	11824	12044	13766	10958	11365
Feb	12073	13062	12269	14020	13901	15137	10662	11163
Mar	15581	23940	15956	26662	15854	29818	11312	24877
Apr	32075	37849	33970	38877	40155	43808	30480	32541
May	40520	44185	40728	43199	42603	45603	31167	37210
June	36998	39544	38448	40354	41152	41039	40331	40275
July	23400	23773	24092	24971	18542	21291	11222	15674
Aug	12663	14329	12825	13737	14323	14812	11655	12556
Sept	14420	21920	12751	19158	13579	19622	11116	17116

TABLE V

REGRESSION COEFFICIENTS FOR VARIOUS LENGTHS OF RECORD AT STATIONS 1645 AND 1945

Station	1645				1945			
Period, Years	15	25	35	39	15	25	35	39
Oct-Nov	-0.01720	0.13725	0.08953	0.13759	0.04457	0.34150	0.26995	0.29692
Nov-Dec	0.47871	0.46802	0.51304	0.37517	0.51275	0.19507	0.25526	0.30744
Dec-Jan	1.12346	0.37445	0.45669	0.46057	0.80680	0.35471	0.51488	0.48089
Jan-Feb	0.40330	1.68120	1.65469	1.67180	0.37026	0.86763	0.91469	0.94474
Feb-Mar	0.95968	0.35288	0.45212	0.44000	0.50356	0.27635	0.36171	0.37105
Mar-Apr	4.15717	2.32507	1.41450	1.42206	2.97324	1.99363	1.52534	1.53405
Apr-May	0.22990	0.22991	0.26087	0.28710	0.22707	0.12897	0.18046	0.21525
May-June	0.63177	0.41979	0.72074	0.66974	0.72923	0.22666	0.43931	0.38300
June-July	0.34321	0.25542	0.52191	0152271	0.14525	0.16427	0.27528	0.28367
July-Aug	0.49894	0.57854	0.23088	0.23429	0.39712	0.52611	0.20571	0.21298
Aug-Sept	0.11046	0.15378	0.21347	0.24720	0.07536	0.16536	0.23307	0.28674
Sept-Oct	1.30251	1.11480	1.17394	0.90383	1.95990	1.71414	1.32028	0.98255

TABLE VI

STANDARD DEVIATIONS FOR VARIOUS LENGTHS OF RECORD AT STATIONS 1645 AND 1945

Station	1645				1945			
Period, Years	15	25	35	39	15	25	35	39
Oct-Nov	6356	7899	11233	10767	22952	29482	31380	30056
Nov-Dec	3114	3159	2997	4559	9415	16295	14977	16458
Dec-Jan	1654	3095	2787	2840	6889	7874	7516	8101
Jan-Feb	2021	2792	2559	2480	9827	9532	8981	8730
Feb-Mar	1176	5649	5043	4898	8010	15018	13519	13195
Mar-Apr	1774	3511	4117	3928	9378	14039	14450	13882
Apr-May	9910	14063	12346	11811	45748	47020	41460	39770
May-June	9532	11392	13756	13675	29928	44100	41731	43747
June-July	8854	8532	12324	11819	41416	33091	36719	35200
July-Aug	4244	7306	11416	10887	9630	21461	33165	31709
Aug-Sept	7639	8690	7654	7292	19271	20573	17925	17167
Sept-Oct	2827	3601	4356	5291	10320	11837	12223	14750

TABLE VII

STANDARD ERROR OF ESTIMATE FOR VARIOUS LENGTHS OF RECORD AT STATIONS 1645 AND 1945

Station	1645				1945			
Period, Years	15	25	35	39	15	25	35	39
Oct-Nov	3230	3031	2866	4369	9712	13088	12537	14016
Nov-Dec	742	2777	2359	2298	5100	7359	6569	6411
Dec-Jan	830	2595	2253	2136	8410	9310	8226	7917
Jan-Feb	879	3212	2780	2641	7406	12806	10896	10438
Feb-Mar	1421	2952	3480	3328	8786	13700	13802	13164
Mar-Apr	6854	11684	11042	10540	37634	38593	35642	34038
Apr-May	9605	11158	13574	13426	29127	44620	41672	43477
May-June	6735	7218	7430	7571	36528	32224	32294	31371
June-July	3074	7123	9573	9085	7804	21207	32062	30500
July-Aug	7922	7757	7294	6923	19600	17568	16825	15994
Aug-Sept	2799	3416	4099	4971	10603	11582	11660	13908
Stp-Oct	5413	6903	10116	9908	11596	21698	27276	27053

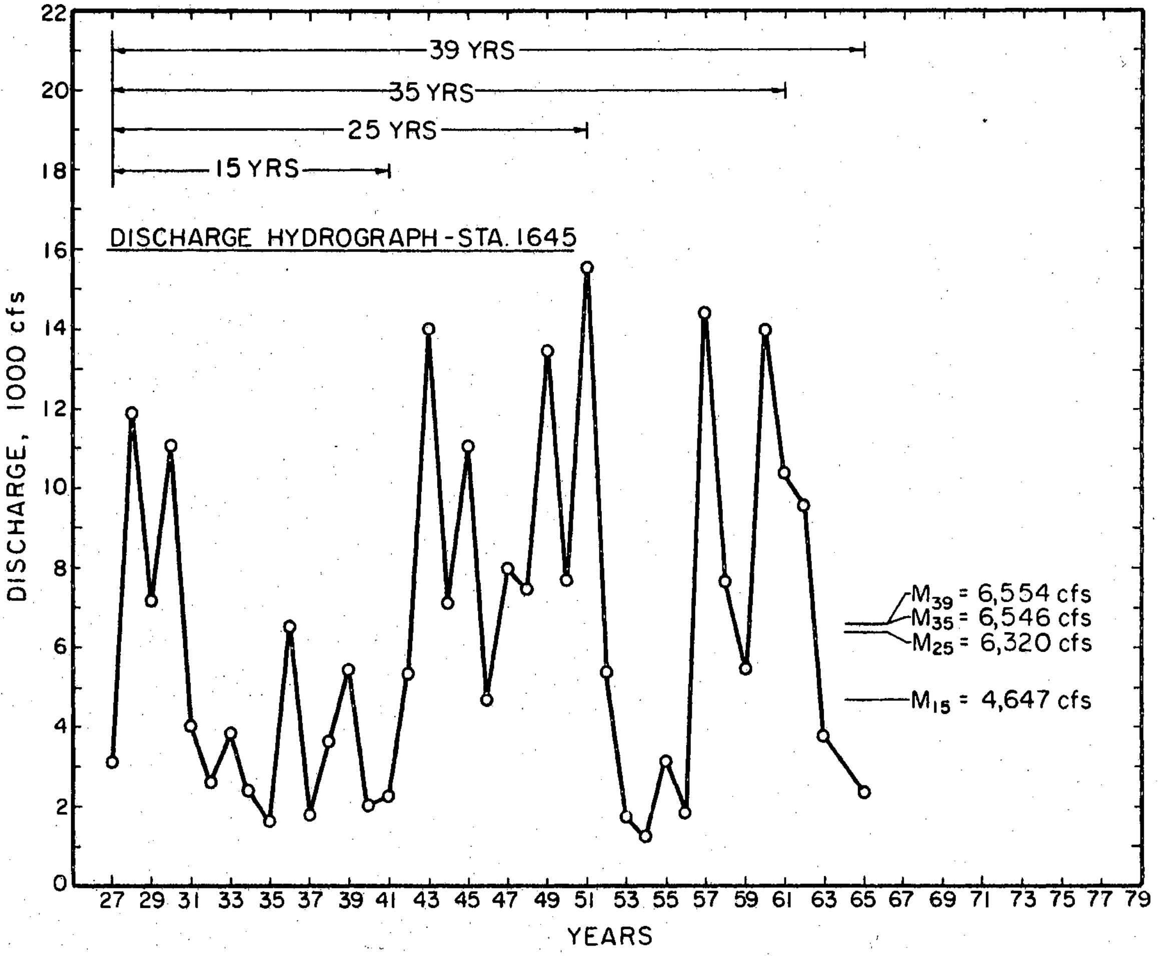

Figure 5 - Station 1645, discharge hydrograph.

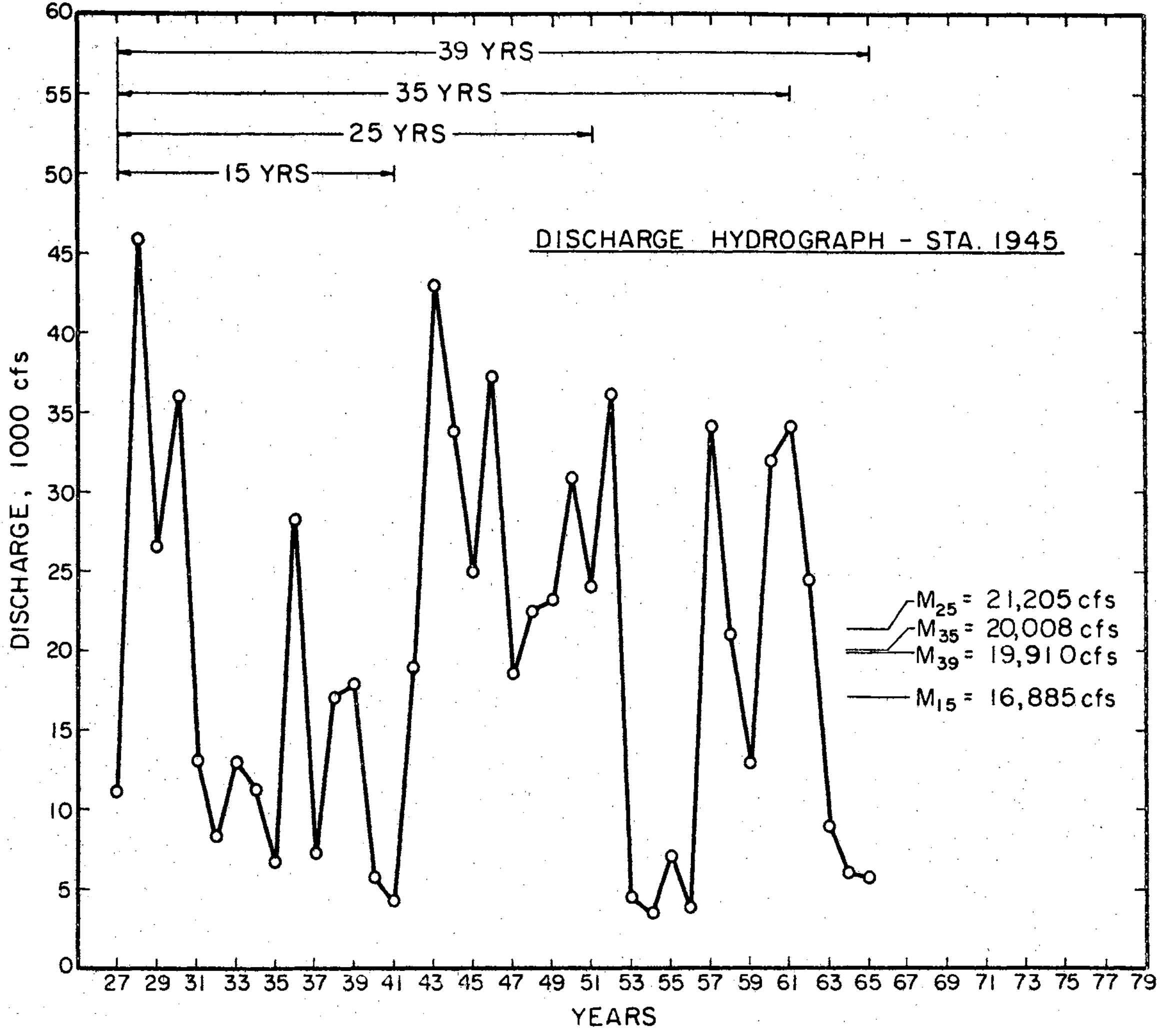

Figure 6 - Station 1945, discharge hydrograph.

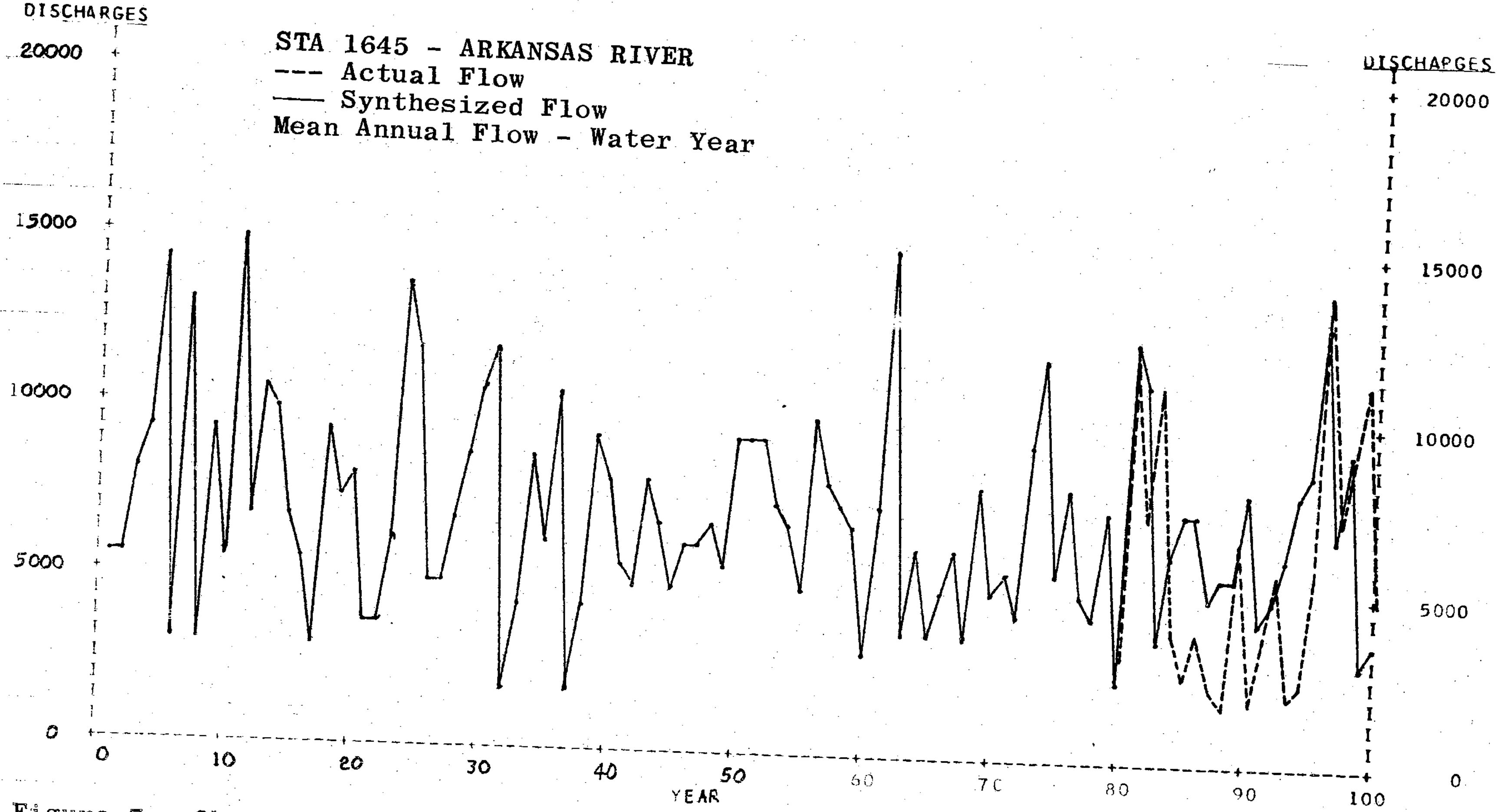

Figure 7 - Observed and synthesized flow in cfs for 25 years of record.

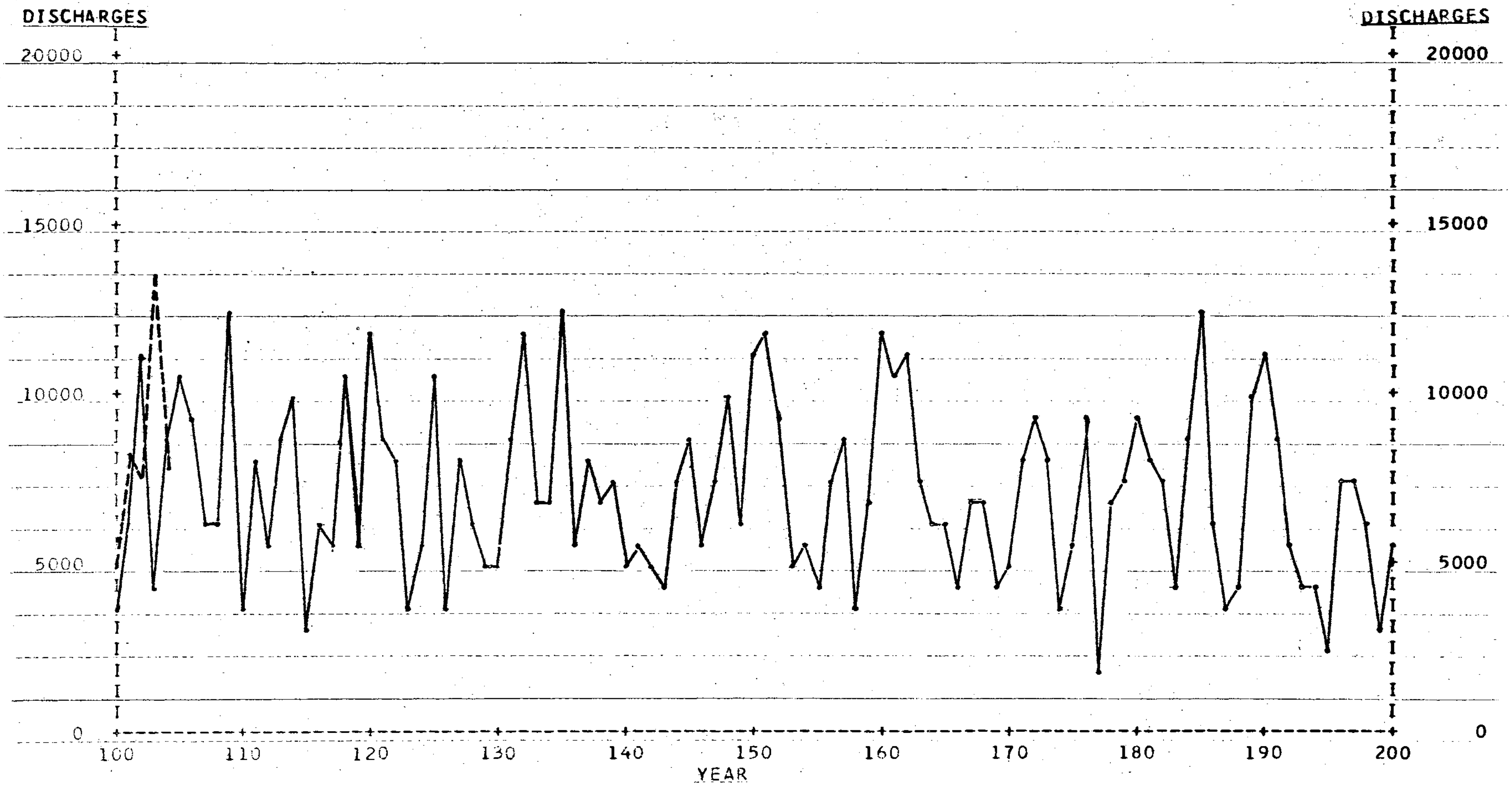

Figure 7 (continued)

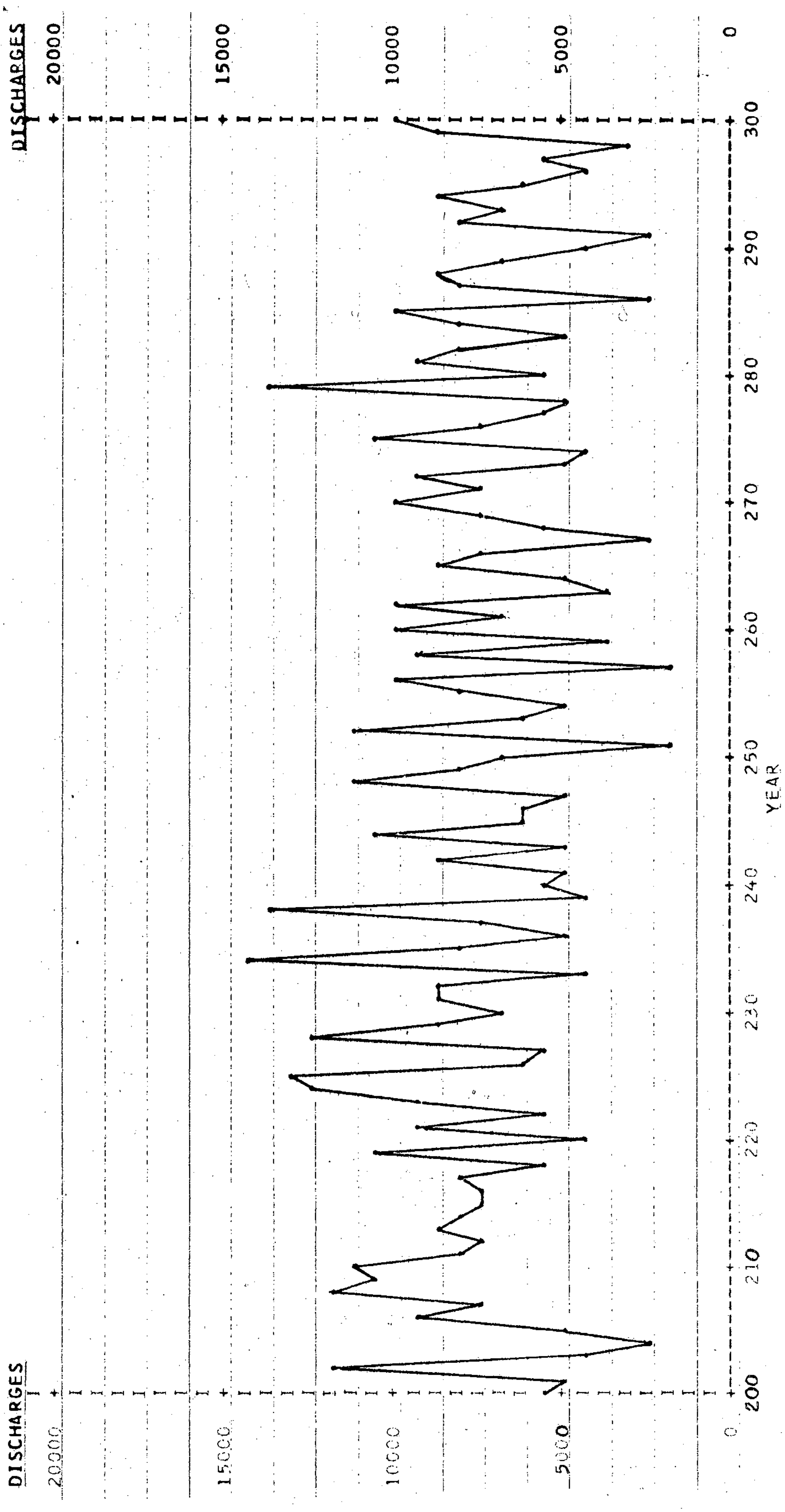

Figure 7 (continued)

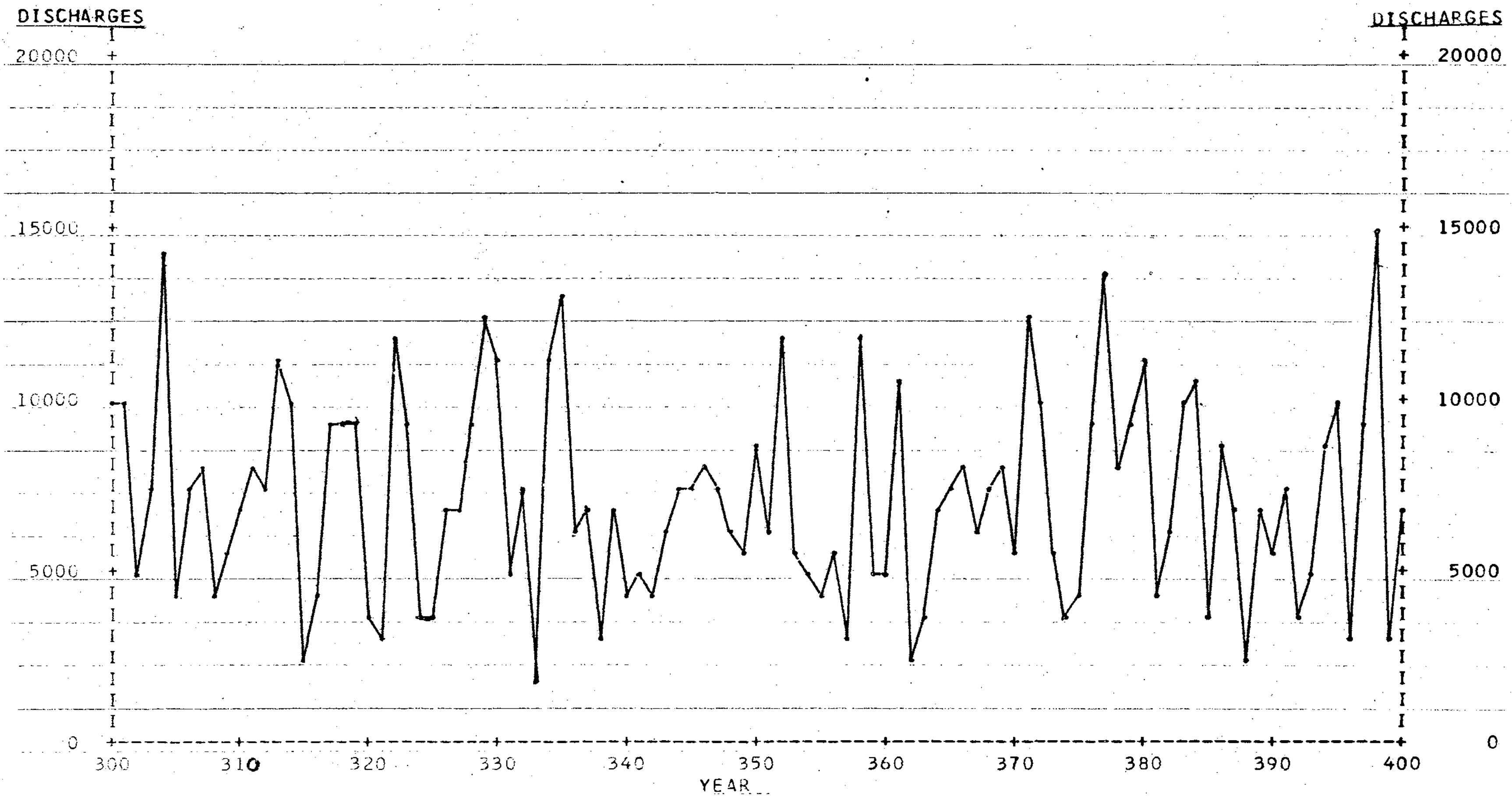

Figure 7 (continued)

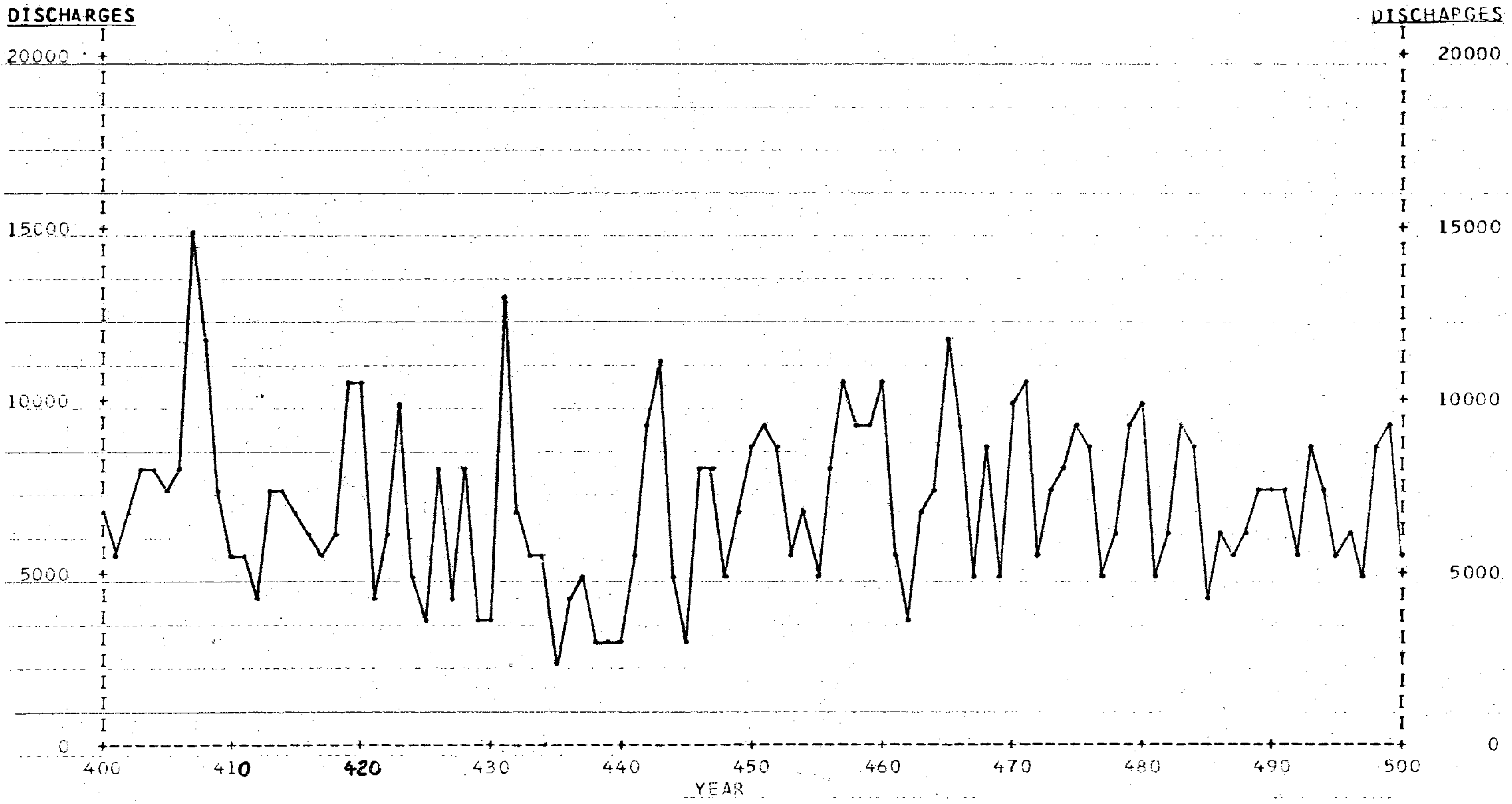

Figure 7 (continued)

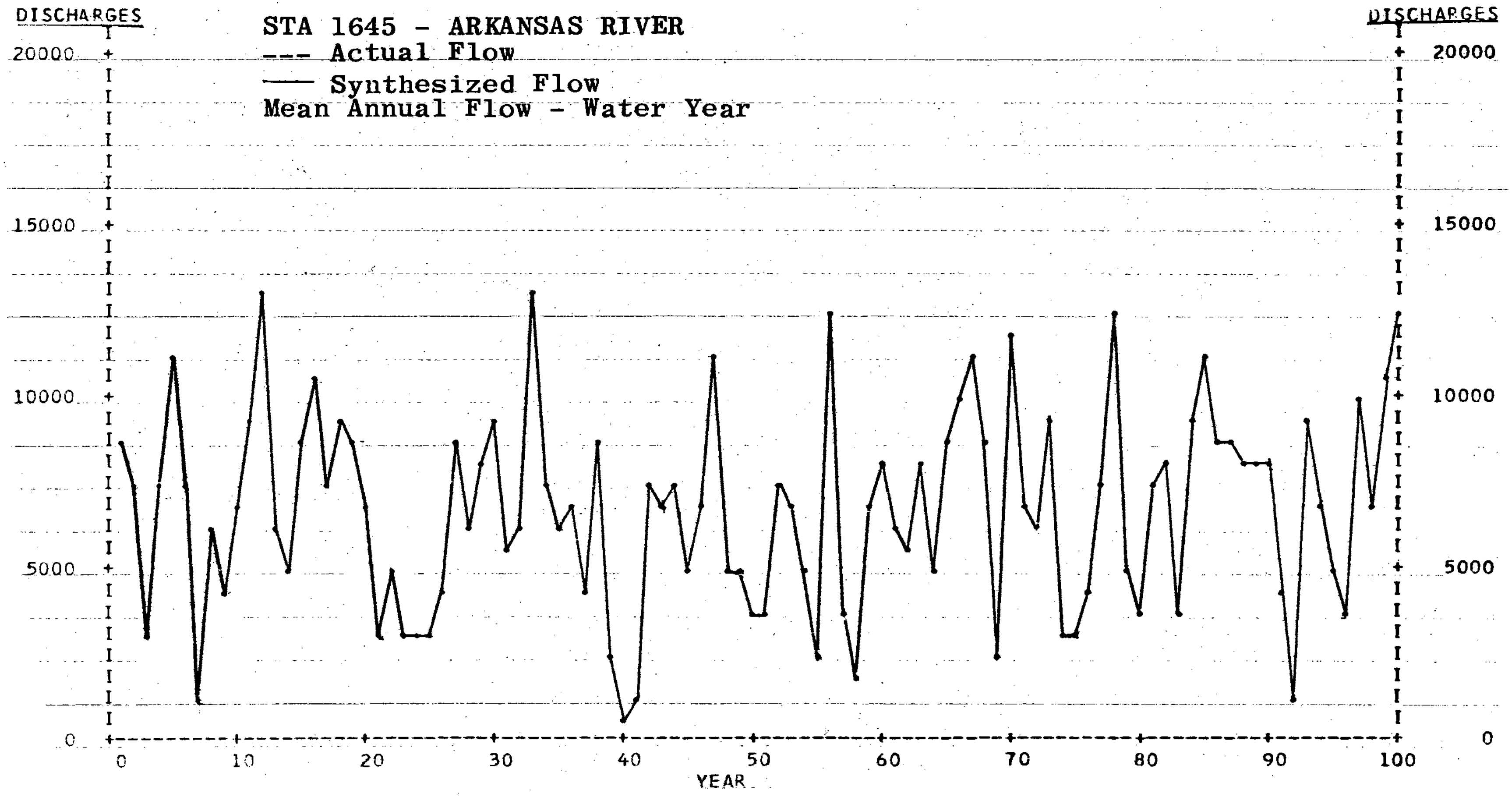

Figure 8 - Observed and synthesized flow in cfs for 35 years of record.

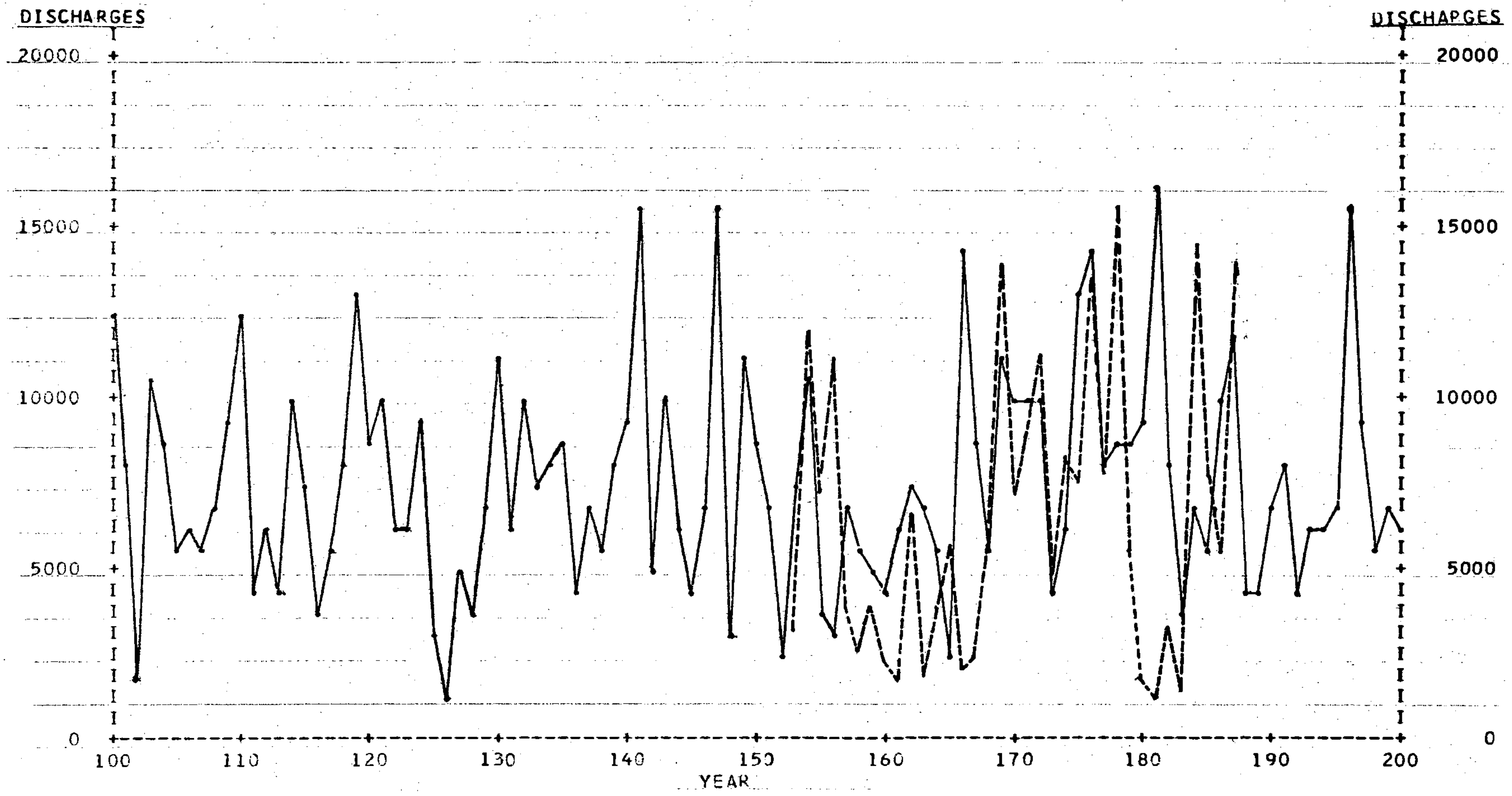

Figure 8 (continued).

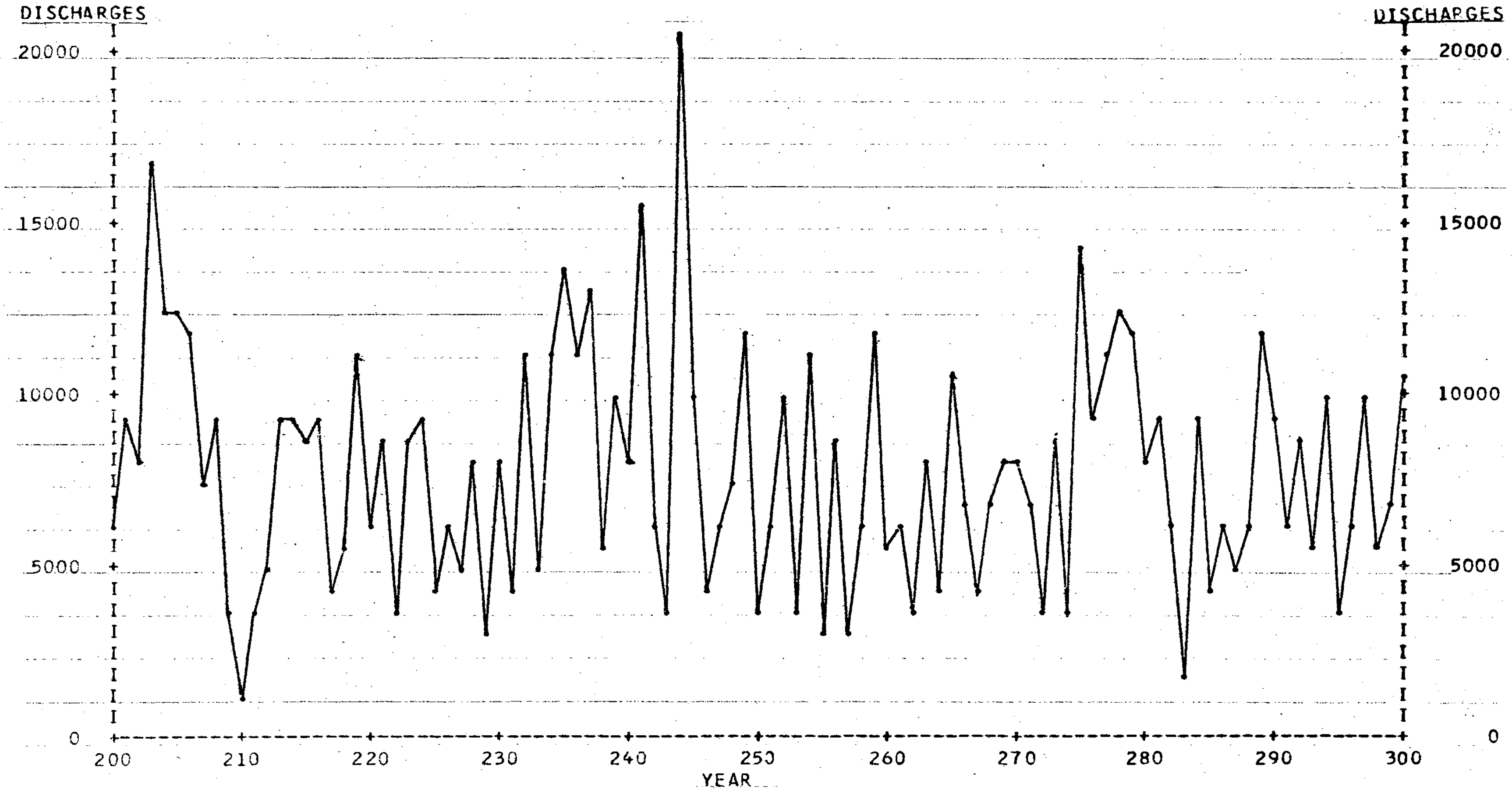

Figure 8 (continued).

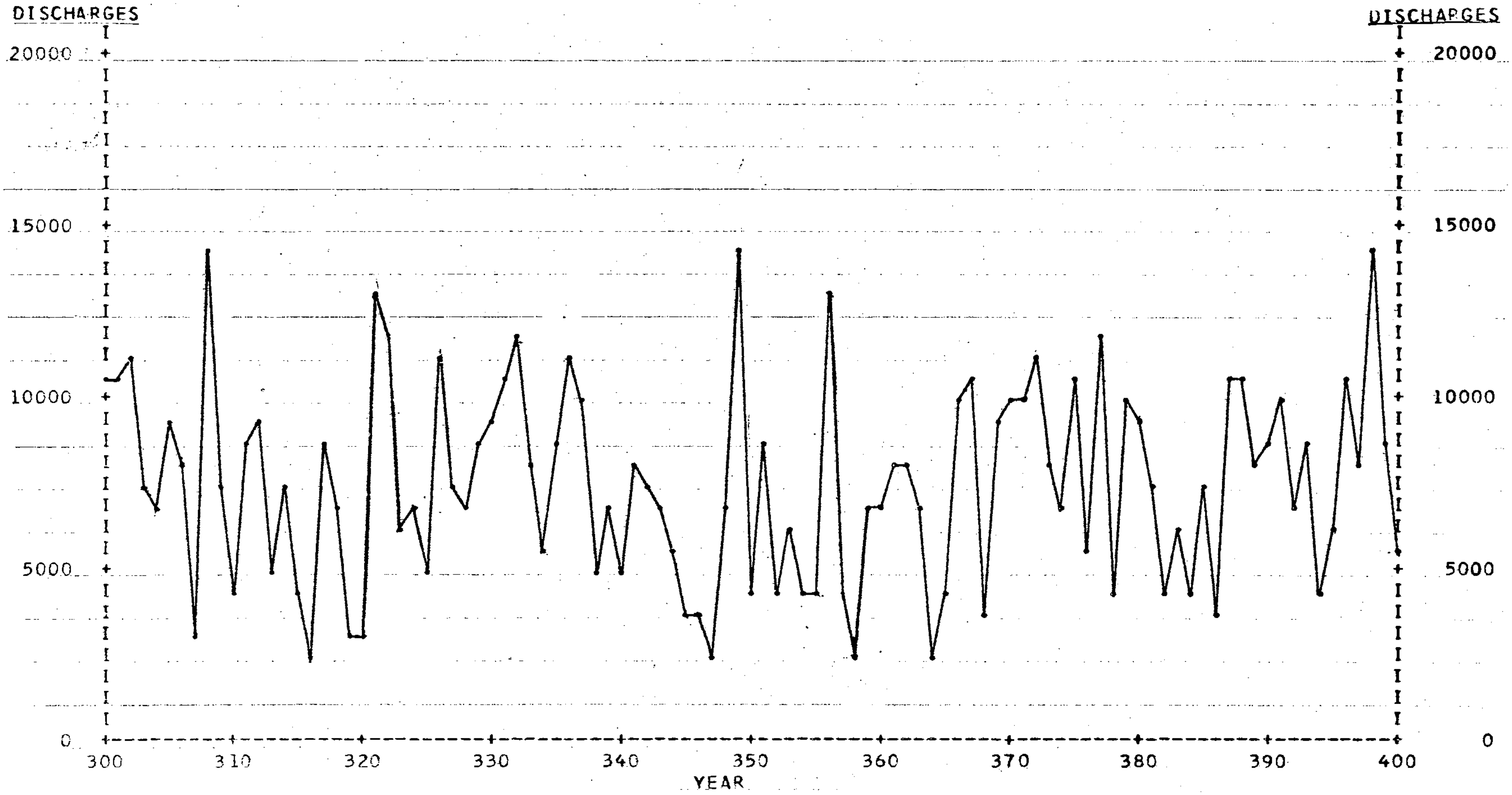

Figure 8 (continued).

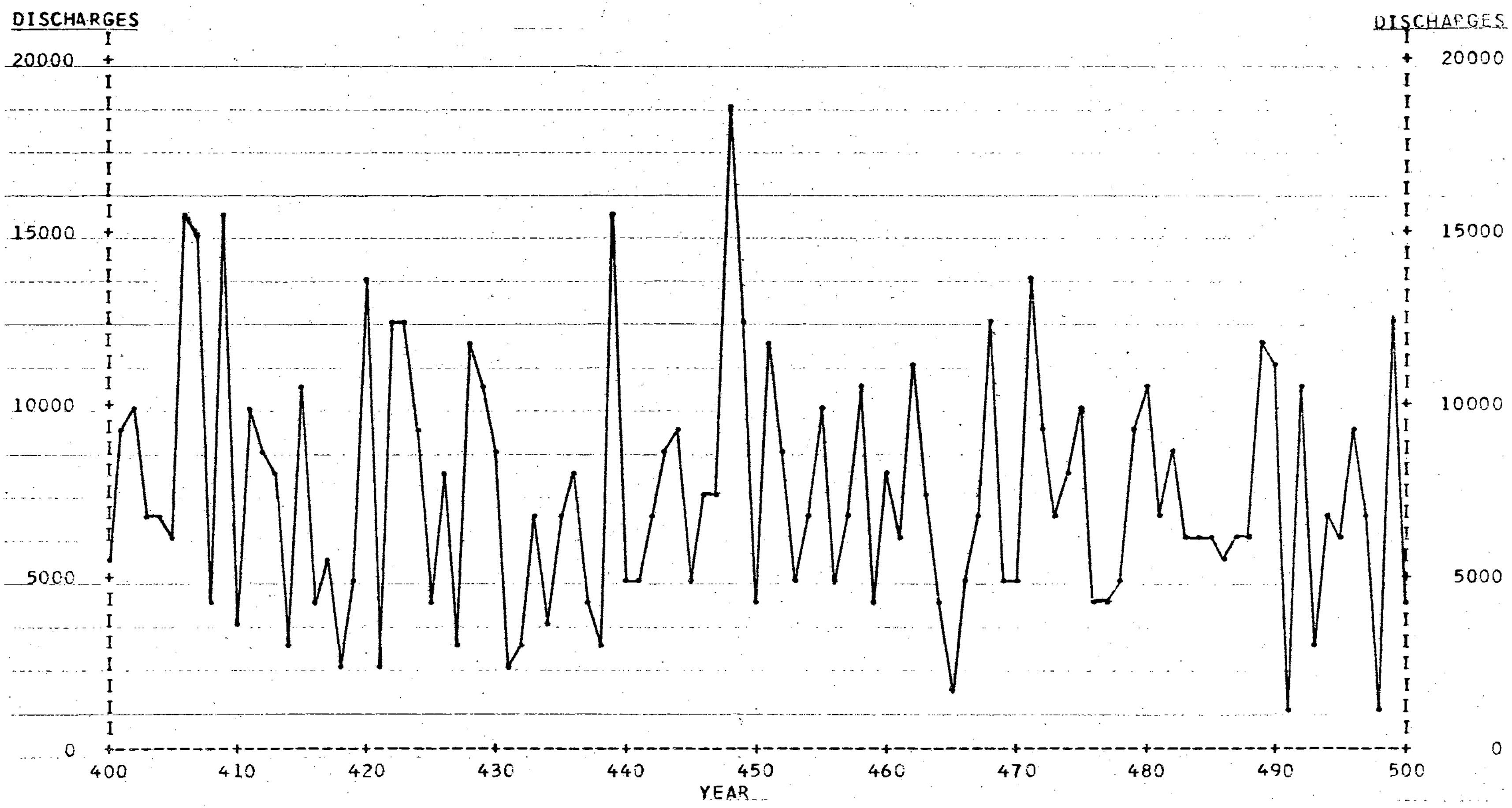

Figure 8 (continued).

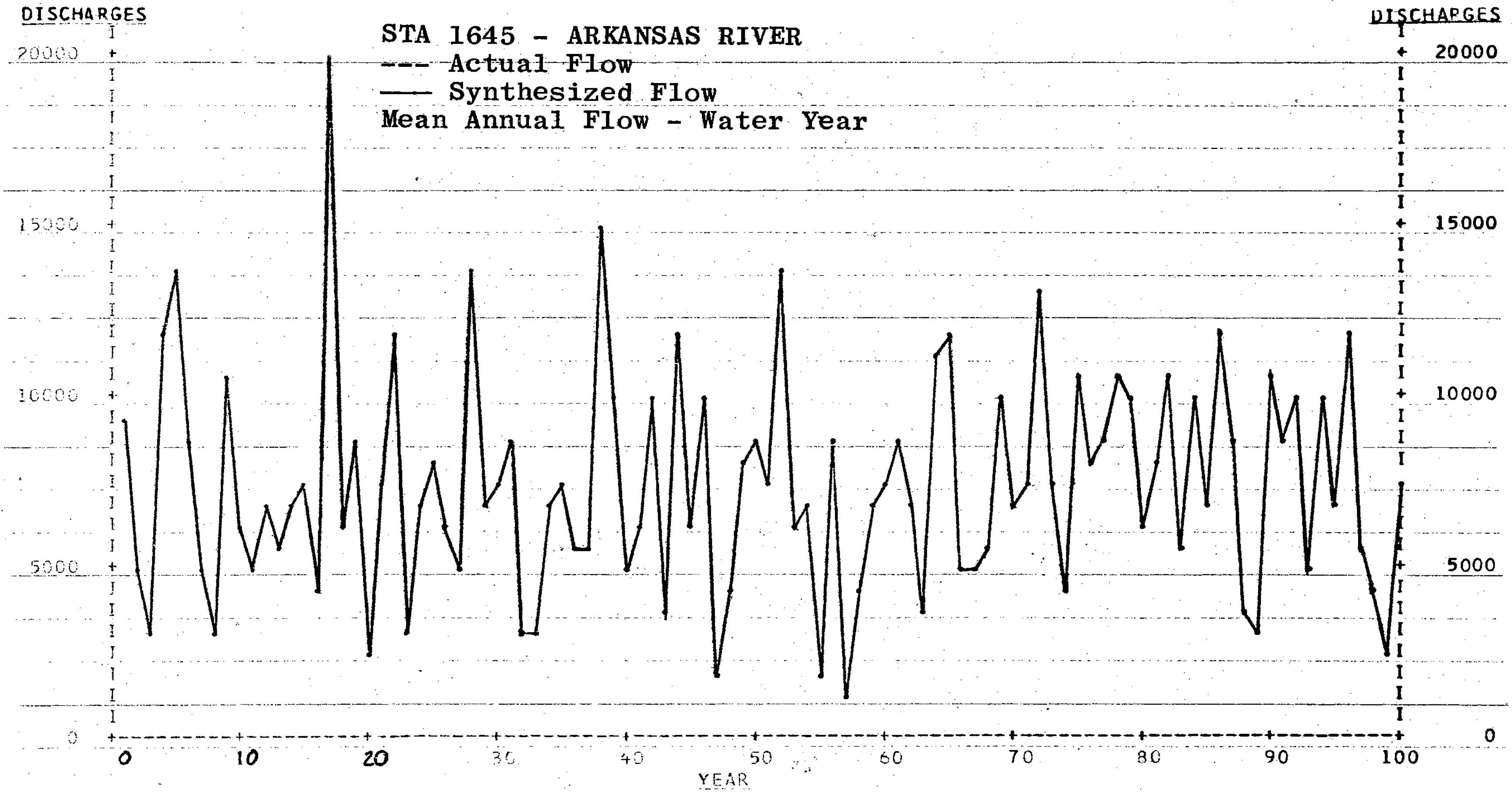

Figure 9 - Observed and synthesized flow in cfs for 39 years of record.

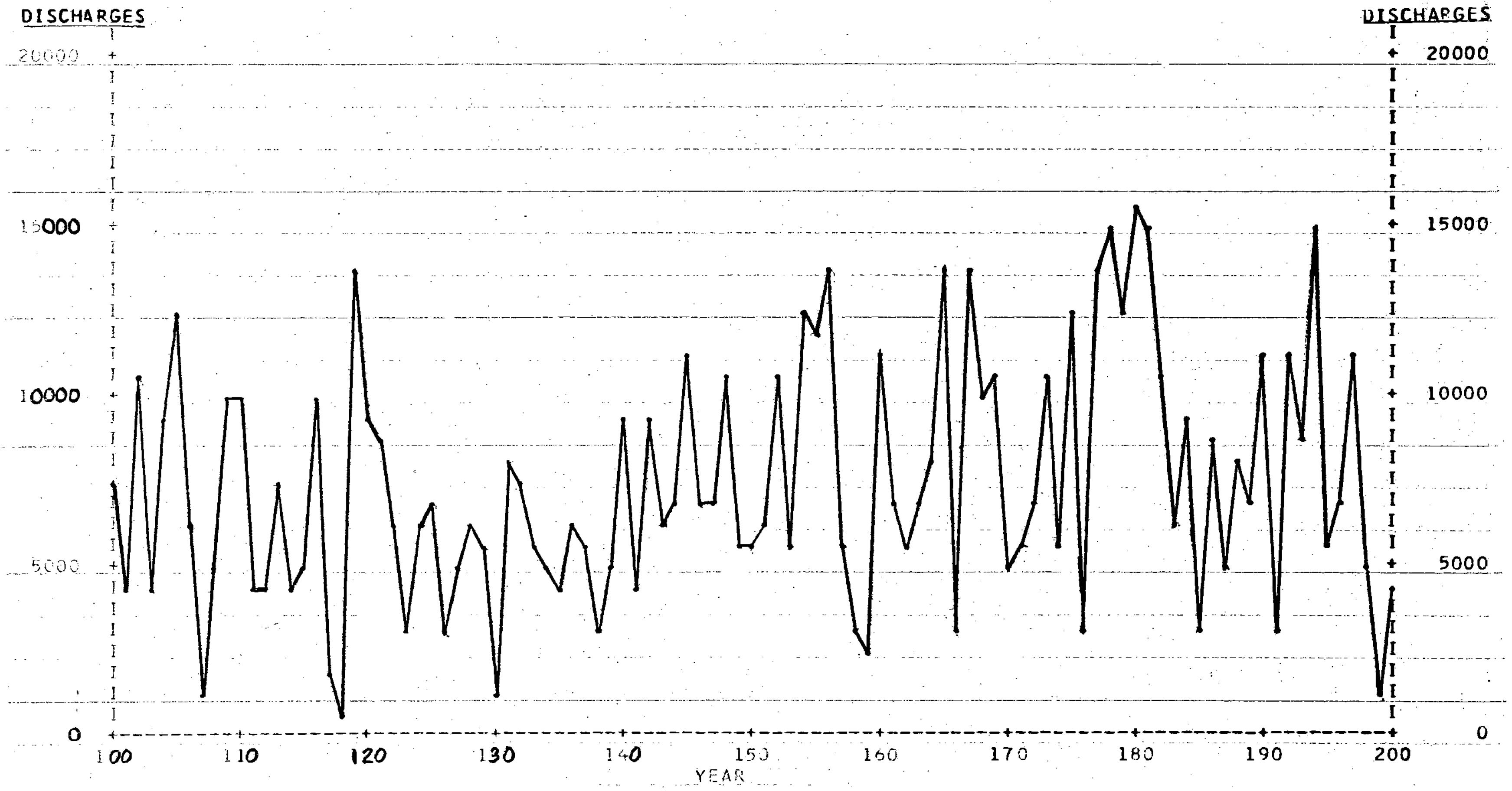

Figure 9 (continued).

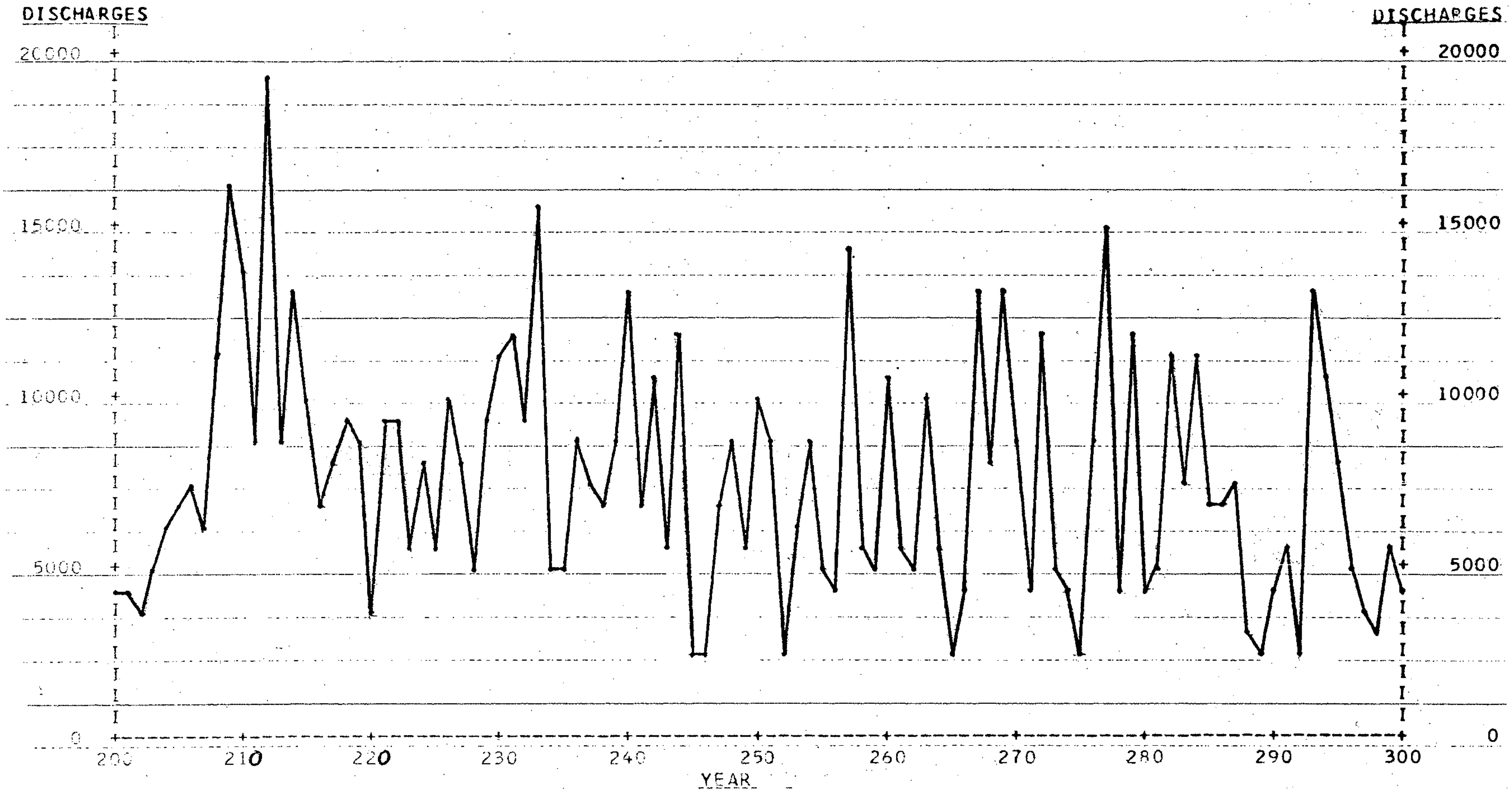

Figure 9 (continued).

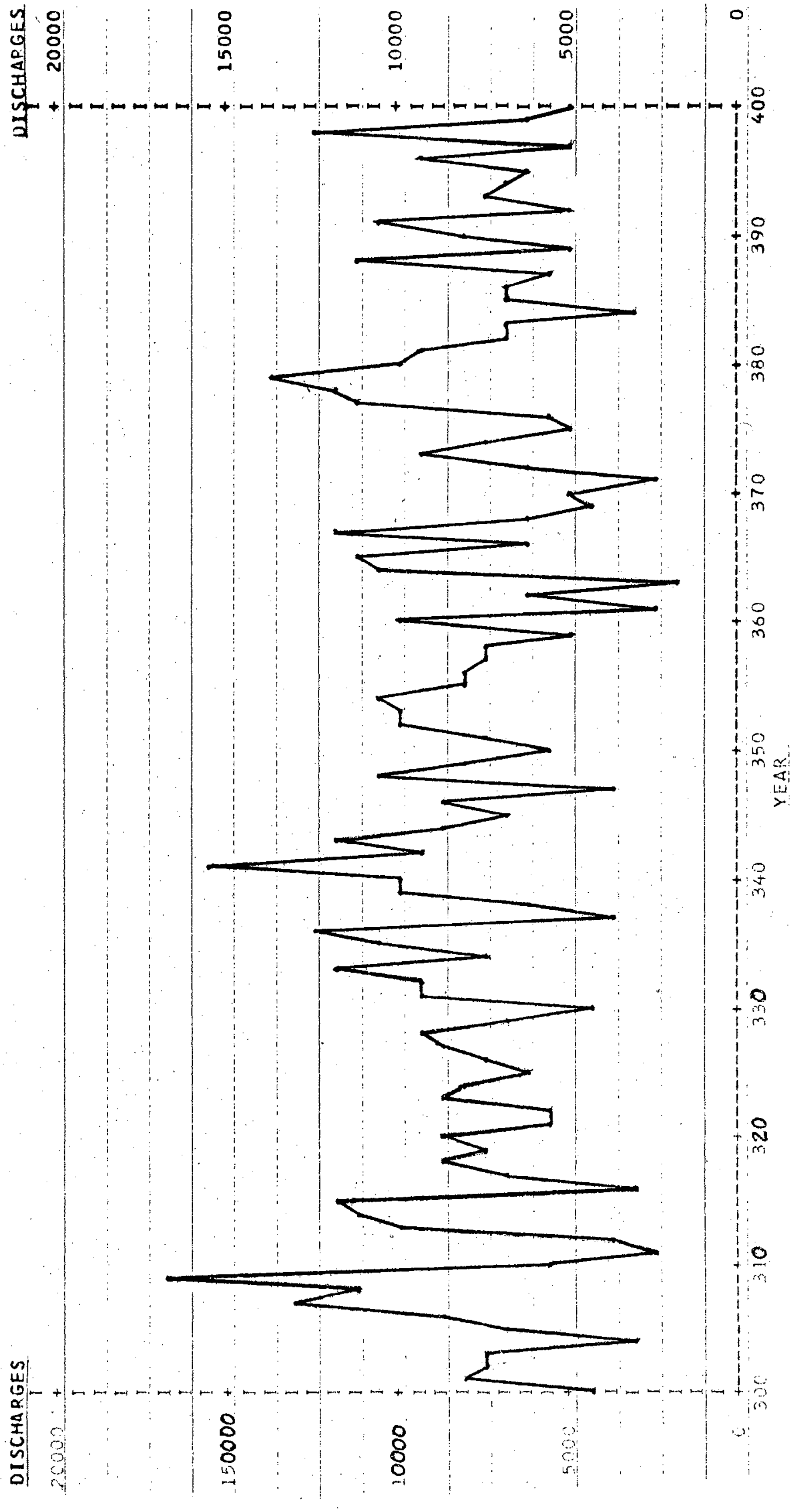

Figure 9 (continued).

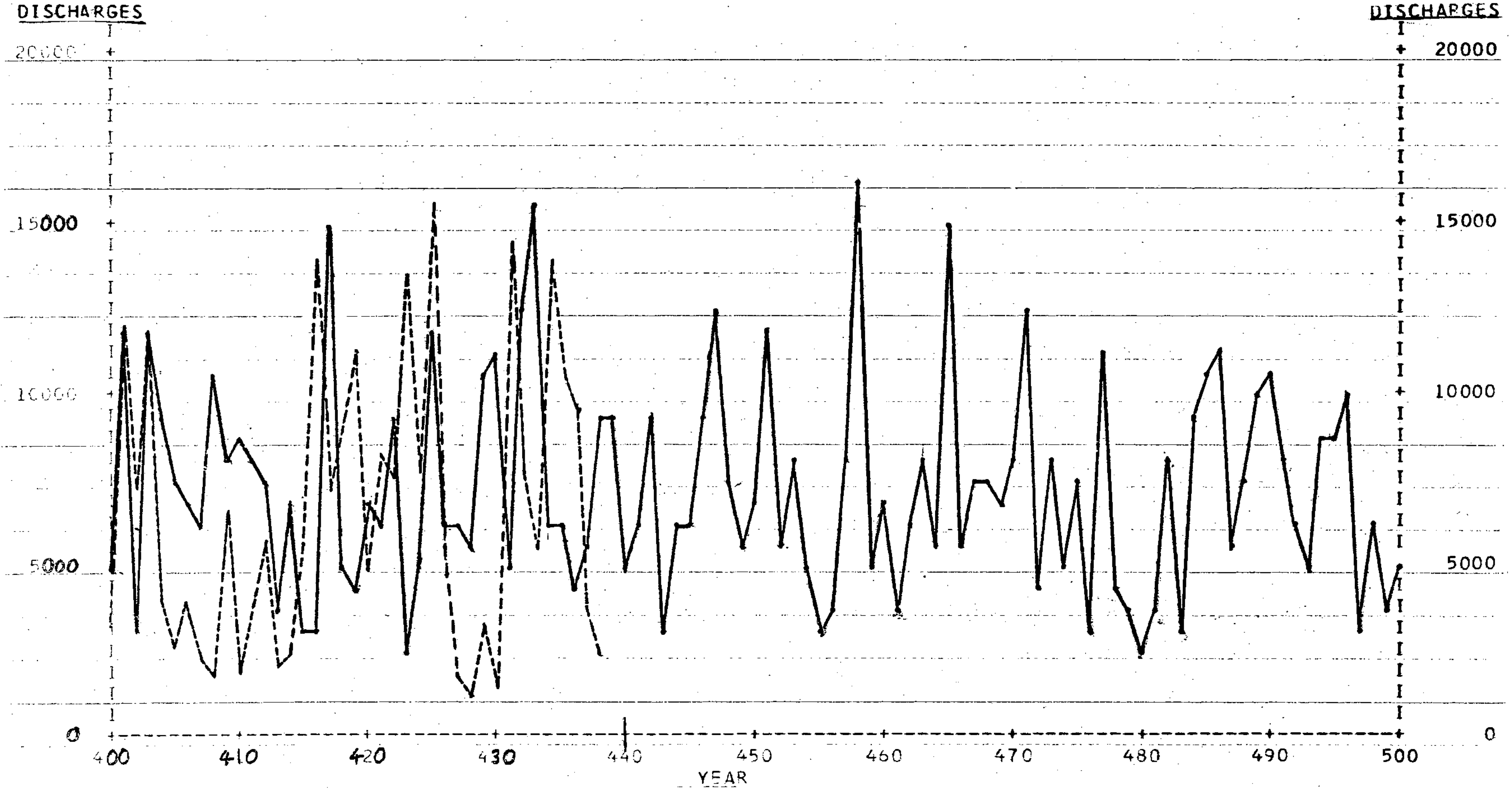

Figure 9 (continued).

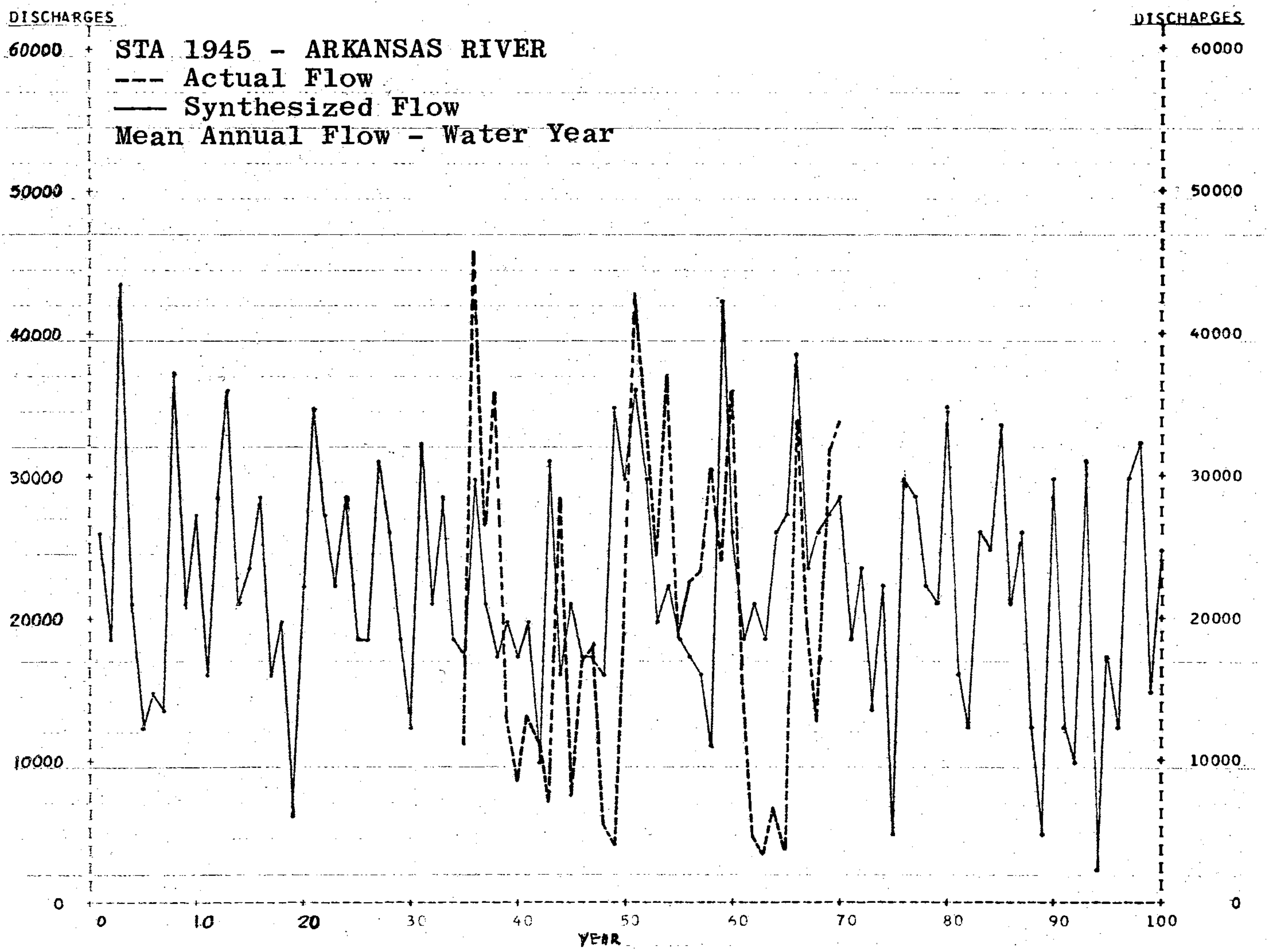

Figure 10 - Observed and synthesized flow in cfs for 35 years of record.

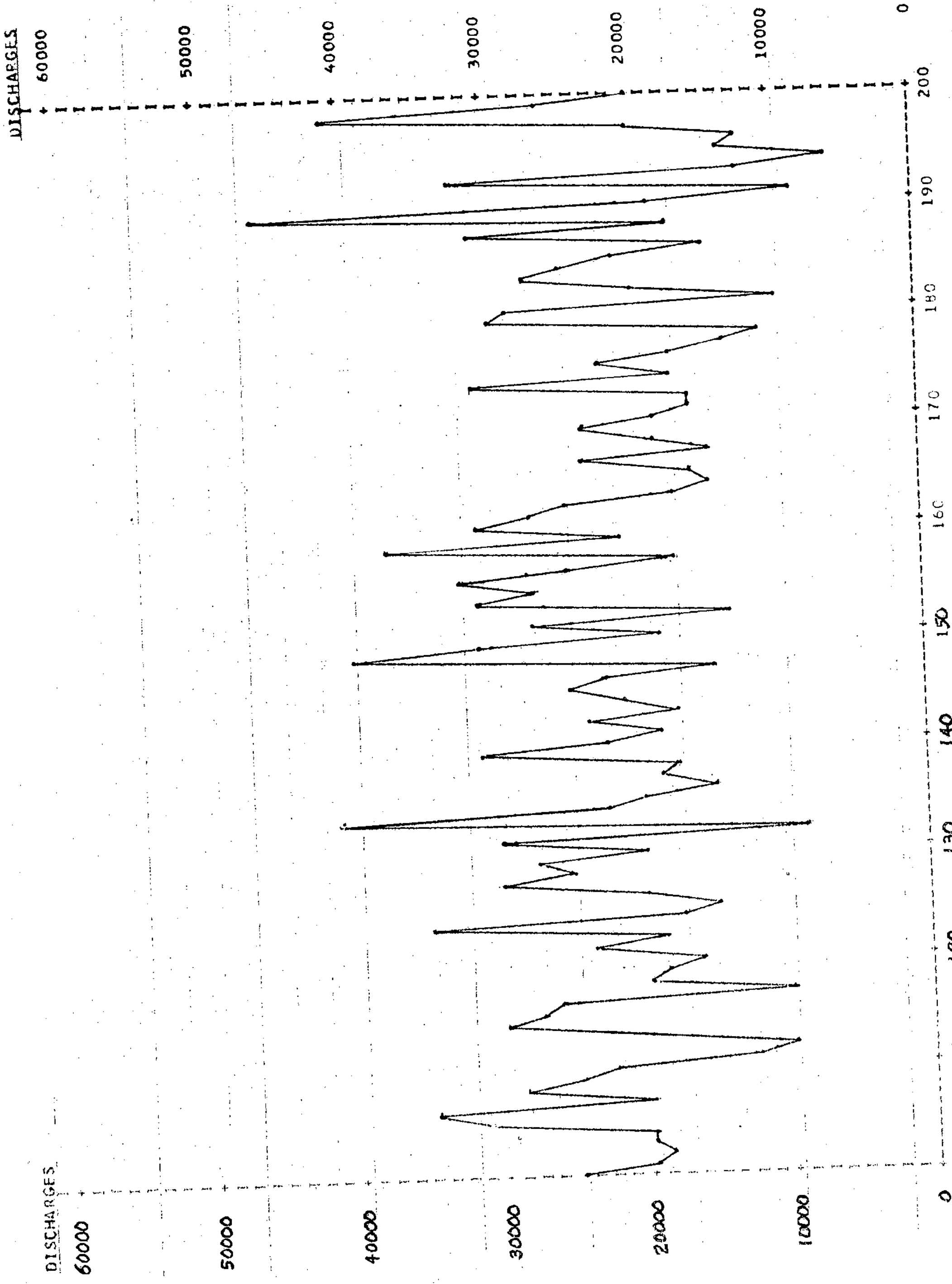

Figure 10 (continued).

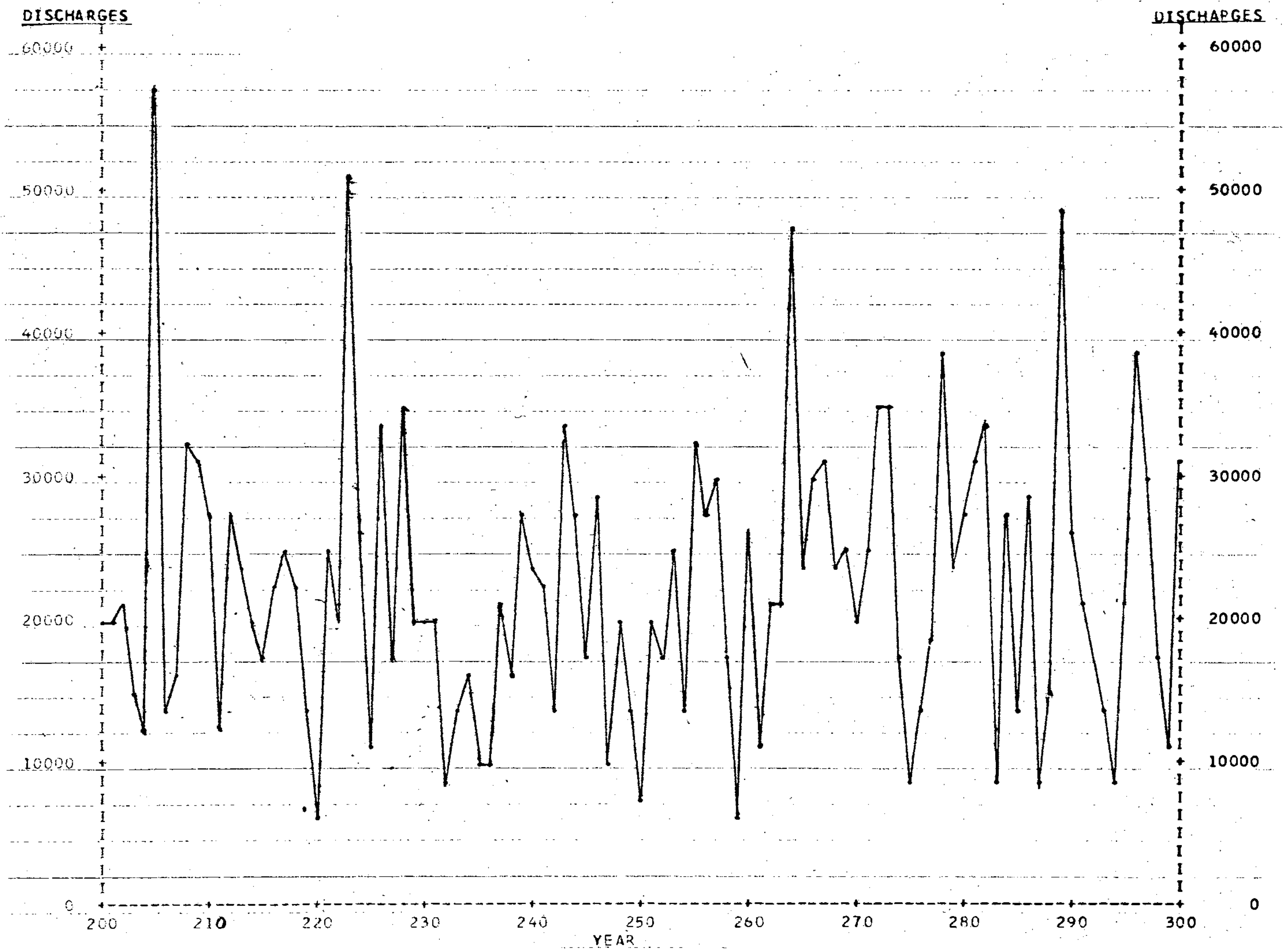

Figure 10 (continued).

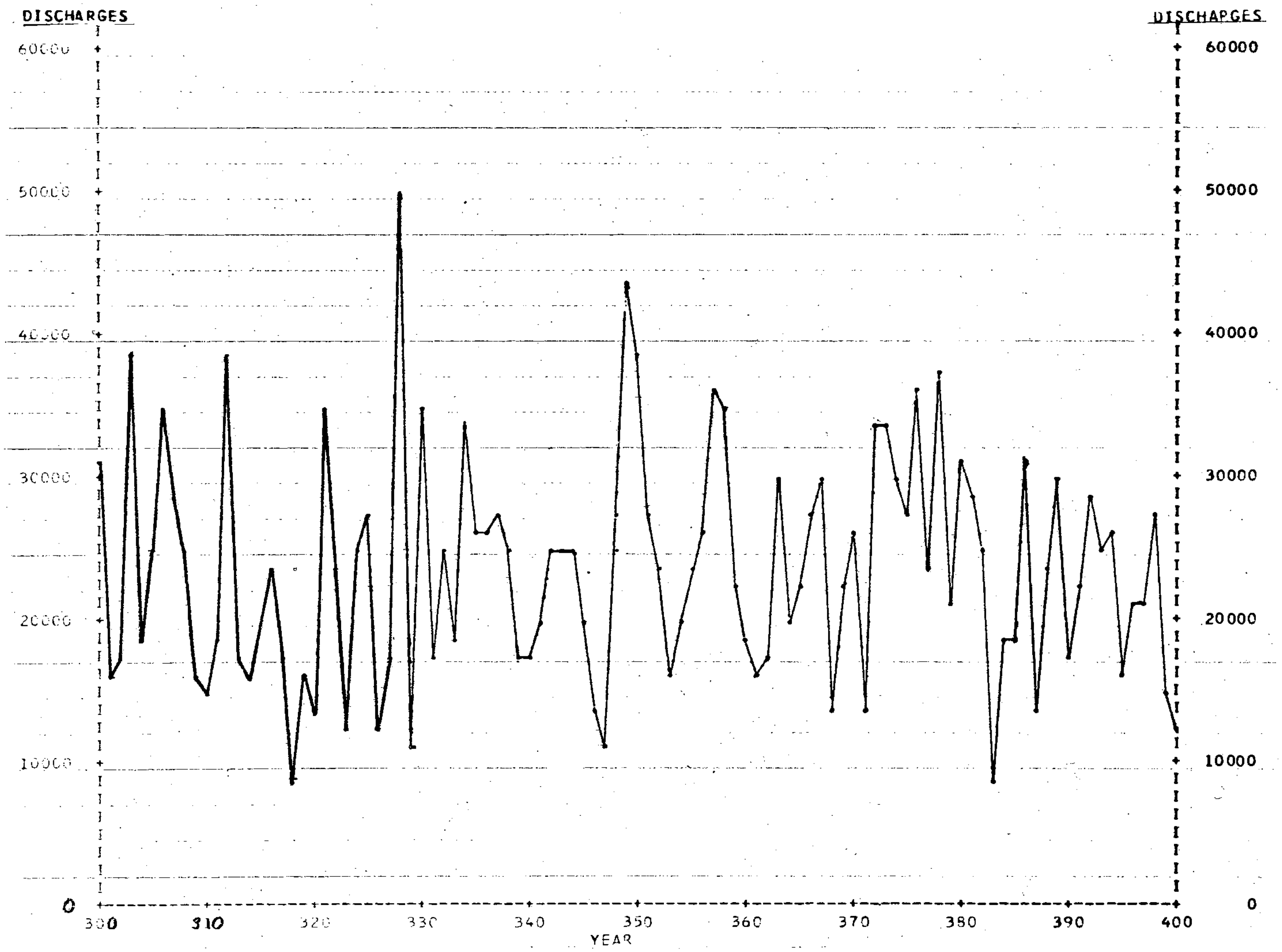

Figure 10 (continued).

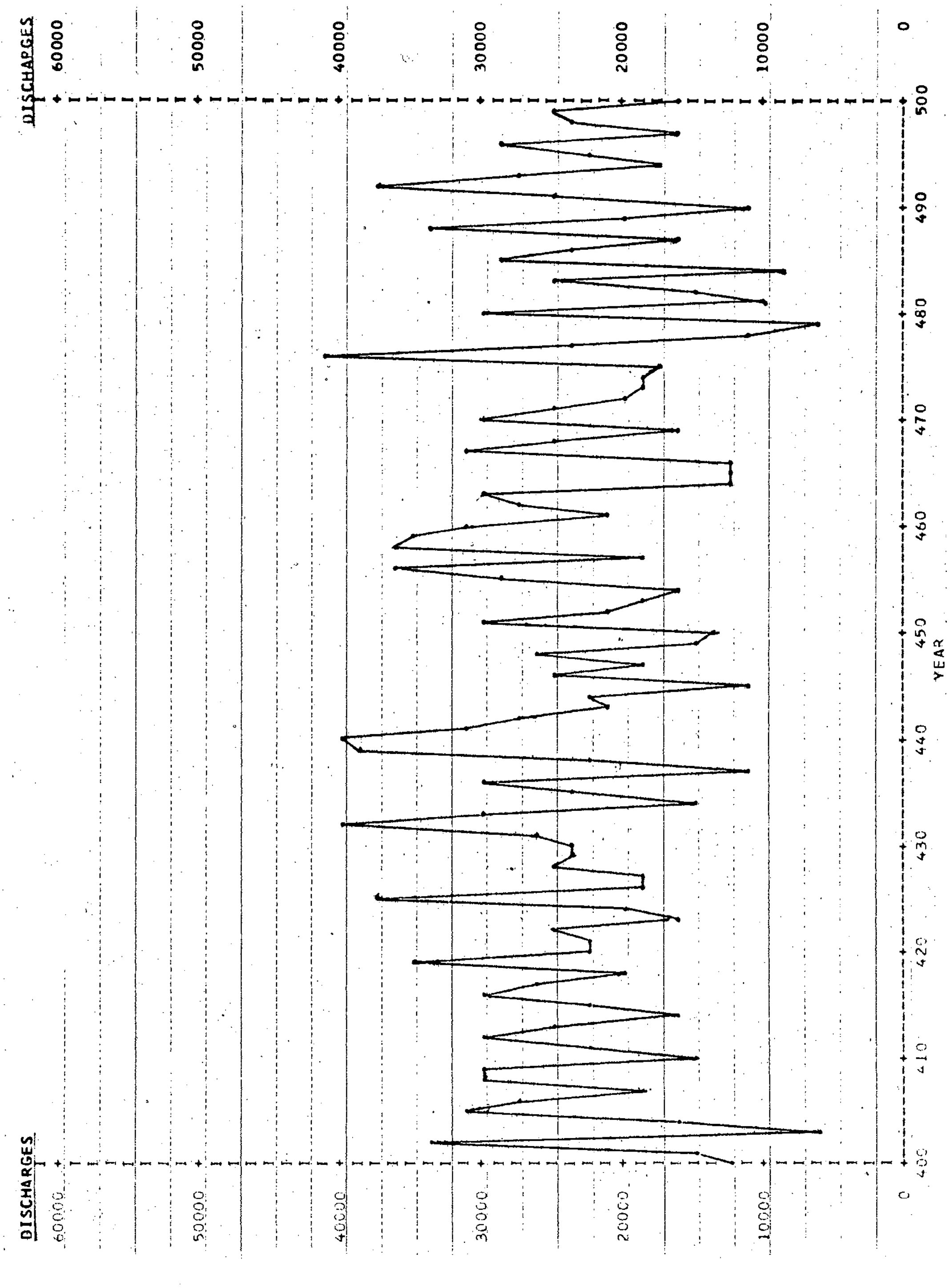

Figure 10 (continued).

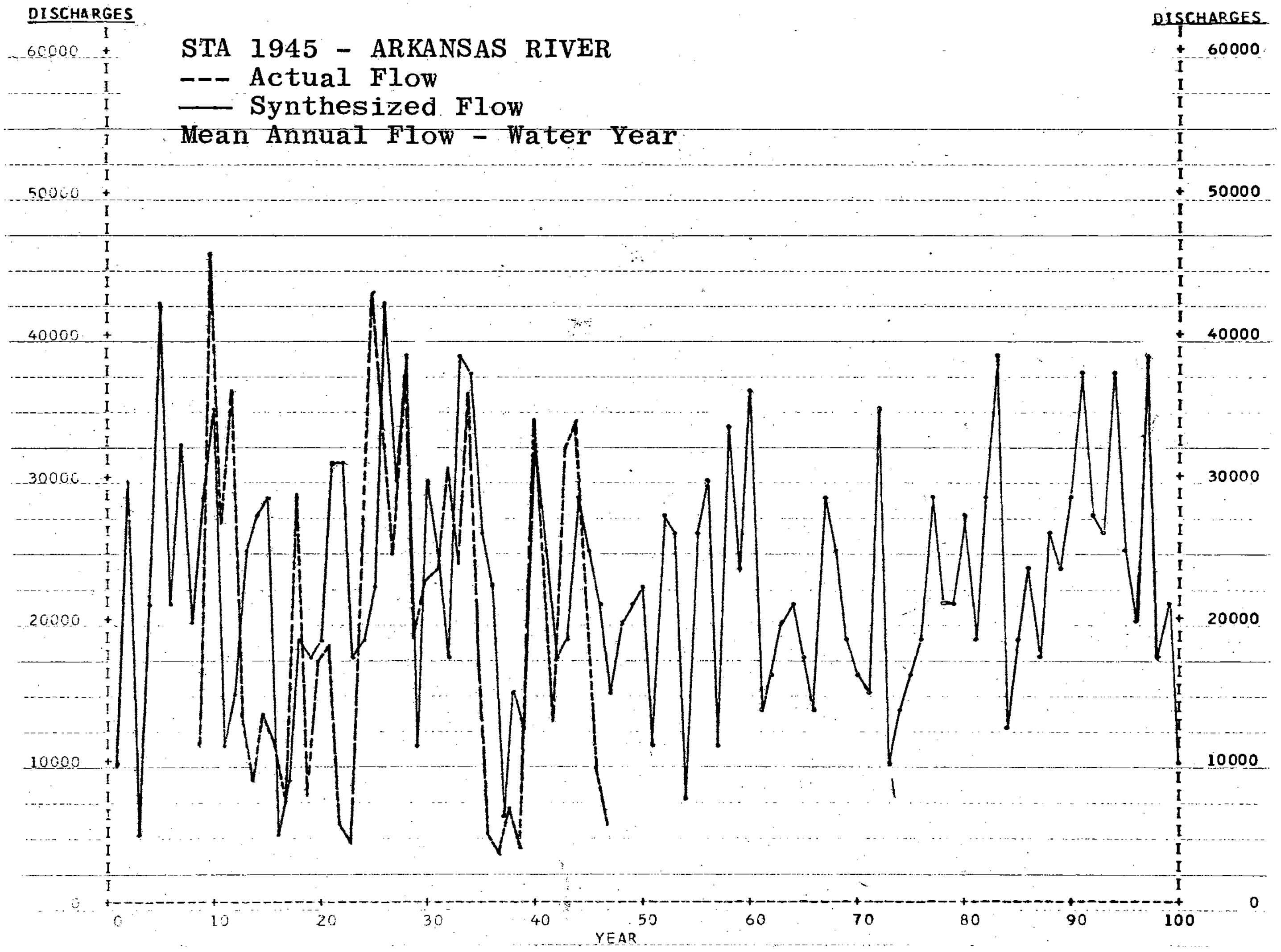

Figure 11 - Observed and synthesized flow in cfs for 39 years of record.

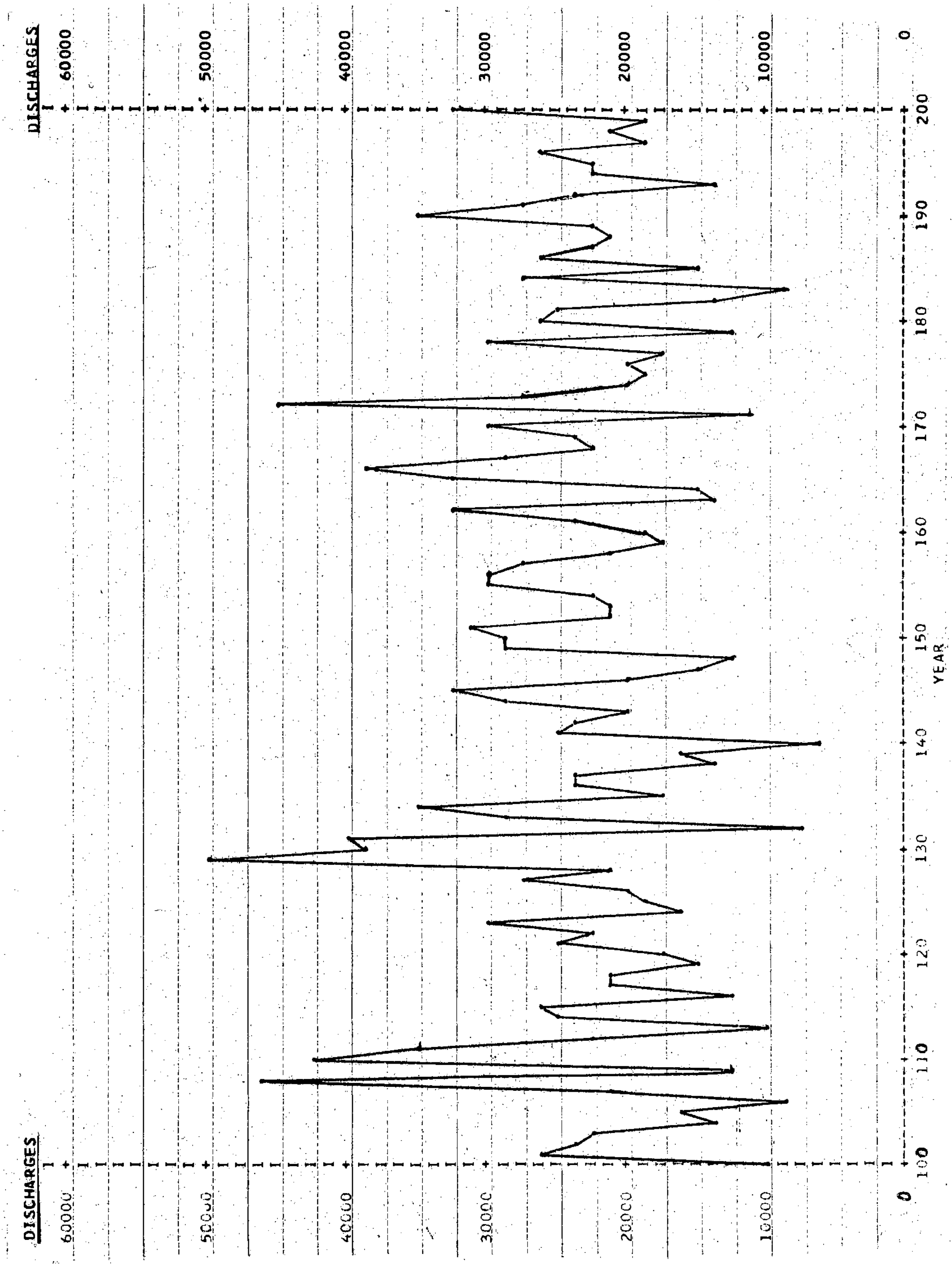

Figure 11 (continued).

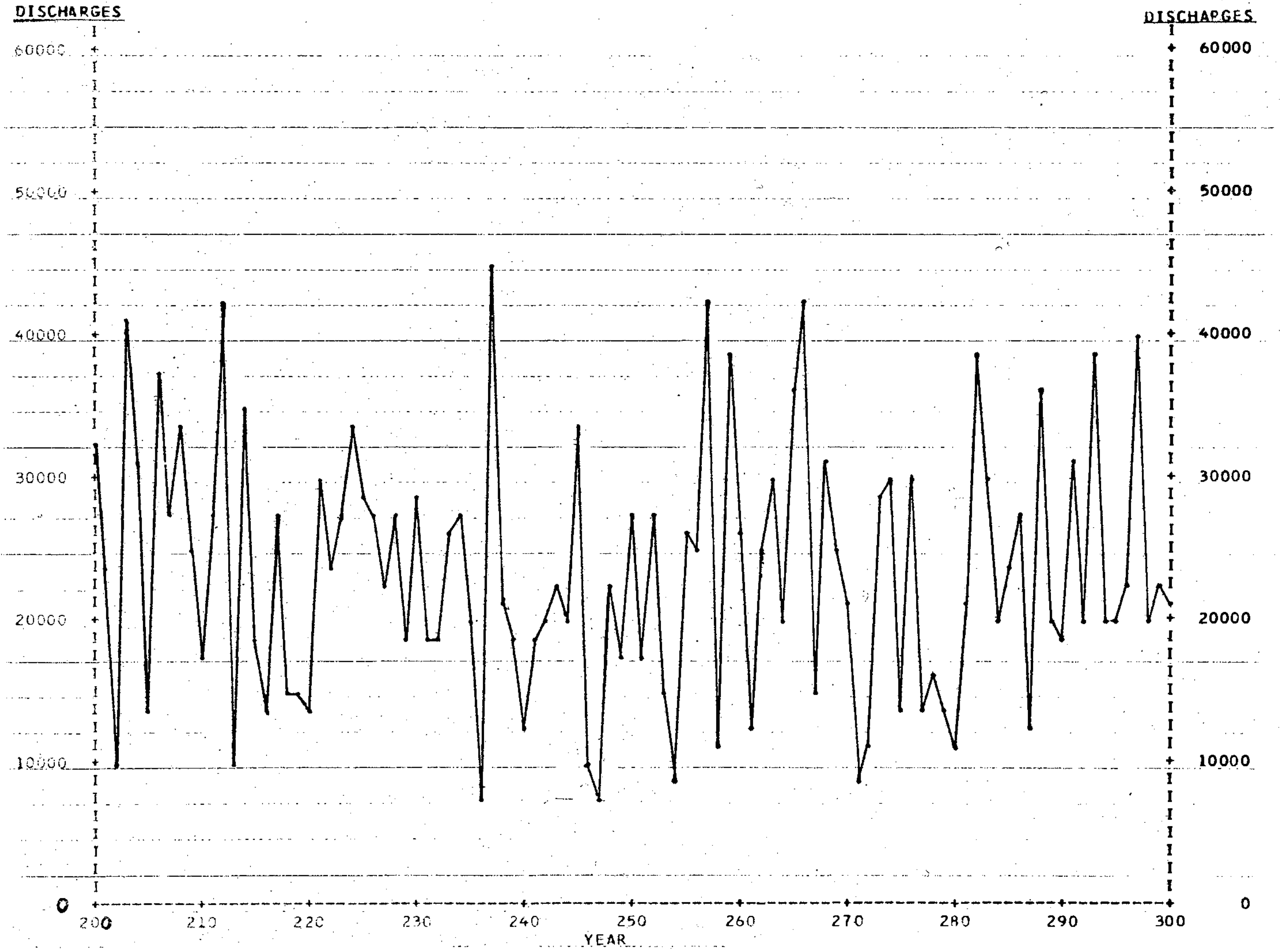

Figure 11 (continued).

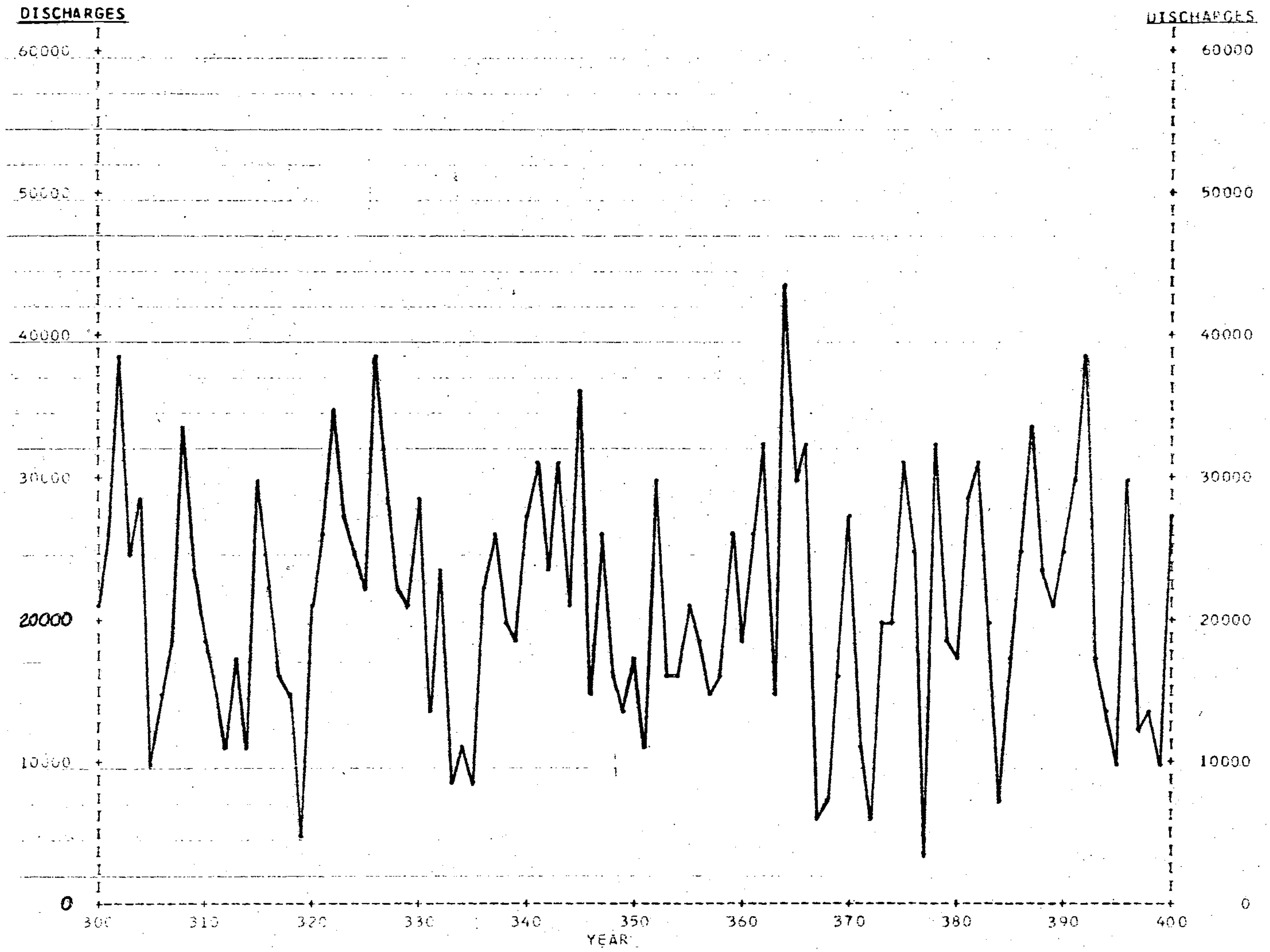

Figure 11 (continued).

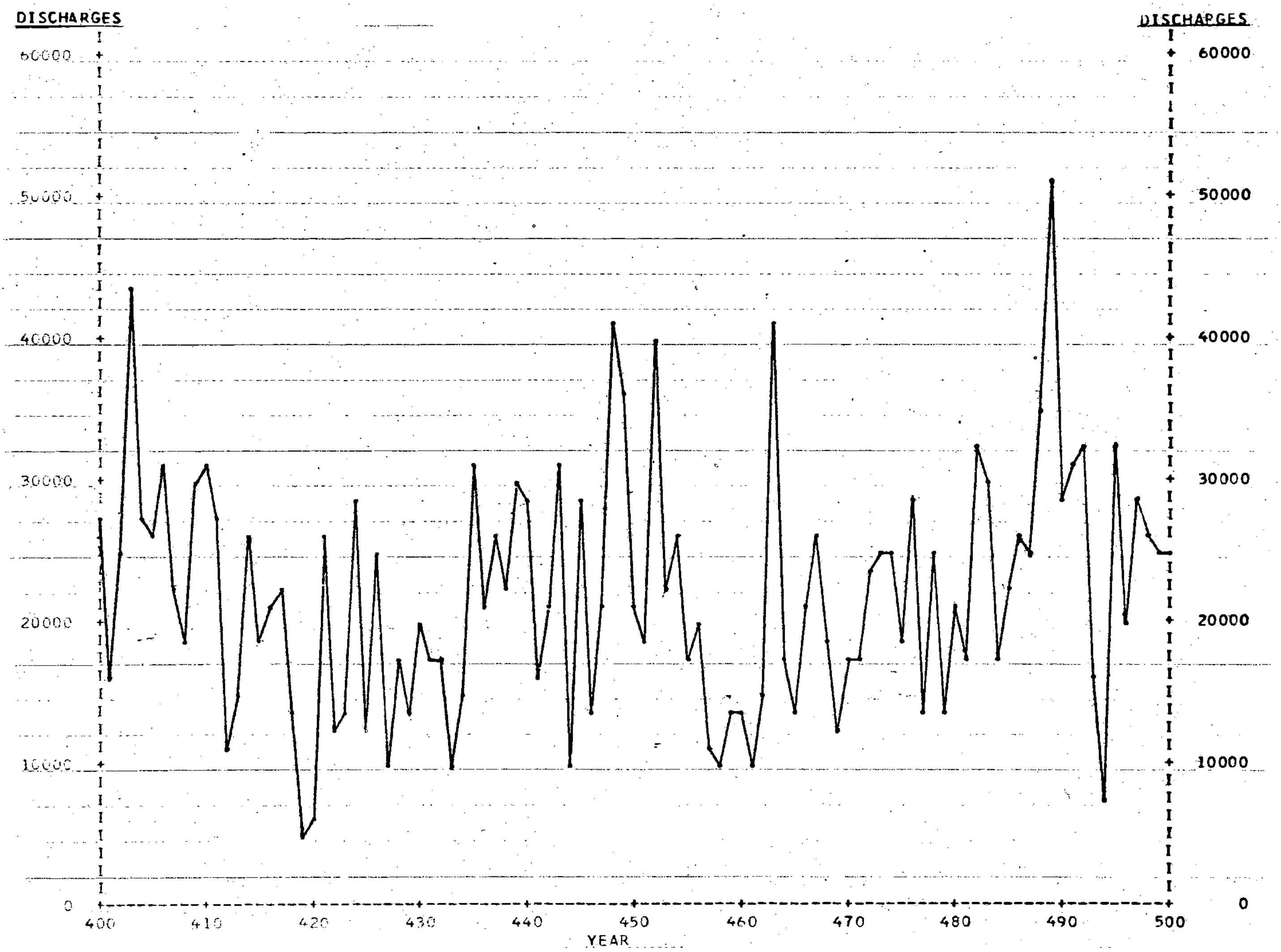

Figure 11 (continued).

synthetic hydrographs in respect to durations and cycles of low annual flows and peak annual flows. They most nearly coincide in this respect for Station 1645 in the years 79 to 104, 132 to 157, and 333 to 358 for the 25-year record, years 116 to 151, 152 to 185, and 389 to 424 for the 35-year record, and in the years 3 to 42, 208 to 247, 292 to 331, 363 to 402, and 399 to 438 for the entire usable 39-year record.

A favorable comparison was also found to exist for Station 1945 in the years 33 to 68, 138 to 173, 205 to 240, and 423 to 458 for the 35-year record, and the years 8 to 47, 202 to 241, 242 to 281, and 431 to 470 for the entire usable record of 39 years. To illustrate this comparison the historical hydrograph for each historical record length is superimposed on the 500-year synthetic hydrograph generated from the corresponding historical record, and are presented in Figures 7 through 11.

CHAPTER VI

DISCUSSION OF RESULTS

A. Illinois River Subbasin at Station 1965

Correlation coefficients for paired months and other statistics (mean monthly flows, regression coefficients, standard deviations, and standard error of estimates) were computed using a computer program. Tabulations of the computed correlation coefficients are reported in Table I. Those correlation coefficients underlined fell below the values for correlation coefficients at the 10 per cent level of significance for various lengths of records, as indicated at the top of the table. Of the twelve paired months for 30 years of record, five correlation coefficients were below the 10 per cent level of significance. For 25 years of record six were below, and for 15 years of record nine were below the 10 per cent level of significance.

Since the hydrologic data for this subbasin provided poor correlations, any attempt to use Thomas and Fiering's stochastic model of the Markov Chain for deriving the synthesized monthly flows would prove unreliable and distort the generated flows. For this reason no further attempt to analyze the effects of record length on the generating technique at this station was made.

Reasons for the poor correlations of paired months at this gaging station are not clear. One possibility is a significant serial correlation between the same months of successive years, i.e., a significant serial correlation between the months of October covering the period of record, and so on for all twelve months of the year. If this were the case, then the t test using Equation (9) would be an invalid means to determine the level of significance of paired months (1). To be sure that successive observations of the same month were independent, all of the months were tested and no significant serial correlatins were found between the same month of successive years. This provides an assurance that the t test is a valid measure of the level significance for the paired monthly correlations.

Another possibility that needs to be substantiated by further study is that the monthly correlations are affected by basin characteristics of size, slope steepness, and quick runoff, thereby causing a low correlation of monthly flows. It may be that either shorter or longer paired intervals such as weeks or seasonal periods would improve the correlation. To further explain the effects of basin characteristics, slope steepness and quick runoff on the monthly correlations, it is necessary to present Table VIII, containing excerpts from the Surface Water Records of the United States Geological Survey. This table clearly illustrates the quick response of the basin to storms and how

TABLE VIII

EXCERPT OF DAILY FLOWS, cfs, FROM SURFACE WATER RECORDS

Day	Feb 1938	Apr 1941	Dec 1942	Feb 1945	Jan 1949	May 1954	July 1960
1	1060	278	1050	187	582	1810	307
2	800	273	971	187	523	5440	289
3	755	283	912	187	482	13000	298
4	630	263	844	187	451	5470	345
5	590	254	816	183	444	2600	350
6	590	254	779	178	523	1870	355
7	515	268	1050	188	459	1520	350
8	480	259	951	174	413	1240	325
9	480	245	902	174	377	1050	298
10	450	240	844	170	363	859	272
11	420	240	807	167	335	746	255
12	396	240	826	167	335	650	247
13	360	231	779	170	328	554	234
14	384	245	726	170	328	481	218
15	847	292	682	170	335	432	210
16	7060	6200	640	187	349	404	206
17	24400	11200	632	252	474	376	199
18	34700	8200	591	2060	709	341	195
19	24800	21800	576	1650	709	327	191
20	8390	30600	552	1180	1080	314	195
21	5000	8940	514	4380	960	294	199
22	3860	4350	536	14200	876	262	221
23	3140	3070	521	7760	917	244	382
24	2640	2430	536	3470	1180	216	1060
25	2320	2050	514	2530	3900	211	2830
26	2000	1700	576	2530	7300	206	23200
27	1790	1490	5550	9220	4760	190	12400
28	1650	1290	22300	6320	6180	190	2810
29		1150	7800	M=2085	6180	181	1950
30		1090	4450		3310	171	1490
31			3360		2510		1210
	M= 4661	M=3648	M=2019		M=1538	M-1349	M-1713

this phenomenon adversely affects the mean monthly flows. For example, the daily flows of April, 1941, were less than 300 cfs for the first fifteen days, but the mean monthly flow exceeded 3600 cfs due to a storm occurring during the second fifteen days. Also, February of 1945 had seventeen days of average daily flow of 200 cfs or less, but the mean monthly was in excess of 2000 cfs due to a storm in the latter part of the month. Thus, these mean monthly flows are not representative and will not yield good correlation coefficients when used in a Markov Chain model. Further study could possibly provide limitations on use of flow intervals for basins fitting certain categorized characteristics for use of the Thomas and Fiering's model of the Markov Chain.

The mean annual flows for historical record lengths ranging from 15 to 30 years remained reasonably stable as they varied from slightly over 1000 cfs to 850 cfs. Even with the stable conditions of having several cycles or durations of low, moderately low, high, and moderately high mean annual flows for the periods of record studied, the correlations of paired months in general were poor, as they were influenced more by the quick runoff characteristics of the small basin. This sets the priorities in using the Thomas and Fiering's stochastic model for streamflow generation. The correlation coefficients have to be acceptable, and after that, an examination of the historical record to shows the stabilizing components of the mean annual flow (low, moderately low, moderately high, and high flows) should be made. Once these two conditions are satisfied, the Thomas and Fiering stochastic model for streamflow generation could be confidently applied for purposes of the historical record.

B. Arkansas River Basin at Stations 1645 and 1945

1. Evaluation of Historical Record Length on Statistics for Mathematical Synthesis

Twelve sets of correlation coefficients computed from various historical record lengths by the least-square method of linear-regression analysis for the purpose of relating the discharge during any month to that in the month immediately preceding it were reported in Table II for Stations 1645 and 1945. The correlation coefficients in this table that are not statistically significant for the various sample sizes (historical record lengths) are underlined, as they fail to exceed the conventionally accepted minimums given in R. A. Fisher's table for testing the statistical significance of sample product-moment correlation coefficients at the 90 per cent level of probability (9). These minimum values are reported as the 10 per cent level of significance values in Table II.

To further illustrate the level of significance for each of the twelve correlation coefficients for sample sizes ranging from 15 to 39, Figures 3 and 4 are presented for Stations 1645 and 1945, respectively. These graphs show which months, for each period of record, have correlations that failed to exceed the 10 per cent level of significance.

As can be seen from Table II and Figures 3 and 4, the April-May correlation for each of the periods of record failed to exceed the 10 per cent level of significance. The change in time and duration of the spring thaw and rains from year to year account for the weak correlations between the monthly spring runoffs. Table VII reports the standard error of estimate for various lengths of record at Stations 1645 and 1945. These values are the random component corresponding to the last term in the stochastic model represented by Equation (3), apart from the random normal deviate e_i. The resulting large standard error of estimate for the April-May monthly pair is a reflection of the large random fluctuations in the observed data for the spring months.

Table II, presenting the correlation coefficients of monthly paired flows, and Figures 3 and 4, presenting the level of significance of the correlation coefficients of monthly paired flows, indicate that as the periods of record decrease from 39 years to 15 years, an increasing number of weak correlations for paired months resulted. It is also noted in Table IV that in general these weak correlated months produced synthetic means from 500 years of generated flows that compared poorly with their corresponding historical mean flows. This is accounted for by the fact that as the historical record is shortened to a certain point, it does not provide enough population to stabilize the mean. The historical record should be long enough to provide for at least a complete cycle of high and low flows which is a characteristic of most hydrologic data. These cycles seem to run in clusters of several years of high flows, then several years of low flows. In the case of Stations 1645 and 1945, the mean annual flows tended to decrease as the historical record was shortened. This is due to the fact that the cycle of high flows was being deleted and the mean annual flows were being more influenced by the cycle of low flows. This illustrates the importance of having a record of adequate length to include both cycles of high and low flows.

With short historical records of less than 30 years in length, the effects of adding two or three additional annual means can affect considerably the overall annual mean of the historical record. This can be readily seen from Table IX. An example of this for Station 1645 is that for 23 years of record the annual mean is 5947 cfs, and with three additional years added, the mean jumped to 6677 cfs. This again illustrates the importance of examining the period of record from a standpoint of cycles of low, moderately low, moderately high, and high flows. It is also noted that Station 1645 had a period of eleven years of low flows followed by a period of eleven years of high flows. Had this total period of 22 years been used, it would have provided an annual mean of 6380 cfs which is close to the overall annual mean for the 39 years of record of 6554 cfs. Station 1645 began to have a stable annual mean from a period of record of 24 years on, as the next cycle of low and high

TABLE IX

EFFECT OF LENGTH OF RECORD ON THE VALUE OF ANNUAL MEAN

		Station 1645		Station 1945	
Year	Length of Record in Years	Mean Annual Flow cfs	Accum. Mean Annual Flow cfs	Mean Annual Flow cfs	Accum. Mean Annual Flow cfs
1940	15	2250	4647	4197	16885
41	16	5448	4697	18990	17017
42	17	13950	5241	42930	18541
43	18	7151	5347	33720	19385
44	19	8960	5537	24910	19675
45	20	11120	5817	37310	20557
46	21	4800	5768	18510	20459
47	22	8096	5874	22480	20550
48	23	7557	5947	23230	20668
49	24	13470	6260	30780	21089
1950	25	7740	6230	2400	21205
51	26	15620	6677	36190	21782
52	27	5444	6632	16740	21595
53	28	1822	6460	4823	20996
54	29	1280	6281	3501	20392
55	30	3227	6180	7274	19955
56	31	1901	6042	3975	19440
57	32	14450	6304	34074	19897
58	33	7761	6348	20980	19930
59	34	5586	6326	13200	19732
1960	35	14020	6546	32190	20088
61	36	10550	6657	34030	20475
62	37	9615	6737	24700	20590
63	38	3923	6663	8991	20284
64	39	2419	6554	5682	19910

flows was short in duration and had little effect on the annual mean for the period of record as additional years were included. Any period of record shorter than 24 years would have a mean annual flow influenced more by the 11-year period of low flows, as illustrated by Table IX. As the period of record is decreased from 26 years to 15 years, the mean decreases from 6677 cfs to 4647 cfs. This can be readily seen from the historical hydrograph in Figure 5.

Station 1945, which is located 66 miles downstream of Station 1645 on the Arkansas River, has a mean annual flow for its 39-year period of record of 19,900 cfs, which is approximately three times as large as the mean annual flow of 6554 cfs at Station 1645. This additional flow is from an additional area that constitutes only 23 per cent more drainage area over Station 1645; however, these flows are highly regulated from this area as a result of multipurpose reservoirs and power development. The mean annual flows for the period of record became stabilized after 18 years of record. The mean annual flow for 18 years of record is 19,385 cfs as compared to 19,910 cfs for the 39-year period of record. The maximum mean annual flow of 21,782 cfs in this interval occurred for the 26-year period of record. This is accounted for by the fact that the station is influenced by the highly regulated area which constitutes approximately two-thirds of the mean annual flow. This regulation of two-thirds of the mean annual flows at this station helped considerably to maintain a stable mean annual flow, and resulted in poor correlation coefficients for paired months at the 25-year record.

After examination of the correlation coefficients and mean annual flows it became apparent that the 15-year period of record was too short to use for generation of synthetic mean monthly flows as the results would be extremely unreliable. The 25-year period of record at Station 1945 yielded six poor correlations of the twelve paired months so that this period of record was also not used for generation of synthetic data and plotting of hydrographs. The poor correlation for the 25-year period of record at Station 1945 is believed to be the result of high regulation of the Grand Neosho River between the 15-year and 25-year periods of record. As more years of record were added with this regulation, the correlations improved to only one poor correlation for the April-May monthly pair for the 35 and 39-year records. Table II illustrates the fact that the introduction of nonhomogeneous flows produces poor correlations and as the length of record increases with these nonhomogeneous flows (regulated flows) the correlations will eventually improve to the extent of being significant. The synthetic data generated from this station, which includes regulated flows, are usable as long as the operational procedure does not change radically for the regulated portion of the flows.

2. Effects of Negative Values Generated in the Sequence of Monthly Flows

One of the distinct characteristics of not using transformed flows in Equation (3) is that a small percentage of the flows will be negative. This percentage of negative flows ranged from slightly over four per cent for one month to 0.01 per cent for another. The negative flows could be eliminated using logarithmic transformations, but since the negative values did not distort the population significantly no transformation was used in the analysis. These negative values were consequently set equal to zero.

CHAPTER VII

CONCLUSIONS

A. General Conclusions

1. Inspection of the historical record is essential. The hydrologist should satisfy himself that the historical record is of adequate length to include low, moderately low, moderately high, and high flows before using the historical records for purposes of streamflow generation. The flow magnitudes should be compared on a regional basis as to low and high flows, i.e., examine surrounding basins with similar basin morphology.

2. The streamflow record should be of sufficient length to provide a stable mean annual flow, i.e., the addition of new annual flows to the record should not change the mean annual flow appreciably for the period of record.

3. The introduction of nonhomogeneous flows into the historical record produces poor correlations, and as the length of record increases with these nonhomogeneous flows, the correlations will eventually improve to the extent of geing significant.

B. Station 1965, Illinois River near Tahlequah, Oklahoma

1. The correlation coefficients of paired months in general failed to exceed the accepted minimum for testing

the statistical significance of product-moment correlation coefficients at the 90 per cent level of probability for all periods of record studied.

2. The use of monthly flow intervals in Thomas and Fiering's stochastic model of the Markov Chain for this station will not produce reliable synthetic data for any of the periods of record studied.

C. Station 1645, Arkansas River at Tulsa, Oklahoma

1. The correlation coefficients in general exceeded the accepted minimum for testing the statistical significance of product-moment correlation coefficients at the 90 per cent level of probability for historical periods of record of 25, 35, and 39 years.

2. The historical record length should be of adequate length to provide at least a cycle each of low and high mean annual flows. This minimum length of record for reliable use at this station should not be less than 22 years, and preferably longer.

3. The hydrograph comparison of the synthesized and the observed flows for 25, 35, and 39 years of record showed that these flows have similar statistical characteristics.

D. Station 1945, Arkansas River near Muskogee, Oklahoma

The correlation coefficients in general exceeded the accepted minimum for testing the statistical significance of product-moment correlation coefficients at the 90 per cent level of probability for historical periods of record of 35 and 39 years.

2. The mean annual flows for periods of record from 18 years on were stable, due to the highly regulated tributaries immediately above this station.

3. The synthetic data generated from this station using periods of record of 35 years or more are usable as long as the operational procedure does not change radically for the presently regulated portion of the flows.

4. The hydrograph comparison of the synthesized and the observed flows for 35 and 39 years of record showed similar statistical characteristics.

CHAPTER VIII

SUGGESTIONS FOR FUTURE WORK

Based on the results of this investigation, the following suggestions are made for future research in the area of generating synthetic data:

1. A study on the use of different intervals of time for computing correlation coefficients to determine the optimum correlations.

2. A study on gaging stations that provide poor linear correlation to determine if non-linear correlations will improve the significance level.

3. A study on arranging the streamflow data in grouped form as in a bivariate frequency distribution, since hydrologic streamflow data is usually in clusters of lows and highs. This may improve poor correlations to an acceptable level of significance.

4. A study on the effects of basin characteristics such as size, slope steepness, and runoff on the correlation of paired mean flows for different intervals of time.

A SELECTED BIBLIOGRAPHY

1. Anderson, R. L., "Distribution of the Serial Correlation Coefficient." Ann. Math. Stat., 13, 1-B (1941).

2. Barnes, F. B., "Storage Required for a City Water Supply." J. Inst. Engrs., Australia, 26, 198-203 (1955).

3. Benson, M. A., "Characteristics of Frequency Curves Based on a Theoretical 1,000-year Record." Tate Dalrymple (ed.), Flood Frequency Analyses, Manual of Hydrology, Pt. 3, Flood Flow Technique, U. S. Geol. Surv. Water Supply Paper 1543-A, 57-64 (1960).

4. Brittan, M. R., "Probability Analysis Applied to the Development of Synthetic Hydrologic for Colorado River." Pt. IV of Past and Probable Future Variations in Stream Flow in the Upper Colorado River, University of Colorado, Bureau of Economic Research, Boulder, Colorado (1961).

5. Chow, Ven Te, "Statistical and Probability Analysis of Hydrologic Data." Chap. 8-4, Handbook of Applied Hydrology, V. T. Chow (ed.) 91-97, 44-45, McGraw-Hill Book Company (1964).

6. Croxton and Cowden, Applied General Statistics, 451-482, 750-751, Prentice-Hall, Inc. (1955).

7. Dowdy, D. R., H. E. Kubik, L. R. Beard, and E. R. Close, "Worth of Streamflow Data for Project Design--A Case Study (1968).

8. Fisher, R. A., "On the Probable Error of a Coefficient of Correlation Deduced from a Small Sample." Metron, 1, 3-32 (1921).

9. Fisher, R. A., Statistical Methods for Research Workers. 13th Ed., 174-209, Hafner Publishing Co., Inc., New York (1958).

10. Haze, Allen, "Storage to be Provided in Impounding Reservoirs for Municipal Water Supply." Trans. Am. Soc. Civil Engrs., 77, 1539-1667 (1914).

11. Julian, P. R., "A Study of the Statistical Predictability of Stream Runoff in the Upper Colorado River Basin." Pt. II of Past and Probable Future Variations in Streamflow in the Upper Colorado River, Bureau of Economic Research, Boulder, Colorado (1961).

12. Kenney, J. R., Mathematics of Statistics, 177-180, D. Van Nastrand Company, Inc., New York (1947).

13. Lindsley, Ray K., and Joseph B. Franzine, Water Resources Engineering, 1-7, McGraw-Hill, Inc. (1964).

14. Spiegel, Murray R., Theory and Problems of Statistics, 141-144, 188-200, Schaum Publishing Company, New York (1961).

15. Steel and Torrie, Principles and Procedures of Statistics, 2-3, 161-193, McGraw-Hill, Inc., New York (1960).

16. Sudler, C. E., "Storage Required for the Regulation of Streamflow." Trans. Am. Soc. Civil Engrs., 91, 622-660 (1927).

17. Thomas, H. A., Jr., and M. B. Fiering, Mathematical Synthesis of Streamflow Sequences for the Analysis of River Basins by Simulation. Chap. 12, in Aruthur Maass, M. M. Hufschmidt, et al., Design of Water Resource Systems, 459-493, Harvard University Press, Cambridge, Mass. (1962).

18. Velz, C. J., "Graphical Approach to Statistics." Water and Sewage Works, 99, R106-35 (1952).

APPENDIX A

```
T VALUES FOR TESTING THE LEVEL OF SIGNIFICANCE FOR CORRELATION OF PARIED
MONTHS
$IBFTC
      DIMENSION SEG(8,A(9),NM(2),AA(21)
      DATA JD/6HHEADIN/
    2 FORMAT(1H0,2A6,9F12.4)
    4 FORMAT(2A6,F5.0,9F7.2)
    3 FORMAT(13A6,/8A6)
      READ(5,3)NM
   70 READ(5,3)AA
      WRITE(6,1)AA
    1 FORMAT(1H1,21A6)
      I=1
   20 READ(5,4) NM,X,A
      IF (NM(1).EQ.JD) GO TO 70
      DO 30 J=1,9
     IF (A(J).EQ.0.0) GO TO 40
   30 A(J)=A(J)/SQRT(1.0-A(J)*A(J)) /(X-2.0))
      N=9
      GO TO 50
   40 N=J-1
   50 WRITE(6,2)             NM      ,(A(L),L=1,N)
      I=I+1
      GO TO 20
  100 STOP
      END
```

APPENDIX B

```
                                   FORTRAN SOURCE LIST
ISN       SOURCE STATEMENT

  0 $IBFTC MAIN
  1       COMMON QI(12),SUMF(12),SUMN(12),SUMC(12),NUMN(12),PCN(12),TSUMN,
         *LNCTR,NTEN,TSUMF
  2       X=0.
  3       LNCTR = 0.
  4       NTEN = 1
  5       DIMENSION        QJ(12),BJ(12),CON(12)
  6       COMMON /DATCOM/DATE (12)
  7    10 FORMAT (12F6.0)
 10    20 FORMAT (12F6.5)
 11       READ (5,10) QJ, CON,QONE,QJONE
 12       READ (5,20) BJ,BJONE
 13       DO 30 I = 1,12
 14       SUMF(I)=0.
 15       SUMN(I)=0.
 16       NUMN(I)=0
 17    30 PCN(I)=0.
 21       QLAST=QONE
 22       QJLAST=QJONE
 23       BJLAST=BJONE
 24       NNUMN = 0
 25       TPCN = 0.
 26       TSUMN= 0.
 27       TSUMC = 0.
 30       TSUMF = 0.
 31       DO 200 K=1,510
 32       DO 100 I=1,12
 33       CALL NORNUM (X)
 34       QI(I)=QJ(I)+BJ(I)*(QLAST-QJLAST)+X*CON(I)
 35       QLAST=QI(I)
 36       QJLAST=QJ(I)
 37   100 BJLAST=BJ(I)
 41       CALL REPORT
 42   200 CONTINUE
 44       WRITE(6,210)
 45   210 FORMAT (36H1   NEGATIVE VALUES FOR STATED MONTH//40H MONTH    NUMBE
         *R   MAGNITUDE   PERCENTAGE)
 46       DO 300 I = 1,12
 47       PCN(I)=(ABS(SUMN(I))/TSUMF   )*100.
 50       WRITE (6,320) DATE (I),NUMN(I),SUMN(I),PCN(I)
 51       NNUMN = NNUMN + NUMN(I)
 52   300 TPCN  = TPCN  + PCN(I)
 54   320 FORMAT (2H0 ,A6,3X,I4,3X,F8.0,8X,F5.2)
 55       WRITE (6,330) NNUMN,TSUMN,TPCN
 56   330 FORMAT (7H0TOTALS,4X,I4,2X,F9.0,7X,F6.2/1H )
 57       WRITE (6,340)
 60   340 FORMAT     (34H-TOTAL FLOW AND MEAN FLOW BY MONTH//18X,9HCORRECTED/
         *35H MONTH      TOTAL      TOTAL      MEAN)
 61       DO 400 I=1,12
 62       SUMC(I) = SUMF(I) - SUMN(I)
 63       AMEAN = SUMC(I)/500.
 64       TSUMC = TSUMC + SUMC(I)
 65       WRITE (6,350) DATE(I),SUMF(I),SUMC(I),AMEAN
 66   350 FORMAT (2H0 ,A6,F8.0,2X,F8.0,4X,F5.0)
 67   400 CONTINUE
 71       WRITE (6,410) TSUMF,TSUMC
 72   410 FORMAT (7H0TOTALS,F10.0,1X,F9.0)
 73       CALL EXIT
 74       END
```

APPENDIX C

SUBROUTINE PROGRAM FOR PLOTTING 500 YEAR HYDROGRAPH OF MEAN ANNUAL FLOWS

```
$IBFTC APLOT   NODECK
      SUBROUTINE APLOT(AA,A,KLA)
      REAL AA(500,4),A(101)
      INTEGER KLA(11)
      DATA BLANK,DEC,EYE/1H ,1H.,1HI/
      DATA DASH,PLUS/1H-,1H+/
      DO 1 I=1,5
      WRITE(6,6)
      DO 3 K=1,56
      DO 2 J=1,100
    2 A(J)=BLANK
      KA=57-K
      DB=FLOAT(KA)/8.
      KB=KA/8
      IF (FLOAT(KB)-DB) 4,5,4
    5 KB=KB*10000
    6 FORMAT(1H1,10HDISCHARGES,107X,10HDISCHARGES/)
      A(1)=PLUS
      A(101)=PLUS
      DO 8 J=1,101
      KK=100*I-100+J-1
      KKK=AA(KK,4)/1250.+.5
    8 IF (KKK.EQ.KA) A(J)=DEC
      IF (I.EQ.1) A(1)=PLUS
      WRITE(6,10) KB,A,KB
   10 FORMAT(5X,I7,2X,101A1,I7)
   11 FORMAT(14X,101A1)
      GO TO 3
    4 A(1)=EYE
      A(101)=EYE
      DO 9 J=1,101
      KK=100*I-100+J-1
      KKK=AA(KK,4)/1250.+.5
    9 IF (KKK.EQ.KA) A(J)=DEC
      IF (I.EQ.1) A(1)=EYE
      WRITE(6,11) A
    3 CONTINUE
      DO 12 J=1,101
      A(J)=DASH
      BK=(J-1)/10
      AK=(FLOAT(J)-1.)/10.
      IF (BK.EQ.AK) A(J)=PLUS
      KK=100*I-100+J-1
   12 IF (AA(KK,4).LT.0.5) A(J)=DEC
      KL=0
      WRITE(6,10) KL,A,KL
   13 FORMAT(13X,11(I3,7X))
      DO 14 J=1,11
   14 KLA(J)=(J-1)*10+100*I-100
      WRITE(6,13) KLA
   15 FORMAT(60X,4HYEAR)
      WRITE(6,15)
    1 CONTINUE
      RETURN
      END
$ENTRY
```

VITA

2

Kenneth L. Perry

Candidate for the Degree of

Master of Science

Thesis: THE EFFECTS OF HISTORICAL RECORD LENGTHS ON GENERATING SYNTHETIC DATA USING A STOCHASTIC MODEL OF THE MARKOV CHAIN

Major Field: Civil Engineering

Biographical:

Personal Data: Born April 21, 1936, at Claremore, Oklahoma, the son of Mrs. Eria M. Randall and the late Bernard L. Perry.

Education: Graduated from Panama High School, at Panama, Oklahoma, in 1955. Graduated from Eastern Oklahoma State Junior College, Wilburton, Oklahoma, in May, 1957. Completed requirements for the Bachelor of Science degree in Civil Engineering from Oklahoma State University in May, 1961; completed requirements for the Master of Science degree in Civil Engineering from Oklahoma State University in May, 1969.

Professional Experience: Completed the Engineer-in-training Program for Oklahoma State Highway Department in May, 1961; worked as a Civil Engineer with the Construction Division, Oklahoma State Highway Department during summer of 1961; Civil Engineer with Design Branch, Engineering Division, U. S. Army Corps of Engineers, Tulsa District Office from September, 1961, to present.

Membership in Honorary and Professional Societies: Registered Professional Engineer, State of Oklahoma; Oklahoma Society of Professional Engineers; National Society of Professional Engineers; American Society of Civil Engineers; American Water Works Association.