AUTOCORRELATION ANALYSIS OF STREAMFLOW SEQUENCES

By

F. T. PAINTER

Bachelor of Science

The University of Newcastle Upon Tyne

United Kingdom

1966

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

AUTOCORRELATION ANALYSIS OF

STREAMFLOW SEQUENCES

Thesis Approved:

Don F. Kincannon
Thesis Adviser

[illegible]

D. D. Durham
Dean of the Graduate College

ACKNOWLEDGMENTS

I wish to thank the following persons and organizations who made the preparation of this thesis possible:

Dr. Hamdy Bechir, my Major Adviser until his resignation in August 1968, who instigated this study and made valuable comments for inclusion in the final work.

Dr. D. F. Kincannon, who became my Major Adviser and provided many helpful suggestions during the writing of the thesis.

Dr. A. F. Gaudy Jr. and Professor Q. B. Graves, who served on my Advisory Committee.

Miss Velda Davis, who accurately typed this manuscript.

Mrs. Grace Wynd, who offerred many suggestions and no little help in the preparation of the manuscript.

Mr. Eldon Hardy, for his accurate drawing of the figures.

The Office of Water Resources Research, Department of the Interior for financial assistance in the form of a Research Assistantship provided by the grant (WRB-006-Okla.) with matching funds from the Oklahoma Water Resources Board.

The Fulbright Foundation and the United States-United Kingdom Educational Commission, who made my visit to the United States possible.

TABLE OF CONTENTS

Chapter		Page
I.	INTRODUCTION	1
II.	THEORY OF ANALYTICAL METHODS	6
	1. Time Series	6
	a. Definition	6
	b. Stationarity	7
	2. Serial Correlation	10
	3. Separation of Deterministic and Random Elements	12
	4. Stochastic Models	16
	a. Standardization of Variables	16
	b. Autoregressive Schemes	17
	c. Skewed Distributions	21
	5. Tests of Goodness of Fit of Autoregressive Scheme	23
	6. Distribution of Variables	25
	a. Introduction	25
	b. Transformations of the Variable X_t	26
	c. Estimation of Parameters	27
	d. Testing Goodness of Fit of Observations	28
	e. Class Boundaries	30
	f. Skewness of Distribution	31
III.	METHODS OF COMPUTATION	32
	1. Autoregressive Schemes	32
	2. Distribution	38
IV.	DRAINAGE BASINS USED IN STUDY	39
V.	RESULTS	47
	1. Autoregressive Models	47
	2. Distribution of Variables and Residuals	47
	3. The Residual ϵ_t	51
	4. Harmonic Removal	58
VI.	DISCUSSION	62
	1. Autoregressive Models	62
	2. Tests of Fit of Autoregressive Scheme	65
	3. Harmonic Removal	67
	4. Comparison of Models	69

Chapter	Page

Chapter	Page
VII. CONCLUSIONS	73
BIBLIOGRAPHY	76
APPENDIX I - LIST OF SYMBOLS	79
APPENDIX II - PARAMETERS OF ACCEPTED MODELS	83
APPENDIX III - CORRELOGRAMS	85

LIST OF TABLES

Table		Page
I.	Gauging Stations Used in Study	40
II.	Physiographic Data for River Basins	42
III.	Computed χ^2 From Test of Autoregressive Scheme	48
IV.	Distribution Parameters (Model A - Untransformed Flows, $\{X_t\}$)	49
V.	Distribution Parameters (Model A - Natural Logs of Flows, $\{X_t\}$)	50
VI.	Distribution Parameters (Model B - Untransformed Flows, Residual $\{Z_t''\}$)	52
VII.	Distribution Parameters (Model B - Natural Logs of Flows, Residual $(Z_t''\})$	53
VIII.	Distribution Parameters (Model A - Untransformed Flows, Residual ϵ_t)	54
IX.	Distribution Parameters (Model A - Natural Logs of Flows, Residual ϵ_t)	55
X.	Distribution Parameters (Model B - Untransformed Flows, Residual ϵ_t)	56
XI.	Distribution Parameters (Model B - Natural Logs of Flows, Residual ϵ_t)	57
XII.	Harmonics Removed	60
XIII.	Models Accepted	65
XIV.	Parameters of Accepted Models	84

LIST OF FIGURES

Figure		Page
1.	Flow Chart of Program System	33
2.	Flow Chart of Model A Analysis	35
3.	Flow Chart of Model B Analysis	37
4.	Location of River Basins in Oklahoma and Arkansas	43
5.	Location of River Basins in Eastern Kansas	44
6.	Location of River Basin in Western Kansas	45

CHAPTER I

INTRODUCTION

The development of techniques to simulate hydrological events has been widely reported in recent years. The paucity of data in many areas where development of water resources systems was desirable has been a major obstacle to efficient design. To overcome this deficiency, hydrologists have found it more and more necessary to attack their problems with the tools of the statistician, to create synthetic data where none or little existed before. Unfortunately, in some circles overemphasis has been placed on synthetic data, which cannot, however sophisticated the techniques of analysis, be more accurate than the original parameters which were used in its generation. This has led in the recent past to a search for more complex methods of analysis than are probably warranted by the original data, or the conclusions which can safely be drawn from the results. However, generation of hydrological data may, if its results are used with caution, be a useful tool for the design engineer.

The analysis of the sequential occurrence of stream flows is based upon the assumption that they form part of a time series, which is considered to be infinite. The first valuable studies of time series were made by Fourier who proposed to his incredulous contemporaries that any series can be described by a process of sums of harmonics, even though the number of harmonics may be very large. However, a successful method to describe the harmonics has had to await the development of spectral analysis in recent years. Other methods have been the use of periodograms, developed by Shuster, and the use of correlograms, each of which attempt to show the significant periods in the harmonic cycle.

For many years research workers attempted to apply the methods of Fourier analysis to many types of time series, with varying degrees of success. However, in a departure from this concept of a series consisting of a sum of pure harmonics, possibly with superposed fluctuations, Yule (1927) considered a system comprising a periodic movement which was affected by true external disturbances. These disturbances would account for changes in phase and amplitude which had been observed by workers attempting harmonic analysis of natural series. Yule's investigation led to a regression equation of the form

$$\omega_t = \beta_1 \, \omega_{t-1} - \beta_2 \, \omega_{t-2} + \epsilon_t \tag{1.1}$$

where ϵ_t is a random variable at time t,

and β_1 and β_2 are constants, given by

$$\beta_1 = \frac{r_1(1 - r_2)}{1 - r_1^2} \tag{1.2}$$

$$\beta_2 = \frac{r_2 - r_1^2}{1 - r_1^2}. \tag{1.3}$$

Here, r_1 is the correlation between successive elements of the series separated by one, and for a general lag k

$$r_k = \frac{\mathrm{cov}[\omega(t) \cdot \omega(t + k)]}{[\mathrm{var}\ \omega(t) \cdot \mathrm{var}\ \omega(t + k)]^{1/2}} \tag{1.4}$$

These correlations were referred to by Yule (1926) as the serial correlation coefficients for the series.

Yule's equation (Equation 1.1), which is known as a process of linear autoregression, was used by Walker (1931) in an analysis of meteorological data. Its use for streamflow sequences has usually been limited to a first order form, ignoring the function of ω_{t-2}.

$$\omega_t = a\, \omega_{t-1} + \epsilon_t \tag{1.5}$$

Julian (1961) used an equation of this type in studies of streamflow sequences where the variable ω_t was the streamflow X_t at time t. The constant a may be shown in this case to be r_1 the first order serial correlation (v. Section 2.4.b). The method was also used by Brittan (1961), using the standardized variable

$$Z_t = \frac{X_t - m}{s}, \tag{1.6}$$

where m and s^2 are respectively the mean and variance of X_t, for the random variable ω_t. The discussion of Section (2.1.b) will show that this is usually a more valid variable.

A model which has received wide attention recently is that of Thomas and Fiering (1962). This is based essentially upon different assumptions to those used above. The standardized monthly flow in the month τ is assumed related to the standardized flow in the month τ-1 by a linear regression b_τ, with the addition of a random component which is a function of r_τ the correlation between these months. The standardized monthly flows are given by

$$Q_t = \frac{X_t - m_\tau}{S_\tau} \tag{1.7}$$

where m_τ and s_τ^2 are respectively the mean and variance of the month τ.

The autoregression equation is given by

$$X_t = \frac{S_\tau\ r_\tau}{S_{\tau-1}}\left(X_{t-1} - m_{\tau-1}\right) + m_\tau + S_\tau\ \eta_t (1 - r_\tau^2)^{1/2} \tag{1.8}$$

where η_t is a standardized normal random variable.

It can be shown that

$$\frac{S_\tau\ r_\tau}{S_{\tau-1}} = b_\tau \tag{1.9}$$

the regression coefficient.

The Thomas and Fiering model assumes that correlation exists between the months τ and τ-1. If correlation does not exist and $r_\tau = 0$, the model cannot be used. It is conceivable that no correlation could exist between months and Thomas and Fiering themselves found that correlations in some months, when tested for significance with the t-test (v. Section 6.4), were not significant. The model was used in studies for the Oklahoma Arkansas Water Planning Study by Perry (1968) and Dunaway (1968). It was successfully applied to river basins with large drainage areas where it was found that insignificant correlations did not arise. However, when applied to basins with small contributing area it was found that often as many as one-half of the correlations were not significant. It was concluded, therefore, that the model could not be used for basins with small areas, and it was hypothesized that rapid run-off after severe storms common in the study area resulted in this poor correlation between monthly flows.

This report is the result of an investigation to attempt to find a model or models which could be used to describe the monthly flows of small river basins which the Thomas and Fiering model had failed to describe. Nine small river basins in and around the study area were selected for analysis, and used to study the applicability of the proposed models. Subsequently, two larger basins were also examined. Statistical tests of the significance of the models, which are not applicable to the Thomas and Fiering model, were also applied.

CHAPTER II

THEORY OF ANALYTICAL METHODS

1. Time Series

a. Definition

Let $\{t\}$ denote a set of points in time and w_t be a variable corresponding to each point t. Such a series of variables is called a time series. The variable w_t may be considered to consist of two parts, one deterministic and a function of t, and the other random, not being a function of t or the deterministic element. Then:

$$\omega_t = \delta_t + \epsilon_t \tag{2.1}$$

where δ_t is a deterministic element

and ϵ_t is a random element.

If the deterministic element is absent, the series is completely random and may be denoted by

$$\omega_t = \epsilon_t \, . \tag{2.2}$$

Similarly, if the random element is absent

$$\omega_t = \delta_t \, . \tag{2.3}$$

The deterministic portion may be found by analysis of the series.

If both δ_t and ϵ_t are present, the time series may be one of two types. If the series can be described by a polynomial of the form

$$\omega_t = a_0 + a_1 t + a_2 t^2 + \dots + a_n t^n , \tag{2.4}$$

the series is described by a process of moving averages. Alternatively, if the series can be described by an expression of the form

$$\omega_t + a_1 \omega_{t-1} + \dots + a_n \omega_{t-n} = \epsilon_t , \tag{2.5}$$

the series is said to be a process of linear autoregression. The process of linear autoregression will be considered in this study.

b. <u>Stationarity</u>

As the process of linear autoregression is to be used only for stationary time series, the series under consideration must be stationary or made stationary by means of a transformation. Stationarity may be defined as follows.

Again, let $\{t\}$ denote a set of points in time and let w_t be a variable corresponding to each point t. The probability distribution function of $w(t_1, t_2, \dots, t_n)$, where $w(t_1, t_2, \dots, t_n)$ is a subset of $\{w_t\}$, is $F(t_1, \dots, t_n; u_1, \dots, u_n)$ and

$$F(t_1, \dots, t_n; u_1, \dots, u_n) = \Pr[w(t_1) < u_1, \dots, w(t_n) < u_n]. \tag{2.6}$$

The set $\{w_t\}$ is termed stationary if, for all $(u_1, \dots, u_n)$, the relationship

$$F(t_1, \ldots, t_n; u_1, \ldots, u_n) = F(t_1 + k, \ldots, t_n + k; u_1, \ldots, u_n) \tag{2.7}$$

is satisfied for all $k < n$ (Wold, 1954), where k is referred to as the lag. Thus, in any subset of the population $\{w_t\}$ statistical parameters obtained from this subset should not vary from those obtained from other subsets by more than is expected by chance. Mathematical expectations (denoted by the symbol E) obtained from this distribution function may be used to describe stationarity in terms of these parameters.

Thus, stationarity of the first order is defined by

$$E[w_t] = \mu = \text{constant} \tag{2.8}$$

where μ is the mean of the population $\{w_t\}$.

Second order stationarity can be defined as

$$E[w_t \cdot w_{t+k}] = \text{constant.} \tag{2.9}$$

As the serial correlation coefficient between w_t and w_{t+k} is defined as

$$\rho_k = \frac{\text{cov}[w_t \cdot w_{t+k}]}{[\text{var}(w_t) \cdot \text{var}(w_{t+k})]^{1/2}} \tag{2.10}$$

and by the hypothesis of stationarity

$$E[w_t] = E[w_{t+k}] = \mu \tag{2.11}$$

and

$$\mathrm{var}(w_t) = \mathrm{var}(w_{t+k}) = \sigma^2 \qquad (2.12)$$

where σ^2 is the variance of the population w_t

$$\rho_k = \frac{E[w_t \cdot w_{t+k}] - \mu^2}{\sigma^2}. \qquad (2.13)$$

Thus,

$$E[\omega_t \cdot \omega_{t+k}] = \rho_k \sigma^2 + \mu^2 = \text{constant} \qquad (2.14)$$

[Roesner and Yevdjevich (1966)]

If an observed series is normally distributed and is stationary to the first and second orders, it is stationary to all higher orders (Matalas, 1967 a). However, such higher orders are beyond the scope of this theoretical discussion.

An observed hydrological sequence $\{X_t\}$, where X_t is the mean monthly flow in the month t, may be considered to be a sample from a population of the form $\{w_t\}$. Such a sequence is rarely found to be stationary, because the period of record is too short for finite subsets to have identical statistical parameters. However, the series may be standardized by means of the transformation

$$Z_t = \frac{X_t - m}{s} \qquad (2.15)$$

where m and s^2 are respectively the mean and variance of $\{X_t\}$, thus

$$m = \frac{1}{n} \sum_{t=1}^{n} X_t \tag{2.16}$$

$$s^2 = \frac{1}{(n-1)} \sum_{t=1}^{n} (X_t - m)^2 \tag{2.17}$$

there being n observations of X_t.

The standardized variable Z_t thus has mean of zero and variance of one and is stationary to the second order.

2. Serial Correlation

The observed series $\{X_t\}$, $(X_1, X_2, \ldots, X_n)$, may be broken down into $(n - 1)$ pairs of series of the form

$$\left.\begin{array}{l} X_1, X_2, \ldots, X_{n-k} \\ \\ X_k, X_{k+1}, \ldots, X_n \end{array}\right\} k < n.$$

The serial correlation coefficient of the series for a lag k is defined as

$$r_k = \frac{\text{cov}[X_t \cdot X_{t+k}]}{[\text{var}(X_t) \cdot \text{var}(X_{t+k})]^{1/2}} \tag{2.18}$$

$$= \frac{E[X_t \ X_{t+k}] - m_t \ m_{t+k}}{[\text{var}(X_t)]^{1/2} \cdot [\text{var}(X_{t+k})]^{1/2}} \tag{2.19}$$

$$= \frac{\sum_{t=1}^{n-k} (X_t X_{t+k}) - \frac{1}{(n-k)} \sum_{t=1}^{n-k} X_t \sum_{t=1}^{n-k} X_{t+k}}{\left[\sum_{t=1}^{n-k} X_t^2 - \frac{1}{(n-k)}\left(\sum_{t=1}^{n-k} X_t\right)^2\right]^{1/2} \left[\sum_{t=1}^{n-k} X_{t+k}^2 - \frac{1}{(n-k)}\left(\sum_{t=1}^{n-k} X_{t+k}\right)^2\right]^{1/2}} \qquad (2.20)$$

This can be seen to be analogous to the correlation between dependent and independent variables. However, by the hypothesis of stationarity $var(X_t) \doteq var(X_{t+k})$. Therefore, Equation (2.20) may be considerably simplified by writing

$$r_k = \frac{\sum_{t=1}^{n-k} (X_t X_{t+k}) - \frac{1}{(n-k)} \sum_{t=1}^{n-k} X_t \sum_{t=1}^{n-k} X_{t+k}}{var\ (X_t)}. \qquad (2.21)$$

As the series $(X_1, X_2, \ldots, X_n)$ is assumed to be a subset of an infinite series $\{X\}$, r_k is an estimate of the population serial correlation coefficient ρ_k, referred to hereafter as the autocorrelation coefficient. For a stationary series $r_k \to \rho_k$ as $n \to \infty$. From the series $\{X_t\}$, $(n-1)$ values of r_k may be computed. However, r_k loses its significance as k increases and ΣX_t and ΣX_{t+k} in Equation (2.21) begin to differ significantly. Although no precise limits for k may be set, Blackman and Tukey (1958) have recommended that $k \ngtr n/10$.

A plot of the serial correlation coefficients r_k against the lag k is called a correlogram. The shape of the correlogram, which may be formed by joining the points of the plot with straight lines (although the graph is not strictly continuous) may reveal the nature of the time series. Kendall (1951) describes four types of correlogram. A random series (Equation 2.2) has a correlogram which is a straight line with $r_k = 0$ for all k, as the serial correlation is zero. A correlogram which oscillates and is not damped is typical of a time series which consists of a sum of harmonic components (Equation 2.3). A correlogram which oscillates but is damped quickly and vanishes is typical of a time series described by a scheme of moving averages (Equation 2.4). A correlogram which is damped but does not vanish is typical of a scheme of linear autoregression (Equation 2.5).

The correlogram of the independent series will only approach a straight line with $r_k = 0$ as $n \to \infty$. Anderson (1942) has shown that if the sample $\{X_t\}$ is normally distributed about its mean with variance of one, r_k may be considered zero if at significance level α

$$\frac{-1 - K_\alpha (n-2)^{1/2}}{(n-1)} < r_k < \frac{-1 + K_\alpha (n-2)^{1/2}}{(n-1)} \qquad (2.22)$$

where K_α is the two-tailed standard normal deviate at significance level α.

If the correlogram falls within these limits, the series is considered to be independent.

3. Separation of Deterministic and Random Elements

As the observed series $\{X_t\}$ is a subset of a population $\{X\}$, in accordance with Equation (2.1)

$$X_t = \delta_t + \epsilon_t. \qquad (2.23)$$

Inspection of the correlogram of the series X_t may reveal the nature of the time series. If $r_k = 0$ at significance level α, the series is random and δ_t will be absent. If the correlogram shows distinct cycles and is not damped, it may be possible to describe the deterministic element as a sum of harmonic components.

The hydrograph of monthly flows from a river basin with typical seasonal flow pattern suggests that the cycle of movement of monthly flows may be described as a periodic function of the form

$$m_t - m_{t-h} = 0 \tag{2.24}$$

where h is the period of the cyclic movement and m_t the mean monthly flow in the month t.

This is a difference equation of order h, whose solution may be written

$$m_t = m + \sum_{p=1}^{n} K_p \sin\left(\frac{2\pi}{h} pt + d_p\right) \tag{2.25}$$

[Wold (1954)]

where p = order of harmonic

n = number of harmonics $< \frac{h}{2}$

d_p = phase of cycle

K_p = constant.

By means of the identity

$$\sin(\alpha + \beta) = \sin\alpha \cdot \cos\beta + \cos\alpha \cdot \sin\beta.$$

Equation (2.25) may be rewritten as

$$m_t = m + \sum_{p=1}^{n} A_p \cos \frac{2\pi}{h} pt + \sum_{p=1}^{n} B_p \sin \frac{2\pi}{h} pt \tag{2.26}$$

where

$$A_p = \frac{2}{h} \sum_{t=1}^{h} X_t \cos \frac{2\pi}{h} pt \tag{2.27}$$

$$B_p = \frac{2}{h} \sum_{t=1}^{h} X_t \sin \frac{2\pi}{h} pt \tag{2.28}$$

[Brooks and Carruthers (1953)].

The constants A_p and B_p are, therefore, determined by the first h terms of the series $\{X_t\}$. However, inspection of the series shows that an annual cycle usually predominates and h is, therefore, chosen to be twelve. The constants A_p and B_p are then formed from the flows of the first year of the observed series, but to make them more representative of the whole series they are chosen to be defined in terms of the mean m_τ of the month $\tau(\tau = 1, 2, \ldots, 12)$ where

$$m_\tau = \frac{1}{N} \sum_{t=1}^{\tau+12(N-1)} X_t \tag{2.29}$$

$t = \tau + 12i$

$i = 0, 1, 2, \ldots, (N-1)$

N = number of complete years in record.

Hence

$$A_p = \frac{2}{12} \sum_{\tau=1}^{12} m_\tau \cos \frac{2\pi}{12} p\tau \tag{2.30}$$

$$B_p = \frac{2}{12} \sum_{\tau=1}^{12} m_\tau \sin \frac{2\pi}{12} p\tau . \tag{2.31}$$

Equation (2.26) then becomes

$$m_t = m + \sum_{p=1}^{n} A_p \cos \frac{2\pi}{12} pt + \sum_{p=1}^{n} B_p \sin \frac{2\pi}{12} pt . \tag{2.32}$$

By a similar derivation, a continuous function of the standard deviation may be constructed, using the standard deviation s_τ of the month τ where

$$s_\tau^2 = \frac{1}{(N-1)} \sum_{t=1}^{\tau+12(N-1)} (X_t - m_\tau)^2 . \tag{2.33}$$

Then

$$s_t = s + \sum_{p=1}^{n} C_p \cos \frac{2\pi}{12} pt + \sum_{p=1}^{n} D_p \sin \frac{2\pi}{12} pt \tag{2.34}$$

where

$$C_p = \frac{2}{12}\sum_{\tau=1}^{12} s_\tau \cos \frac{2\pi}{12} p\tau \qquad (2.35)$$

$$D_p = \frac{2}{12}\sum_{\tau=1}^{12} s_\tau \sin \frac{2\pi}{12} p\tau. \qquad (2.36)$$

4. Stochastic Models

a. Standardization of Variables

The series $\{X_t\}$ or some transformation of $\{X_t\}$ may then be standardized by means of Equation (2.15).

Then

$$Z_t = \frac{X_t - m}{s} \qquad (2.37)$$

will be called the standardized series of $\{X_t\}$.

Alternatively, the continuous functions m_t and s_t of Equations (2.32) and (2.34) may be used, whence

$$Z_t'' = \frac{X_t - m_t}{s_t}. \qquad (2.38)$$

However, this series does not necessarily have mean of zero and variance of one, and must be standardized by means of the expression

$$Z_t' = \frac{Z_t'' - m_Z}{s_Z} \qquad (2.39)$$

where m_z and s_z are respectively the mean and variance of Z_t''. This expression, Equation (2.39), will be called the fitted series.

b. Autoregressive Schemes

As the series $\{Z_t\}$ and $\{Z_t'\}$ are stationary they may be described by a process of linear autoregression. From Equation (2.5), for a first order autoregressive scheme

$$Z_t + a_1 Z_{t-1} = \epsilon_t . \tag{2.40}$$

This represents the regression of Z_t on Z_{t-1}, the term ϵ_t being a residual error. The constant a_1 may be found by regression, and is seen to be $-r_1$ the first order serial correlation coefficient. Thus,

$$Z_t - r_1 Z_{t-1} = \epsilon_t \tag{2.41}$$

is the first order autoregressive scheme for generating Z_t.

For the second order scheme, from Equation (2.5)

$$Z_t + \beta_1 Z_{t-1} + \beta_2 Z_{t-2} = \epsilon_t'. \tag{2.42}$$

The coefficients β_1 and β_2 which may be determined by regression are

$$\beta_1 = \frac{r_1 - r_1 r_2}{1 - r_1^2} \tag{2.43}$$

$$\beta_2 = \frac{r_2 - r_1^2}{1 - r_1^2} \tag{2.44}$$

Thus

$$Z_t - \left(\frac{r_1 - r_1 r_2}{1 - r_1^2}\right) Z_{t-1} + \left(\frac{r_2 - r_1^2}{1 - r_1^2}\right) Z_{t-2} = \epsilon_t' \tag{2.45}$$

is the expression for the second order autoregressive scheme.

The residual ϵ_t in the above expressions is independent of Z and ϵ. Considering the first order scheme, let

$$\eta_t = \frac{\epsilon_t}{\lambda} \tag{2.46}$$

where $\lambda^2 = \text{var}\ (\epsilon_t)$.

η_t is then a standardized independent variable. Further, as

$$\text{var}\ (Z_t) = \text{var}\ (Z_{t-1})$$

and using the expansions of variance and covariance

$$\begin{aligned}
\text{var}\ (\epsilon_t) &= \text{var}\ (Z_t - r_1\ Z_{t-1}) \\
&= 1 + r_1^2 - 2\ \text{cov}\ (Z_t,\ r_1\ Z_{t-1}) \\
&= 1 + r_1^2 - 2\ \text{cov}\ (r_1\ Z_{t-1} + \epsilon_t,\ r_1\ Z_{t-1}) \\
&= 1 + r_1^2 - 2[\text{cov}(r_1\ Z_{t-1},\ \epsilon_t) + r_1^2] \\
&= 1 + r_1^2 - 2r_1^2
\end{aligned}$$

as Z_{t-1} and ϵ_t are independent.

$$\therefore\ \lambda^2 = 1 - r_1^2 \tag{2.48}$$

$$\text{and } \epsilon_t = \eta_t (1 - r_1^2)^{1/2}. \tag{2.49}$$

Thus, the autoregressive scheme of Equation (2.41) becomes

$$Z_t = r_1 Z_{t-1} + \eta_t (1 - r_1^2)^{1/2}. \tag{2.50}$$

If the standardized series of $\{Z_t\}$ (Equation 2.37) is used, Equation (2.50) becomes

$$\left(\frac{X_t - m}{s}\right) - r_1\left(\frac{X_{t-1} - m}{s}\right) = \eta_t (1 - r_1^2)^{1/2} \tag{2.51}$$

whence

$$X_t = r_1 X_{t-1} + m(1 - r_1) + s\eta_t (1 - r_1^2)^{1/2}. \tag{2.52}$$

This is a form of the simple first order expression widely used in hydrological studies and will be referred to as Model IA.

If the Fitted Series referred to above (Equation 2.39) is used, Equation (2.41) gives

$$X_t = \frac{s_t}{s_{t-1}} r_1 \ (X_{t-1} - m_{t-1}) + m_t - s_t (1 - r_1) m_Z + s_Z s_t \eta_t (1 - r_1^2)^{1/2}. \tag{2.53}$$

This expression will be referred to as Model IB.

Instead of using the Standardized or Fitted Series of $\{Z_t\}$, a model may be formed using the series

$$Y_t = \frac{X_t - m_\tau}{s_\tau} \tag{2.54}$$

where as above m_τ and s_τ^2 are the mean and variance of the month τ. Equation (2.41) then becomes

$$\left(\frac{X_t - m_\tau}{s_\tau}\right) - r_1 \left(\frac{X_{t-1} - m_{\tau-1}}{s_{\tau-1}}\right) = \eta_t (1 - r_1^2)^{1/2} \tag{2.55}$$

and

$$X_t = \frac{s_\tau \, r_1}{s_{\tau-1}} (X_{t-1} - m_{\tau-1}) + m_\tau + s_\tau \, \eta_t (1 - r_1^2)^{1/2}. \tag{2.56}$$

This is similar to the autoregressive scheme used by Thomas and Fiering (1962), Equation (1.8). However, in their expression r_τ was the correlation coefficient between the months τ and $\tau-1$, there being twelve values of r_τ. In that case, it can be shown that

$$\frac{r_\tau \, s_\tau}{s_{\tau-1}} = b_\tau \tag{2.57}$$

where b_τ is the regression coefficient used by Thomas and Fiering in their expression.

c. Skewed Distributions

Models IA and IB were constructed for series having insignificant skewness. If, however, the observed series $\{X_t\}$ is drawn from a population $\{X\}$ with skewness γ_X, the skewness g_X of $\{X_t\}$ is an estimate of γ_X, where

$$\gamma_X = \frac{\mu_3}{\mu_2^{3/2}} \tag{2.58}$$

μ_3 = third moment about mean = $\frac{1}{n}\sum (X - \mu)^3$

μ_2 = second moment about mean = σ^2.

Accordingly, the standardized variable η_t of Equation (2.50) must also be skewed. Denoting such a skewed variable by ξ_t, with skewness γ_ξ,

$$\xi_t = \frac{2}{\gamma_\xi}\left(1 + \frac{\gamma_\xi \eta_t}{6} - \frac{\gamma_\xi^2}{36}\right) - \frac{2}{\gamma_\xi} \tag{2.59}$$

[Matalas (1967b)]

where as before $\eta_t \sim n(0,1)$.

Thomas and Fiering (1963) have shown that the skewness of ξ_t is related to the skewness of $\{X\}$ by

$$\gamma_\xi = \frac{(1 - \rho_1^3)}{(1 - \rho_1^2)^{3/2}} \cdot \gamma_X. \tag{2.60}$$

As g_X, expressed by

$$g_x = \frac{n^2}{(n - 1)\ (n - 2)} \cdot \frac{m_3}{s^3} \tag{2.61}$$

where $\quad m_3 = \frac{1}{n} \sum_{t=1}^{n} (X_t - m)^3$

is an estimate of γ_x, and r_1 an estimate of ρ_1, the estimate g_ξ of γ_ξ may be represented as

$$g_\xi = \frac{(1 - r_1^3)}{(1 - r_1^2)^{3/2}} \cdot g_x \tag{2.63}$$

and thus

$$\xi_t = \frac{2}{g_\xi}\left(1 + \frac{g_\xi\ \eta_t}{6} - \frac{g_\xi^2}{36}\right) - \frac{2}{g_\xi}. \tag{2.64}$$

Equation (2.50) then becomes

$$Z_t = r_1\ Z_{t-1} + \xi_t (1 - r_1^2)^{1/2} \tag{2.65}$$

and Model IIA is the expression

$$X_t = r_1\ X_{t-1} + m(1 - r_1) + s\ \xi_t (1 - r_1^2)^{1/2} \tag{2.66}$$

and Model IIB the expression

$$X_t = \frac{s_t\ r_1}{s_{t-1}} (X_{t-1} - m_{t-1}) + m_t - s_t (1 - r_1) m_Z + s_Z s_t \xi_t (1 - r_1^2)^{1/2}. \tag{2.67}$$

5. Tests of Goodness of Fit of Autoregressive Schemes

The expression representing the first order autoregressive scheme is Equation (2.41) which may be rearranged thus

$$\epsilon_t = Z_t - r_1 Z_{t-1} . \tag{2.68}$$

If this expression describes the process, ϵ_t should be stochastically independent (Roesner and Yevdjevich, 1966). This independence may be tested by constructing the correlogram of ϵ_t, and determining whether the serial correlation coefficients of ϵ_t fall within the confidence limits of Equation (2.22),

$$\text{C.L.}(\alpha) = \frac{-1 \pm K_\alpha (n-2)^{1/2}}{n-1} . \tag{2.69}$$

As mentioned above, this test is only applicable if ϵ_t is normally distributed with variance one. If the residual ϵ_t is found to be independent in this way the hypothesis that Equation (2.41) represents the autoregressive scheme is accepted. Similarly ϵ_t' for the second order scheme (Equation 2.45) may be tested in the same way.

Alternatively, a large sample χ^2-test proposed by Quenouille (1947) may be used. The statistic R_{p+k} which is $\sim n[0, \sigma^2(R_p)]$ is used to construct a large sample χ^2-test of the form

$$\chi^2_{\ell-p} = \sum_{k=1}^{\ell-p} \frac{R_{p+k}^2}{\sigma^2(R_p)} \tag{2.70}$$

to test whether the autoregressive scheme is of order p. Here ℓ is the number of lags used to estimate ρ_k and $(\ell - p)$ is the number of degrees of freedom. The hypothesis that the autoregressive scheme is of order p is rejected if $\chi^2_{\ell-p} > \chi^2(\alpha)$, the value of χ^2 with $(\ell - p)$ degrees of freedom at significance level α.

To test whether the autoregressive scheme is of order $p = 1$, Matalas (1967a) has shown that

$$R_{1+k} = r_{k+1} - 2\, r_1\, r_k + r_1^2\, r_{k-1} \tag{2.71}$$

and

$$\sigma^2(R_1) = \frac{1}{(n-k)} \cdot (1 - r_1^2)^2. \tag{2.72}$$

Similarly, for the second order scheme, where $p = 2$

$$R_{2+k} = r_{k+2} - 2\beta_1 r_{k+1} + (\beta_1^2 - 2\beta_2) r_k + 2\beta_1\beta_2 r_{k-1}$$

$$+ \beta_2^2\, r_{k-2} \tag{2.73}$$

$$\sigma^2(R_2) = \frac{1}{(n-k)} [(1 + \beta_1^2 - \beta_2^2 - 2\beta_1 r_1)^2 +$$

$$2(r_1 - \beta_1 - \beta_2\, r_1^2)] \tag{2.74}$$

where $\beta_1 = \dfrac{r_1 - r_1 \cdot r_2}{1 - r_1^2}$ (2.75)

$$\beta_2 = \frac{r_2 - r_1^2}{1 - r_1^2}. \tag{2.76}$$

6. Distribution of Variables

a. Introduction

As mentioned above, the analytical methods used in this study are applicable only to time series which are stationary to the second order. Only if the distribution of the variables is normal can the series be considered stationary to all higher orders (Matalas, 1967a). Hydrological sequences have been found to be approximately normally distributed, or may be transformed to a normal distribution. Other distributions, particularly those of Pearson and Gumbel have been found to fit hydrological sequences (Matalas, 1963 ; Markovic, 1965) but detailed consideration of these distributions is beyond the scope of this study.

A variable X having probability density function

$$f(x) = \frac{1}{b(2\pi)^{1/2}} e^{-\frac{1}{2}\left(\frac{x - a}{b}\right)^2} \tag{2.77}$$

is said to be normally distributed with parameters a and b, and is denoted by $X \sim n(a, b)$. To test whether an observed series $\{X_t\}$ is normally distributed, it is necessary to compare the frequency distribution of the variables with that of a normal distribution. The sample may be divided into k mutually exclusive classes and the relative frequency of events in each class compared with that of a normal distribution. Frequently, this analysis is done with class intervals of equal length and variable probability. The choice of class intervals of equal probability has been shown by Markovic (1965) to lead to simplicity in computation. This method is based upon the analysis of Mann and Wald (1942). As there is no theoretical method for determining the most suitable number of class intervals statisticians have formulated numerous rules. It is generally accepted that too few classes may obscure the main characteristics of the distribution, and that too many classes may overemphasize chance variations. A commonly accepted rule is that the expected class frequency f_j of any class j,

$$E[f_j] \geq 5 \tag{2.78}$$

[Hald (1952)].

b. Transformations of the Variable X_t

It was decided to analyze the distribution of $\{X\}$ and $\{\log_e X\}$. The distribution functions of the transformations considered are derived from Equation (2.77) as follows.

i. Normal Distribution

$$f(x) = \frac{1}{\sigma(2\pi)^{1/2}} \, e^{-\left(\frac{X - \mu}{\sigma}\right)^2} . \tag{2.79}$$

ii. Log-normal Distribution

$$f(x) = \frac{1}{\sigma_1 (2\pi)^{1/2}} \, e^{-\left(\frac{\log_e X - \log_e \mu_1}{\sigma_1}\right)} \tag{2.80}$$

where $\log_e \mu_1$ and σ_1 are respectively the mean and variance of $\{\log_e X\}$.

c. Estimation of Parameters

As the parameters μ and σ^2 of the population are not known they must be estimated. If a random sample having parameters m and s is taken from the population, it can be shown by the Method of Maximum Likelihood that

$$\mu = m \tag{2.81}$$

and

$$\sigma^2 = \frac{(n-1)}{n} \cdot s^2 \tag{2.82}$$

[Hald (1952)].

However, for n large, $\frac{(n-1)}{n} \rightarrow 1$; thus,

$$\sigma^2 = s^2 . \tag{2.83}$$

The sample parameters may then be used as estimates of the population parameters, and, as before

$$\mu = m = \frac{1}{n} \sum_{t=1}^{n} X_t \tag{2.84}$$

$$\sigma^2 = s^2 = \frac{1}{(n-1)} \sum_{t=1}^{n} (X_t - m)^2 \tag{2.85}$$

Similarly

$$\log_e \mu_1 = \log_e m_1 = \frac{1}{n}\sum_{t=1}^{n} (\log_e X_t) \qquad (2.86)$$

$$\sigma_1^2 = s_1^2 = \frac{1}{(n-1)}\sum_{t=1}^{n} (\log_e X_t - \log_e m_1)^2 . \qquad (2.87)$$

d. Testing Goodness of Fit of Observations

The χ^2-test may be used to test the goodness of fit of the observed series with the theoretical probability distributions of Equations (2.79) and (2.80). The test, developed by Pearson (1900) may be summarized as follows.

If a set of random variables x_i $(i = 1, 2, \ldots, n)$ are stochastically independent, it can be shown that the statistic

$$Q_{n-1} = \sum_{i=1}^{n} \frac{(x_i - n\pi_i)^2}{n\pi_i} \qquad (2.88)$$

where π_i is the probability of occurrence of the event x_i is approximately distributed as χ^2 with $(n - 1)$ degrees of freedom, the approximation being more valid if $n\pi_i > 5$ (Equation 2.78). It is assumed thus far that the probabilities π_i are known. However, in a random sample drawn from a population with parameters μ and σ^2, only the estimates p_i are known, as are the estimates m and s^2 of the population parameters. Fisher (1924) has shown that the number of degrees of freedom $(n - 1)$ of Equation (2.88) must be reduced by c, the number of population parameters which are estimated from the sample. Equation (2.88) then becomes

$$\chi^2_{n-c-1} = \sum_{i=1}^{n} \frac{(x_i - np_i)^2}{np_i} \tag{2.89}$$

$$= \sum_{i=1}^{n} \frac{(x_i^2 - 2nx_i p_i + n^2 p_i^2)}{np_i}$$

$$= \frac{1}{n} \sum_{i=1}^{n} \frac{x_i^2}{p_i} - 2 \sum_{i=1}^{n} x_i + n \sum_{i=1}^{n} p_i .$$

But $(x_1 + x_2 + \ldots + x_n) = n$

and $(p_1 + p_2 + \ldots + p_n) = 1.$

Therefore,

$$\chi^2_{n-c-1} = \frac{1}{n} \left(\sum_{i=1}^{n} \frac{x_i^2}{p_i} \right) - n. \tag{2.90}$$

If the sample is grouped into k classes and the frequency of events in the j^{th} class is f_j, then

$$\chi^2_{k-c-1} = \frac{k}{n} \left(\sum_{j=1}^{n} f_j^2 \right) - n. \tag{2.91}$$

The hypothesis that the observed sequence agrees with the theoretical distribution is rejected if $\chi^2_{k-c-1} > \chi^2(\alpha)$ at (k-c-1) degrees of freedom for a given significance level α.

e. Class Boundaries

Class boundaries for k classes of equal probability may be determined from the normal probability distribution functions of Equations (2.79) and (2.80). The solution may be simplified by standardizing the variables X_t and $\log_e X_t$ by writing

$$Y = \frac{X_t - m}{s} \tag{2.92}$$

or

$$Y_1 = \frac{\log_e X_t - \log_e m_1}{s_1}. \tag{2.93}$$

Then, for both Y and Y_1

$$Pr[-\infty < Y < b_j] = f(b_j) = \int_{-\infty}^{b_j} \frac{1}{2\pi} e^{-\frac{y^2}{2}} dy \tag{2.94}$$

for any class boundary b_j ($j = 1, 2, \ldots, k$) in the standardized series. No definite solution to this integral exists, but approximate values are tabulated. From the class boundaries b_j of the standardized series, the boundaries B_j of the observed series are obtained from the expression

$$B_j = m + b_j s. \tag{2.95}$$

f. Skewness of Distribution

The skewness g_x of X_t is estimated by the expression (Equation 2.60)

$$g_x = \frac{n^2}{(n - 1)(n - 2)} \cdot \frac{m_3}{s^3}. \tag{2.96}$$

The coefficient g_x is $\sim n\left[0, \frac{6}{(n + 3)}\right]$ (Snedecor and Cochran, 1967). Confidence limits for g_x are then

$$C.L. = \pm K_\alpha\left[\frac{6}{(n + 3)}\right]^{1/2}. \tag{2.97}$$

If g_x falls within these confidence limits, the hypothesis that g_x is zero is accepted at significance level α.

CHAPTER III

METHODS OF COMPUTATION

1. Autoregressive Schemes

The complexity of the calculations involved in the procedures described in Chapter II made the use of a computer essential. A program system was set up to analyze the record from any station given its mean monthly flows. According to the controls chosen, this program analyzed both models described in Chapter II: Model A (Equation 2.52)

$$X_t = r_1 X_{t-1} + m(1 - r_1) + s\eta_t (1 - r_1^2)^{1/2} \tag{3.1}$$

and Model B (Equation 2.53)

$$X_t = \frac{s_t}{s_{t-1}} r_1 (X_{t-1} - m_{t-1}) + m_t - s_t (1 - r_1) m_z + s_z s_t \eta_t (1 - r_1^2)^{1/2} . \tag{3.2}$$

The program also tested the possible validity of the second order scheme of the form of Equation (2.45). The program examined the untransformed flows, the logarithmic flows, or both. A flow chart of the program showing its possible operation combinations is shown in Figure 1. The operations conducted in Models A and B are discussed below. Many of these operations were carried out by subprograms under

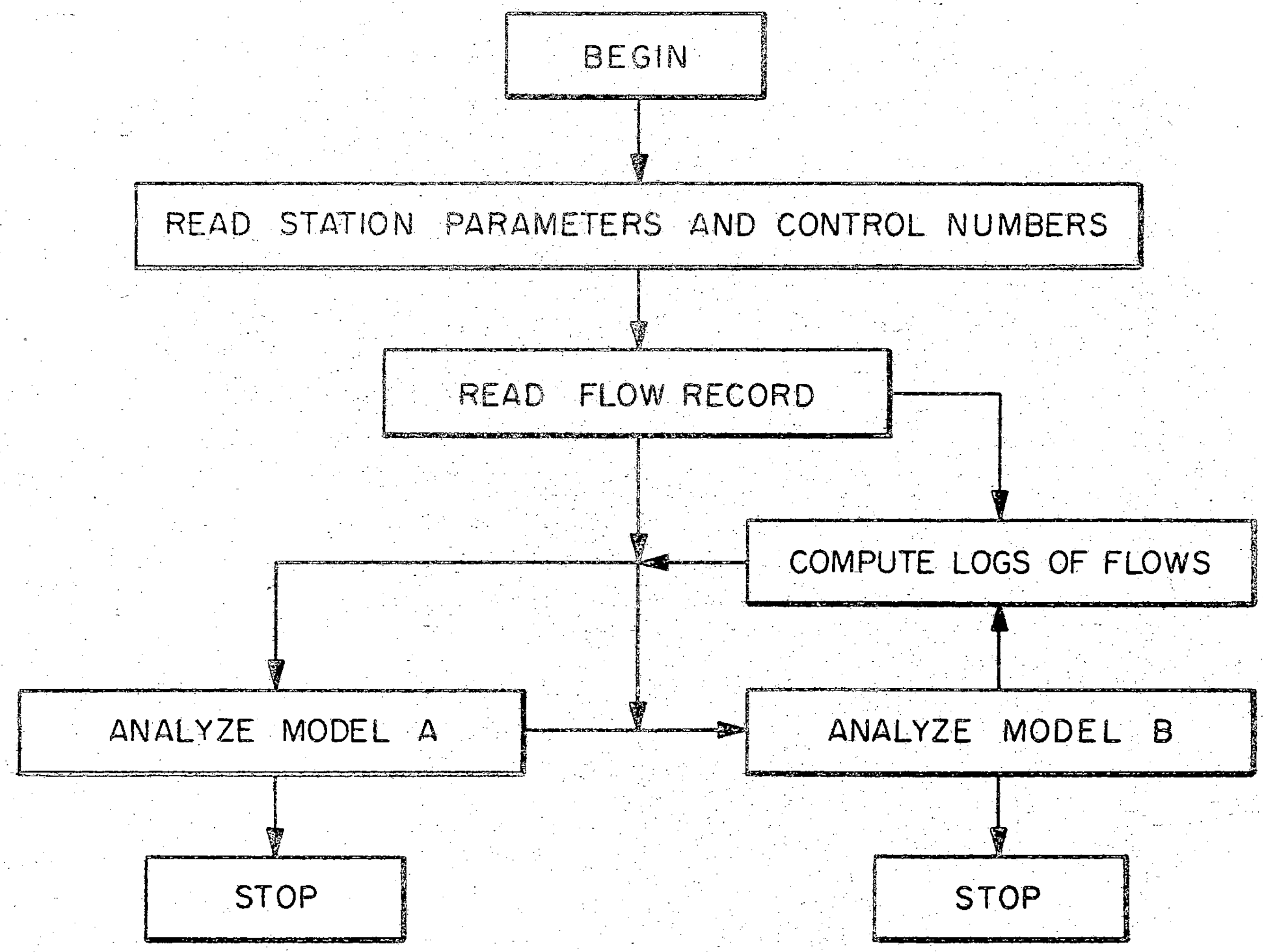


Figure 1. Flow Chart of Program System

the control of these models, which were themselves subprograms under the control of the main program.

The subprogram for analysis of Model A is shown in flow chart form in Figure 2. The program computed the mean variance and skewness of $\{X_t\}$. The skewness was tested for significance by means of the confidence limits of Equation (2.97). If the skewness was significant, the skew parameter g_ξ of Equation (2.63) was calculated, and the program wrote the mean and variance, skew coefficient and skew parameter. If the skewness was not significant, the program wrote the mean and variance, and the message that the skewness was insignificant. The program then analyzed the distribution of $\{X_t\}$. The analysis of distribution is more fully discussed below. The standardized series of Equation (2.37) was then formed and correlations computed on the variable $\{Z_t\}$. Twenty-five lag correlations were calculated using Equation (2.21), and used to test both first and second order models by use of Equation (2.70). The computed values of χ^2 were compared with χ^2 at significance level 0.05. For the first order model, the number of degrees of freedom was 24, and for the second order model 23. Either model was accepted if the computed χ^2 was less than $\chi^2_{0.05}$ at the stated number of degrees of freedom.

The series ϵ_t was also constructed in accordance with Equation (2.68) and correlations for 25 lags calculated. To determine whether the residual had a normal distribution and variance of one, in order to be tested by the confidence limits of Equation (2.69), its mean and variance were calculated and distribution examined. The computed correlations were punched on cards for subsequent examination.

The flow chart of the subprogram to analyze Model B is shown in

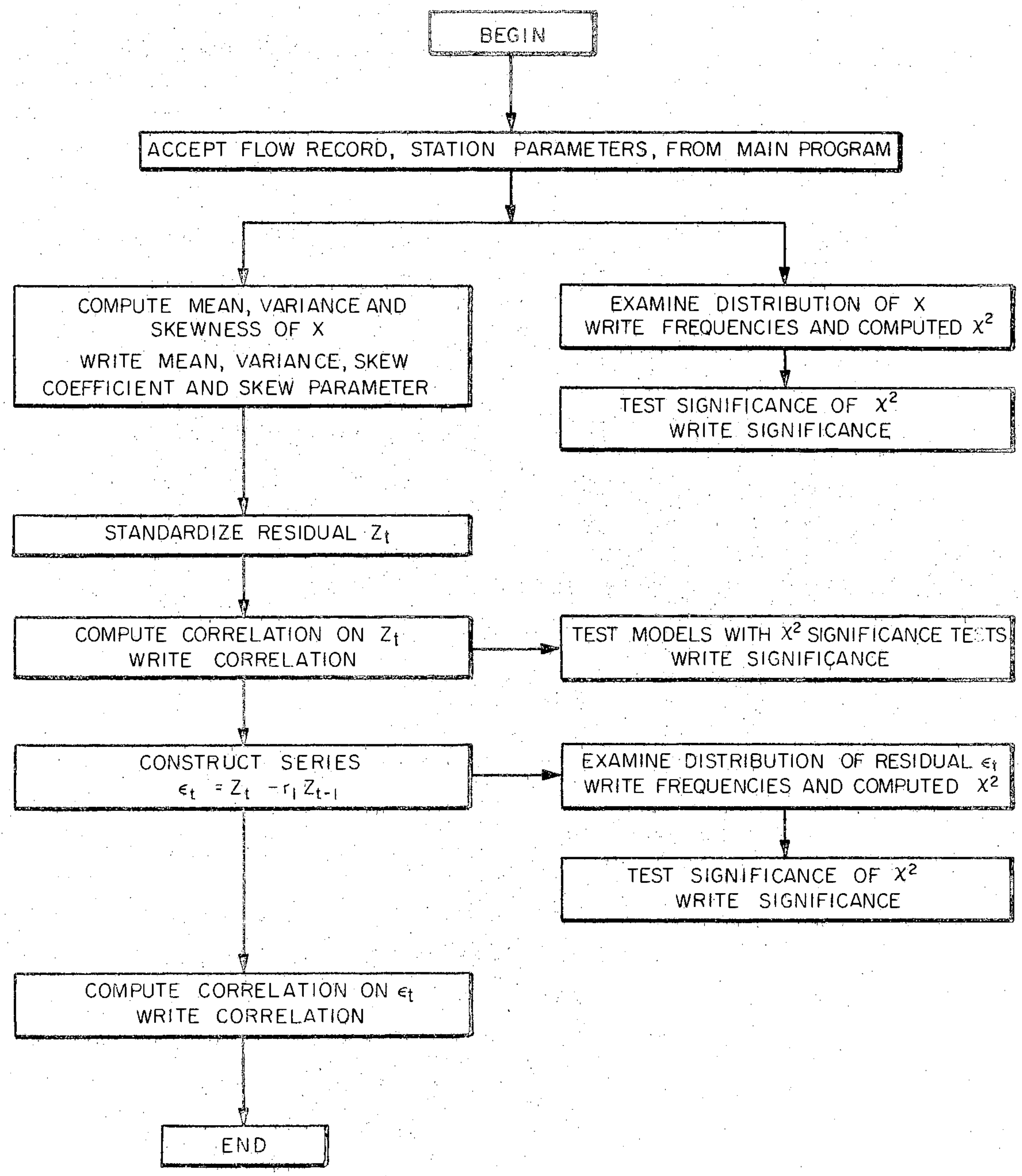


Figure 2. Flow Chart of Model A Analysis

Figure 3. The first part of this program was set up to remove harmonics from the time series in accordance with the method of Section 2.3 until the residual series had correlation which was insignificant at a chosen probability level, or until six harmonics had been removed. After calculation of the confidence limits of Equation (2.69), correlations of twenty-five lags were first run on the original series and the correlogram thus produced examined for significance. If such significance existed, the first harmonic was removed, correlation was repeated and the correlogram again examined. It was found that the test of significance applied in the program was too rigid, and in all cases six harmonics were automatically removed, and significant residual correlation indicated. However, in certain circumstances the residual had become insignificant prior to this point. This was determined by subsequent inspection of the correlograms (plotted by means of a separate subprogram) from the serial correlation coefficients calculated after each harmonic removal. The number of harmonics to be removed was selected by inspection and the program rerun. This aspect is more fully discussed below in Chapter V.

When the desired number of harmonics had been removed, the constants A_p, B_p, C_p, D_p of Equations (2.30), (2.31), (2.35), and (2.36) were tabulated. The program then analyzed the residual $\{Z_t''\}$ remaining after the removal of the continuous functions of m_t and s_t as in Equation (2.38). The mean variance and skewness of $\{Z_t''\}$ were calculated as for Model A, and the significance of the skewness tested in the same way. The skew parameter g_ξ was calculated if necessary and the program wrote the parameters of the series as for Model A. The distribution of $\{Z_t''\}$ was also examined. The program then constructed the fitted

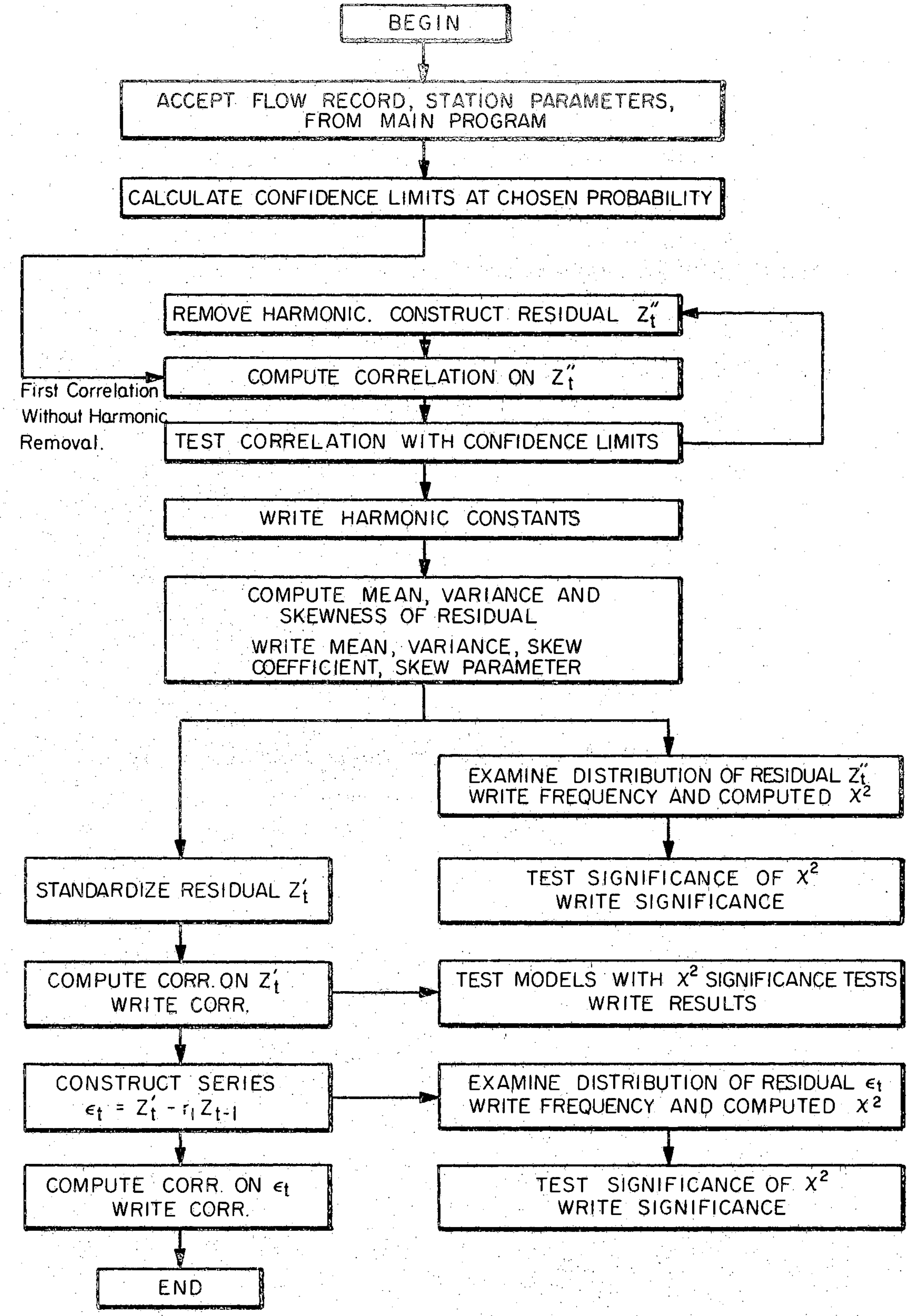


Figure 3. Flow Chart of Model B Analysis

series of Equation (2.39) and twenty-five lag correlations were calculated. These were used to test the validity of both first and second order schemes as described for Model A. Similarly, the series $\{\epsilon_t\}$ was produced and its parameters calculated and distribution examined. Twenty-five lag correlations calculated on $\{\epsilon_t\}$ were punched on cards for subsequent examination.

2. Distribution

The distributions of the variables $\{X_t\}$ and the residuals $\{Z_t''\}$ and $\{\epsilon_t\}$ were analyzed by a separate subprogram using the method of Section 2.6. Twenty classes were chosen for the analysis and the boundaries of the standardized normal distribution of Equation (2.94) were obtained from tabulated values of Fisher and Yates (1963). The boundaries of the series under examination were calculated by Equation (2.95) using the computed mean and variance of the variable. The program then counted the frequency f_j of the events in each of the twenty classes and computed χ^2 according to Equation (2.91). The computed χ^2 was then tested for significance at 17 degrees of freedom, and the hypothesis that the distribution of the variable was normal was rejected if $\chi^2 > \chi^2_{0.05}$, the value of χ^2 at significance level 0.05 for 17 degrees of freedom.

CHAPTER IV

DRAINAGE BASINS USED IN STUDY

River basins having relatively small areas and long periods of record whose flow had not been significantly affected by regulation, diversion or abstraction were initially selected for the analysis. It was important also that the location of the gauging station had not been changed substantially during the period of record. The stations were selected from records in the U.S.G.S. Water Supply Papers prior to 1960, and subsequently from Surface Water Records for Oklahoma and Kansas. The criteria for selection had to be somewhat flexible as in most of the basins considered some minor interference with the natural flow of the stream, such as construction of farm ponds in the upper reaches, or abstraction for water supply or irrigation, was reported in the records. The record for a basin was rejected if a major reservoir was operated in the basin during the period of record, if the record indicated substantial diversion, or if the gauging station had been moved to include or exclude a substantial drainage area.

Nine stations meeting these criteria were chosen for the analysis. Three were in Oklahoma, one was on the Oklahoma-Arkansas border, and five were in Kansas. It was subsequently decided to analyze also two stations with large contributing area on the Arkansas River, which had been analyzed by Perry (1968). The locations of these eleven stations, and the period of record available for the study are shown in Table I.

TABLE I

GAUGING STATIONS USED IN STUDY

U.S.G.S. Station No.		Period of Record (Water Years)
7-1478	Walnut River, at Winfield, Kansas	1921-1966
7-1645	Arkansas River, at Tulsa, Oklahoma	1925-1964*
7-1705	Verdigris River, at Independence, Kansas	1922-1948*
7-1945	Arkansas River, near Muskogee, Oklahoma	1925-1964*
7-1965	Illinois River, near Tahlequah, Oklahoma	1937-1966
7-3325	Blue River, near Blue, Oklahoma	1937-1966
7-3365	Kiamichi River, near Belzoni, Oklahoma	1926-1966
7-3390	Mountain Fork River, near Eagletown, Oklahoma	1930-1966
6-8680	Saline River, near Wilson, Kansas	1930-1963*
6-8905	Delaware River, at Valley Falls, Kansas	1923-1966
6-8915	Wakarusa River, near Lawrence, Kansas	1930-1966

*Regulation of the river basin commenced.

Table II shows a summary of the major physiographic features of the basins studied. Figures 4, 5, and 6 show the location of the river basins. The gauging stations are hereafter referred to without their prefix numbers.

The mean monthly flows available in the records were compiled from daily flow measurements except in the few years when only estimated mean monthly flows were available. The accuracy of the records in the stations selected is reported as generally "good", indicating an estimated error in the record of ±5%. Occasionally records in some winter months, particularly during periods of ice cover, are reported as "fair" or "poor", indicating a poorer standard of accuracy. However these periods were found to be infrequent and were not considered to have substantially affected the over-all accuracy of the records.

The records also indicated if regulation or other interference with the natural streamflow had occurred, or that the gauging station had been moved, during the period of record. Minor interference or insignificant movement of the gauge were considered to be acceptable in considering the streamflow record to be a continuous and homogeneous sample. Station 1705 on the Verdigris River at Independence, Kansas was reported as having abstraction above the gauge for municipal water supply which is returned to the stream from the sewage treatment plant below the gauge. This was not considered to have a significant effect upon the record. The construction of the Fall Reservoir, where regulation began in 1949, limited the use of the record to 1948 as shown in Table I. The gauge at Station 8905 on the Delaware River, Valley Falls, Kansas was reported to have been moved but the slight change in location was not considered to have affected the record. Records at Station 8680,

TABLE II

PHYSIOGRAPHIC DATA FOR RIVER BASINS

Station No.	Area mi^2	Length mi	Average Slope ft/mi	Average Slope Lower Reaches ft/mi	Average Slope Upper Reaches ft/mi
1478	1840	96.4	3.3	0.4	25.0
1645	74615	-	-	-	-
1705	2892	156.0	5.7	1.8	9.6
1945	96674	-	-	-	-
1965	959	94.2	1.1	3.3	36.1
3325	478	112.4	10.0	2.0	200.0
3365	1423	121.3	8.3	2.5	97.0
3390	787	87.5	9.2	16.9	69.0
8680	1900	234.5	8.9	5.6	13.3
8905	922	67.1	6.9	4.2	9.8
8915	458	80.4	4.0	3.0	5.0

- denotes information not available

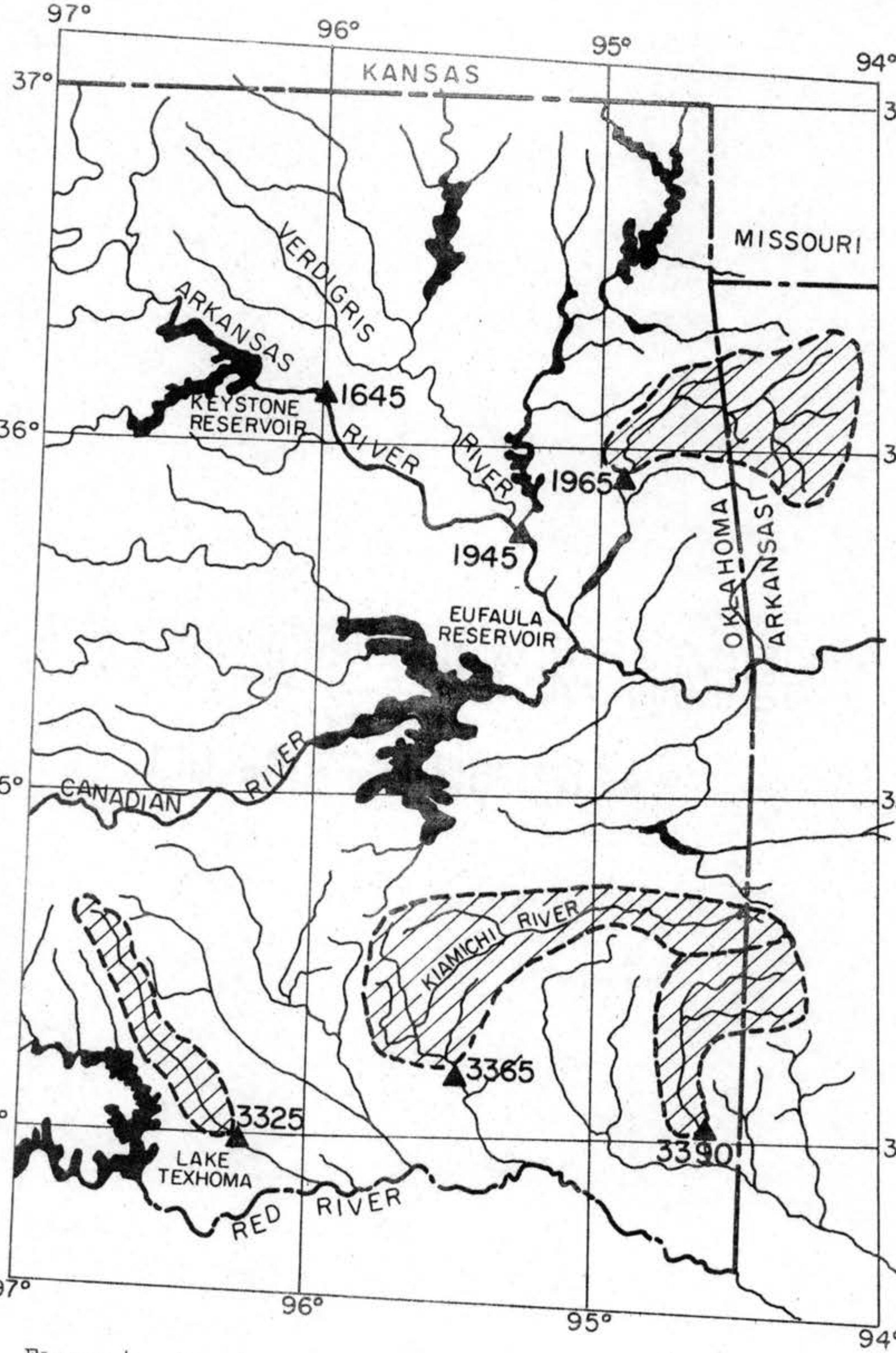


Figure 4. Location of River Basins in Oklahoma and Arkansas

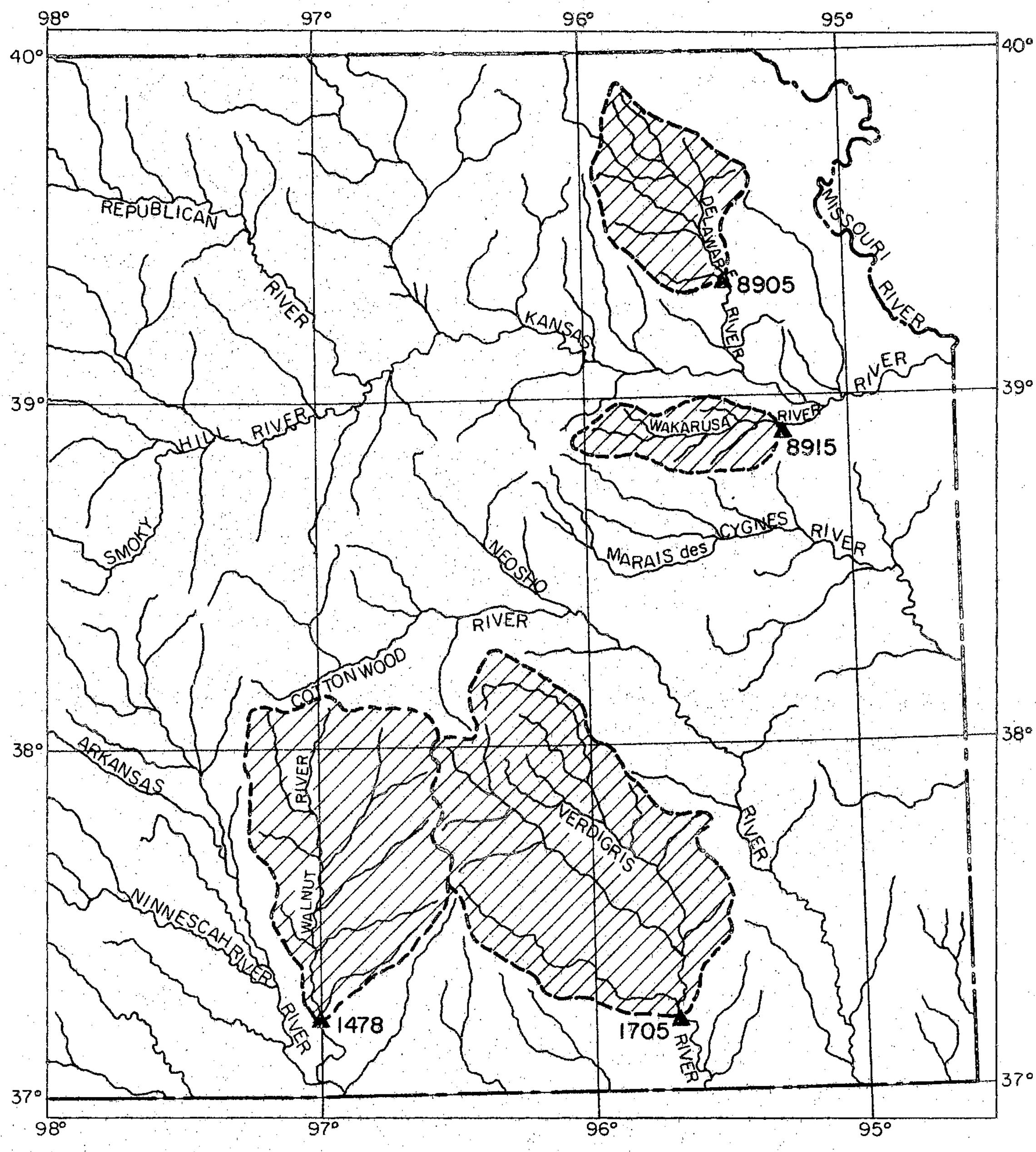


Figure 5. Location of River Basins in Eastern Kansas

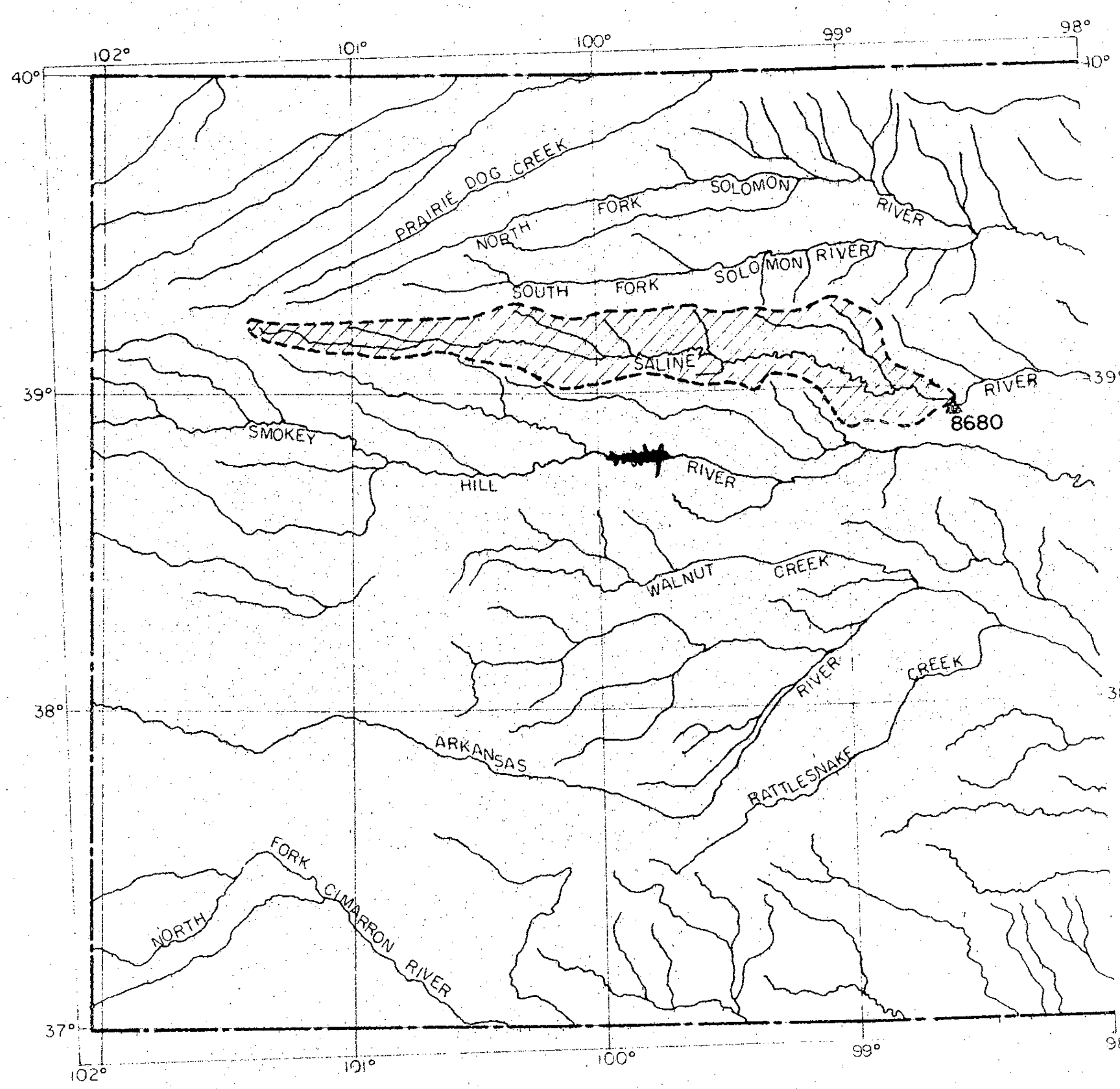


Figure 6. Location of River Basin in Western Kansas

Saline River, near Wilson, Kansas, are usable up to 1963, when regulation at the Wilson Dam commenced. The records at the seven other smaller stations had no reported interference with the natural streamflow.

Records at Station 1645 on the Arkansas River at Tulsa, Oklahoma, were considered to be homogeneous by Perry up to 1964 after which time regulation at Keystone Reservoir commenced. Prior regulation at John Martin Reservoir in Colorado and Great Salt Plains Reservoir in Oklahoma was considered to be insignificant. Station 1945, at Muskogee, Oklahoma, is influenced by an additional 22,000 sqaure miles of drainage area from the rivers Neosho and Verdigris both of which are highly regulated by multiple reservoir development. However, this station was selected to afford a comparison with Perry's results.

CHAPTER V

RESULTS

1. Autoregressive Models

A summary of the results of the analysis of the autoregressive models is shown in Table III. The table shows the value of χ^2 computed by the χ^2-test of the autoregressive scheme (Section 2.5) and the values which are accepted at significance level 0.05 at the respective number of degrees of freedom are underlined. It is seen that Model A was applicable to only a few stations, while Model B satisfied more. Further, the second order scheme was not frequently accepted, and there was little difference between the acceptance of untransformed and logarithmic flows. These results must be viewed in conjunction with the results of the distribution of the variables and residuals as described below.

2. Distribution of Variables and Residuals

Tables IV and V show the distribution of the untransformed and logarithmic flows for the eleven stations examined, which were used in the Model A analysis. The untransformed flows shown in Table IV had distributions which were positively skewed, some highly, and none of the distributions could be accepted as approximately normal. Table V shows the distribution of the logarithmic transformation of the flows. All but one of the distributions which when untransformed were

TABLE III

COMPUTED X^2 FROM TEST OF AUTOREGRESSIVE SCHEME

Station No.	MODEL A				MODEL B			
	Untransformed		Logs of Flows		Untransformed		Logs of Flows	
	First Order	Second Order	First Order	Second Order	First Order	Second Order	First Order	Second Order
1478	26.641	30.740	43.400	119.810	-	-	54.426	146.824
1645	56.582	97.425	65.677	274.110	41.594	63.441	34.425	108.933
1705	39.509	45.616	34.490	142.416	31.433	32.285	25.086	50.910
1945	50.384	65.243	51.758	185.166	32.256	43.899	35.320	89.716
1965	48.439	61.404	91.959	346.441	41.206	59.887	36.124	60.368
3325	33.286	38.844	57.123	151.657	22.012	23.970	34.287	70.243
3365	82.605	81.007	130.739	369.668	23.464	21.953	33.035	41.354
3390	103.877	145.850	117.367	342.608	26.154	31.439	30.295	40.003
8680	45.819	95.474	101.345	293.110	37.739	39.141	49.584	98.889
8905	70.519	66.276	80.945	151.915	36.977	33.701	83.949	104.962
8915	24.934	34.264	79.871	309.095	15.805	17.389	79.761	121.458

- denotes station not examined

TABLE IV

DISTRIBUTION PARAMETERS

(MODEL A - UNTRANSFORMED FLOWS, $\{X_t\}$)

Station No.	Mean Discharge cfs	Standard Deviation cfs	Coefficient of Skewness	Computed χ^2	Acceptance of Normal Distribution
1478	758.2	1388.9	3.278	>999.000	R
1645	6543.1	9090.8	2.955	860.017	R
1705	1640.9	2893.8	3.076	717.728	R
1945	19900.3	27467.6	3.212	708.068	R
1965	864.8	1136.4	3.215	520.333	R
3325	281.0	431.4	3.410	713.111	R
3365	1690.1	2257.0	2.178	721.821	R
3390	1303.3	1548.1	2.062	640.324	R
8680	165.6	352.0	6.048	>999.000	R
8905	386.1	678.3	4.027	>999.000	R
8915	172.4	356.3	4.384	>999.000	R

R denotes rejected

TABLE V

DISTRIBUTION PARAMETERS

(MODEL A - NATURAL LOGS OF FLOWS, $\{\log_e X_t\}$)

Station No.	Mean Discharge cfs	Standard Deviation cfs	Coefficient of Skewness	Computed χ^2	Acceptance of Normal Distribution
1478	5.28732	1.98085	-1.444	28.296	R
1645	8.09282	1.20470	*	19.179	A
1705	5.85637	2.50473	-2.167	54.642	R
1945	9.22163	1.20773	*	11.316	A
1965	6.12847	1.18877	-0.374	23.333	A
3325	4.84248	1.31153	*	18.444	A
3365	6.05085	2.56575	-2.127	157.106	R
3390	5.97518	2.45792	-2.453	179.514	R
8680	4.17482	1.28808	0.433	23.863	A
8905	4.79631	1.70113	-0.302	34.878	R
8915	2.56645	3.71200	-1.429	174.144	R

A denotes accepted

R denotes rejected

* denotes insignificant skewness

positively skewed had negative or insignificant skewness. In three stations, the skewness was accepted as insignificant at the 0.05 significance level using the confidence limits of Equation (2.97). The negative skewness was, in general, less than the positive skewness of the untransformed flows, and five of the stations had distributions which could be accepted as approximately normal.

Tables VI and VII show the distribution of the residual $\{Z_t''\}$ constructed after harmonic removal in Model B. These are the variables used in the construction of the autoregressive scheme and are analogous to the variables $\{X_t\}$ of Model A. Table VI shows that the untransformed residuals are all positively skewed, generally to the same degree that the parent sample $\{X_t\}$ was skewed. None of the untransformed residuals could be accepted as approximately normally distributed. In Table VII it can be seen that the residuals of the logarithmic transformations also had distributions approximately the same as their parent samples. However, in general, the skewness was less, and five of the samples had skewness which was insignificant at the 0.05 significance level. Six of the stations were found to have residuals which were approximately normally distributed.

3. The Residual ϵ_t

The distribution of the residual ϵ_t formed in each model was examined and results are shown in Tables VIII, IX, X, and XI. The residuals from the untransformed variables, both in Model A and Model B, had in no case distributions which could be considered approximately normal. The variances shown in Tables VIII and X are seen to be only approximately equal to one. The residuals from the logarithmic variables shown in

TABLE VI

DISTRIBUTION PARAMETERS

(MODEL B - UNTRANSFORMED FLOWS, RESIDUAL $\{Z_t''\}$)

Station No.	Mean Discharge cfs	Standard Deviation cfs	Coefficient of Skewness	Computed χ^2	Acceptance of Normal Distribution
1478	-	-	-	-	-
1645	0.02650	1.08106	3.53928	457.983	R
1705	0.07730	1.26894	3.56132	583.284	R
1945	0.03049	1.08077	2.66490	433.111	R
1965	0.02309	1.11134	2.94235	339.222	R
3325	-0.00255	1.00070	2.89633	412.333	R
3365	0.00928	1.04573	2.34738	366.293	R
3390	0.00505	1.01849	1.93300	227.712	R
8680	0.13115	1.58483	6.87809	868.961	R
8905	0.00631	1.07037	3.34920	771.470	R
8915	0.01847	1.11879	4.12966	925.459	R

R denotes rejected

- denotes station not examined

TABLE VII

DISTRIBUTION PARAMETERS

(MODEL B - NATURAL LOGS OF FLOWS, RESIDUAL $\{Z_t''\}$)

Station No.	Mean Discharge cfs	Standard Deviation cfs	Coefficient of Skewness	Computed χ^2	Acceptance of Normal Distribution
1478	-0.00493	1.02276	-0.88108	25.926	A
1645	-0.00141	1.01066	*	16.359	A
1705	0.00027	0.99818	-1.34490	31.185	R
1945	0.00062	0.99399	*	18.838	A
1965	0.00007	0.99123	*	19.333	A
3325	-0.00066	1.00865	*	20.556	A
3365	0.00393	1.00562	-0.90517	42.634	R
3390	0.00259	1.00450	-1.16477	45.099	R
8680	0.00059	1.00645	*	17.392	A
8905	0.00021	1.00795	-0.27878	32.000	R
8915	0.00358	0.99401	-1.38399	153.027	R

A denotes accepted

R denotes rejected

* denotes insignificant skewness

TABLE VIII

DISTRIBUTION PARAMETERS

(MODEL A - UNTRANSFORMED FLOWS, RESIDUAL ϵ_t)

Station No.	Mean Discharge cfs	Standard Deviation cfs	Computed χ^2	Acceptance of Normal Distribution
1478	0.00099	1.00066	>999.000	R
1645	0.00114	1.00077	790.088	R
1705	0.00128	0.96726	810.313	R
1945	0.00076	0.91768	833.086	R
1965	0.00014	0.92795	690.972	R
3325	-0.00067	0.91911	640.498	R
3365	-0.00001	0.95160	556.861	R
3390	0.00080	0.92196	327.700	R
8680	0.00070	0.87935	>999.000	R
8905	0.00078	0.94795	>999.000	R
8915	0.00064	0.92714	>999.000	R

R denotes rejected

TABLE IX

DISTRIBUTION PARAMETERS

(MODEL A - NATURAL LOGS OF FLOWS, RESIDUAL ϵ_t)

Station No.	Mean Discharge cfs	Standard Deviation cfs	Computed χ^2	Acceptance of Normal Distribution
1478	0.00242	0.99934	28.978	R
1645	0.00142	0.70996	57.411	R
1705	0.00128	0.70744	20.963	A
1945	0.00084	0.73056	43.707	R
1965	-0.00240	0.74397	39.942	R
3325	-0.00232	0.77592	27.574	A
3365	-0.00158	0.81204	54.866	R
3390	-0.00029	0.81377	92.214	R
8680	0.00361	0.76015	62.828	R
8905	0.00304	0.78742	30.306	R
8915	-0.00058	0.67789	58.806	R

A denotes accepted

R denotes rejected

TABLE X

DISTRIBUTION PARAMETERS

(MODEL B - UNTRANSFORMED FLOWS, RESIDUAL ϵ_t)

Station No.	Mean Discharge cfs	Standard Deviation cfs	Computed χ^2	Acceptance of Normal Distribution
1478	-	-	-	-
1645	0.00065	0.87699	257.325	R
1705	0.00133	0.98358	564.368	R
1945	0.00063	0.90015	321.822	R
1965	-0.00211	0.93481	443.173	R
3325	-0.00152	0.94042	409.858	R
3365	-0.00167	0.98088	381.383	R
3390	0.00037	0.97288	225.487	R
8680	0.00041	0.96618	947.152	R
8905	0.00061	0.94364	688.598	R
8915	0.00040	0.94803	867.293	R

R denotes rejected

- denotes station not examined

TABLE XI

DISTRIBUTION PARAMETERS

(MODEL B - NATURAL LOGS OF FLOWS, RESIDUAL ϵ_t)

Station No.	Mean Discharge cfs	Standard Deviation cfs	Computed χ^2	Acceptance of Normal Distribution
1478	0.00118	0.75812	69.720	R
1645	0.00099	0.70548	41.051	R
1705	0.00105	0.80658	25.668	A
1945	0.00046	0.73974	31.544	R
1965	-0.00256	0.81417	50.192	R
3325	-0.00304	0.82531	44.621	R
3365	-0.00144	0.92394	10.385	A
3390	-0.00049	0.92091	30.724	R
8680	0.00058	0.75889	20.174	A
8905	0.00233	0.82044	27.649	R
8915	-0.00095	0.77374	57.722	R

A denotes accepted

R denotes rejected

Tables IX and XI are seen to have distributions which in a number of cases are accepted as being approximately normal, but the variances of these residuals differ widely from one. As the residual correlation test of the autoregressive model (Equation 2.69) assumes that the distribution of the residual ϵ_t is normal with variance one, and this condition was not satisfied in any case, the test was not used.

4. Harmonic Removal

As described in Chapter III, the method used to remove harmonics and test the residual correlation was not completely satisfactory. It was found that the test in the harmonic removal subprogram which rejected a residual and proceeded to the next harmonic if one correlation fell outside the confidence limits was too severe. This test also suffered from the limitation described above for the residual ϵ_t, in that the residual should have had a normal distribution with variance of one. No check was made for this condition; no other test was available if the condition was not satisfied.

In no case, even after the removal of six harmonics was the test satisfied. In all cases the first serial correlation coefficient r_1 lay outside the confidence limits, and in many cases succeeding values of r_k. However, it was found in some cases that from the point where one r_k fell inside the confidence limits all subsequent r_k also fell inside. This condition was then accepted as satisfying the test and residual correlation was considered to be insignificant. However, this required visual inspection of the correlogram, which was plotted subsequently, and defeated the purpose of the test in the program which was designed to eliminate this step. After a residual was found to be independent in this way, the program was rerun with the chosen number of harmonics for the results of the station to be obtained.

Only in seven of the twenty-two cases examined was this modified condition, which will be referred to as Condition 1, satisfied. A second type of result was also observed. In this case, the residual almost satisfied Condition 1, but one or two of the twenty-five correlations calculated fell outside the confidence limits. It was frequently found that their departure from the confidence limits increased rather than decreased as more harmonics were removed. In such cases, harmonic removal was terminated where the best condition was obtained. This "best" condition was judged somewhat arbitrarily, but it was usually the point where the departure from the confidence limits was a minimum. This condition will be referred to as Condition 2.

A third condition was also observed. Here harmonic removal was found to remove perhaps one or two distinct cycles from the correlogram, but thereafter, although there was highly significant correlation in all twenty-five calculated r_k, further harmonic removal had no effect upon the correlogram. This condition will be referred to as Condition 3. Table 12 shows the results of harmonic removal: the number of harmonics removed for each station using the above criteria, and the condition satisfied. Station 1478 was unique. For the untransformed variables, harmonic removal was not found to produce significant change to the correlogram to satisfy any of the above conditions. Therefore, no harmonics were removed: in effect, only Model A was examined.

TABLE XII

HARMONICS REMOVED

Station No.	Unt.	Logs
1478	0*	1**
1645	2**	1**
1705	2*	2+
1945	2*	2**
1965	1*	2**
3325	1+	1+
3365	1+	2+
3390	1*	2+
8680	2*	1**
8905	1+	1**
8915	1*	1**

+Condition 1

*Condition 2

**Condition 3

Correlograms typical of the three conditions described above are shown in Appendix III. These plots were produced by computer print-out, and the accuracy of the position of the points is limited by the spacing of the lines of print on the printer. However, they are sufficiently accurate for illustration.

The correlograms shown in Appendix III for Station 3365 (Logs) illustrate Condition 1. The correlogram without harmonic removal shows a distinct cycle of twelve months, and this is removed by the first harmonic. The resulting correlogram shows fluctuations with a period of approximately six months which are significant at the 0.05 level when tested by the confidence limits. Removal of the period produces a correlogram in which the first six correlations are significant, but thereafter, correlation is insignificant and the independence of the residual is accepted. Harmonic removal was then terminated.

Condition 2 is illustrated by correlograms for Station 1705 (Untrans). Before harmonic removal, the correlogram shows both twelve and six month cycles. After removal of the twelve month cycle, the six month period becomes more distinct. This is then removed and a weak three month cycle appears. However, removal of this cycle does not reduce the correlation at significant points but increases it, and it increases further with further harmonic removal. Harmonic removal was, therefore, terminated after three harmonics.

Correlograms for Station 8915 (Logs) illustrate Condition 3. Here, after removal of a distinct twelve month cycle, the correlogram does not change with removal of subsequent harmonics; nowhere does the correlation fall within the confidence limits. Harmonic removal was therefore terminated after one harmonic.

CHAPTER VI

DISCUSSION

1. Autoregressive Models

For the eleven stations examined, an autoregressive model was found to describe the hydrological sequence at an acceptable level of significance. In one case, only Model A was applicable; in two other cases, both Models A and B were applicable. In five cases, one for Model A and four for Model B, both the untransformed and logarithmic flows gave an acceptable model. It was found that the first order autoregressive scheme was more widely accepted than the second and in only one case, Station 8905, Model B (Untrans) was only the second order scheme accepted at the 0.05 significance level, where the first order scheme was accepted at the 0.01 significance level. In one case, Station 8680, no model was accepted at the 0.05 significance level; Model B (Untrans) was acceptable at the 0.01 significance level. Table III shows the values of X^2 computed in the significance test; those accepted at the 0.05 significance level are underlined in full, whereas those accepted at the 0.01 significance level are underlined with a broken line.

The choice of an autoregressive model when more than one gives an acceptable result depends upon the distribution of the variables upon which the model is operating. As described in Section (2.1.b), the autoregressive model is applicable only to stationary time series, and the transformations used in the analysis make the series stationary only to the second order, unless the series is normally distributed. Thus, the record generated by the autoregressive model corresponds with the original record only in the first and second moments, the mean and standard deviation. The introduction of the skewness component g_ξ in the random ξ_t of Model IIA or Model IIB (Equations 2.66 and 2.67) extends this correspondence to the third moment about the mean, the skewness. However, there is no correspondence at higher moments, and the distribution of the synthesized record will only agree with the original record in the mean, standard deviation and coefficient of skewness. If, however, the series is normally distributed and is stationary to the second order, it is stationary to all higher orders (Matalas, 1967a). This may be shown by the definition of stationarity (Equation 2.7). In that case, the synthetic record generated by the model corresponds in all its moments with the original record and the distribution of the synthesized record is the same as that of the original record.

For Model A the variable used in the autoregressive scheme is the original series $\{X_t\}$, or $\{\log_e X_t\}$ if the logarithmic transformation is used. The distribution of the untransformed flows is summarized in Table IV where it is seen that none of the distributions could be accepted as normal. Table V shows the distribution of the logarithmic transformations, and it is seen that five of the stations had distributions which were accepted as normal at the 0.05 significance level. The distribution of the residuals $\{Z_t''\}$ of Model B are shown in Tables VI and VII, which show that none of the untransformed residuals could be accepted as normally distributed, while six of the logarithmic transformations had distributions which were accepted.

The criteria, therefore, which were used for selecting a model are as follows. If any or all of the variables (or residuals) were normally distributed, only these models were retained. From these Model A was selected for its simplicity in preference to Model B; the first order model was selected for its simplicity in preference to the second order model. If the choice lay between normally distributed untransformed and logarithmic flows, the distribution with the lowest computed X^2 was accepted. If none of the residuals or variables were normally distributed, Model A was chosen for its simplicity in preference to Model B. If the skewness as defined in Equation (2.96) was significant, the model (II) was selected in preference to the model (I) (Equations 2.66 and 2.67).

Models selected for the eleven stations examined using these criteria are summarized in Table XIII, which shows that the most desirable combination was not always obtained. For example, although the residual of the logarithmic transformation for Station 1478 was normally distributed the autoregressive model using this residual was rejected. In this case Model A was then considered, and Model IIA (Untrans) selected. Similarly, for Station 8680, although the logarithmic residuals were normally distributed, none of the autoregressive models were accepted at the 0.05 significance level. However, Model IIB (Untrans) was accepted at the 0.01 significance level. Appendix II summarizes the parameters required to describe the models listed in Table XIII.

TABLE XIII

MODELS ACCEPTED

Station No.	Model	
1478	Model IIA	Untransformed
1645	Model IB	Logs
1705	Model IIB	Logs
1945	Model IB	Logs
1965	Model IB	Logs
3325	Model IB	Logs
3365	Model IIB	Untransformed
3390	Model IIB	Untransformed
8680	Model IIB	Untransformed
8905	Model IIB	Untransformed
8915	Model IIA	Untransformed

2. Tests of Fit of Autoregressive Scheme

It was intended that both the tests described in Section (2.5) should be used to test the adequacy of the proposed autoregressive schemes. It was hoped that a comparison could be made between the effectiveness of the tests. Such a comparison had not been found reported in the literature. The residual correlation test was used by Roesner and Yevdjevich (1966) in the analysis of some 140 run-off stations in the western United States. No reported use of Quenouille's χ^2-test for hydrological sequences was found.

The residual correlation test as described by Anderson (1942) assumes that the variable ϵ_t is normally distributed with variance of one. However, as reported above in Section (5.2), none of the forty-four sets of residuals obtained in this study (from Models A and B, untransformed and logarithmic variables) satisfied this condition. The test, therefore, is not strictly applicable to the stations examined in this study. Roesner and Yevdjevich did not report whether this condition had been investigated. As five of the stations which they examined were also examined in this study, where it was found that in none of them was the test strictly applicable, the validity of results obtained in other stations examined by them may also be questioned.

Although the test is not strictly valid, an attempt was made to apply it. The same problems encountered in determining the significance of the residual correlation after harmonic removal were found in the use of the test. The same three conditions (q.v. Section 5.4) were observed, and in the few stations examined the model was rejected by the test because of the significance of perhaps one correlation, even under Condition 1 which was the usual condition found, while the χ^2-test accepted the model. The test was found to be cumbersome and tedious, because the correlograms had to be examined, and after initial failures was discarded in favour of the χ^2-test which was performed by the program system during execution of the model analysis on the r_k obtained from the residual $\{Z_t\}$. No results from the residual correlation test have, therefore, been reported, and all testing of the adequacy of the models was made with the χ^2-test.

No criteria have been found to judge the adequacy of the χ^2-test. However, its author, Quenouille (1947) used it to test the adequacy of published autoregressive models. He found that not all reported models which had hitherto been accepted satisfied the test.

3. Harmonic Removal

Although the method of harmonic removal produced adequate solutions for the stations examined, it was cumbersome to use in the form presented here. It was unnecessary to assume that harmonics be removed until the residual was independent. This would be true for a stochastic model which consisted only of a harmonic component and a random element. However it was found that after removal of one or two harmonics the series, as judged by the shape of the correlogram, was transformed from a harmonic series to one which could be described by linear autoregression. Kendall's (1951) description of the shape of the correlogram for various stochastic processes was discussed in Chapter II. In practice the program removed harmonics up to the limit which had been selected, six, and the resulting residual was tested for the application of an autoregressive scheme. This would be acceptable, but it was felt that too many harmonics were being indiscriminately removed, leading to an unnecessary number of constants, and the the method of examining correlograms described in Chapter V was adopted.

This study has thus far omitted reference to spectral analysis, a method of removal of significant cycles from a time series recently used in hydrology by Roesner and Yevdjevich (1966) and Quimpo (1968). The method gives a spectrum of the frequency of cycles in the series from which significant cycles and their periods may be detected. The method is complex, but yields accurate results and is more sensitive than the methods used in this study. However, Roesner and Yevdjevich reported that 12, 6, 4, ... month cycles were removed, and nowhere did they report cycles which did not have periods of 12/n, where n was an integer and did not exceed six. These same six harmonics were removed in this study without the use of the spectral analysis technique. The only advantage which the method would bring is in the detection of significant cycles, which the method of this study failed to do adequately. However, as mentioned above, the removal of cycles until the residual is independent is not necessary. Quimpo (1968) found that the results obtained from spectral analysis did not justify the effort involved, and questioned the applicability of this sophisticated technique to series as imprecise and short as hydrological sequences.

An alternative method of performing the analysis which was not appreciated until all the results were available could be as follows. Instead of testing the autoregressive scheme at the end of harmonic removal, the scheme could be tested after each harmonic removal, and analysis stopped as soon as an acceptable model was produced. This method could also combine the analysis of Model A and Model B, as Model A is essentially Model B without harmonic removal. One harmonic would be removed from the series, the residual correlated and the correlations used in the χ^2-test to test the adequacy of the model. If the model was not accepted at this stage a further harmonic would be removed and the process repeated. Only when sufficient harmonics were removed for the model to be accepted would the mean and variance of the residual be calculated and its distribution analyzed. If this method had been used, it is possible that some of the models which were judged to require removal of two harmonics would have been accepted with only one, and that other models which failed with the selected number of harmonics would have been accepted with more. The analysis of Station 1645, one of the last stations examined, led to this method. Using Condition 3 (described in Section 5.4) to judge harmonic removal, the model failed with one harmonic. It also failed with two and three, and not until four harmonics were removed was the model accepted.

4. Comparison of Models

As discussed in Chapter I, this investigation sought to describe a model or models which could be used where the model of Thomas and Fiering (1962) was found to be unsatisfactory. Perry (1968) and Dunaway (1968) reported that the Thomas and Fiering equation (Equation 1.8) was not applicable to stream basins with small drainage areas, although it was applicable to large areas. The model requires the calculation of the correlation r_τ between the months τ and $\tau-1$, and requires that this correlation is not zero. The correlation must be tested for significance, using small sampling theory, by use of Student's-t (Fisher, 1958). Perry found that for Station 1965 five of these correlations were not significant at the 0.10 significance level, which meant that the hypothesis that the correlation was not zero could not be accepted. He concluded that the method could not be used for Station 1965, although it was successfully used for Stations 1645 and 1945. This problem had also been alluded to by Thomas and Fiering who found in their original investigation that there was a tendency for correlations in some months to be insignificant. They reported that in April and May, the months of the spring thaw in the station they examined, correlation was not significant.

The two models presented in this study did not have this limitation. As reported above, an autoregressive model, Model IB (Logs), has been shown to describe the hydrological sequence of Station 1965, which Perry had concluded could not be described by the Thomas and Fiering model. The significance of the serial correlation coefficients calculated from the series $\{Z_t'\}$ cannot be tested by the t-test described by Fisher. This test assumes that the two variables being correlated, denoted for this description by X and Y, are each stochastically independent: all values of X are independent of all other values of X; similarly for Y. However, by the assumption of an autoregressive model, X is dependent upon preceding values of X (v. Equation 2.5). The mean monthly flows of the hydrological sequence are not therefore stochastically independent, and the t-test is not applicable. However, no such conclusion need be made about the sequences of flows for given months used in the Thomas and Fiering model. An analysis of the monthly sets of Station 1965 for serial correlation, using the test described by Anderson (1942) showed that eleven of the twelve sets had insignificant serial correlation and could, therefore, be considered to be independent. The correlation between the sets is thus valid and the t-test may be used to test its significance.

A further disadvantage of the Thomas and Fiering model is that Quenouille's χ^2-test cannot be used to test the adequacy of the model. The model is not based upon the assumption of an autoregressive scheme describing a continuous series $\{X_t\}$, but upon twelve combined sub-series consisting of sets of months. The χ^2-test could be conceivably applied to these sub-series, but as mentioned above, examination of Station 1965 showed that the sets were independent and could not be described by linear autoregression. Furthermore, the application of the test to the sub-series does not determine the adequacy of the model as a whole.

The residual correlation test could be applied to the residual produced by the Thomas and Fiering model, but the test was found to be impracticable, even if applicable, when used in this study. It would presumably be no more successful when used with Thomas and Fiering's model. They reported only the results of comparison of the synthesized record with the actual record, in order to demonstrate the applicability of the model. Only the mean and variance could be compared by this method as the series is stationary only to the second order. Harms and Campbell (1967) reported fairly good agreement when using this comparison. However, higher moments will not necessarily agree and Dunaway (1968) found large discrepancies in the frequency distribution, presumably because the model assumed that the untransformed flows were normally distributed, which in this study was invariably found not to be the case.

Of the two models investigated in this study, Model A was more simple to apply. It required only four parameters: the mean and variance of the variable, the skewness parameter and the first order serial correlation coefficient. Model B required at least eight parameters: the four required in Model A, and four for each successive harmonic removed. However, the Thomas and Fiering model requires forty-eight parameters: twelve monthly means, twelve monthly standard deviations, twelve correlation coefficients from month-to-month, and twelve regression coefficients from month-to-month.

Model A and Model B referred to above are first order autoregressive schemes. Although the second order model was investigated in this study, it was felt that it would be hard to justify the use of a second order model in preference to a first order model if both were acceptable. Although the sample being investigated is comparatively small, it is assumed to be fully representative of the population from which it was drawn. However, this may not be true, and a model more complicated than the most simple acceptable could not be justified. Therefore, when only a second order model was acceptable at the 0.05 significance level, as in Station 8905, a first order model acceptable at the 0.01 significance level was preferred.

CHAPTER VII

CONCLUSIONS

1.) A simple first order linear autoregressive model of the form

$$\omega_t = r_1 \omega_{t-1} + \epsilon_t \tag{7.1}$$

was found to describe weakly stationary transformations of hydrological sequences in river basins with small contributing areas. Two versions of this model, one without and one with removal of a harmonic component from the sequence, were examined. The version without harmonic removal was found to be less widely accepted than the version with harmonic removal, which required more parameters for its description. The linear first order model was compared with the model of Thomas and Fiering (1962)

$$\left(\frac{X_t - m_\tau}{s_\tau}\right) = r_\tau\left(\frac{X_{t-1} - m_{\tau-1}}{s_{\tau-1}}\right) + \eta_t (1 - r_\tau^2)^{1/2} \tag{7.2}$$

which had been found not to be applicable to drainage basins with small contributing areas. The linear autoregressive model was found to be more simple to apply, requiring fewer parameters and simpler computation for its analysis.

2.) Two tests of the adequacy of the autoregressive model were compared. One tested the assumption that the residual ϵ_t of Equation (7.1) was independent, which it would be if the series described a process of linear autoregression. The independence of the residual was tested by examining the significance of its serial correlation coefficients. This test is only strictly applicable to a stationary residual. The other test was a large sample χ^2-test proposed by Quenouille (1947) using the serial correlation coefficients from the stationary time series. The former test, the residual correlation test, was found to be cumbersome to apply. Moreover, none of the residuals examined in the study were stationary and the test was not therefore strictly valid. The χ^2-test was found to be simple to apply and was adopted as the criterion for selecting a model.

3.) Removal of significant cycles from the time series was attempted with harmonics with a fundamental period of twelve months. Correlograms were used in an attempt to define significant cycles, and their significance was tested by examining the serial correlation coefficients using a method proposed by Anderson (1942). The method of harmonic removal was not very successful. However, it was concluded that it was not necessary to determine the significance of cycles, but simply to investigate whether the residual after their removal could be described by the linear autoregressive model of Equation (7.1). Quenouille's χ^2-test was used to test the adequacy of the model in this way.

4.) The distribution of the variables and residuals was also examined. The distribution of an observed series was compared with the theoretical normal probability distribution using the χ^2-test proposed by Pearson (1900). It was found that in no case could the sequence of mean monthly flows be considered to be approximately normally distributed, although for some sequences examined the distribution of the natural logarithms of the mean monthly flows was found to be approximately normal. The distribution of the residuals formed after removal of the harmonic component was found in all cases to be similar to that of the original sequence.

The distribution of the residual ϵ_t (Equation 7.1) was also examined. Few sequences were found to have a residual which could be considered to be approximately normally distributed. None were normally distributed with variance one, the assumption upon which the test of the significance of the residual was based.

BIBLIOGRAPHY

Anderson, R. L. (1942) Distribution of the Serial Correlation Coefficient, Ann. Math. Stat., 13, 1-13.

Blackman, R. B., and Tukey, J. W. (1958) The Measurement of Power Spectra, Dover, New York.

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

Brooks, C. E. P., and Carruthers, N. (1953) Handbook of Statistical Methods in Meteorology, Meteorological Office Publication No. 538, H. M. S. O., London.

Dunaway, L. E. (1968) Analysis of Low Flows by Statistical Methods, unpublished M. S. Thesis, Oklahoma State University.

Fisher, R. A. (1924) The Conditions Under Which X^2 Measures the Discrepancy Between Observation and Hypothesis, Journ. Roy. Stat. Soc, 87, 442-49.

Fisher, R. A. (1958) Statistical Methods for Research Workers, 13th Edition, Hafner, New York.

Fisher, R. A., and Yates, F. (1963) Statistical Tables for Biological Agricultural and Medical Research, 6th Edition, Hafner, New York.

Hald, A. (1952) Statistical Theory With Engineering Applications, John Wiley, New York.

Harms, A. A., and Campbell, T. H. (1967) An Extension to the Thomas and Fiering Model for the Sequential Generation of Streamflow, Water Resources Research, 3, No. 3, 653-661.

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

Kendall, M. G. (1951) The Advanced Theory of Statistics II, Hafner, New York.

Mann, H. B., and Wald, A. (1942) On the Choice of the Number of Class Intervals in the Application of the X^2-test, Ann. Math. Stat., 13, 306-317.

Markovic, R. D. (1965) Probability Functions of Best Fit to Distributions of Annual Precipitation and Runoff, Hydrological Paper No. 8, Colorado State University, Fort Collins, Colorado.

Matalas, N. C. (1963) Probability Distribution of Low Flows, U. S. G. S. Prof. Paper No. 434-A.

Matalas, N. C. (1967a) Time Series Analysis, Water Resources Research, 3, No. 3, 817-829.

Matalas, N. C. (1967b) Mathematical Assessment of Synthetic Hydrology, Water Resources Research, 3, No. 4, 937-945.

Pearson, K. (1900) On a Criterion That a Given System of Deviations From the Probable in the Case of a Correlated System of Variables is Such That it can be Reasonably Supposed to Have Arisen From Random Sampling, Phil. Mag., Series V, 50, 157-75.

Perry, K. L. (1968) The Effects of Historical Record Lengths on Generating Synthetic Data Using a Stochastic Model of the Markov Chain, unpublished M. S. Thesis, Oklahoma State University.

Quenouille, M. H. (1947) A Large Sample Test for the Goodness of Fit of Autoregressive Schemes, Journ. Roy. Stat. Soc., 110, 123-129.

Quimpo, R. G. (1968) Autocorrelation and Spectral Analysis in Hydrology, J. Hyd. Div. A. S. C. E., 94, HY2, 363-373.

Roesner, L. A., and Yevdjevich, V. M. (1966) Mathematical Models for Time Series of Monthly Precipitation and Monthly Runoff, Hydrological Paper No. 15, Colorado State University, Fort Collins, Colorado.

Thomas, H. A., and Fiering, M. B. (1962) Mathematical Synthesis of Streamflow Sequences for the Analysis of River Basins in Simulation, Chapter 12 in Maass, A. et al., Design of Water Resources Systems, Harvard University Press, Cambridge, Mass.

Thomas, H. A., and Fiering, M. B. (1963) Statistical Analysis of Reservoir Storage Yield Relations, Chapter 1: Operations Research in Water Quality Management in Final Report to Bureau of State Services, Div. of Water Supply and Pollution Control, Public Health Service, U. S. Dept. of Health, Education and Welfare.

Snedecor, G. W., and Cochran, W. G. (1967) Statistical Methods, 6th Edition, Iowa State University Press, Ames, Iowa.

Walker, G. (1931) On Periodicity in Series of Related Terms, Proc. Roy. Soc., A131, 518.

Wold, H. (1954) A Study in the Analysis of Stationary Time Series, Almqvist and Wiksell, Stockholm.

Yule, G. U. (1926) Why Do We Sometimes Get Nonsense-Correlations Between Time Series? - A Study in Sampling and the Nature of Time Series, Journ. Roy. Stat. Soc., 89, 1-64.

Yule, G. U. (1927) On a Method of Examining Periodicities in Disturbed Samples With Special Reference to Wolfer's Sunspot Numbers, Phil. Trans., A226, 267.

APPENDIX I

LIST OF SYMBOLS

Symbol	Definition
a, b	constants
A_p, B_p	constants to describe m_t
C_p, D_p	constants to describe s_t
b_τ	regression coefficient (Thomas and Fiering model)
b_j	boundary of class j in standardized series
B_j	boundary of class j in observed series
d_p	phase of harmonic
e	exponential
E	Mathematical Expectation
f_j	frequency of occurrence in class j
F(t)	Probability Distribution Function
g_x	estimated skewness of $\{X_t\}$
g_ξ	estimated skewness of ξ_t
h	period of harmonic
k	lag
k	number of classes in distribution analysis
K_α	standard normal deviate at significance level α
K_p	constant
ℓ	number of lags used in χ^2-test of autoregressive scheme
m	mean of $\{X_t\}$
m_1	mean of $\{\log_e X_t\}$
m_3	third moment about mean
m_t	continuous function of monthly mean
m_z	mean of $\{Z_t''\}$

Symbol	Definition
m_τ	mean of months τ
n	number of harmonics
n	number of observations $\{X_t\}$
N	number of years of record
p	order of harmonic
p	order of autoregressive scheme
p_1	estimate of π_1
r_k	serial correlation coefficient for lag k
R	statistic for χ^2-test of autoregressive scheme
s^2	variance of $\{X_t\}$
s_1^2	variance of $\{\log_e X_t\}$
s_t^2	continuous function of monthly variance
s_Z^2	variance of $\{Z_t''\}$
s_τ^2	variance of months τ
t	time
u	constant
X_t	mean monthly discharge in month t
$\{X_t\}$	set of observations X_t
$\{X\}$	set of observations of which $\{X_t\}$ is a sample
$\{Z_t\}$	set of standardized $\{X_t\}$
$\{Z_t''\}$	set of $\{X_t\}$ after harmonic removal
$\{Z_t'\}$	set of standardized $\{Z_t''\}$
α	significance level
β_1, β_2	constants
γ_X	skewness of $\{X\}$
γ_ξ	skewness of $\{\xi\}$

Symbol	Definition
δ_t	deterministic component of element
ϵ_t	random component of element
η_t	standardized normal random variable at time t
λ^2	variance of ϵ_t
μ	mean of population $\{X\}$
μ_1	mean of population $\{\log_e X\}$
μ_2	second moment about mean
μ_3	third moment about mean
ξ_t	skewed standardized random variable at time t
$\{\xi\}$	set of ξ_t
π_i	probability of event i
ρ_k	autocorrelation coefficient for lag k
σ^2	variance of $\{X\}$
σ_1^2	variance of $\{\log_e X\}$
τ	index of month τ ($\tau = 1, 2, \ldots, 12$)
ω_t	variable at time t

APPENDIX II

PARAMETERS OF ACCEPTED MODELS

XIV

PARAMETERS OF ACCEPTED MODELS

Station	Model	Transformation	Mean of Variable	Standard Deviation of Variable	Skewness Parameter g_{ξ}	r_1	No. Harmonics Removed	Constants			
								A_p	B_p	C_p	D_p
1478	IIA	U	758.214	1388.915	3.62402	0.28464	-	-	-	-	-
1645	IB	L	-0.00141	1.01066	*	0.42082	4	-0.0389	-0.6575	0.1503	-0.0676
								-0.1663	0.2158	0.0875	0.0764
								0.0578	0.0970	0.0095	0.0063
								0.0458	0.0062	-0.0566	0.0713
1705	IIB	L	0.00027	0.99818	-2.03646	0.59213	2	-0.7663	-1.0077	0.3225	0.9254
								-0.2128	0.5809	-0.0746	-0.0602
1945	IB	L	0.00062	0.99399	*	0.66544	2	-0.2240	-0.5747	0.1185	0.0171
								-0.1561	0.2360	0.0788	0.0715
1965	IB	L	0.00007	0.99123	*	0.58085	2	-0.8057	-0.3344	0.1513	0.0649
								-0.1103	0.2238	0.0973	-0.0119
3325	IB	L	-0.00066	1.00865	*	0.56227	1	-0.7508	-0.2818	0.1696	-0.0019
3365	IIB	U	0.00928	1.04573	2.47133	0.19651	1	-1355.112	-191.826	-900.670	-245.993
3390	IIB	U	0.00505	1.01849	2.07846	0.23570	1	-1088.298	139.352	-711.115	45.655
8680	IIB	U	0.13115	1.58483	7.51543	0.26215	2	32.996	-166.852	64.666	-241.677
								-77.793	-6.535	-100.592	-25.686
8905	IIB	U	0.00631	1.07037	3.84682	0.33281	1	-35.229	-259.853	41.506	-309.559
8915	IIA	U	172.413	356.318	5.20169	0.37704	-	-	-	-	-

U Untransformed

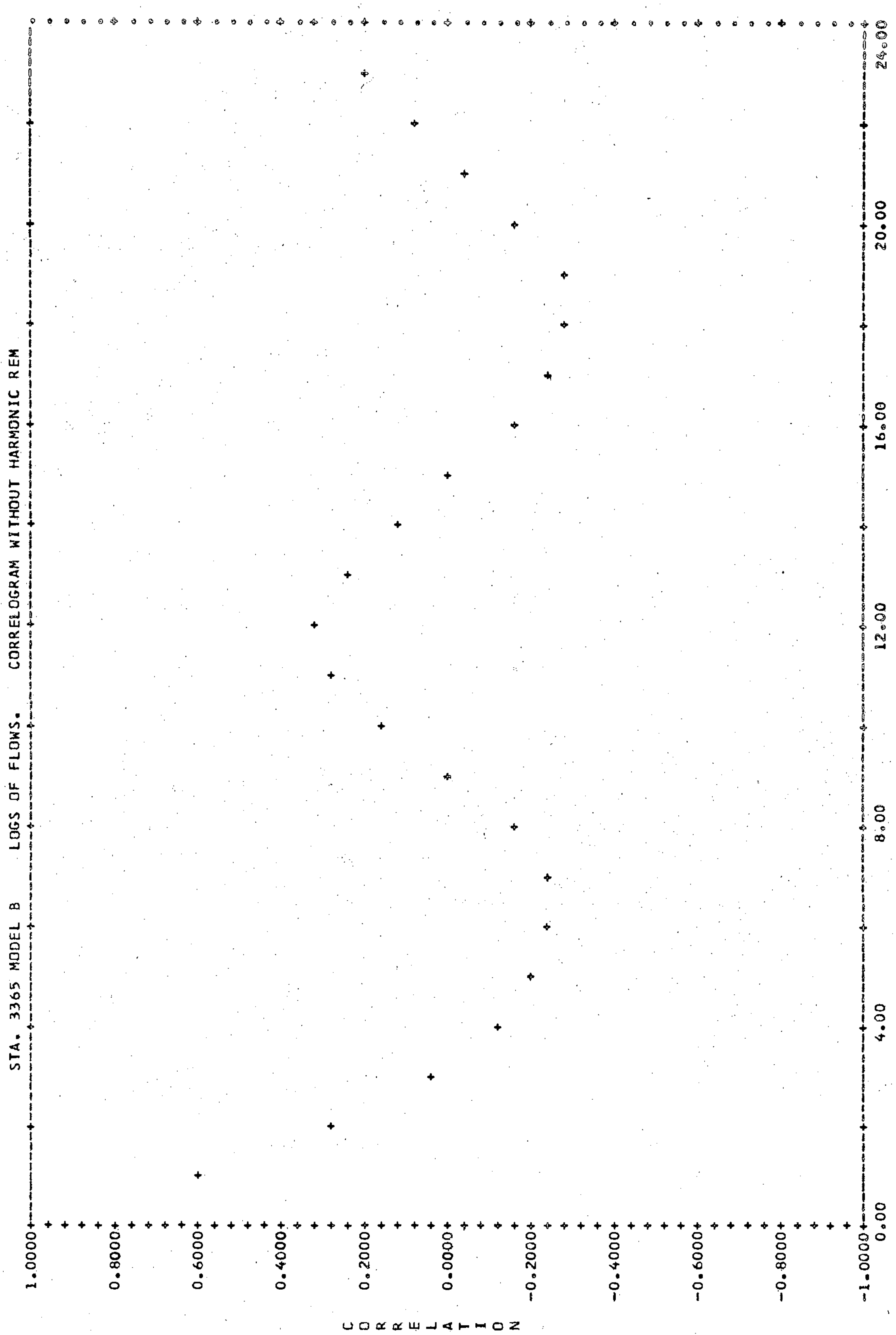
L Logs

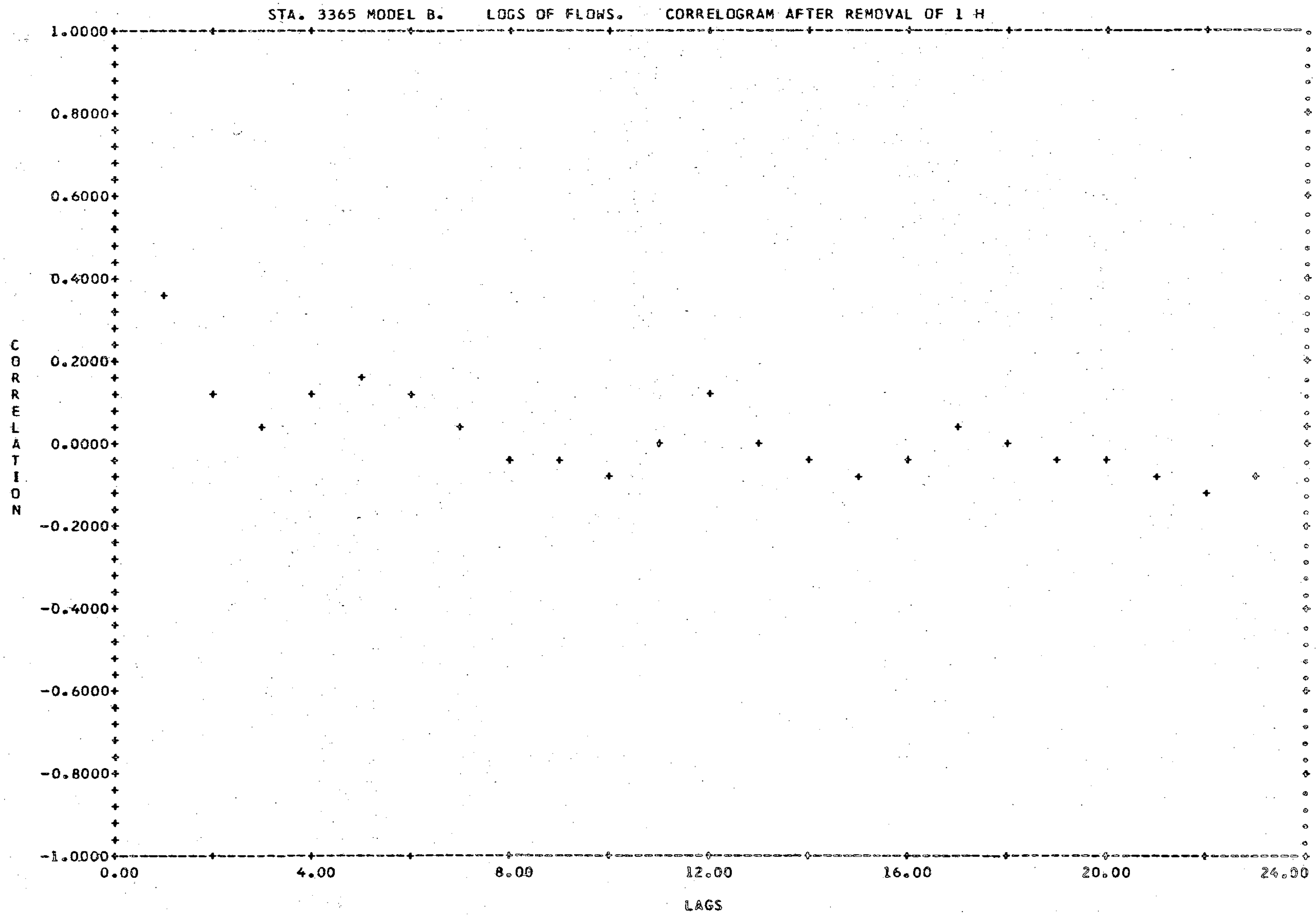
* Skewness insignificant

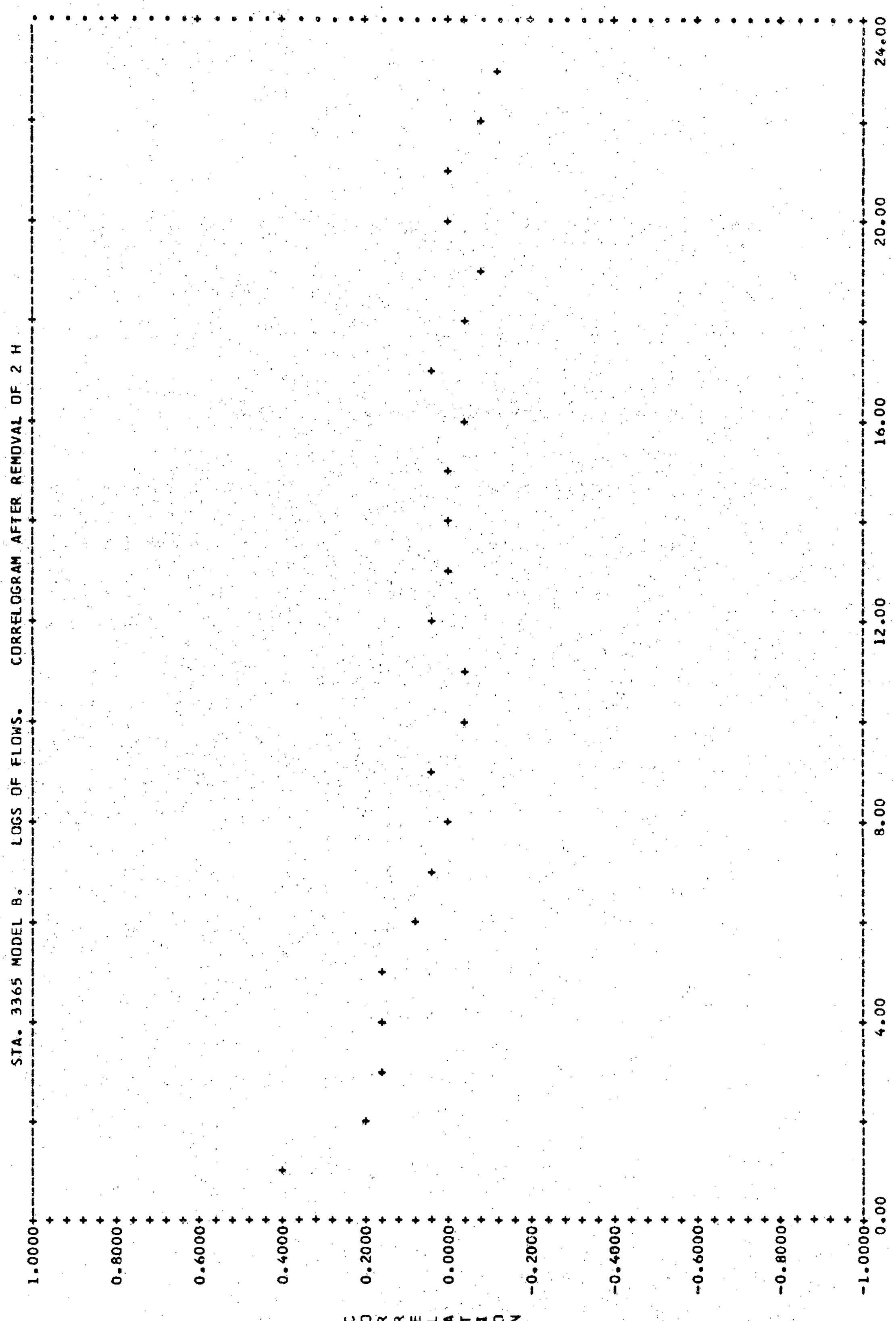
- denotes parameter not required

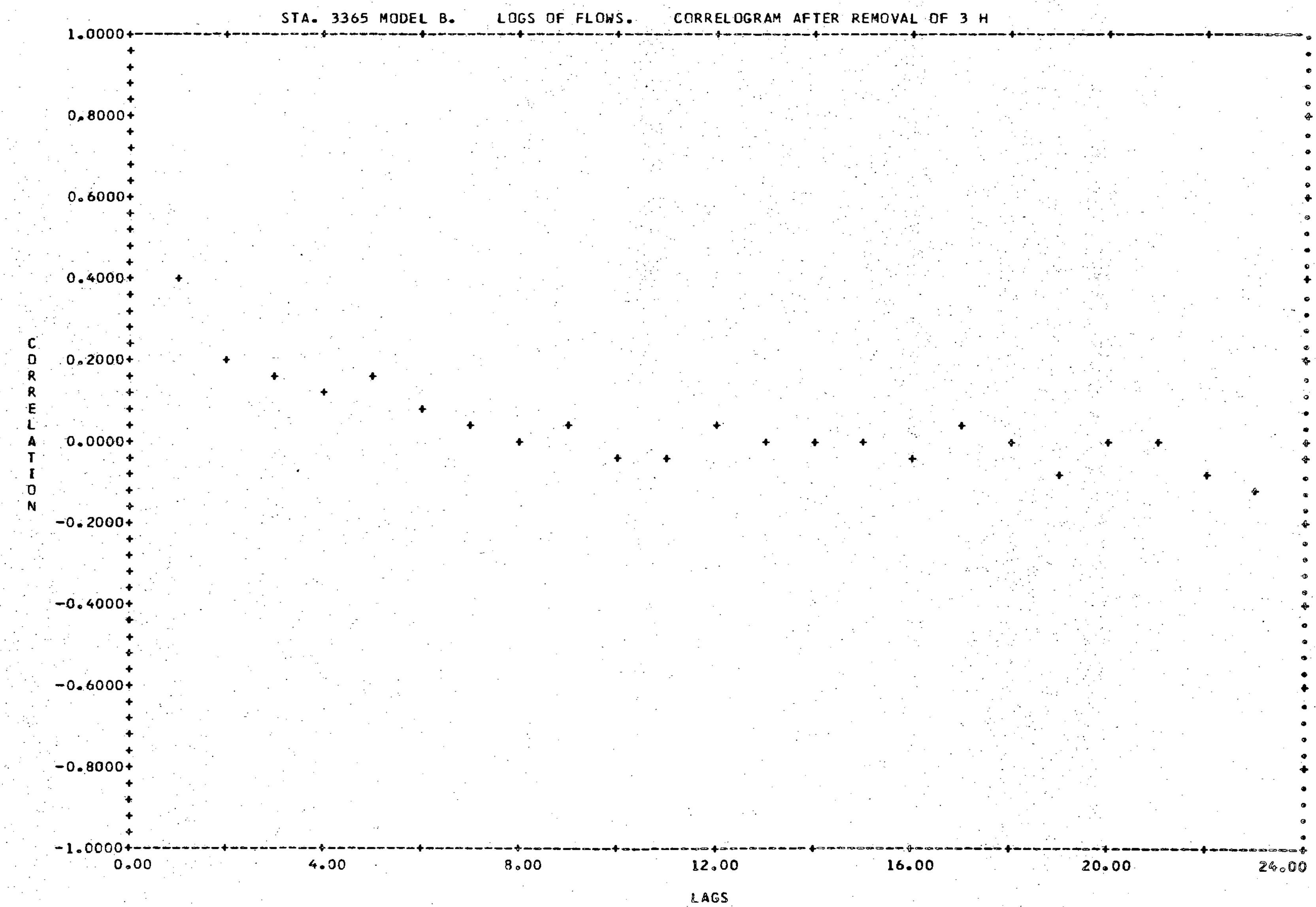
APPENDIX III

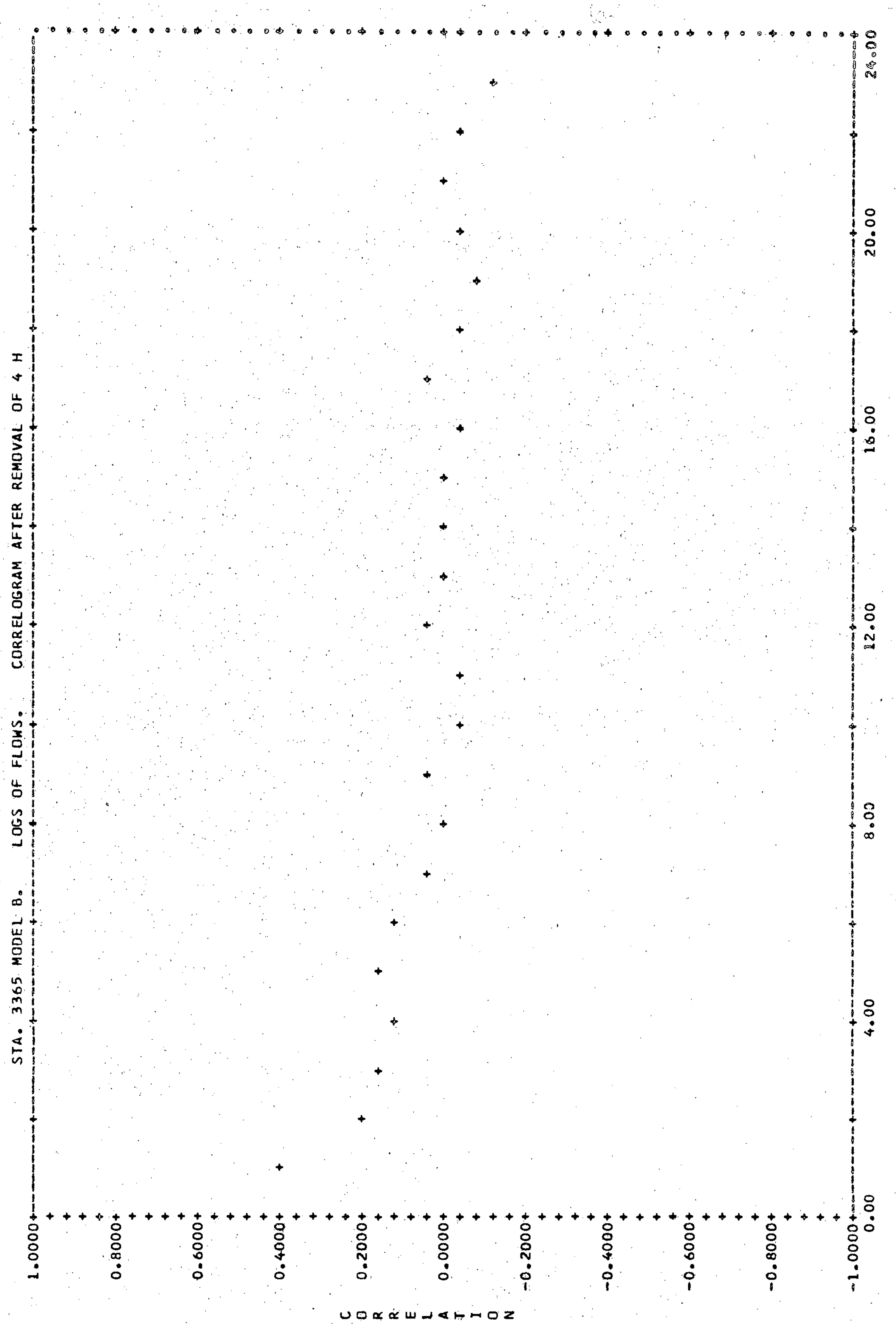
CORRELOGRAMS

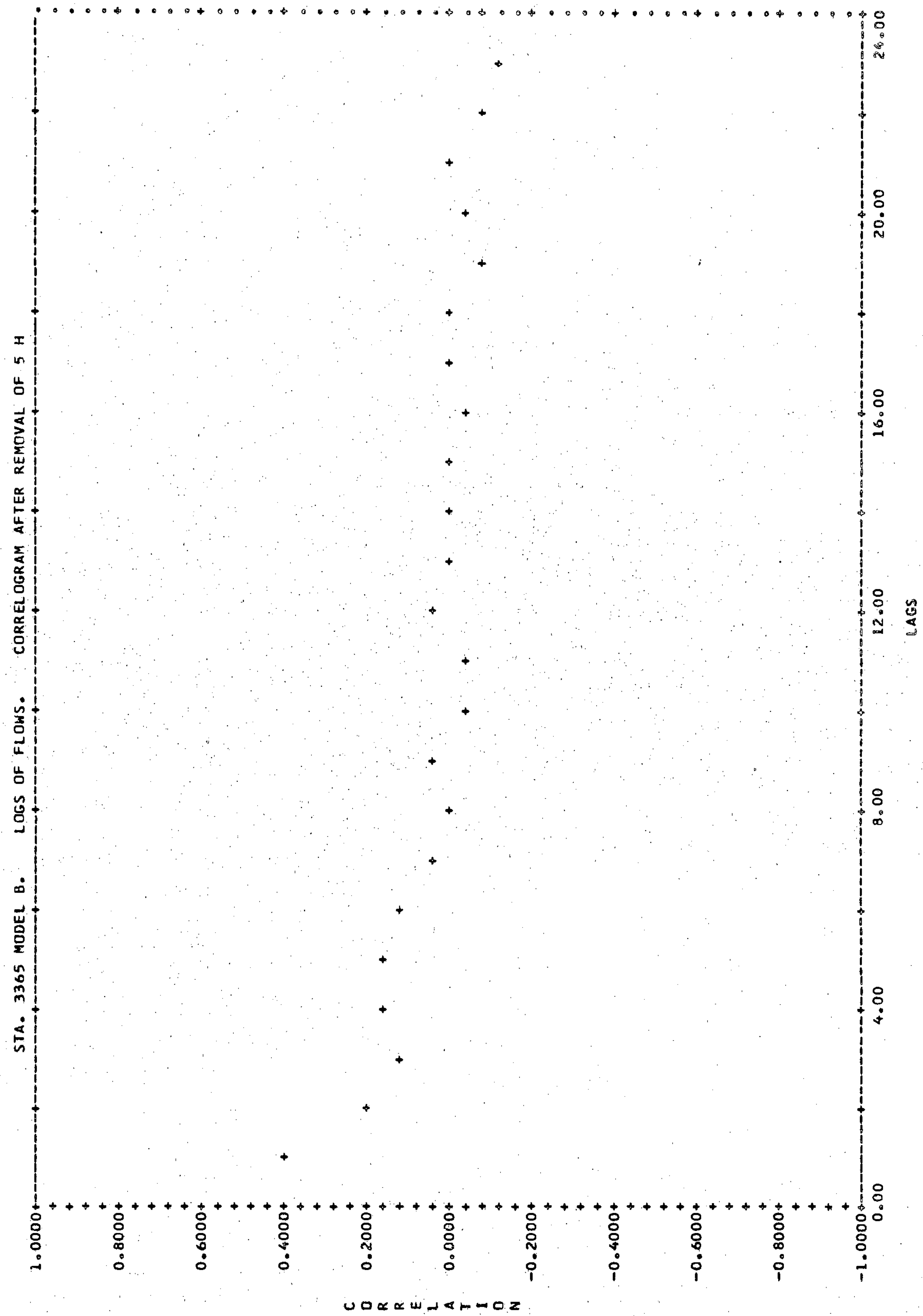

STA. 3365 MODEL B LOGS OF FLOWS. CORRELOGRAM WITHOUT HARMONIC REM
1.0000
0.8000
0.6000
0.4000
0.2000
0.0000
-0.2000
-0.4000
-0.6000
-0.8000
-1.0000
CORRELATION
0.00
4.00
8.00
12.00
16.00
20.00
24.00
LAGS

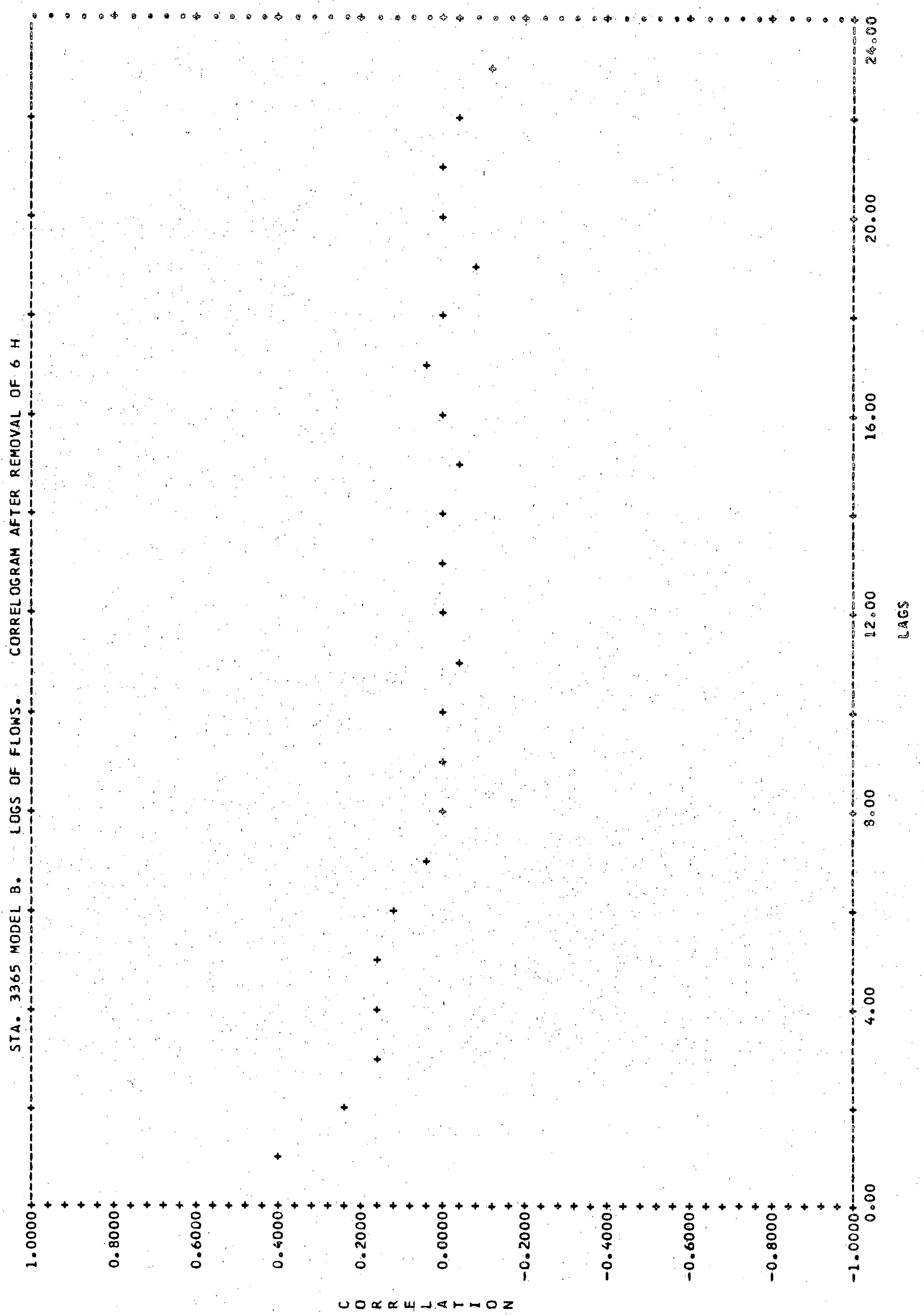

STA. 3365 MODEL B.   LOGS OF FLOWS.   CORRELOGRAM AFTER REMOVAL OF 1 H
CORRELATION
1.0000
0.8000
0.6000
0.4000
0.2000
0.0000
-0.2000
-0.4000
-0.6000
-0.8000
-1.0000
0.00
4.00
8.00
12.00
16.00
20.00
24.00
LAGS


STA. 3365 MODEL B. LOGS OF FLOWS. CORRELOGRAM AFTER REMOVAL OF 2 H
1.0000
0.8000
0.6000
0.4000
0.2000
0.0000
-0.2000
-0.4000
-0.6000
-0.8000
-1.0000
CORRELATION
0.00
4.00
8.00
12.00
16.00
20.00
24.00
LAGS

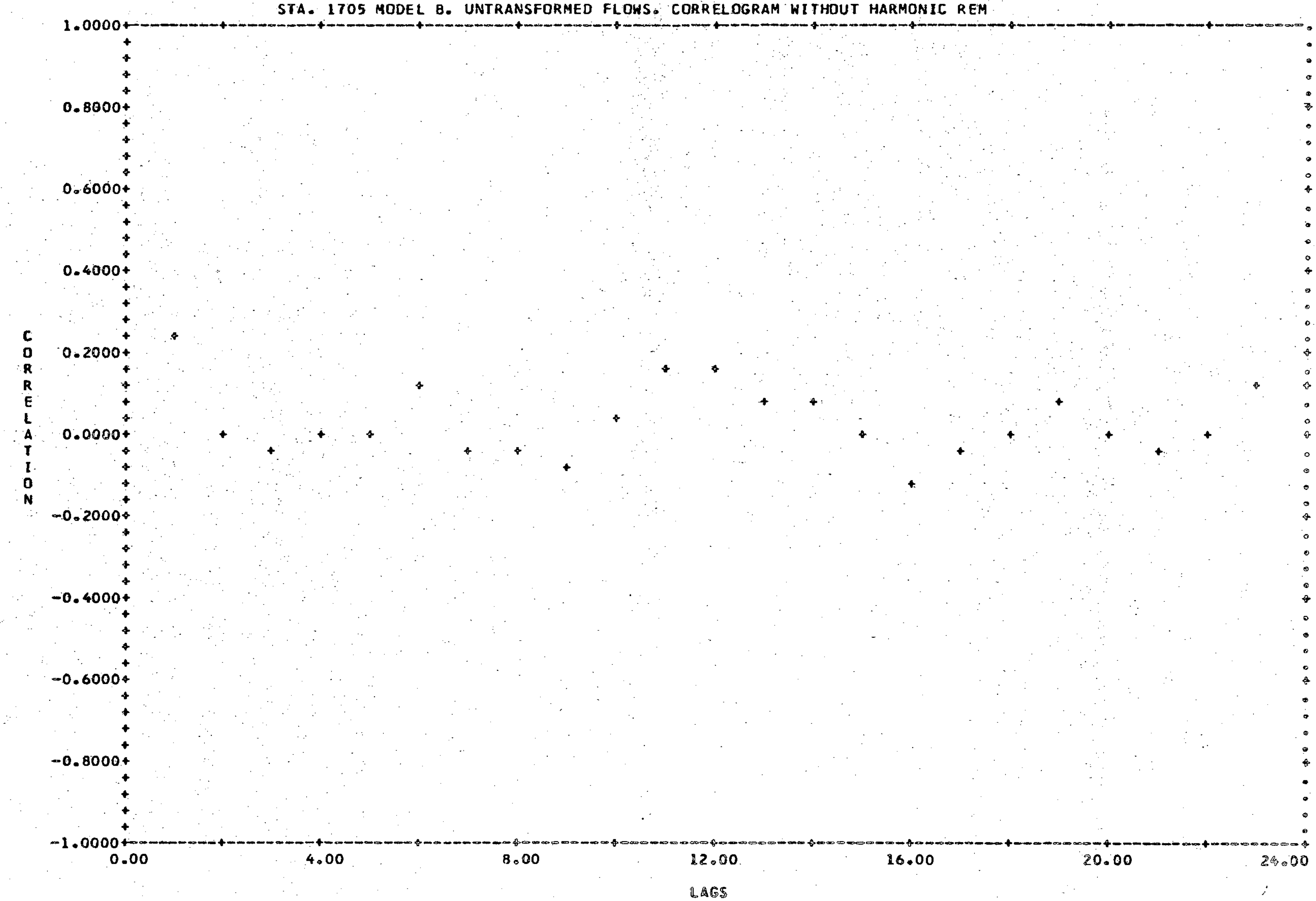

STA. 3365 MODEL B.   LOGS OF FLOWS.   CORRELOGRAM AFTER REMOVAL OF 3 H
1.0000
0.8000
0.6000
0.4000
0.2000
0.0000
-0.2000
-0.4000
-0.6000
-0.8000
-1.0000
CORRELATION
0.00
4.00
8.00
12.00
16.00
20.00
24.00
LAGS

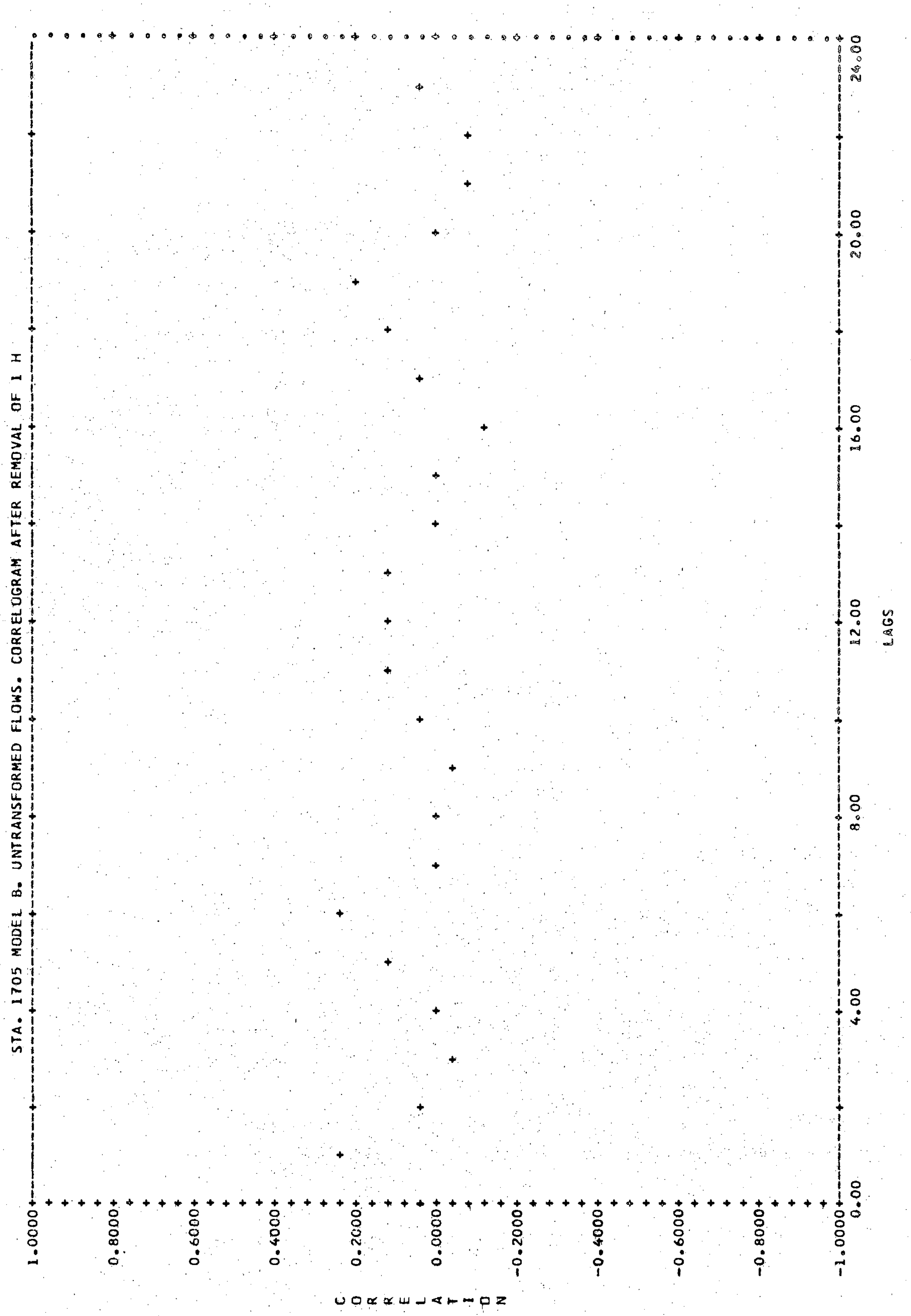

STA. 3365 MODEL B. LOGS OF FLOWS. CORRELOGRAM AFTER REMOVAL OF 4 H
1.0000
0.8000
0.6000
0.4000
0.2000
0.0000
-0.2000
-0.4000
-0.6000
-0.8000
-1.0000
CORRELATION
0.00
4.00
8.00
12.00
16.00
20.00
24.00
LAGS

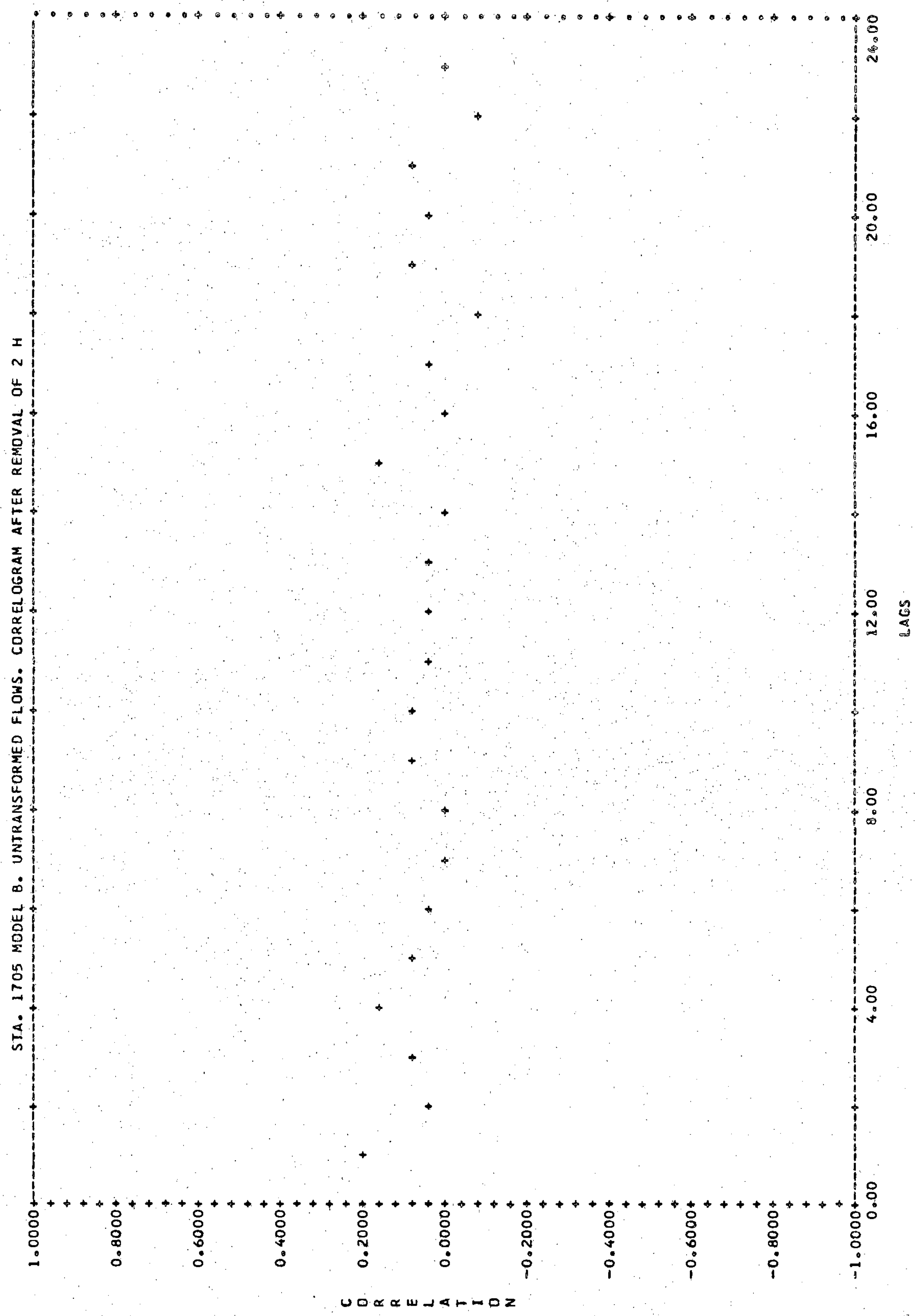

STA. 3365 MODEL B. LOGS OF FLOWS. CORRELOGRAM AFTER REMOVAL OF 5 H
1.0000
0.8000
0.6000
0.4000
0.2000
0.0000
-0.2000
-0.4000
-0.6000
-0.8000
-1.0000
CORRELATION
0.00
4.00
8.00
12.00
16.00
20.00
24.00
LAGS

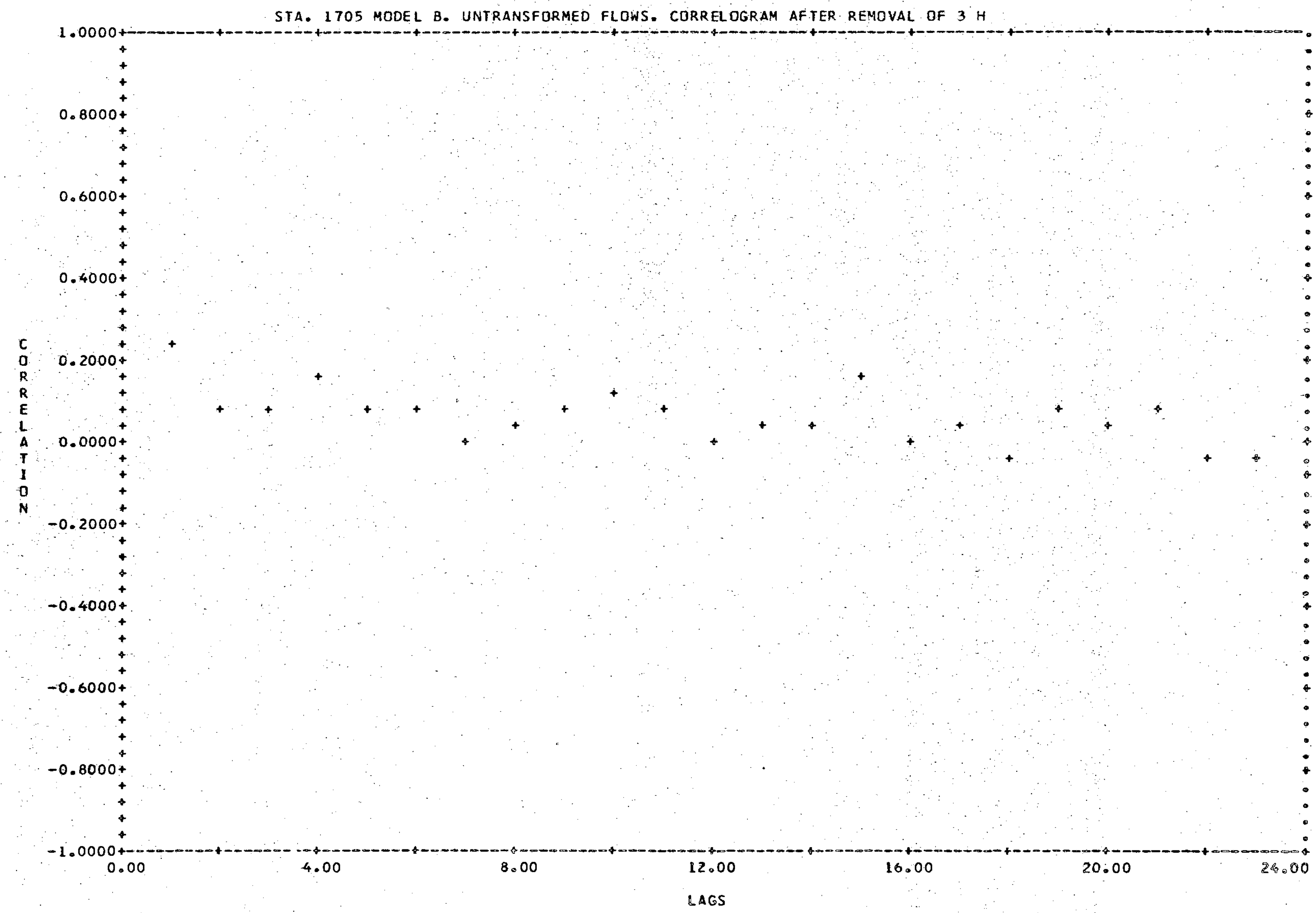

STA. 3365 MODEL B. LOGS OF FLOWS. CORRELOGRAM AFTER REMOVAL OF 6 H
1.0000
0.8000
0.6000
0.4000
0.2000
0.0000
-0.2000
-0.4000
-0.6000
-0.8000
-1.0000
CORRELATION
0.00
4.00
8.00
12.00
16.00
20.00
24.00
LAGS

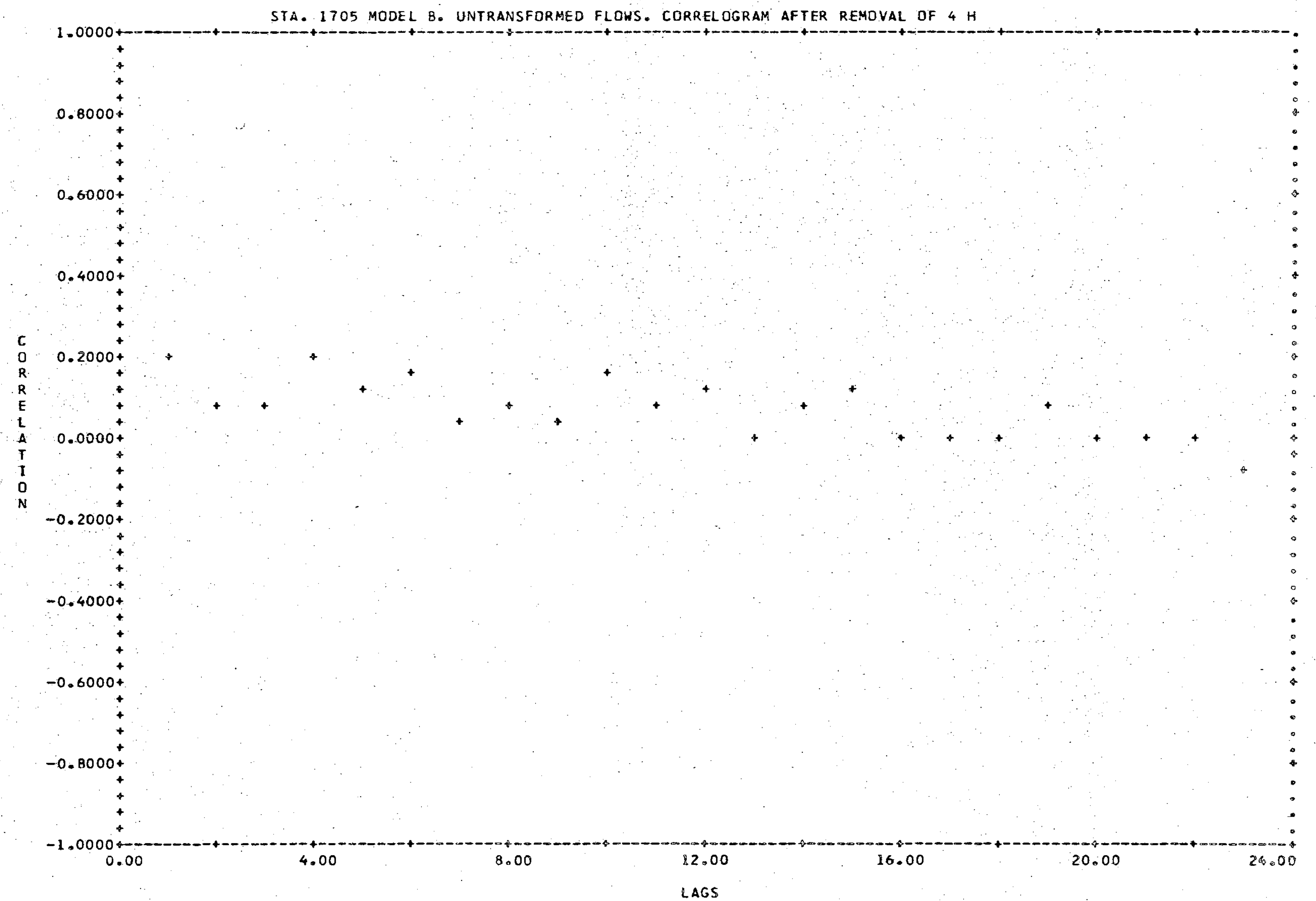
STA. 1705 MODEL B. UNTRANSFORMED FLOWS. CORRELOGRAM WITHOUT HARMONIC REM

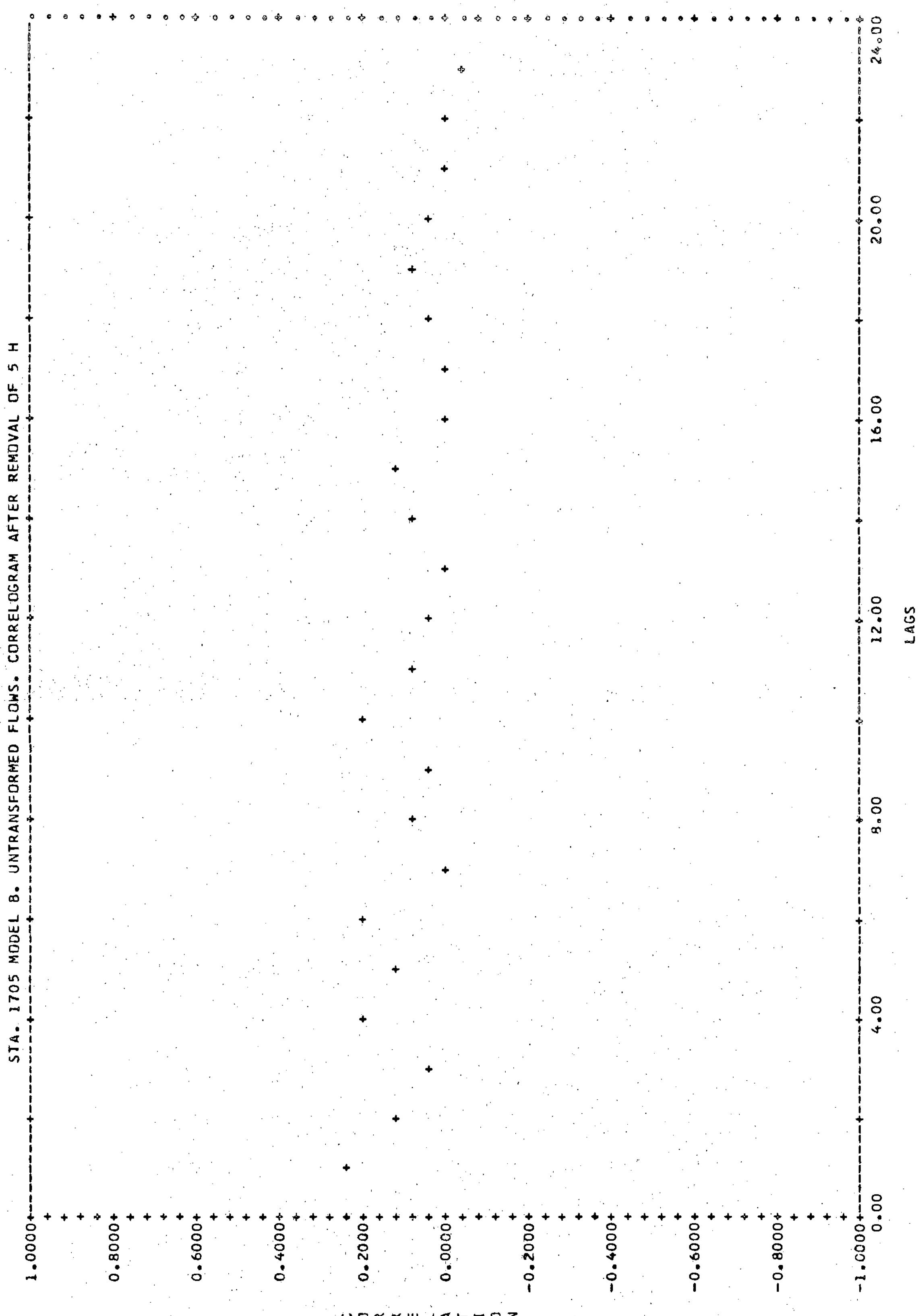


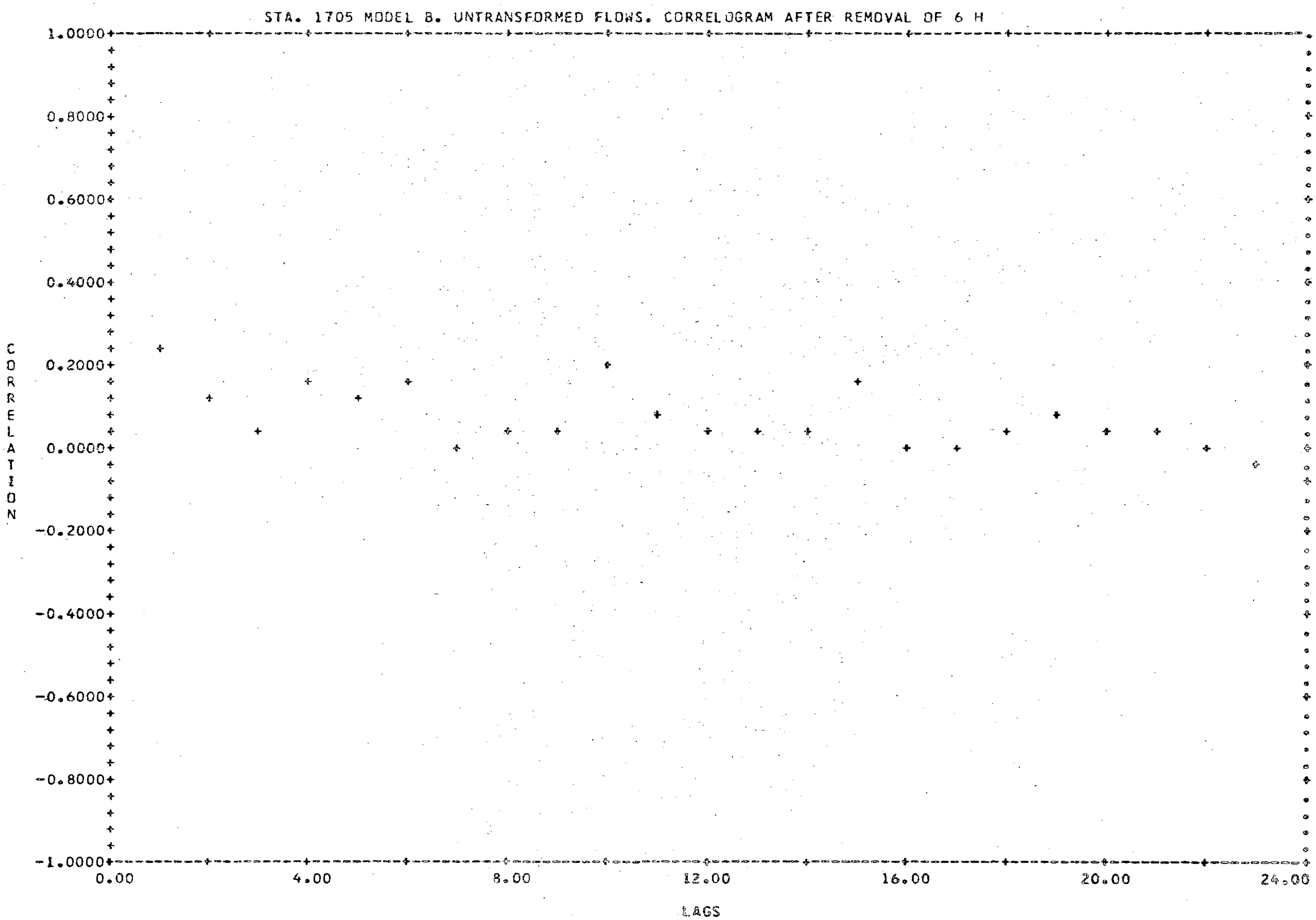

STA. 1705 MODEL B. UNTRANSFORMED FLOWS. CORRELOGRAM AFTER REMOVAL OF 1 H
1.0000
0.8000
0.6000
0.4000
0.2000
0.0000
-0.2000
-0.4000
-0.6000
-0.8000
-1.0000
CORRELATION
0.00
4.00
8.00
12.00
16.00
20.00
24.00
LAGS

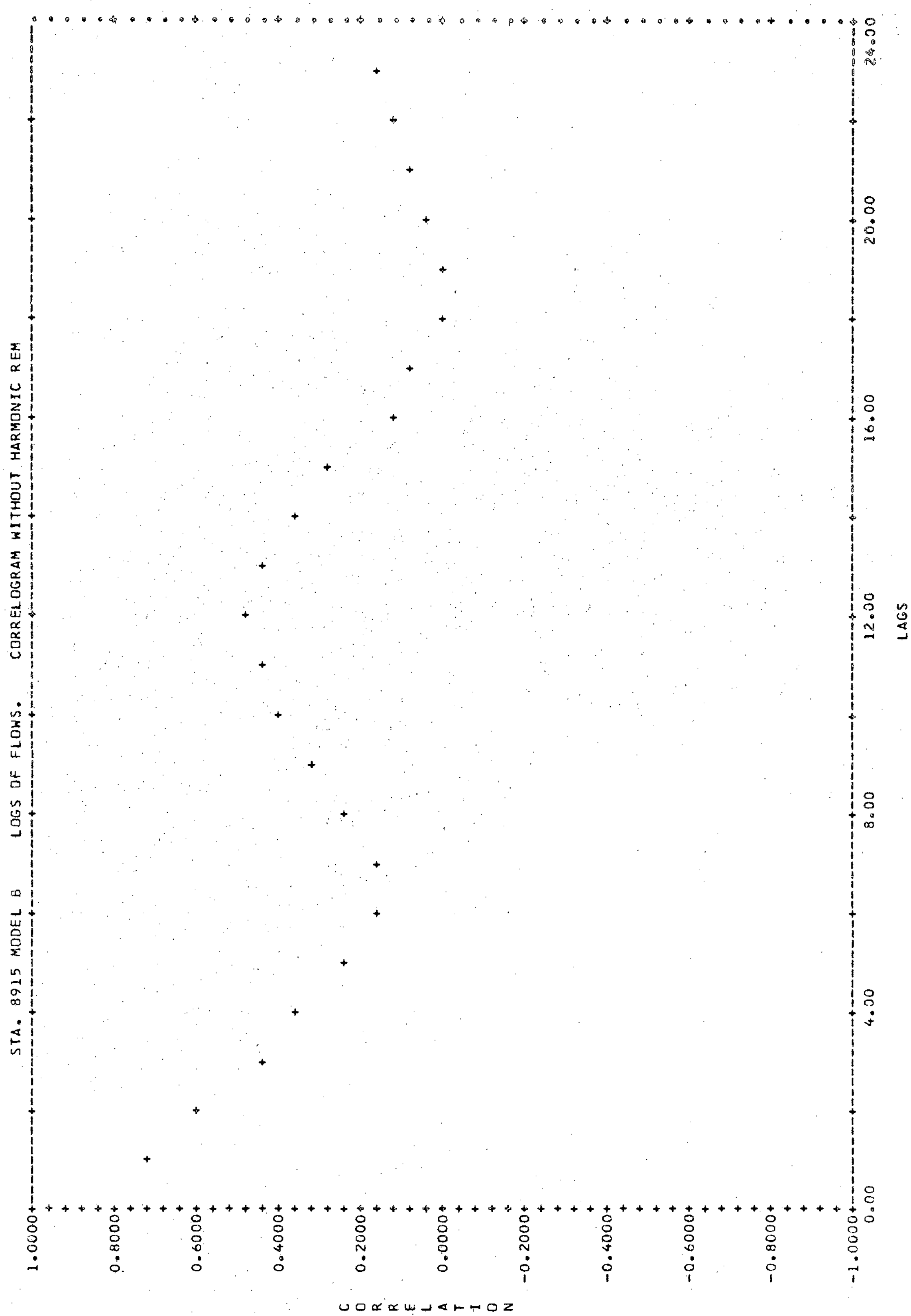

STA. 1705 MODEL B. UNTRANSFORMED FLOWS. CORRELOGRAM AFTER REMOVAL OF 2 H
1.0000
0.8000
0.6000
0.4000
0.2000
0.0000
-0.2000
-0.4000
-0.6000
-0.8000
-1.0000
CORRELATION
0.00
4.00
8.00
12.00
16.00
20.00
24.00
LAGS

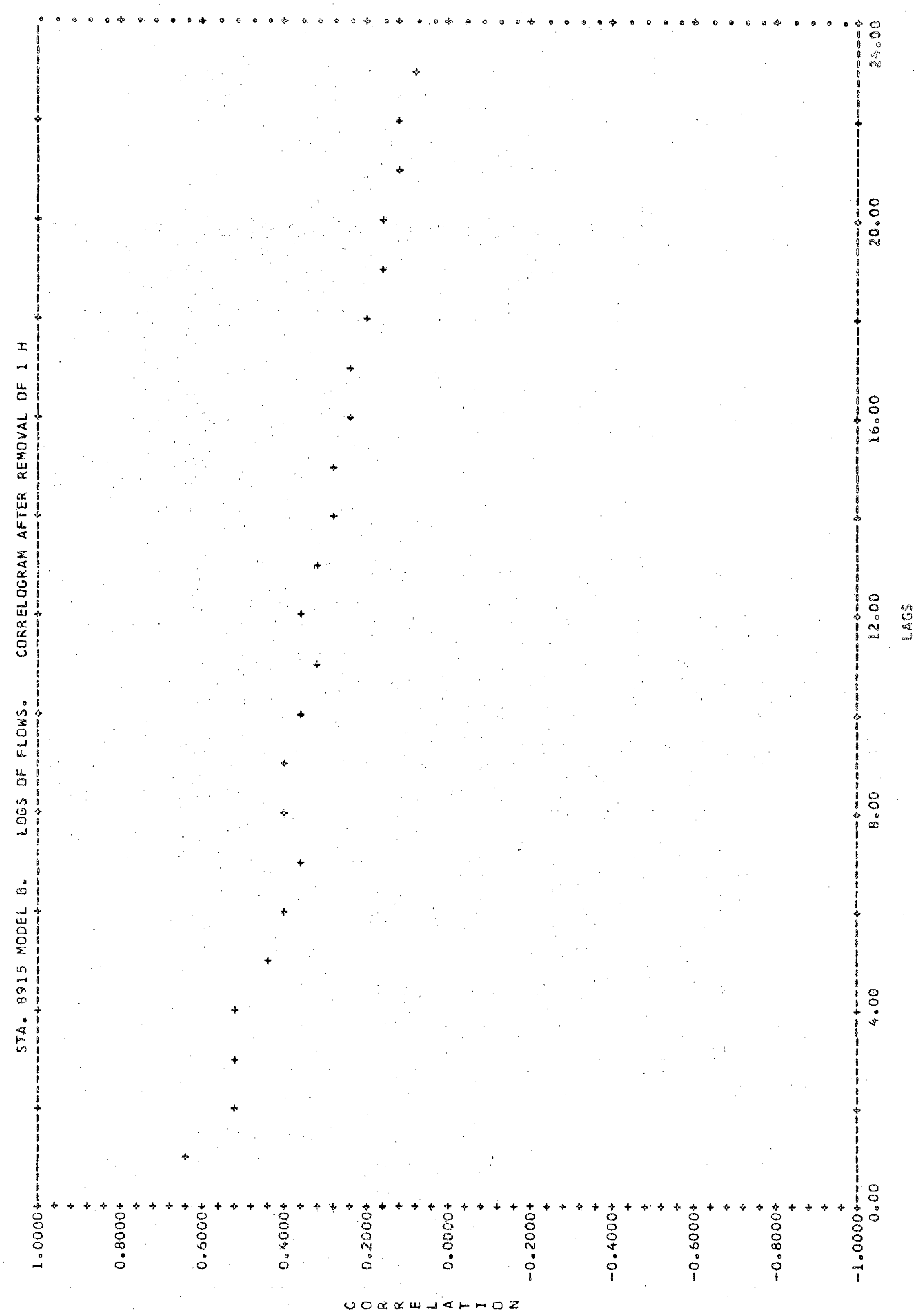

STA. 1705 MODEL B. UNTRANSFORMED FLOWS. CORRELOGRAM AFTER REMOVAL OF 3 H
CORRELATION
1.0000
0.8000
0.6000
0.4000
0.2000
0.0000
-0.2000
-0.4000
-0.6000
-0.8000
-1.0000
0.00
4.00
8.00
12.00
16.00
20.00
24.00
LAGS

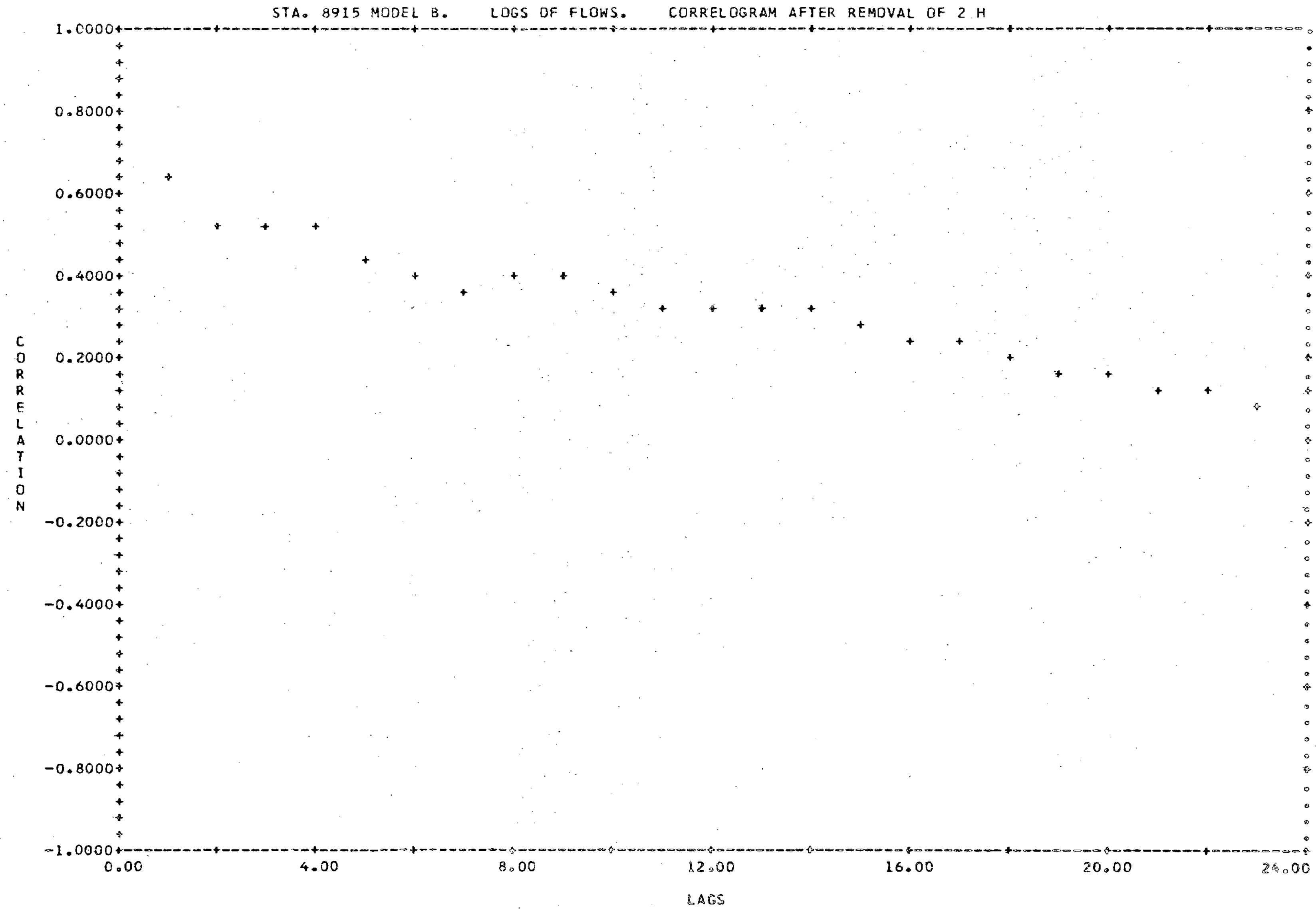

STA. 1705 MODEL B. UNTRANSFORMED FLOWS. CORRELOGRAM AFTER REMOVAL OF 4 H
1.0000
0.8000
0.6000
0.4000
0.2000
0.0000
-0.2000
-0.4000
-0.6000
-0.8000
-1.0000
CORRELATION
0.00
4.00
8.00
12.00
16.00
20.00
24.00
LAGS

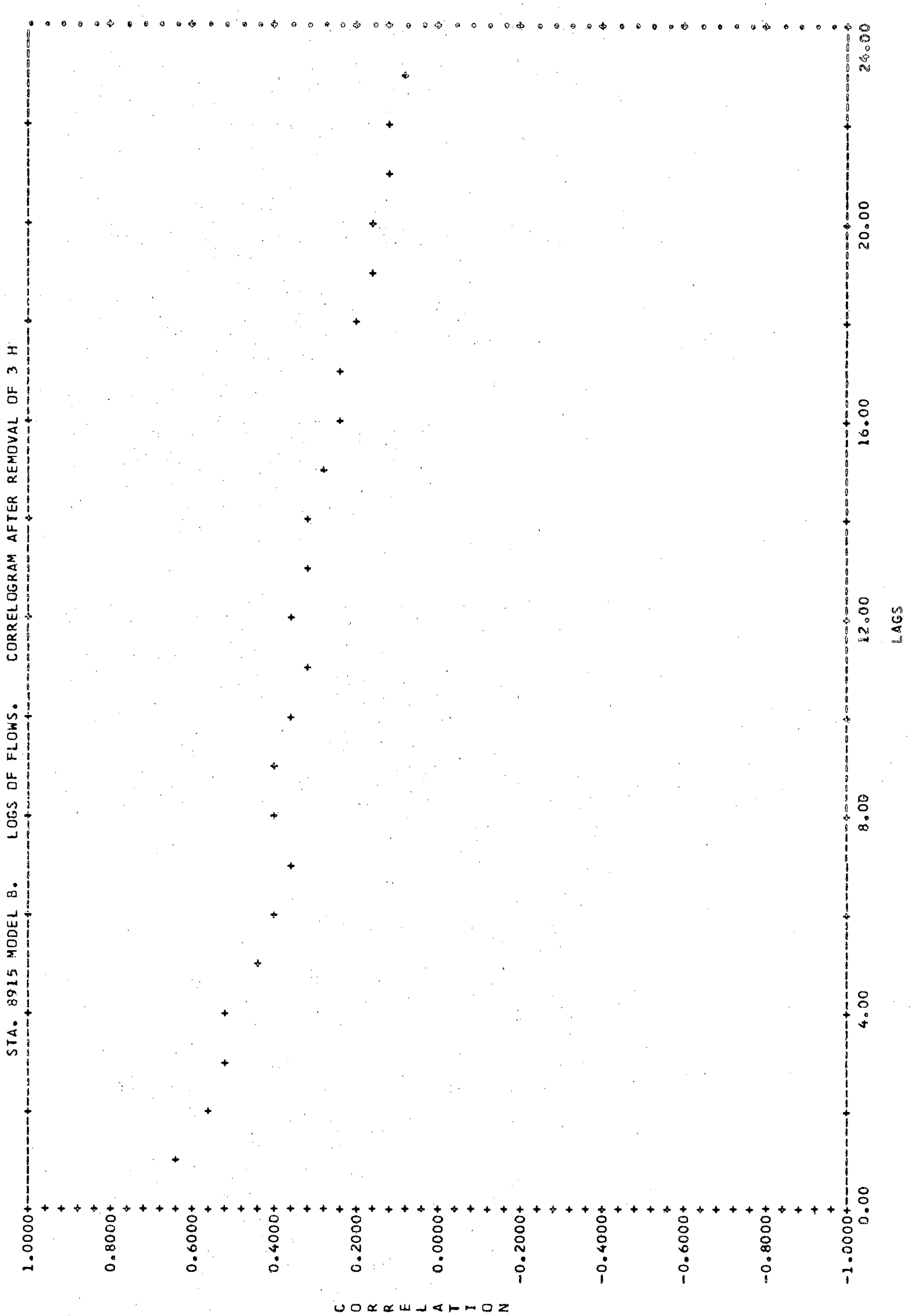

STA. 1705 MODEL B. UNTRANSFORMED FLOWS. CORRELOGRAM AFTER REMOVAL OF 5 H
1.0000
0.8000
0.6000
0.4000
0.2000
0.0000
-0.2000
-0.4000
-0.6000
-0.8000
-1.0000
CORRELATION
0.00
4.00
8.00
12.00
16.00
20.00
24.00
LAGS

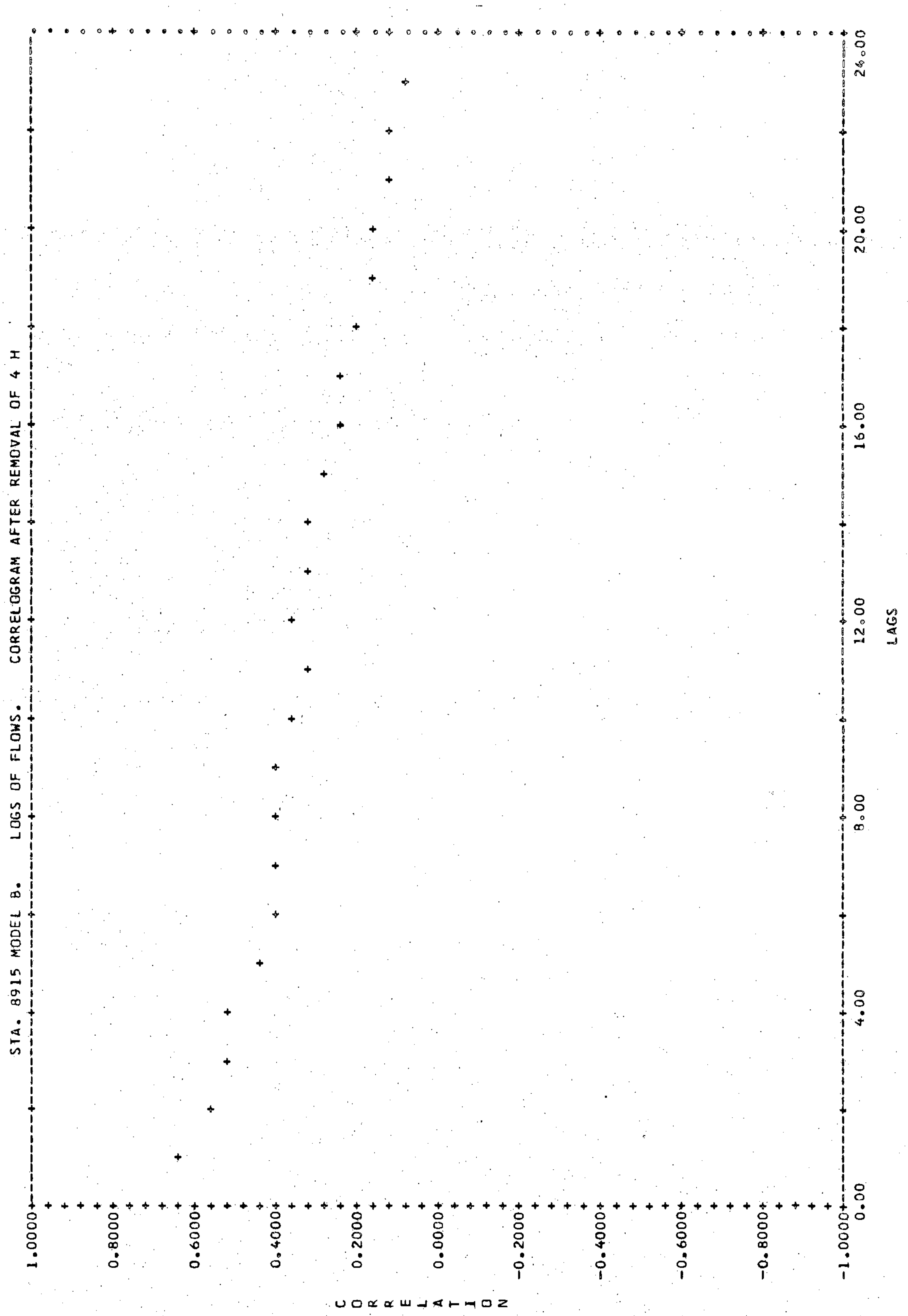

STA. 1705 MODEL B. UNTRANSFORMED FLOWS. CORRELOGRAM AFTER REMOVAL OF 6 H
1.0000
0.8000
0.6000
0.4000
0.2000
0.0000
-0.2000
-0.4000
-0.6000
-0.8000
-1.0000
CORRELATION
0.00
4.00
8.00
12.00
16.00
20.00
24.00
LAGS

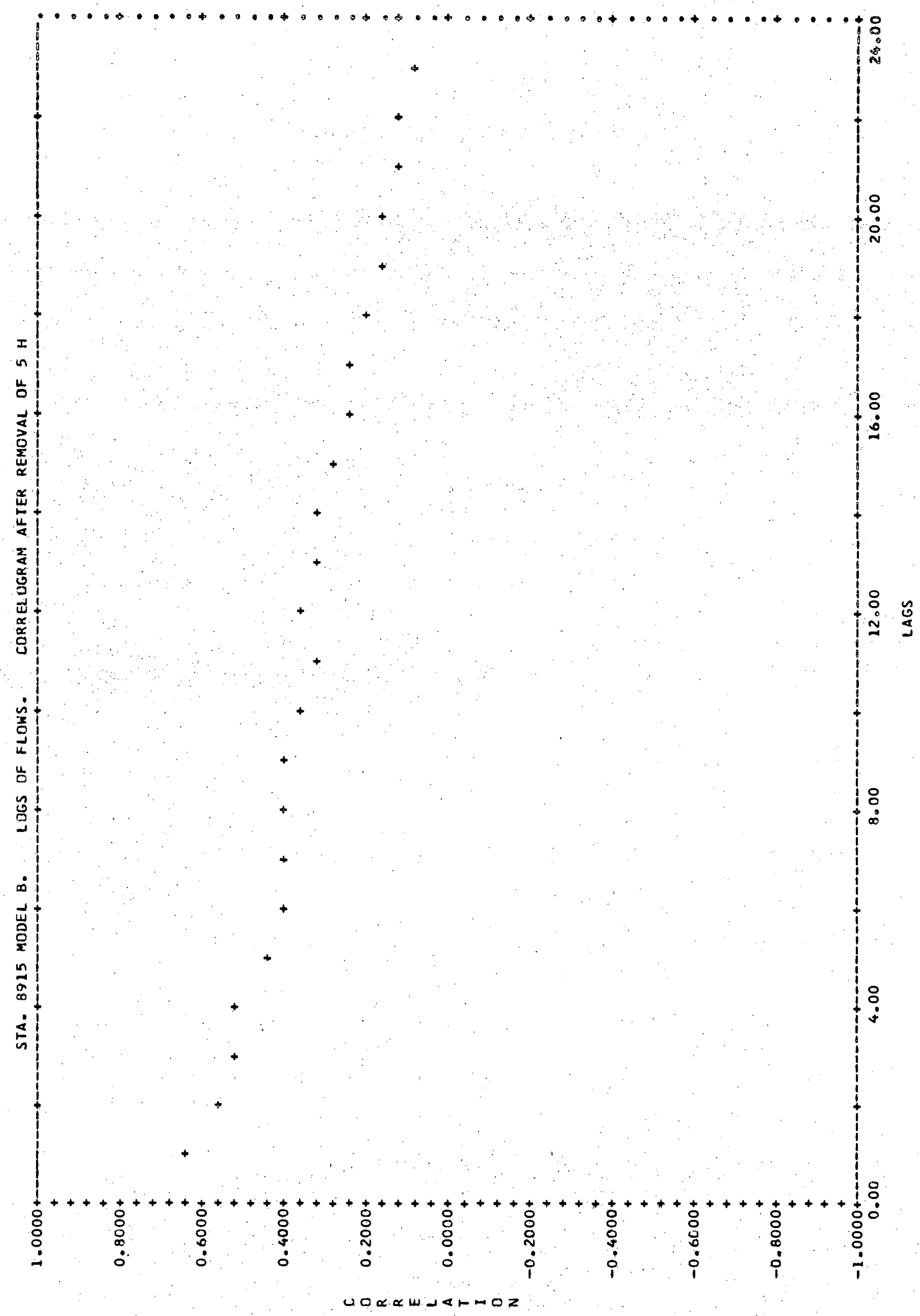

STA. 8915 MODEL B   LOGS OF FLOWS.   CORRELOGRAM WITHOUT HARMONIC REM
1.0000
0.8000
0.6000
0.4000
0.2000
0.0000
-0.2000
-0.4000
-0.6000
-0.8000
-1.0000
CORRELATION
0.00
4.00
8.00
12.00
16.00
20.00
24.00
LAGS

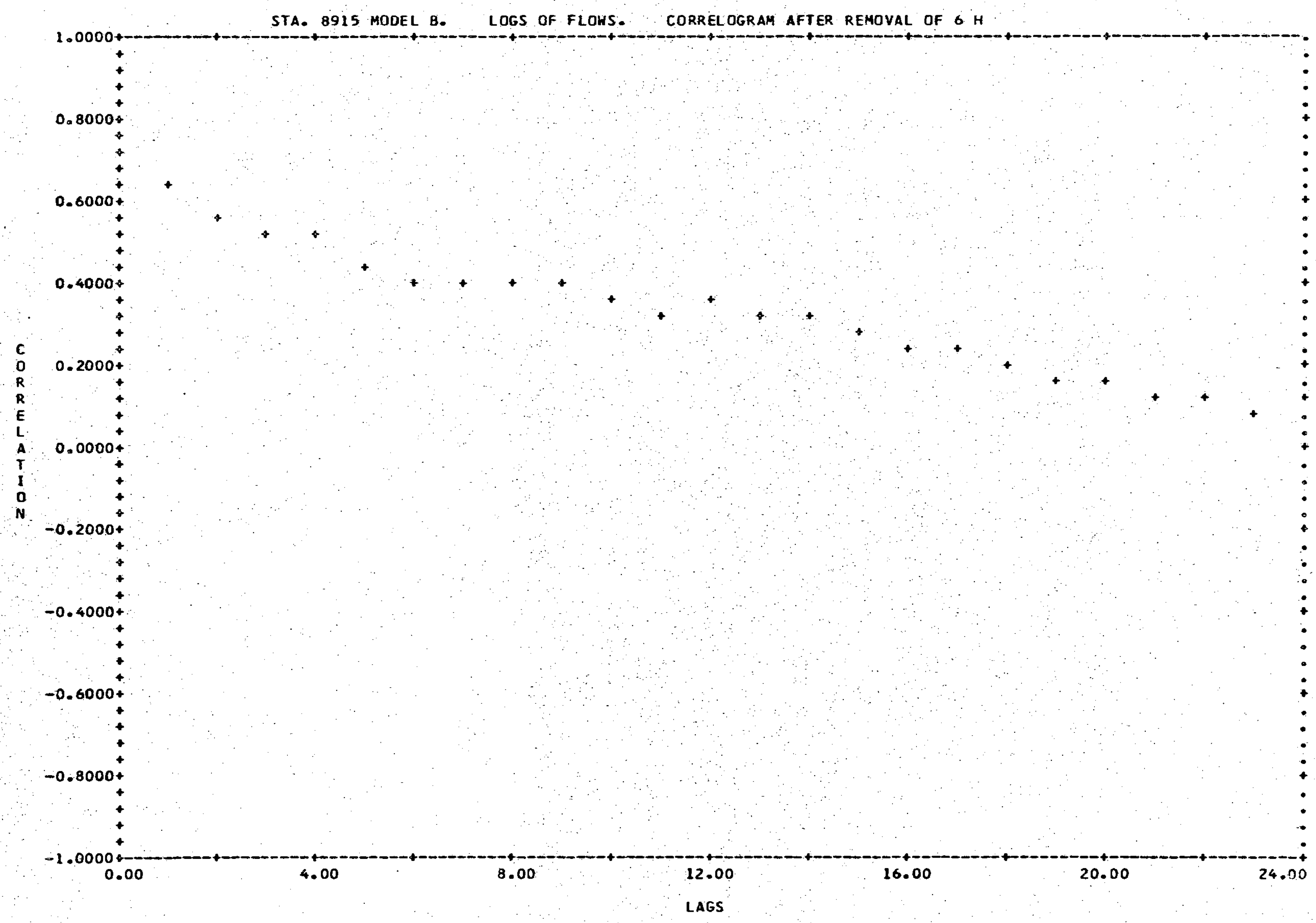

STA. 8915 MODEL B. LOGS OF FLOWS. CORRELOGRAM AFTER REMOVAL OF 1 H
1.0000
0.8000
0.6000
0.4000
0.2000
0.0000
-0.2000
-0.4000
-0.6000
-0.8000
-1.0000
CORRELATION
0.00
4.00
8.00
12.00
16.00
20.00
24.00
LAGS


STA. 8915 MODEL B. LOGS OF FLOWS. CORRELOGRAM AFTER REMOVAL OF 2 H
CORRELATION
1.0000
0.8000
0.6000
0.4000
0.2000
0.0000
-0.2000
-0.4000
-0.6000
-0.8000
-1.0000
0.00
4.00
8.00
12.00
16.00
20.00
24.00
LAGS


STA. 8915 MODEL B. LOGS OF FLOWS. CORRELOGRAM AFTER REMOVAL OF 3 H
1.0000
0.8000
0.6000
0.4000
0.2000
0.0000
-0.2000
-0.4000
-0.6000
-0.8000
-1.0000
CORRELATION
0.00
4.00
8.00
12.00
16.00
20.00
24.00
LAGS


STA. 8915 MODEL B. LOGS OF FLOWS. CORRELOGRAM AFTER REMOVAL OF 4 H
CORRELATION
1.0000
0.8000
0.6000
0.4000
0.2000
0.0000
-0.2000
-0.4000
-0.6000
-0.8000
-1.0000
0.00
4.00
8.00
12.00
16.00
20.00
24.00
LAGS


STA. 8915 MODEL B. LOGS OF FLOWS. CORRELOGRAM AFTER REMOVAL OF 5 H
CORRELATION
1.0000
0.8000
0.6000
0.4000
0.2000
0.0000
-0.2000
-0.4000
-0.6000
-0.8000
-1.0000
0.00
4.00
8.00
12.00
16.00
20.00
24.00
LAGS



STA. 8915 MODEL B. LOGS OF FLOWS. CORRELOGRAM AFTER REMOVAL OF 6 H
1.0000
0.8000
0.6000
0.4000
0.2000
0.0000
-0.2000
-0.4000
-0.6000
-0.8000
-1.0000
CORRELATION
0.00
4.00
8.00
12.00
16.00
20.00
24.00
LAGS

VITA

F. T. Painter

Candidate for the Degree of

Master of Science

Thesis: AUTOCORRELATION ANALYSIS OF STREAMFLOW SEQUENCES

Major Field: Civil Engineering

Biographical:

Personal Data: Born August 13, 1944, at Swansea, Wales, the son of Mr. and Mrs. John Painter.

Education: B.Sc. (Hons.), The University of Newcastle Upon Tyne, United Kingdom, June 1966. Completed requirements for the Degree of Master of Science at Oklahoma State University in May 1969.

Professional Experience: Engineering Assistant, John Taylor and Sons, Chartered Civil Engineers, London, U. K. 1966-67; Research Assistant, Oklahoma State University, 1967-68.

Membership of Professional Societies: Associate Member of the Institution of Civil Engineers, London.

Publications: Painter, F. T. (1966) Water Resources in Great Britain, J. Proc. Inst. Sew. Purif., (6), 581-584.