TIME SERIES ANALYSIS OF WATER QUALITY DATA

by

NAVIN KUMAR BHARGAVA

B.Sc. (Engg.), University of Delhi, Delhi, India, 1969

A MASTER'S REPORT

submitted in partial fulfillment of the

requirements for the degree

MASTER OF SCIENCE

Department of Industrial Engineering

KANSAS STATE UNIVERSITY

Manhattan, Kansas

1974

Approved by:

Major Professor

THIS BOOK CONTAINS NUMEROUS PAGES WITH THE ORIGINAL PRINTING ON THE PAGE BEING CROOKED.

THIS IS THE BEST IMAGE AVAILABLE.

THIS BOOK CONTAINS NUMEROUS PAGES WITH DIAGRAMS THAT ARE CROOKED COMPARED TO THE REST OF THE INFORMATION ON THE PAGE.

THIS IS AS RECEIVED FROM CUSTOMER.

ACKNOWLEDGEMENTS

I offer sincere appreciation to my major professor, Dr. E. Stanley Lee for the guidance and inspiration throughout both this report and my graduate studies.

I also thank Dr. L. E. Grosh for his assistance in various aspects of this project.

I am grateful to the faculty members of the Industrial Engineering Department at Kansas State University who helped me in various ways during my graduate program.

I would like to thank Mrs. M. T. Jirak for typing the report.

TABLE OF CONTENTS

CHAPTER I

INTRODUCTION

Great attention has been devoted in recent years to building water quality models so as to gain greater understanding and insight of the underlying phenomena of water pollution and for purposes of prediction of future behaviour of the pollutants. The most common parameters studied for water quality analysis are

a) Temperature

b) Dissolved Oxygen

c) Biochemical Oxygen Demand

d) Chloride Contamination

e) Flow rate and others.

The purpose of water quality management system is to control the aforesaid factors with a view to maintain the water quality within certain acceptable standards. For any such system to be effective, it is necessary to have knowledge of the pollution phenomenon and its future behaviour.

Various approaches have been suggested in the past years to build mathematical models for these parameters. Among the important water quality indicators, dissolved oxygen and biochemical oxygen demand relationship has been studied most extensively. The dissolved oxygen (DO) serves as a surrogate variable indicating the general 'health' of the stream and its ability to maintain and propagate a balanced ecological system [28]. From an analytical point of view, the DO system is quite complex and

reflects interrelationships between the chemistry and biology of the stream, together with the man imposed effects of waste discharge. The general approach towards building a suitable model for DO system has been to consider the relationship between the DO level and the discharges of oxidizable organic matter. Various models differ in the complexity of the assumptions lying behind them.

Streeter and Phelps [1], Dobbins [2], O'Connor [4], Thomann [3] presented some of the models with varying degree of complexity in terms of the sources and sinks of dissolved oxygen and flow type.

Streeter and Phelps [1] model was formulated considering that the changes in the concentrations of DO and BOD in a stream is affected by two processes only (a) the source of DO is natural aeration (b) the primary sink of DO is biodegradable organic matter which uses DO in its stabilization. The system was assumed to be uniform and steady state.

Due to its apparent limitations as far as the sources and sinks of oxygen are concerned, this model was modified by Dobbins [2] who extended the Streeter's model to include more sources and sinks of oxygen. It was pointed out that certain additional processes should be considered as sources and sinks of oxygen such as removal of BOD by sedimentation or absorption, removal of oxygen by respiration of acquatic plants etc. O'Connor [16] proposed a model for the oxygen balance of an estuary wherein it was assumed that the movement of the organic impurities is caused by tidal action and a distributed source such as land runoff. The dissolved oxygen profile depends on the concentration of the organic material, its rate of oxidation and the resulting rate of reaeration. A set of differential

equations was obtained for steady state which were solved assuming constant coefficients. This model was applied for analysis at the Delaware Estuary and the Lower James River which are subjected to tidal action of the Chesapeake Bay. The DO profiles were found to be consistent and in better agreement with the experimental results than the BOD decay profiles.

Another approach by segmenting the whole estuary into a number of sections, where each segment is considered to be completely mixed volume and no gradients are permitted within the section, was proposed by Thomann [3, 26]. The mass transport of material (waste discharge) across any section by the net river flow over a tidal cycle is written as the resultant of material brought in from the previous section and given out to the next section. A solution was obtained for steady state conditions. Later on a solution for non-steady state was obtained by Pence, Jeglic and Thomann [5]. In another publication, an approach for providing the spatial and temporal distribution of dissolved oxygen was given by O'Connor [4].

Since reaeration rate plays an important role in the study of dissolved oxygen behaviour, its value should be determined quite accurately. A study [18] was conducted to ascertain the effect of water temperature on stream reaeration rate. An experimental investigation was done which showed that the rate of reaeration increases at the geometric rate of 2.41% per ^{o}C throughout the range of temperature found in a natural stream. This study pointed out a definite relationship between DO and temperature. O'Connor and Dobbins [19] proposed two formulae

for the prediction of aeration coefficient. Some other formulations have also been put forward to study the reaeration coefficient.

Li [17] obtained a model for DO defecit at any cross section of a polluted stream under the assumptions that (i) the stream discharge is steady at any cross-section of the stream but may vary along the course (ii) the sewage is well mixed vertically and laterally (iii) the effect of longitudinal turbulent diffusion is negligible. Three cases were solved using this model (a) steady state BOD and DO loading (b) BOD discharge loading fluctuations as a function of time (c) BOD and DO discharge fluctations of one cycle/day.

Earlier models usually considered the DO - BOD System as a deterministic one. Information on the time variability of water quality parameters is being generated at an ever increasing rate due primarily to the installation of continuous water quality monitoring stations. This provides basis to analyse the time varying properties of the water pollutants.

A stochastic model to describe the behaviour of BOD/DO in streams was given by Thayer and Krutchkoff [6]. This model is essentially based on Dobbin's model [2] with the mechanisms affecting BOD and DO modified so as to be random in nature. The changes in the levels of BOD and DO were considered in intergal units according to a Poisson birth and death process. An important result of the model was that the variance of dissolved oxygen vs. time or distance increases with decreasing mean DO. This brings out a serious shortcoming of the deterministic standards of water quality as even if the mean DO is slightly above standards at

a point, it is the point where the DO has the maximum variability and hence chances of violation. Later on this model was modified for use in estuaries by Custer and Krutehkoff [7]. They also considered both temporal and spatial variations in BOD and DO concentrations.

Thomman, O'Connor and Di'Tora [29] discussed the time varying aspects of water quality indicators. Using Fourier analysis and least squares estimation method; certain periodicities, corresponding to cyclic fluctuations in water quality indicators were incorporated in a model. It was seen that DO shows several periodicities such as annual, semiannual, triannual and semitriannual etc. These cyclic fluctations could be given physical interpretation based on information about the system. Further, a time varying model was formulated and used to study the variation of DO, chloride contents of Potomac estuary.

Moving average analysis and linear regression analysis was used by Anderson and Zogro. ski [30] to study the long term trends in water quality parameters in Passaic river basin, New Jersey. Moving averages were used as they tend to dampen the extremes of short term fluctuations. Though moving average analysis can be used to indicate macroscopic trends, they should not be used for microscopic variations or for prediction purposes. Correlation analysis should be carried out along with this to obtain an effective model.

A multiple regression analysis approach was used by Tirabassi [31] to investigate within station and interstation relationships between various water quality parameters. Predictive models were obtained for 18 water quality indicators for Passaic River which explained the data

satisfactorily. These models seem to provide a basis for obtaining mathematical relationship between several water quality parameters using statistical methods alone.

Another approach to analyse the hydrologic system is the use of time series analysis which makes use of the feature that the hydrologic events are not independently distributed in time i.e. the behaviour of the system is governed according to laws of probability as well as the sequential relationship between the events. Matalas [10], Julian [11], Thomann [8,12], Gunnerson [22], Wastler [9], Wallace [14], DeMayo [15] have initiated the use of this technique in water pollution domain.

Gunnerson [22] applied spectral analysis technique to analyse dissolved oxygen data for the Potomac River and Raritan Bay at the mouth of Raritan River in an effort to gain useful insight into various physical and chemical processes in an estuary and to optimise the sampling interval for data collection. Periodicities at 24 hr., 12 hr. and 14 days were observed which were attributed to photosynthesis, semidurinal tides and linear fortrightly, respectively. It was concluded that a sampling interval of 2 hrs. was sufficient as no more useful information could be extracted from a more frequent data. But a more frequent sampling may be necessary near the dominant source of pollution in an estuary. With increasing distance downstream, mixing and stabilization processes and dilution from tributaries result increasing homogeneity which can be described by less frequent sampling.

Thomann [8] analyzed temperature and dissolved oxygen data from

the Delaware Estuary using harmonic analysis in conjunction with spectral analysis. It was concluded that temperature has a dominant annual cyclic variation while dissolved oxygen showed annual, semi-annual peaks. Dinrnal variability in DO due to photosynthesis was also identified. In another publication, Thomann [12] reported a study about the variability of waste treatment plants. Simple spectral analysis along with cross spectral analysis was performed to study the relationship of flow, influent BOD and effluent BOD from a waste treatment plant. The results indicated a high degree of variability in secondary effluent BOD as measured by the coefficient of variation. A low coherency coefficient showed that the secondary effluent variability is not significantly influenced by the raw influent variability.

Wastler and Walter [9] studied the behaviour of the Charleston Harbor as regards to the river flow and chloride concentration. The purpose of the study was to predict the estuary's behaviour under conditions of reduced inflow. Sampling at a frequency of 4 hrs. was done at 10 stations both at the surface and at a depth of 20 - 25 ft. The total sampling period was one month. Sampling at two depths was done to ascertain the stratification of the river. Using cross spectral analysis, it was concluded from the results of the surveys that below discharges of about 16000 cfs, the river would behave as completely unstratified.

Cross spectral analysis has been used to study evaporation, rainfall and run off by Yu (24) and Nordin [23].

Use of parameteric time series models for water quality purposes has been suggested by Mimichael and Hunter [13]. Forecast functions

were developed for the Ohio River temperature and flow data.

The purpose of this report was to investigate the behaviour of water quality parameters for the Potomac and the Ontario River using time series analysis.

Spectral analysis was used for the identification of the causal phenomena behind individual pollutant variation in Chapter II. Using stepwise regression procedure in conjunction with these results, a forecast model was formed.

The correlation among different pollutants in the Ontario River was studied with the help of cross-spectral analysis in Chapter III. Chapter IV of the report deals with the parametric modeling of Ontario river data using autoregressive-moving average models.

Chapter V concerns the analysis of temperature, DO, BOD and Chloride data at Potomac river using spectral analysis. Interpollutant and interstation correlational analysis among different pollutant records at different stations on Potomac river was conducted using cross spectral analysis. Prediction models were formed for each pollutant using parameter time series modeling.

CHAPTER II

SPECTRAL ANALYSIS

2.1 Introduction

Water quality is described by the quantitative determination of several physical, chemical and biological parameters such as temperature, dissolved oxygen etc. Several models that have been proposed to determine and/or forecast the water quality parameters can be broadly classified into two catagories. One class of models such as proposed by Streeter and Phelps [1], Dobbins [2] etc. are based on the exact knowledge of the causal water polluting phenomena. The other class of models is based on the statistical methods and requires only the general ideas of the causal phenomena. Time series analysis of water quality data falls in the second catatory of models. One of the principal advantages of this technique is that useful quantitative results can be obtained with only a general knowledge about the underlying causative phenomena and their relationship to water quality parameters. This technique can thus be used advantageously when the system under study is dominated by highly variable and complex processes.

The ultimate quality of water is the result of interaction of several physical, chemical and biological processes, either natural such as rainfall or man-made such as sewage disposal. Due to these interactions, water quality of a stream shows a marked variation over time and space.

Temperature has a profound effect on the water quality. Seasonal changes in stream temperature may be expected due to the seasonal changes

in the air temperature. Kothandaraman [32] obtained a model for stream temperatures by regressing it with air temperature. A model of the form

$$(T_w)_i = \frac{A_0}{2} + A_1 \frac{\cos 2\pi i}{N} + B_1 \frac{\sin 2\pi i}{N}$$

$$+ \beta_1 (R_a)_i + \beta_2 (R_a)_{i-1} + \beta_3 (R_a)_{i-2}$$

was obtained

where $(T_w)_i$ = predicted daily mean temperature on i^{th} day.

$(R_a)_i$ = Air temperature for i^{th} day

Waste heat from the several industrial plants or electric power plants also produces considerable effect on water thermal pollution. Variation in temperature causes variability in dissolved oxygen accordingly. A diurnal variation in dissolved oxygen may be expected due to photosynthesis. The stream flow rate and biochemical oxygen demand exert proportionate influence on variation of dissolved oxygen. As a result of these interactions, the variability of the pollution parameters may occur at constant time intervals as well as randomly. One important aspect of time series analysis is the spectral analysis, which is concerned with the resolving of the total variance of a time series record of a pollutant into its component parts at different frequencies.

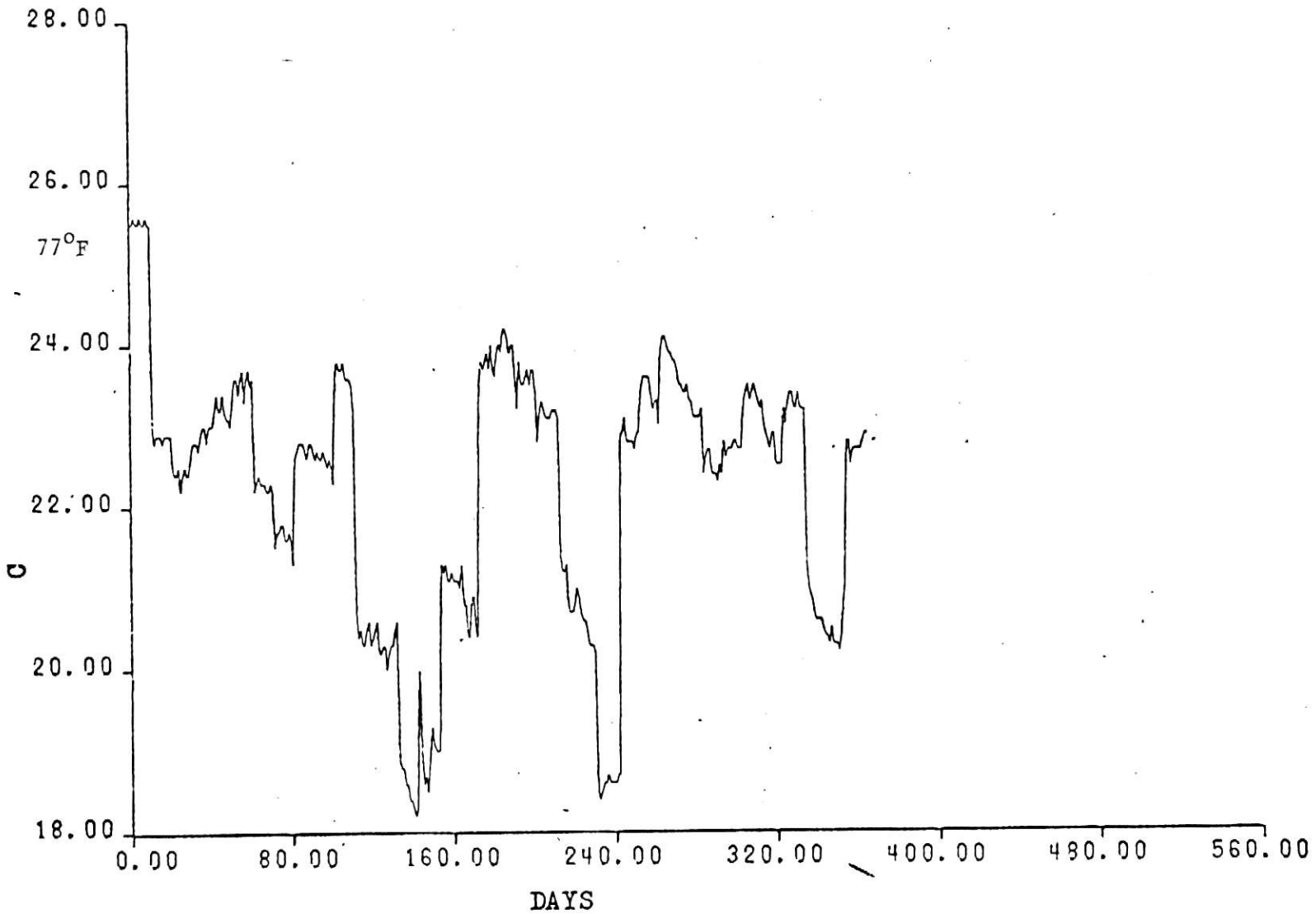

Fig. 2.1 Temperature record - Ontario river.

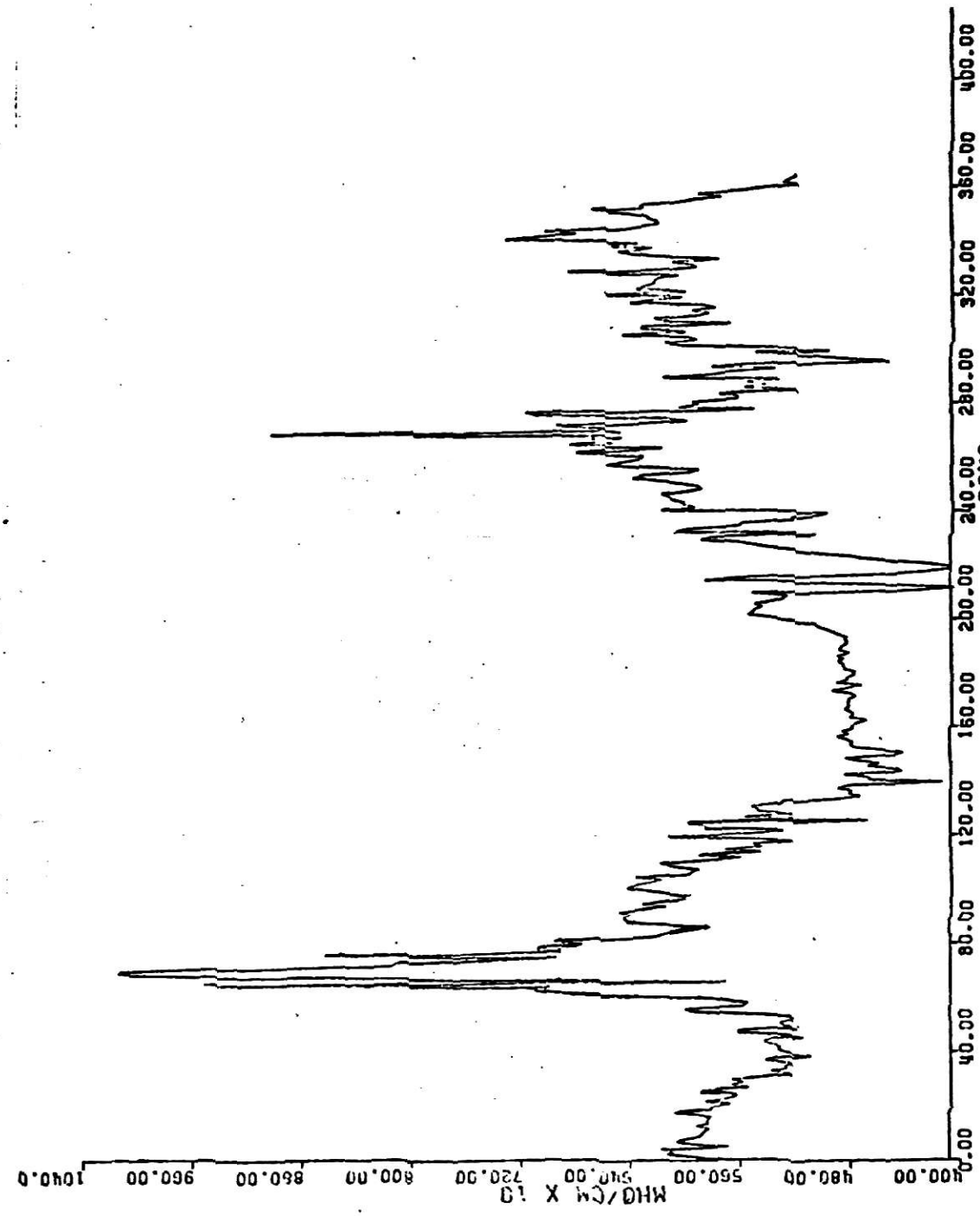

Fig. 2.2 Specific conductance record - Ontario river.

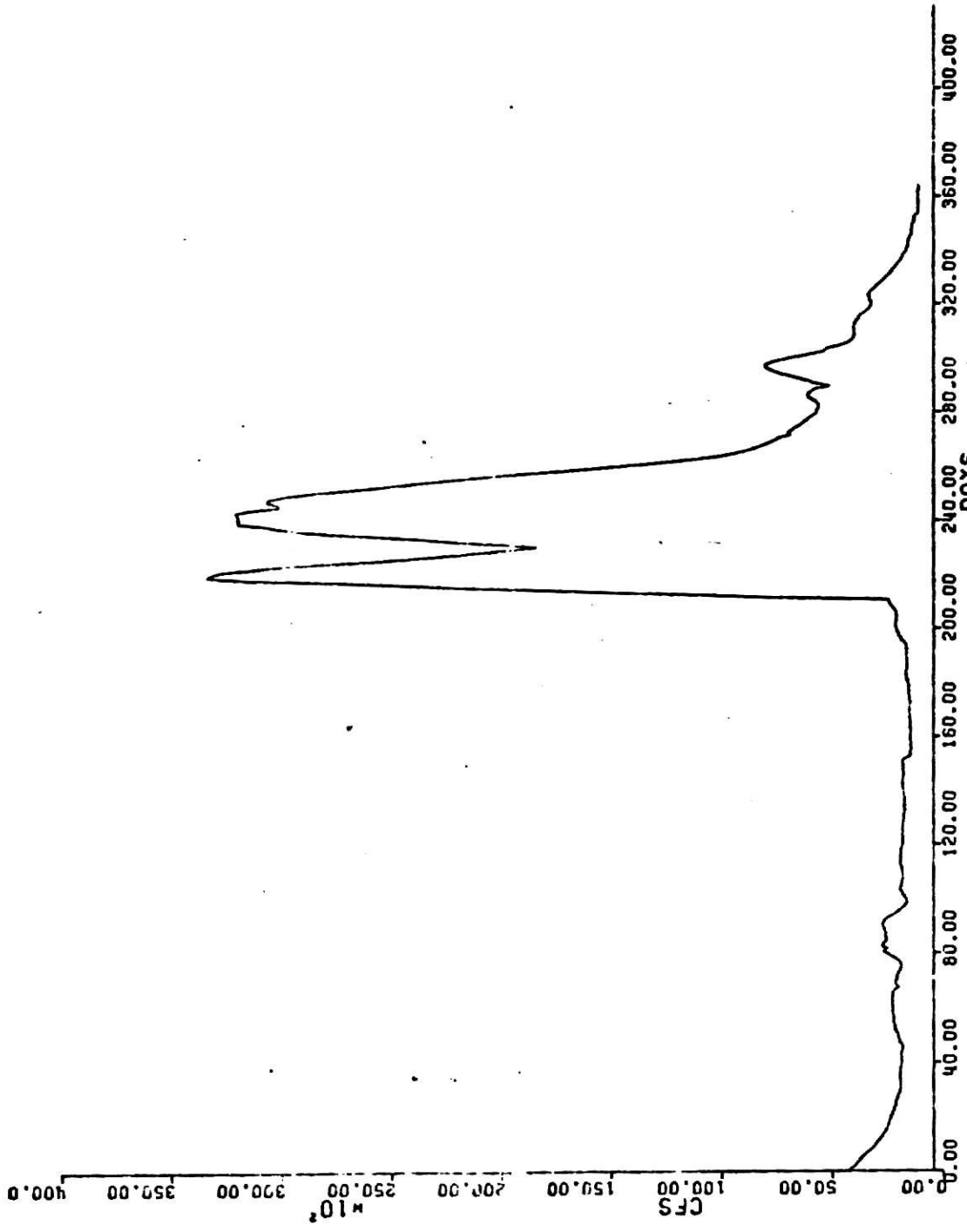

Fig. 2.3 Flow rate record - Ontario river.

In this chapter first the concept of spectral analysis will be introduced and then this technique will be used for analyzing the temperature, specific conductance and flow records for the Ontario River, Canada.

2.2 Data Acquisition

This set of data consists of the records of temperature, specific conductance and flow rate obtained by sampling at station MA10-12 on the Ontario River. Daily samples for a period of one year (from 9/1/1966 to 8/31/1967) were obtained. No missing observations were encountered in the whole record. For the purpose of this study, this data was obtained from a publication by DeMayo [16].

2.3 Analysis of Data

Figures 2.1, 2.2 and 2.3 show the plots of daily records of temperature, specific conductance and flow rate respectively. Temperature seems to fluctuate around a mean value of 22^{o} with maximum temperature of 25.6^{o} in the beginning of September and minimum temperature of 18^{o} in January. The presence of some cyclic fluctuations is also indicated though the entire variability does not seem to be due to only a single cyclic fluctuation. Some short period fluctuations may be superimposed on the large annual fluctuation.

The specific conductance plot (Fig. 2.2) also indicates the presence of some compound cyclic fluctuations. No definite information can be obtained by the visual inspection of flow rate data.

As the data indicate the presence of cyclic fluctuations, it would

be desirable to remove them from the record and study the residuals separately. Several filters are available which can remove a desired band of frequencies from the data [33]. One type of analysis that can be performed to extract information about a periodic component is the Fourier or harmonic analysis. Prior to performing spectral analysis, it is instructive to conduct harmonic analysis to get some information about the behaviour of the observed data.

2.4 Harmonic analysis:

Harmonic analysis is a very effecient tool for the purposes of analysis of deterministic data which indicates the presence of some cyclic terms. Though it should not be used to analyze stochastic time series yet it gives a useful amount of initial information to proceed with spectral analysis. The time series data may be analyzed for the harmonics of the dominant frequency; these harmonics may be removed and the residuals may then be analyzed by spectral analysis. One of the reasons prohibiting the use of Fourier analysis for stochastic time series is that it is based on the assumption of fixed amplitudes, frequencies and phases, whereas time series are subjected to random changes of frequencies, amplitudes and phases [33].

The Fourier representation of a time series X(t) is given by

$$X(t) = X_0 + \sum_{m=1}^{N/2} \{A_m \cos m\omega t + B_m \sin m\omega t\} \qquad (2.1)$$

where,

X_0 = mean of the data

N = total data points of the time series

$\omega = 2\pi f$, where $f = \frac{1}{N\Delta}$ is called the fundamental frequency of the data and it corresponds to a period equal to the length of the record.

m = mth integer multiple (harmonic) of fundamental frequency.

Δ = sampling interval

In practice, however, all the harmonics are not of interest and hence need not be calculated. If only M harmonics are to be calculated where $M \leq N/2$, then the Fourier representation is given by

$$X(t) = X_0 + \sum_{m=1}^{M} \{A_m \cos m\omega t + B_m \sin m\omega t\} + \text{Residual} \qquad (2.2)$$

The coefficients A_m and B_m can be calculated using the expressions,

$$A_m = \frac{2}{N} \sum_{r=1}^{N} x_r \cos \frac{2\pi rm}{N\Delta} \qquad (2.3)$$

$$B_m = \frac{2}{N} \sum_{r=1}^{N} x_r \sin \frac{2\pi rm}{N} \qquad (2.4)$$

A equivalent expression for a Fourier representation of a time series is

$$X(t) = R_0 + \sum_{m=1}^{N/2} R_m \cos (m\omega t - \phi_m) \qquad (2.5)$$

where R_m = amplitude of m^{th} harmonic

$$= \sqrt{A_m^2 + B_m^2}$$

Table 2.1 Harmonic Analysis - Temperature Ontario River

Mean = 22.2°C

Source	Amplitude (°C)	Phase (in degrees)	Percentage contribution to total variance
fundamental	0.41	-19.7	13.29
2nd harmonic	0.26	-84.2	5.25
3rd harmonic	0.37	5.38	10.68
4th harmonic	0.60	36.24	27.46
5th harmonic	0.30	17.77	7.21
7th harmonic	0.39	28.95	11.65

Table 2.2 Harmonic analysis - specific conductance Ontario river

Mean = 576.05 Mho/Cmx10^{-2}

Source	Amplitude (Mho/Cmx10^{-2})	Phase (in degrees)	% contribution to total variance
fundamental	28.56	5.97	21.28
2nd Harmonic	27.42	-5.73	19.60
3rd Harmonic	14.63	-83.79	5.59
4th Harmonic	21.65	-43.88	12.22
6th Harmonic	14.11	-83.85	5.19
7th Harmonic	10.81	-54.82	3.05
8th Harmonic	10.51	35.08	2.88

Table 2.3 Harmonic analysis - flow rate Ontario river

Mean = 5895.58 cfs

Source	Amplitude (cfs)	Phase (in degrees)	% contribution to total variance
Mean	5895.58	0.0	33.55
fundamental	3544.91	65.25	36.51
2nd Harmonic	2967.2876	-56.33	25.58
3rd Harmonic	2189.25	-4.63	13.93
4th Harmonic	1657.12	42.73	7.98
5th Harmonic	1410.37	-84.96	5.78
6th Harmonic	802.20	-18.63	1.87

and ϕ_m = phase of m^{th} harmonic

$$= \arctan \frac{B_m}{A_m} .$$

The expression for obtaining the variance contributed by each harmonic is

$$\left.\begin{aligned} \sigma_m^2 &= \frac{R_m^2}{2} \qquad m < N/2 \\ \sigma_m^2 &= R_m^2 \qquad m = N/2 \end{aligned}\right\} \tag{2.6}$$

and if the total variance is available, the percentage variance due to each harmonic may be obtained.

It may be noted here that the values of the estimators of A_m and B_m obtained by least squares estimation procedure are quite efficient and this method was used in this study to remove these harmonics from the data. But for purposes of identifying the important harmonics, expressions (2.3) and (2.4) were used for estimating A_m and B_m.

In the harmonic analysis of the Ontario River data, it was found that the contribution to the mean square [33] by the mean alone accounts for about 95-98% of the total mean square value. In order to study the importance of individual harmonics more effectively, it was decided to use the contribution of each harmonic towards total variance instead of total mean square value. Variance (σ^2) is the mean square value of the signal (x(t)) about the mean (x_0).

i.e. $$\sigma^2 = \frac{1}{N} \sum_{t=1}^{N} (x(t) - x_0)^2$$

Tables (2.1), (2.2) and (2.3) show the results of harmonic analysis for the temperature, specific conductance and flow rate data of Ontario River. It is seen that for temperature and specific conductance, the mean accounts for 95-99% of total mean square value. The remainder of the variance is accounted for by the first eight harmonics. The first harmonic has a period of 365 days, second harmonic has 182 days and so on. These harmonics may be expected since the temperature is known to have annual cyclic fluctuations. Harmonic Analysis for flow data indicates that mean accounts for about 33.5% of total mean square value. The remainder of the variance is distributed in the first six harmonics. About 30% of the residual variance is due to the fundamental frequency of 1 cycle/ 365 days. This indicates that flow has a strong tendency to follow annual cyclic fluctuations. This is emphasized by the gradual reducing effects of subsequent harmonics.

After having identified the harmonics in the data, these are then removed using a least squares procedure and spectral analysis is carried out on the residuals. Spectral analysis would point out any additional cyclic variations or random variations in the data.

2.5 Spectral Analysis:

Spectral analysis is a useful method for the analysis of a time series. Water quality data collected sequentially over a period of time constitutes a time series.

The theory of spectral analysis has been very well explained by several authors [33], [34], [35], [36] and will not be covered in details here. However, the basic concepts and the terms which shall be used in the explanation of the data are described in the report.

One of the properties of the spectral analysis is that it can be applied only to stationary time series. It means that the statistical properties of the time series are unaffected by the change of time origin. More properties of the stationary time series will be examined in the later part of this section. However, nonstationarity of the time series does not present any special problems in its analysis as it can be reduced easily to a stationary series using filters.

The computation of the individual power spectrum involves several steps. First step involves the estimation of autocovariance and auto-correlation functions. These functions determine how one observation of the data is related to any other observation of the data. The covariance between two data points separated by k intervals of time is called the <u>auto-covariance</u> at lag k.

$$c_k = \text{cov}\,[x(t), x(t+k)]$$

The stationarity property of the time series implies that the auto-covariance is a function of lag k only and does not depend upon the origin of time

$$c_k = \text{cov}\,[x(t), x(t+u)]$$

$$= \text{cov}\,[x(t), x(t-u)]$$

It has been shown [33] that the most satisfactory estimate of auto-covariance at k^{th} lag is given by

$$c_k = \frac{1}{N} \sum_{k=1}^{N-k} (x_t - \bar{x})(x_{t+k} - \bar{x}) \qquad k = 1,2, \ldots, M \qquad (2.7)$$

where, M = maximum no. of lags

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N}$$

In practice, instead of auto-covariance function, <u>autocorrelation function</u> is generally used which is given by

$$\rho_k = \frac{c_k}{c_0} \qquad (2.8)$$

where c_0 = auto-covariance at zero lag

= variance of the whole record.

One useful property arising out of the stationarity assumption is that the auto-covariance function and the autocorrelation function are both even functions of lag k implying

$$\rho_k = \rho_{-k}$$

and hence these functions need be calculated only for positive lags.

One disadvantage of the auto-correlation function is that it considers only the amplitude of fluctuation and disregards any phase difference between them. Thus, all the periodic functions having the same harmonic amplitudes but differing in initial phase angles will have the same auto-correlation function.

An equivalent description of a stationary stochastic process is provided by the Fourier transform of the auto-covariance function which is called a power spectrum. The theoretical power spectrum of a stationary time series is defined as

$$\Gamma_{xx}(f) = \int_{-\infty}^{\infty} \gamma_{xx}(k)\ e^{-j2\pi fk}\ dk \tag{2.9}$$

where $\gamma_{xx}(k)$ is the theoretical auto-covariance at lag k.

Inverse transform of (2.9) provides

$$\gamma_{xx}(k) = \int_{-\infty}^{\infty} \Gamma_{xx}(f)\ e^{j2\pi fk}\ df \tag{2.10}$$

In particular for k = 0,

$$\gamma_{xx}(0) = \int_{-\infty}^{\infty} \Gamma_{xx}(f)\ df \tag{2.11}$$

hence $\Gamma_{xx}(f)$ shows the distribution of the variance over frequency.
The computational formula for the estimation of raw spectrum for a discrete case is given by

$$S(f) = 2\Delta\{c(0) + 2 \sum_{k=1}^{M-1} c(k) \cos 2\pi fk\Delta\},$$

$$0 \leq f \leq \frac{1}{2\Delta t} \tag{2.12}$$

where Δt = sampling interval

$f = \frac{1}{2\Delta t}$ is the Nyquist's frequency.

c(k) = estimation of theoretical auto-covariance at lag k.

If the auto-correlation function is used instead of the auto-covariance function, the corresponding spectral estimate is called the spectral density function and is defined as

$$R(f) = 2\Delta\{1 + 2 \sum_{k=1}^{M-1} \rho_k \cos 2\pi fk\} \qquad 0 \leq f \leq \frac{1}{2\Delta} \tag{2.13}$$

Although, the auto-correlation (auto-covariance) function gives useful information about the time series, yet in actual practice its fourier transform; the spectral density function is generally preferred. One of the main reasons for this is that the neighbouring values of the auto-correlations (auto-covariance) have strong correlation among themselves if the correlation in the original series is fairly strong. This may lead to misinterpretation of these plots. On the other hand the estimates of spectrum at neighbouring frequencies are approximately independent [33]. Secondly, the spectral function is in the frequency domain and hence easier to interpret than the auto-correlation function which is in the time domain.

The sample spectrum as given by expressions (2.12) or (2.13) is not a consistent estimator of the true spectrum in the sense that its distribution does not tend to cluster more closely about the true spectrum as the sample size increases. Moreover, in practice it is not possible to obtain a infinite sample size. To obtain spectral estimates with lower variance, use of spectral windows is made. The purpose of

smoothing by the spectral windows (or lag windows) is to modify the values of c(k) differently for different lags. The windows that are generally used are

(1) Bartlett window

(2) Parzen window

(3) Tukey Hanning window

(4) Hamming window

The requirements of a spectral window are two fold

(i) There should be negligible leakage from one frequency band to another to reduce the possibility of distortion of the spectrum from remote frequencies.

(ii) The greatest weight should be given to the estimate at the principal frequency.

Figure 2.4 shows typical behaviour of Parzen and Tukey Hanning windows. To meet the first requirement mentioned above, the side lobes should be as small as possible and for the second requirement, main lobe should be as high as possible. It is seen the the main lobe can be made high by increasing the number of lags. But this introduces another problem that the variance of spectral estimates increases as the number of lags increases. Table 2.4 shows the properties of some of the lag windows [33].

A subjective judgement is often made for selection of the M value depending upon the resolution needed and the variance arising out of it. A process known as window closing can often be used for this purpose. It consists of using a low M and increasing it in steps till a value is

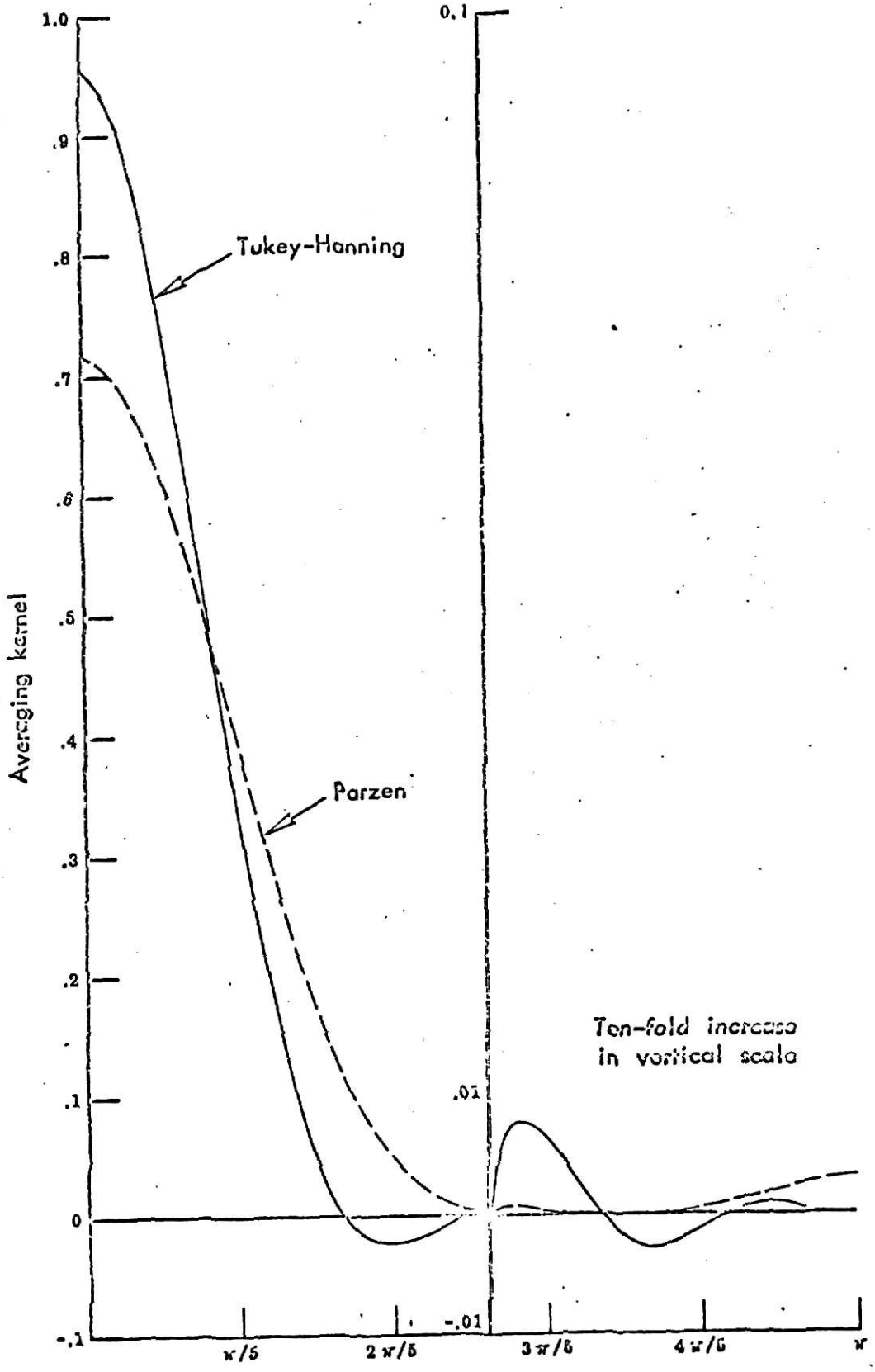

Fig. 2.4 Tukey - Hanning and Parzen windows (35).

Table 2.4 Some lag windows and their properties

Description	Lag window	Variance	Bandwidth
Bartlett	$1 - \frac{\lvert k\rvert}{M}$, $\lvert k\rvert \leq M$ 0, $\lvert k\rvert > M$	$0.667\,\frac{M}{N}$	$\frac{1.5}{M}$
Parzen	$1 - 6\,(\frac{k}{M})^2 + 6\,(\frac{\lvert k\rvert}{M})^3$, $\lvert k\rvert \leq \frac{M}{2}$ $2(1 - \frac{\lvert k\rvert}{M})^3$, $\frac{M}{2} < \lvert k\rvert \leq M$ 0 $\lvert k\rvert > M$	$0.537\,\frac{M}{N}$	$\frac{1.86}{M}$
Tukey & Hamming	$\frac{1}{2}\,(1 + \cos\frac{\pi k}{M})$, $\lvert k\rvert \leq M$ 0, $k > M$	$\frac{0.75M}{N}$	$\frac{1.333}{M}$

reached beyond which no more significant detail is observed in the spectrum. [33]

In this study the Tukey-Hanning lag window was used. Several values of M were tried such as 45, 60, 75, 90 and 120 for a record of 365 data points. It was found that 75 lags provided sufficient resolution of frequencies.

Generally, the spectral estimates are plotted on a logarithmic scale so that the variation in the spectrum can be accomodated. In this report also, the natural logarithmic scale was used. It has another advantage in that the confidence interval for the logarithm of the spectral estimate can be given by two horizontal lines for all frequencies. It can be shown that $\nu\bar{S}(f)/\Gamma_{xx}(f)$ is distributed according to chi square distribution with ν degrees of freedom,

where $\nu = \frac{2N}{M} \cdot b_1$

b_1 = standardized bandwidth (1.33 for Tucky Hanning window).

$\bar{S}(f)$ = smoothed spectral estimate.

A confidence interval for $\Gamma_{xx}(f)$ can be given by

$$\ell n\ \bar{S}(f) + \ell n \frac{\nu}{\chi^2_\nu\ (1 - \frac{\alpha}{2})}\ ,\ \ell n\ \bar{S}(f) + \ell n \frac{\nu}{\chi^2_\nu\ (\frac{\alpha}{2})} \qquad (2.14)$$

where α = confidence level.

Some other considerations which should be borne in mind while doing spectral analysis are as below:

(1) The choice of a suitable sampling interval is very important. As shown earlier in expression (2.12), the Nyquist's frequency is given by $\frac{1}{2\Delta t}$ where Δt is the sampling interval. Hence the longest frequency that can be analyzed by this procedure is corresponding to 1 cycle/2 time units. The sampling interval should be such that the highest frequency expected in the process is equal to or less than Nyquist's frequency. As a thumb of rule, the sampling interval should be about one third of the lowest expected period [37].

(2) The resolution power and the longest periodicity indicated by the spectrum are a function of lags. Both increase as the number of lags increase. But the increase in lags also increases the variance thus reducing the precision of estimation. Thus a balance between precision of estimation and resolving power is needed. Generally, the number of lags are taken as 0.1N to 0.4N.

(3) Due to a finite record length, appearance of a negative spectral estimate is possible with a Tukey-Hanning lag window. This should be interpreted as a very small power.

(4) Any linear trend in the record or any periodic components which are too long to be detected by this record length appear as a zero frequency spectral estimate.

2.6 Prewhitening: It was seen in this study that in most cases, there was a strong concentration of variance at low frequencies. As discussed earlier, this may be due to the presence of a trend or long cyclic fluctuations.

It may occur that low frequency bands distort the spectrum at high frequency bands due to leakage. To analyze the data more effectively a high frequency bands, prewhitening of the data was carried out. Prewhitening consists of filtering the original data to remove low frequency components. In this analysis simple differencing filter was used.

$$y_t = x_t - \alpha x_{t-1}, \quad \text{where } 0 < \alpha < 1 \tag{2.15}$$

for this analysis $\alpha = 0.99$.

The spectral estimate of the filtered data and that of original data are related as [35]

$$\bar{s}_y(f) = (1 - 2\alpha \cos 2\pi f + \alpha^2)\, \bar{s}_x(f) \tag{2.16}$$

This relationship can also be used to recolor the estimate of $\bar{s}_x(f)$. An estimate of the frequencies which this filter attenuates is given by

$$f = \frac{1}{2\pi} \cos^{-1} (\alpha/2) \tag{2.17}$$

This simple differencing procedure can be extended to nth order differences. By a judicious choice of α and n it is possible to flatten the spectrum in the low frequency range i.e. reduce the spectral power of all frequencies less than an arbitrary threshold and essentially not reduce the spectral power of those frequencies above the threshold frequency. Another method of removing trend from the data is to fit a least squares polynomial model. This approach was carried out for the development of the predictive model.

In section 2.4, harmonic analysis was done for the temperature, the specific conductance and the flow rate records of Ontario River. After some general information about the behaviour of the pollutants has been obtained by harmonic analysis, we proceed to spectral analysis for a more accurate investigation.

2.7 Spectral analysis of Ontario river data:

(a) Temperature: As discussed earlier, 75 lags were used in all calculations. Figure 2.5 shows the auto-correlation plot of temperature data. High positive correlation exists for all lags upto 75. Such high autocorrelation for larger lags may be taken as an indication of non-stationarity in the time series. The spectral plot of the raw data indicated a high variance at zero frequency. This indicates that either a trend or a long range frequency is present in the data. To study the series more effectively, prewhitening was carried out. A simple differencing filter was used to remove the low frequencies.

As given by (2.17), it attenuates the frequencies upto

$$f = \frac{1}{2\pi} \cos^{-1} (0.495) \text{ cycles/day}$$

$$= 0.09 \text{ cycles/days}$$

Figure 2.6 shows the prewhitened spectrum for the same data. Two points may be noted about this spectrum. Firstly, it effectively removes any low range frequency from the series and secondly, no high frequency fluctuation seems to be present. Recolored spectra was obtained from this prewhitened spectra and is shown in Figure 2.7. This shows a high variance at low range frequencies. It is known that the temperature of a river exhibits an annual frequency. Hence it may be safe to assume that a position of large concentration of variance at zero frequency is due to annual cyclic fluctuation. In order to confirm this, more data is needed. As discussed earlier in section 2.4, harmonic analysis

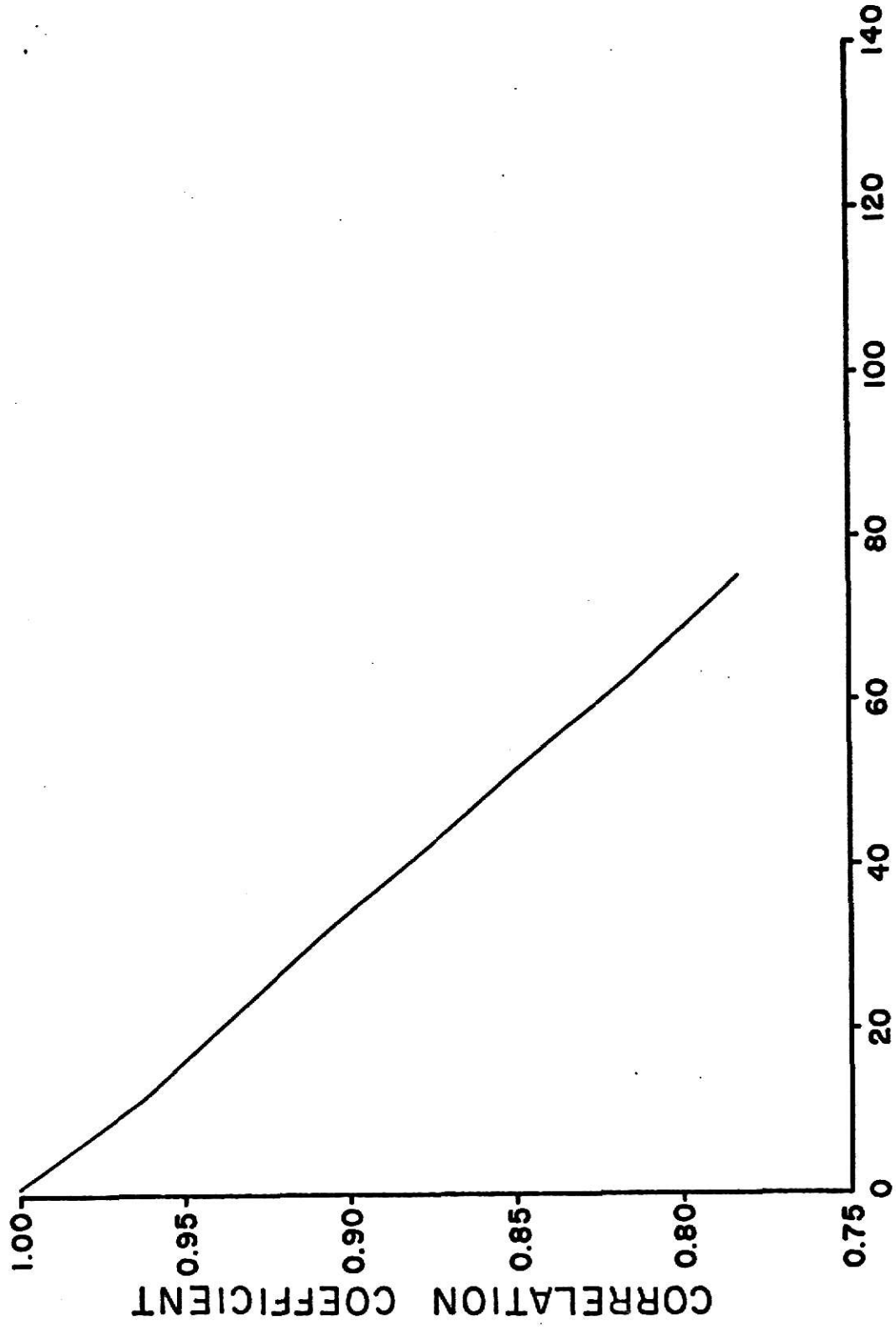

Fig. 2.5 Autocorrelation of temperature data - Ontario river.

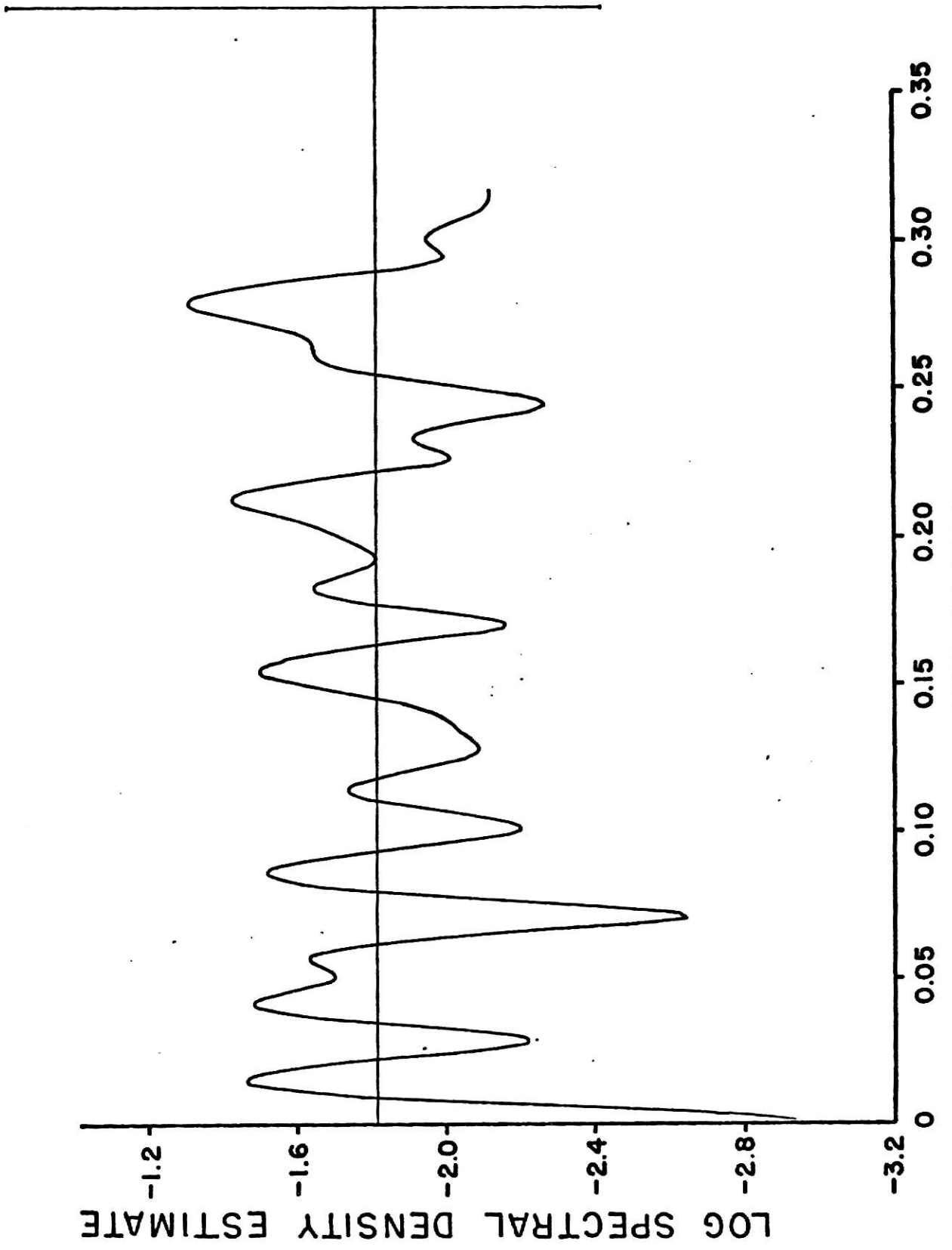

Fig. 2.6 Spectral density estimate for temperature - Ontario river. (prewhitened)

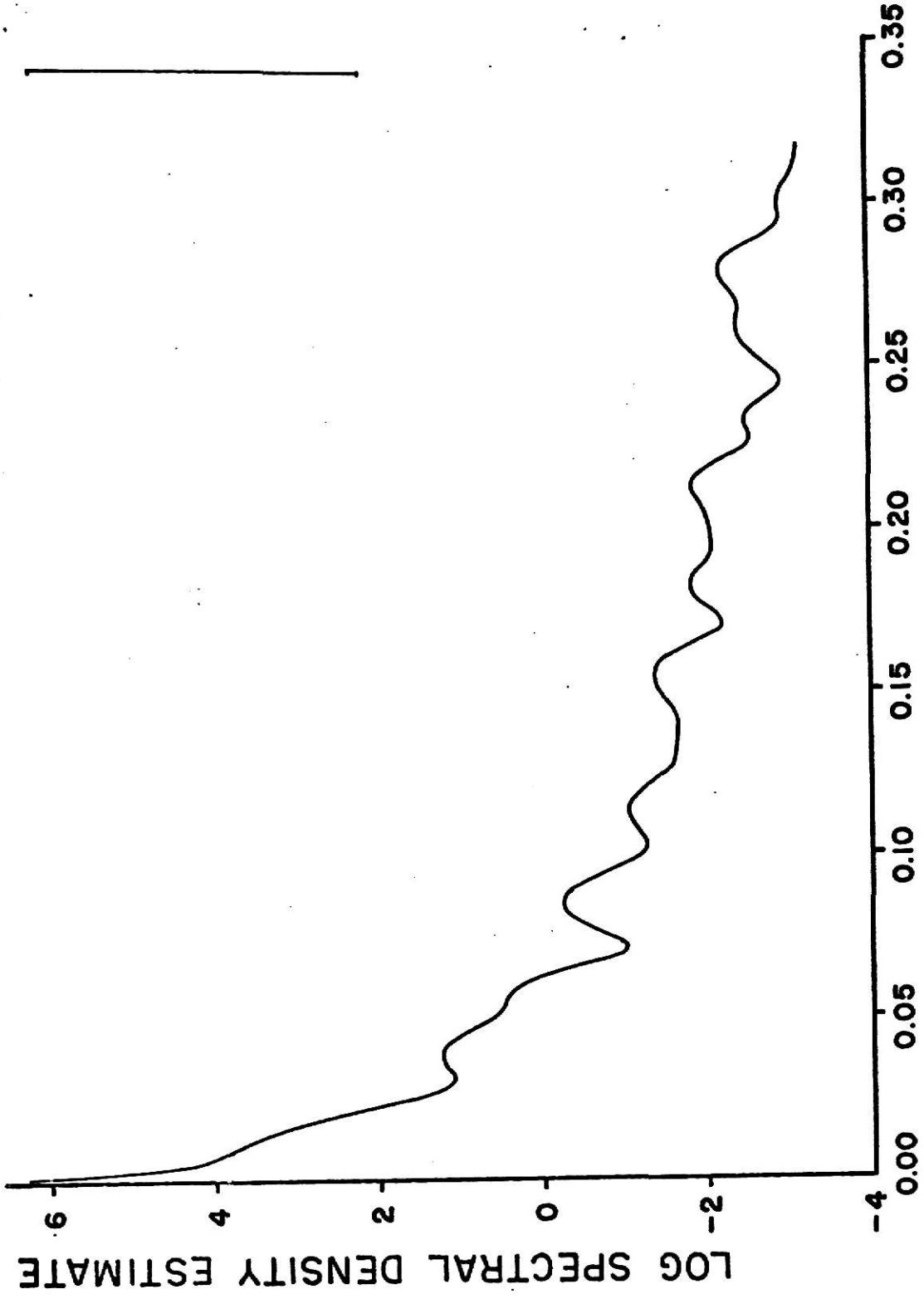

Fig. 2.7 Spectral density estimate for temperature - Ontario river. (recolored)

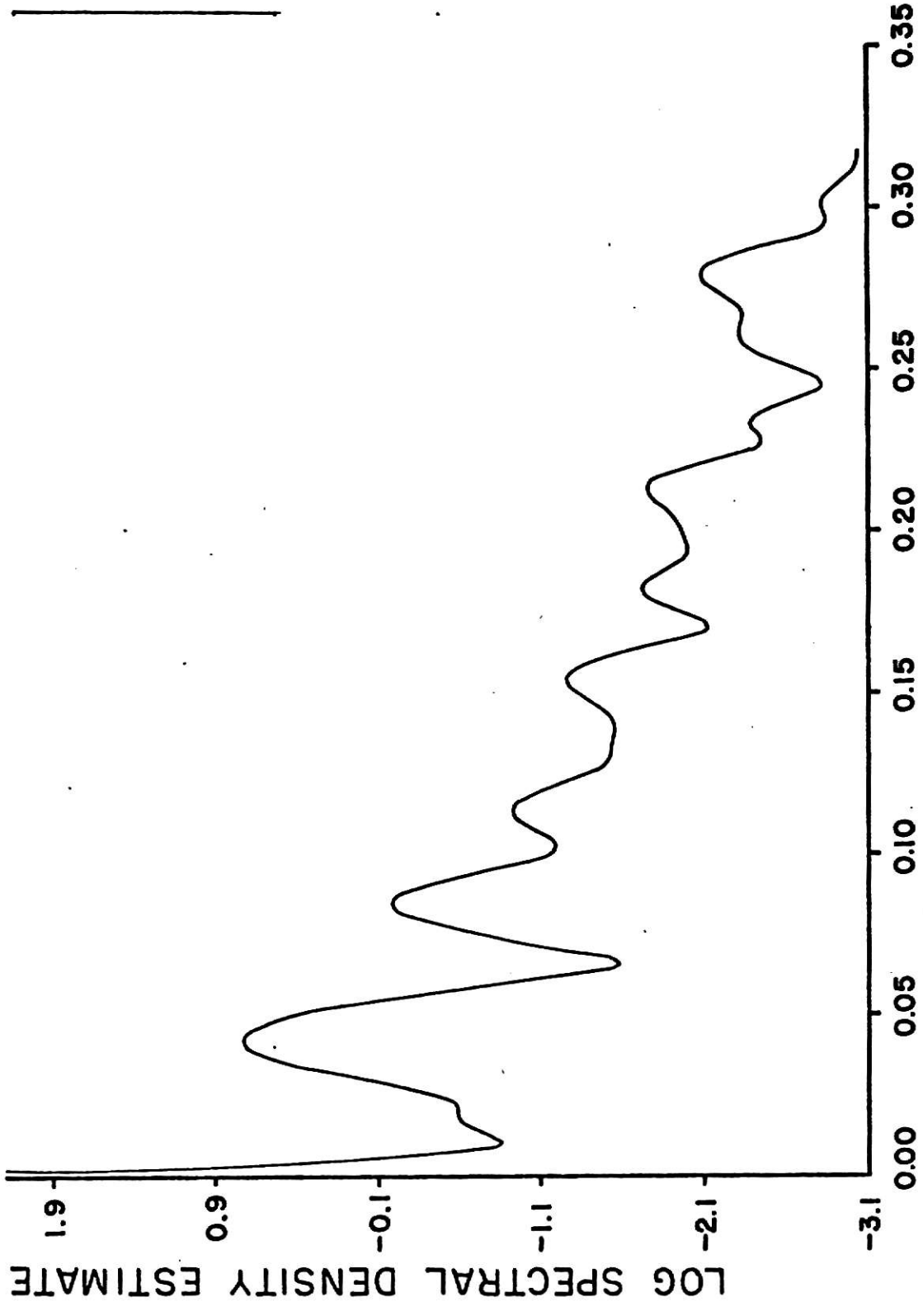

Fig. 2.8 Spectral density estimate for temperature - Ontario river. (residuals)

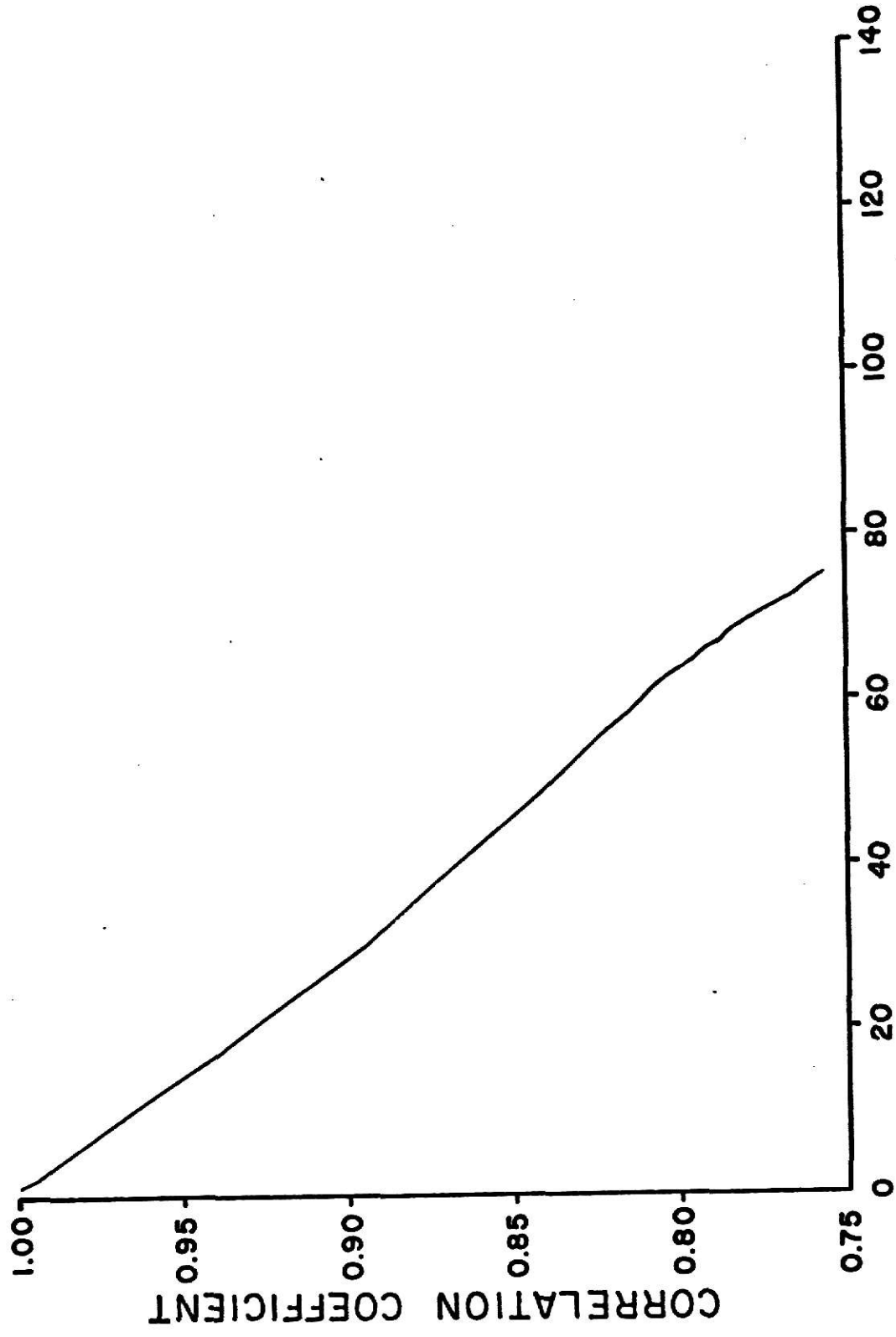

Fig. 2.9 Autocorrelation of specific conductance - Ontario river.

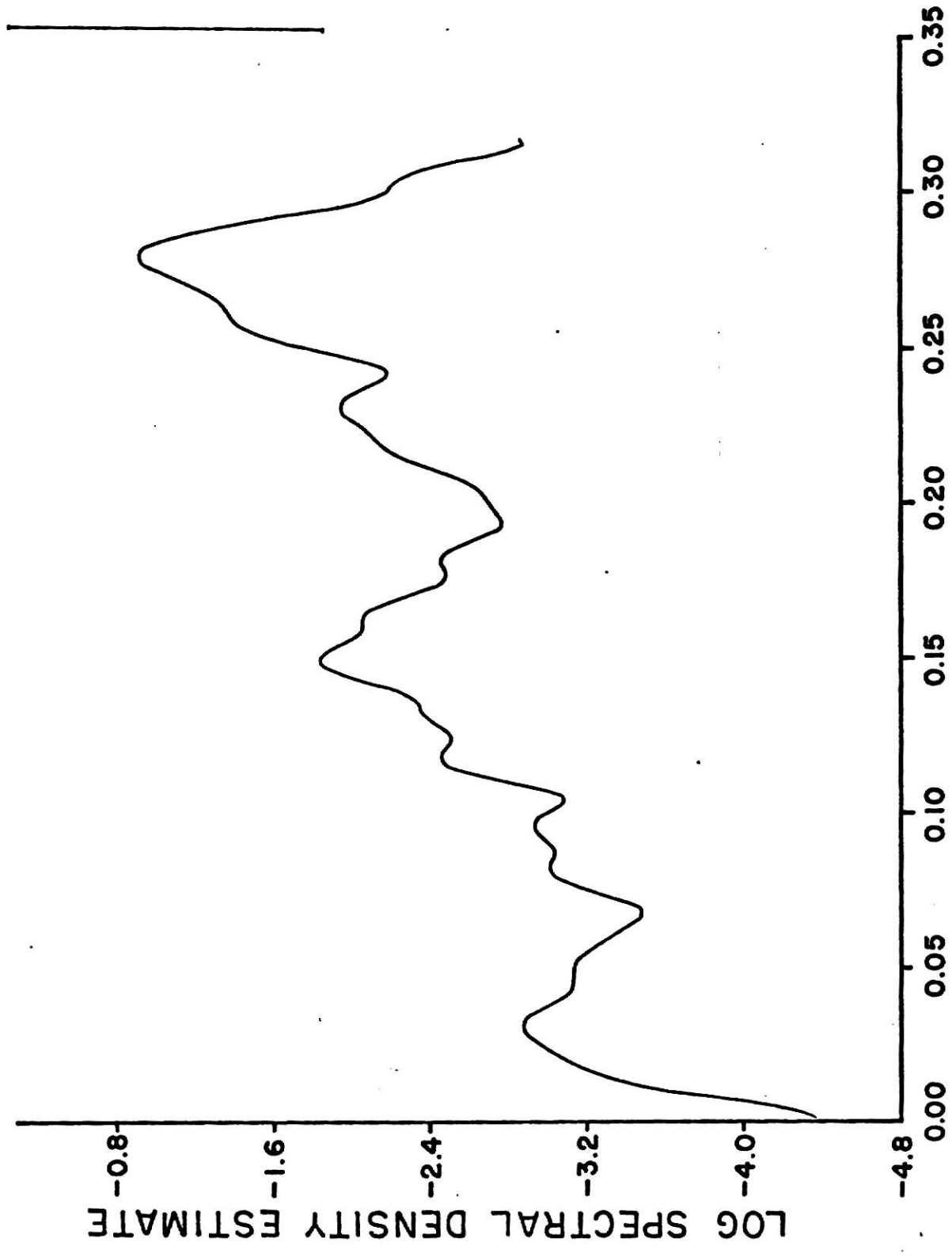

Fig. 2.10 Spectral density estimate for specific conductance - (prewhitened)

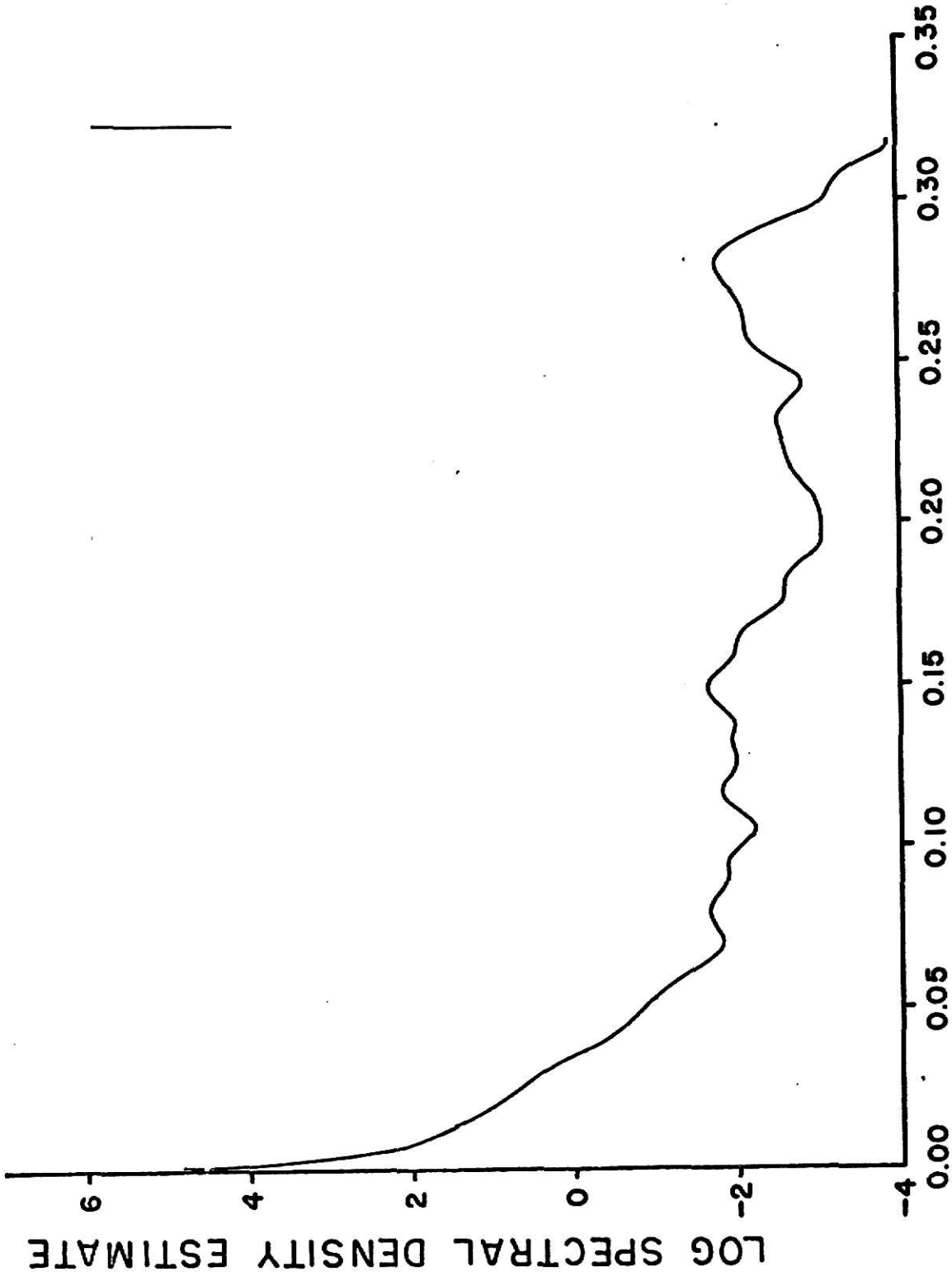

Fig. 2.11 Spectral density estimate for specific conductance- (recolored)

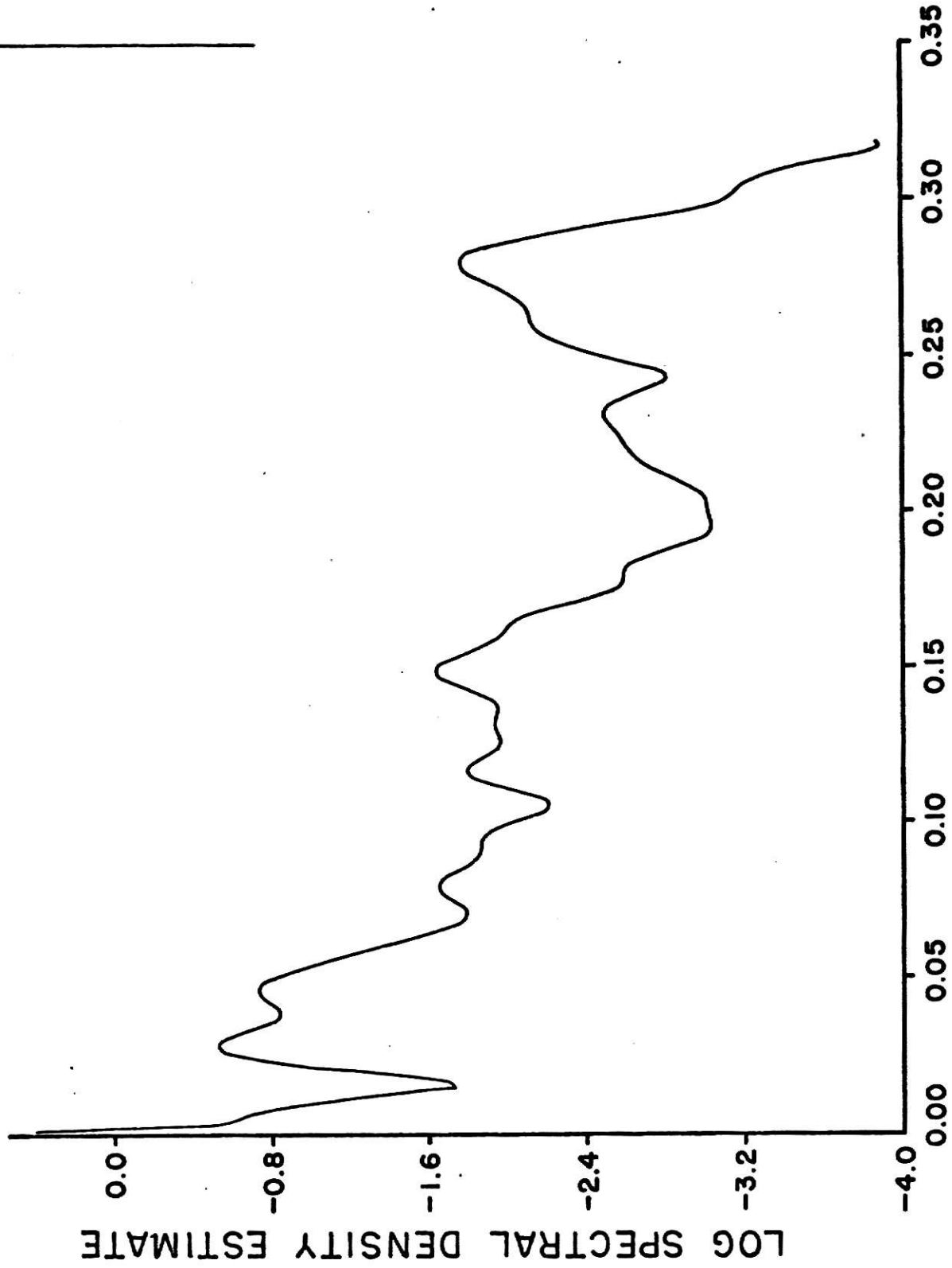

Fig. 2.12 Spectral density estimate for specific conductance - (residuals)

indicates the presence of semi-annual, 120 days, 90 days and 30 days period terms. All these terms were added in a regression model and the residual thus obtained were subjected spectral analysis again. Figure 2.8 shows the spectral plot of the residuals. It is observed that the variance at low frequencies has been reduced considerably as compared to the spectrum of the raw data shown in Figure 2.7. No other peak is evident in the spectrum. The confidence interval was drawn and it was found that almost all the points on the plot were within the confidence level.

b) Specific Conductance:

Auto-correlation plot of raw specific conductance data is shown in Figure 2.9. Again a very high positive correlation is observed for all lags. High positive auto-correlation may either indicate that for all lags up to 75, a high (low) value of specific conductance tends to follow a high (low) value or it may again be taken a tendency of series to be non-stationary. The spectral plot of the raw data consisted of high power at zero frequency, hence it was decided to prewhiten the spectral estimates. In this case also a simple difference filter was used. Figure 2.10 shows the prewhitened spectra of the raw data. The plot does not indicate the presence of any dominant high frequency component. The corresponding recolored spectrum is shown in Figure 2.11. All the variance is concentrated at the low frequency component indicating the presence of either a trend or a long periodic fluctuation. The trend can be explained being due to growth in the concentration of salts in the water. An annual cycle may be correlated with the annual variation

in flow rate. The dominant harmonics as indicated by harmonic analysis were added in a regression model and removed from the series. Figure 2.12 shows the spectral plot after all the harmonics have been removed. Comparing this plot with Figure 2.11 reveals that the variance at low frequencies has been considerably reduced and no more cyclic fluctuation seems to be present. A confidence bend for 95% confidence level was drawn and it was found that most of the fluctuations in the spectrum were within this band.

(c) Flow: Auto-correlation plot of raw flow data is shown in Figure 2.13. The auto-correlation function remains positive for all lags, though it decreases gradually. It implies that a high (low) flow rate tends to be followed by another high (low) flow rate even upto 75 days lag. As before, the spectral plot of raw data indicated need for prewhitening. Figure 2.14 shows the prewhitened spectra for the same data. This shows the presence of some dominant low frequency cyclic fluctuations. Inspite of attenuation at low frequencies, a peak corresponding to 120 days period is evident in the spectrum. The corresponding recolored spectrum is shown in Figure 2.15. The recolor spectrum shows large variance associated with low frequencies. This may be expected as the flow is known to have an annual cycle and other seasonal variations. Harmonic analysis (Table 2.3) had indicated the dominant effect of fundamental frequency, second harmonic, third harmonics corresponding to 365 days, 182 days, 120 days period. The first six harmonics were added to a regression model and spectral analysis was carried out on the residuals. The corresponding plot is shown in Figure 2.16. This plot

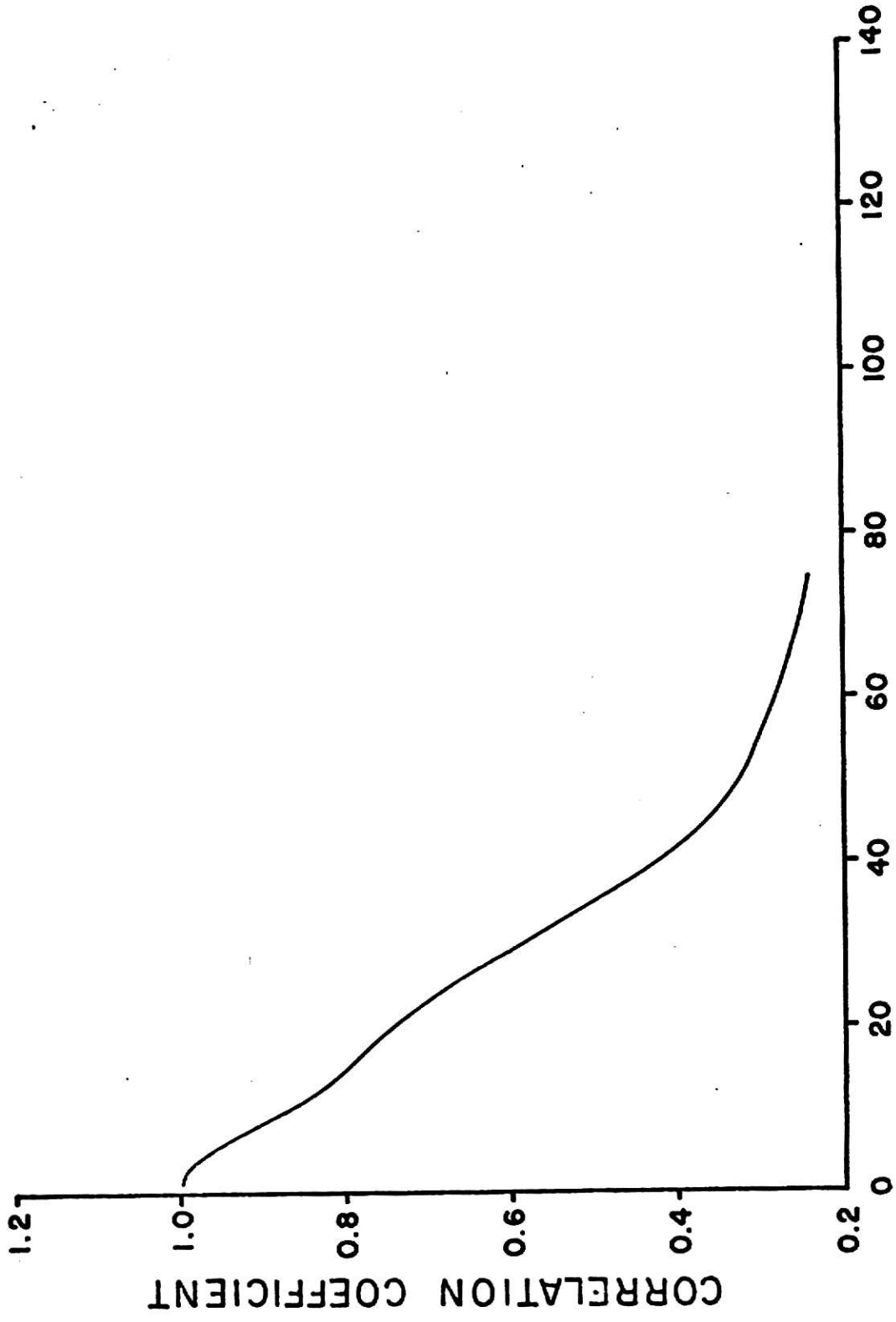

Fig. 2.13 Autocorrelation of flow rate - Ontario river.

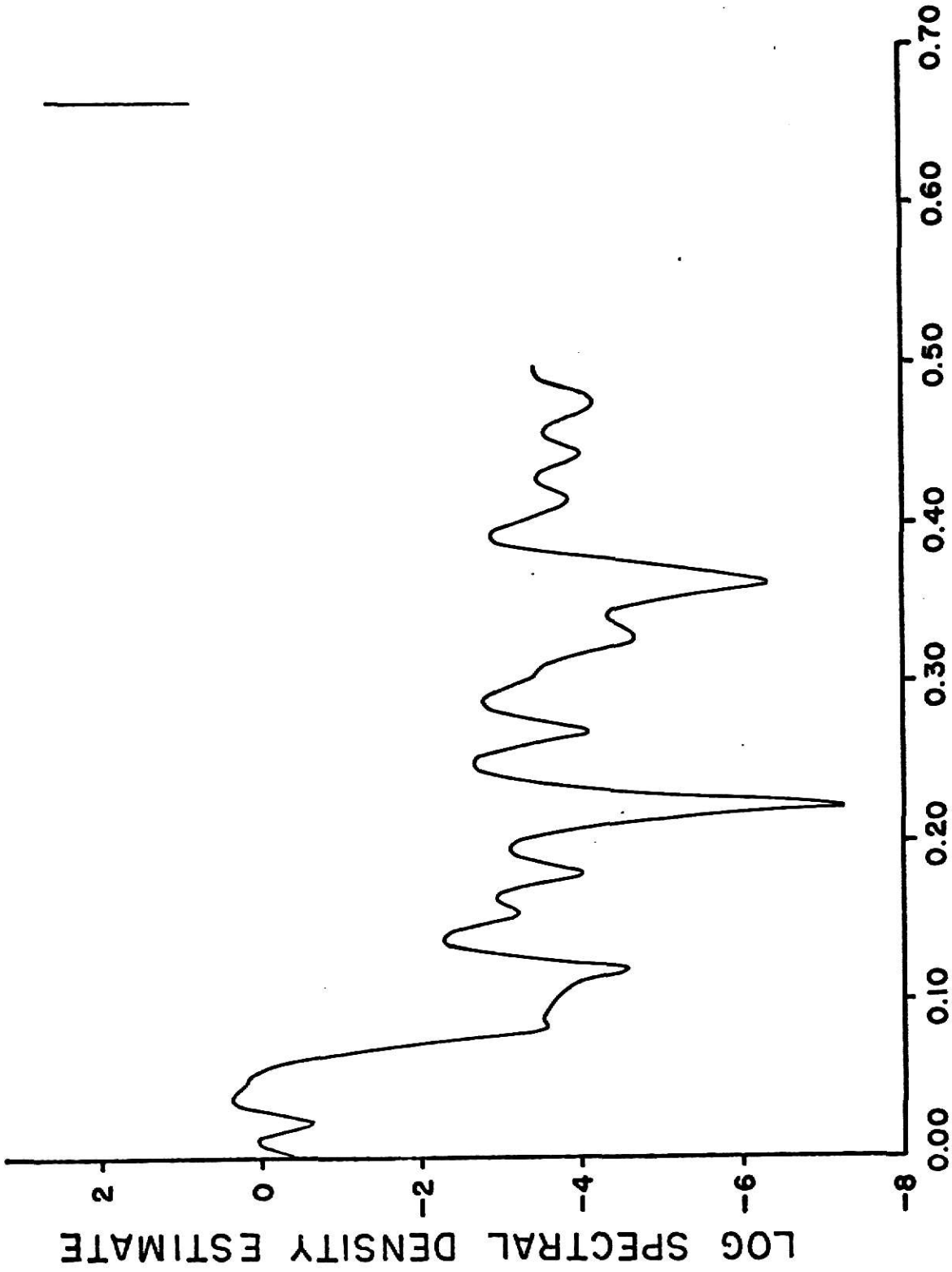

Fig. 2.14 Spectral density estimate for flow rate - Ontario river. (prewhitened)

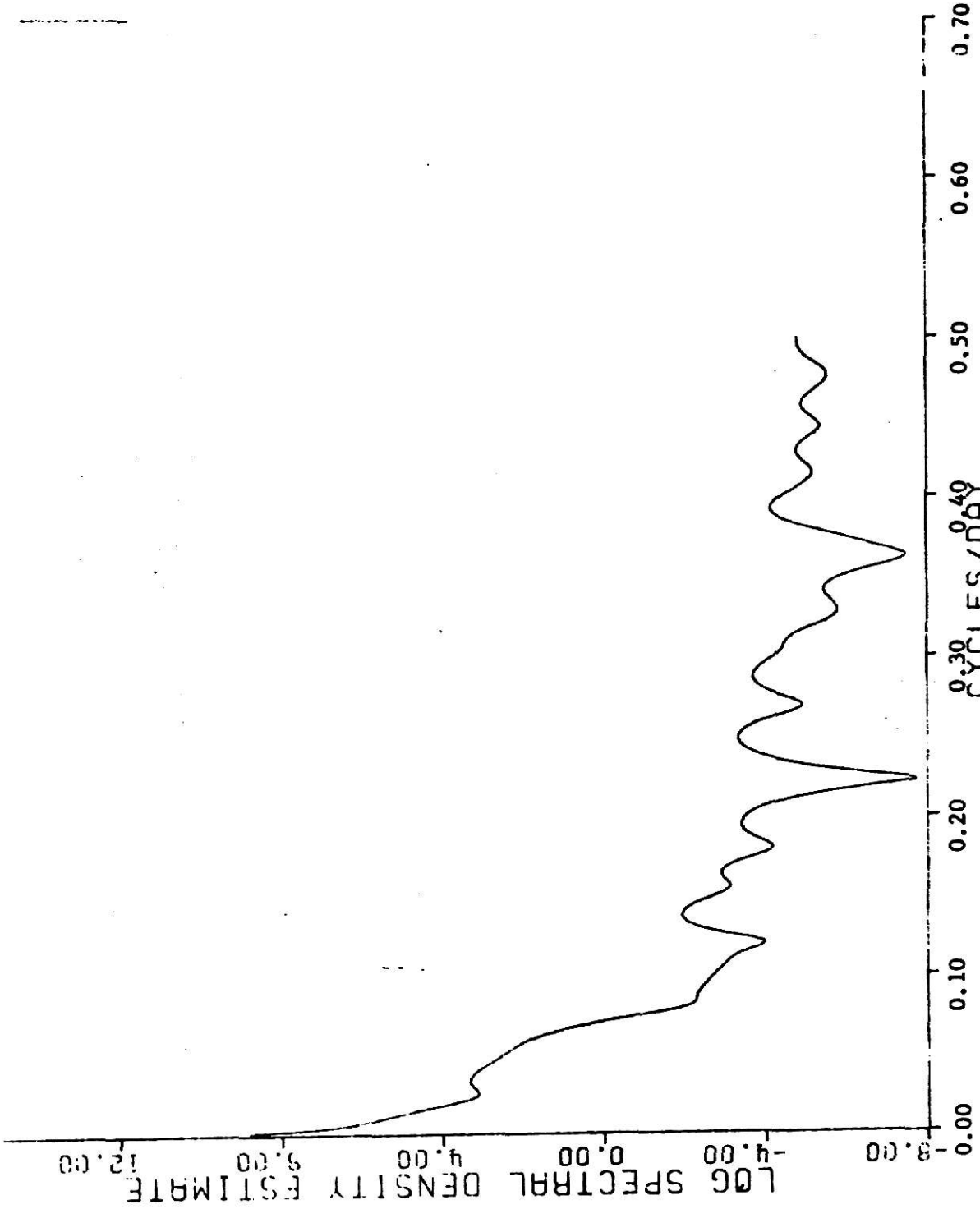

Fig. 2.15 Spectral density estimate for flow rate - Ontario river. (recolored)

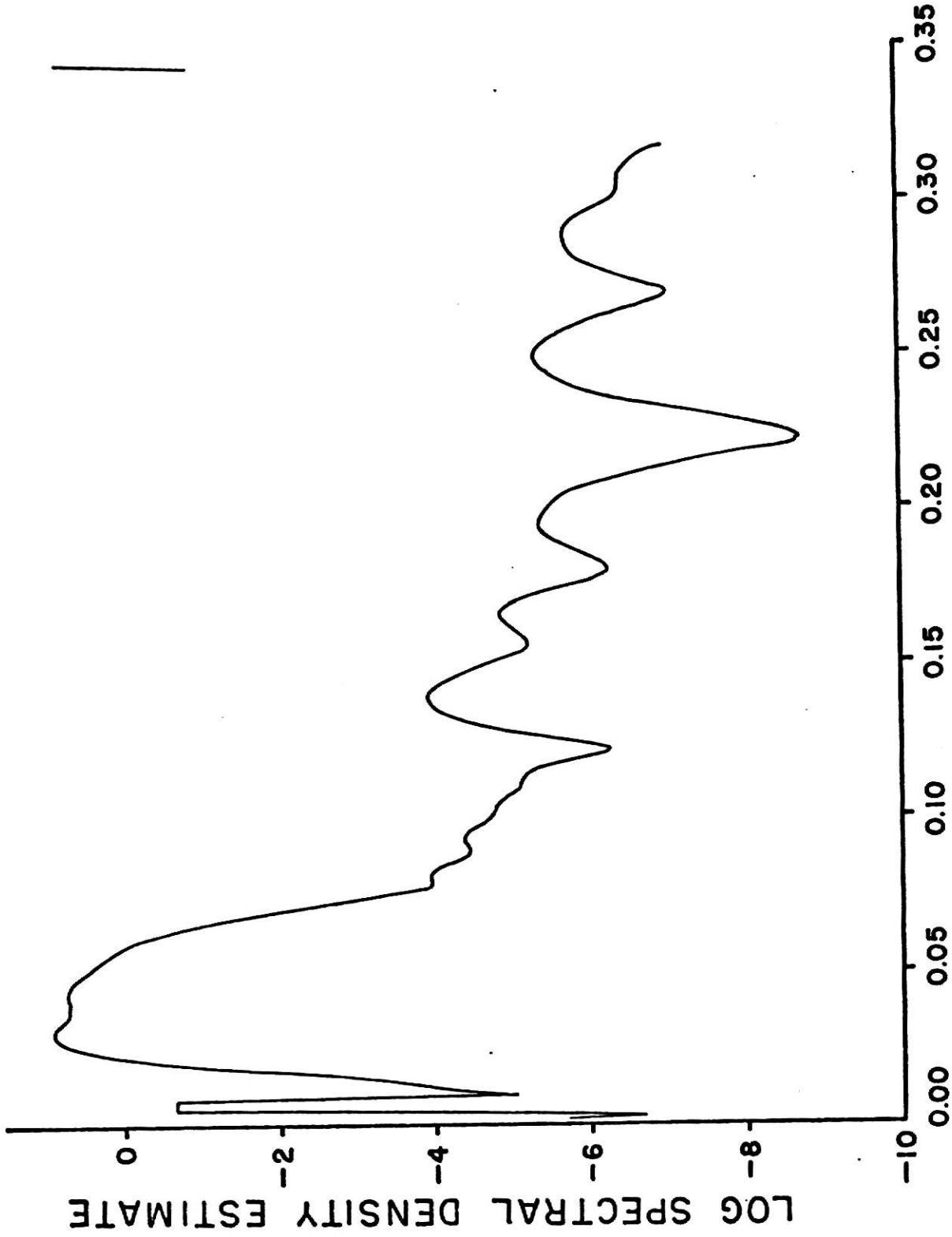

Fig. 2.16 Spectral density estimate for flow rate - Ontario river. (residuals)

shows that still some low frequency is present in the series but it could not be detected by this analysis, though the magnitude of variance at low frequencies has reduced considerably as compared to the raw data spectrum, Figure 2.15.

The analysis of Ontario river data brings forth two main points

(i) Since all the three polutants are known to follow long period cyclic fluctuations, hence a longer record is needed for obtaining more reliable information.

(ii) Temperature is expected to follow a daily cycle but due to the sampling interval of one day; this fluctuation can not be detected by this analysis.

2.8 Development of a prediction model

A predictive model for each pollutant at a station was developed through regression analysis using the least squares method. The general form of a linear regression model is given by

$$Y = X\beta + \varepsilon \tag{2.18}$$

where Y is a (nxl) vector representing dependent variable

X is a (nxp) matrix representing the independent variables

β is a (pxl) vector of parameters

ε is a (nxl) vector of errors. It is assumed that the error "ε" is a random variable with zero mean and σ^2 variance, and ε_i and ε_j are uncorrelated, $i \neq j$. The purpose of the regression analysis is to obtain estimates b of the parameters which minimize the error sum of squares.

$\sum_{i=1}^{N} (Y_i - \hat{Y}_i)^2$ where $\hat{Y}_i$ is estimated value of Y_i

In this analysis stepwise regression procedure was used to develop the prediction model [38]. It's a process wherein variables are inserted in turn until a satisfactory regression equation is obtained. Each time a new variable is entered a check is also made for variables already in equation to see whether they are still significant. It may happen that the variable which was best in some earlier stage may become superfluous because of its relationships with other variables now in regression. This is checked by performing partial F-test for each variable in regression equation. This F-value is checked with a preselected percentage point of the F-distribution. Any variable which fails this test is removed from regression equation. This process is continued until no more variables can be entered in or removed from the equation.

One important consideration in the use of stepwise regression is the specification of F-values for the addition and the removal of a variable from the model. In this study both values were taken as zero so that all the variables may be entered in the model to account for all cyclic fluctuations.

The square of the multiple correlation coefficient was used as the criterion for the acceptance of a satisfactory model. It is defined as

$$R^2 = \frac{\text{Sum of squares due to regression corrected for mean}}{\text{Total corrected sum of squares}} = \frac{\sum_{i=1}^{N} (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{N} (Y_i - \bar{Y})^2} \qquad (2.19)$$

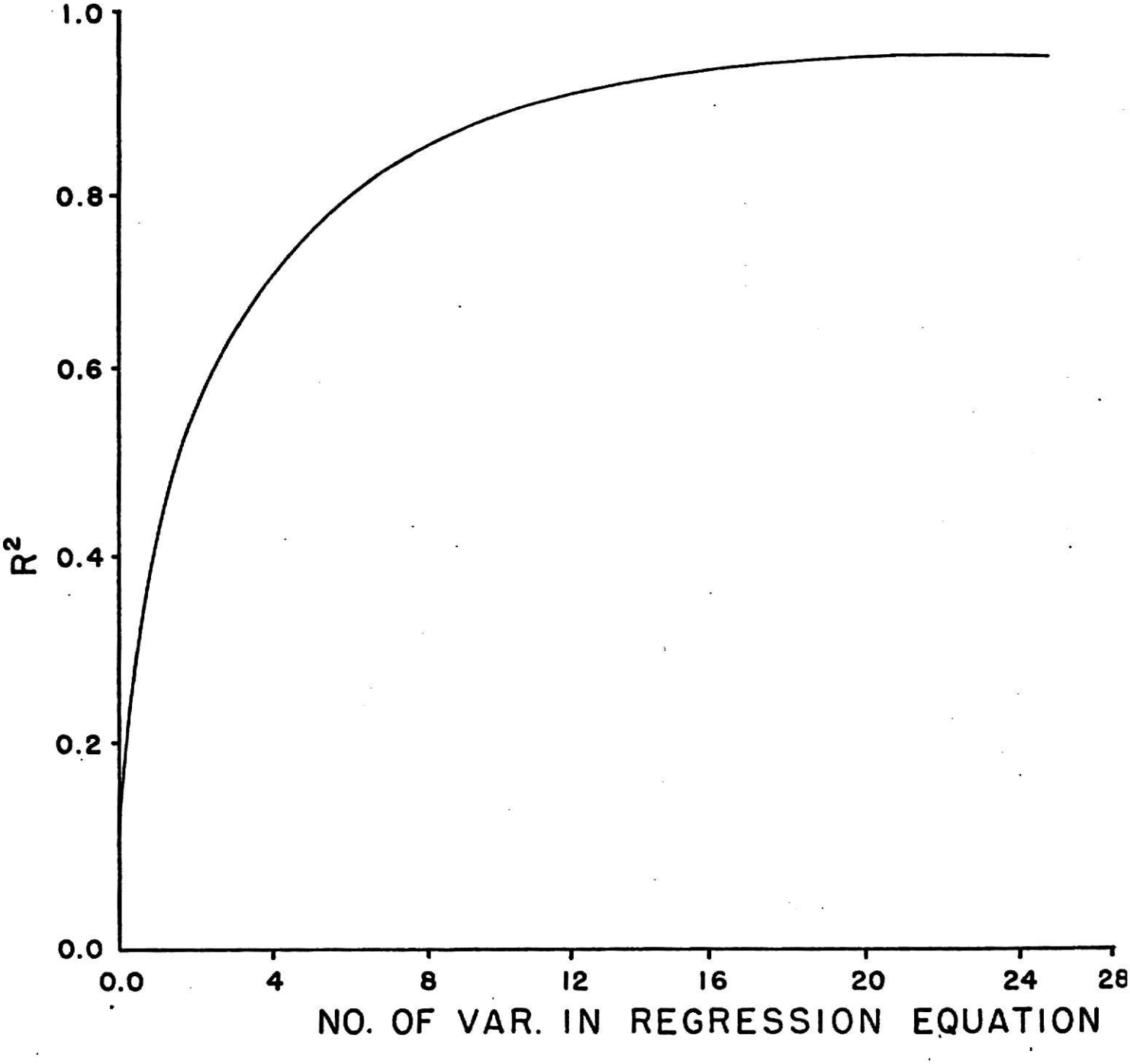

Fig. 2.17 R^2 vs. No. of variables in regression equation

where $\bar{Y}$ = mean of the dependent variable.

Its value lies between 0 and 1. A larger value of R^2 shows that the fitted equation explains the variation in the data satisfactorily.

In the stepwise regression procedure, R^2 was calculated at each step. It is seen that initially R^2 increases rapidly but tends to flatten afterwards indicating that not much improvement is made in the predictive capability of the model by the addition of more variables. A graph between R^2 and the number of variables entered for temperature data at Ontario river is shown in Figure 2.17.

2.8 Prediction models for Ontario river data:

Fitting a predictive model to Onterio River temperature data involved the regression of 25 independent variables and one dependent variable. The first independent variable was a linear term to account for any linear trend in the data. The other 24 terms corresponded to the cyclic variations as suggested by the harmonic analysis and the spectral analysis. R^2 increases rapidly upto addition of 13 variables and no significant improvement is made by addition of other 12 variables. These variables raise R^2 from 0.918 to 0.942. Considering the computational effort involved by using all 25 variables, it may be worth-while to use only 13 variables.

The model thus obtained is

$$T_t = 22.20 + 0.2989 \sin \frac{2\pi t}{30} - 0.7935 \cos \frac{2\pi t}{120}$$

$$+ 0.7605 \cos \frac{2\pi t}{365} - 0.3204 \sin \frac{2\pi t}{365}$$

$$+ 0.9920 \cos \frac{2\pi t}{91} + 0.5680 \sin \frac{2\pi t}{91}$$

$$+ 0.7216 \cos \frac{2\pi t}{52} + 0.2599 \sin \frac{2\pi t}{17}$$

$$- 0.2734 \cos \frac{2\pi t}{24} - 0.6573 \cos \frac{2\pi t}{73}$$

$$+ 0.3813 \sin \frac{2\pi t}{61} + 0.5195 \sin \frac{2\pi t}{182.5}$$

$$+ 0.3121 \sin \frac{2\pi t}{40} \qquad (2.20)$$

with $R^2 = 91.3\%$

$F(13,351) = 135.6229.$

Similar prediction models were obtained for specific conductance and flow data.

Model for Specific Conductance:

$$SC_t = 576.25 + 56.89 \cos \frac{2\pi t}{365} - 27.69 \sin \frac{2\pi t}{120}$$

$$- 55.04 \cos \frac{2\pi t}{182} + 33.89 \cos \frac{2\pi t}{90}$$

$$- 29.28 \sin \frac{2\pi 6}{90} + 26.25 \sin \frac{2\pi t}{60}$$

$$+ 20.96 \sin \frac{2\pi t}{52} - 16.21 \cos \frac{2\pi t}{45}$$

$$- 23.15 \sin \frac{2\pi t}{40} - 17.70 \cos \frac{2\pi t}{28}$$

with $R^2 = 85.6\%$

$F(11,353) = 88.37$

Model for Flow data

$$F_t = 5859.34 - 3212.40 \cos \frac{2\pi t}{365}$$

$$- 6351.24 \sin \frac{2\pi t}{365} + 4433.74 \cos \frac{2\pi t}{120}$$

$$- 3275.62 \cos \frac{2\pi t}{182} + 5116.85 \sin \frac{2\pi t}{182}$$

$$- 2168.54 \cos \frac{2\pi t}{90} - 2411.54 \sin \frac{2\pi t}{90}$$

$$+ 2718.60 \sin \frac{2\pi t}{73} + 1603.75 \cos \frac{2\pi t}{60}$$

with $R^2 = 95.7\%$

$F(9,355) = 435.98$

CHAPTER III

CROSS-SPECTRAL ANALYSIS

Spectral analysis as discussed in Chapter II is applied only to a individual record. But in certain cases, it may be desirable to study the interaction of two time series. These may be input-output of a system or two inputs to a system or two outputs from a system. Spectral analysis can be extended to analyze a pair of time series; this extended form being called cross-spectral analysis.

It is known that dissolved oxygen in a stream is affected by several variables such as temperature, photosynthesis, biochemical oxygen demand etc. Using individual spectral analysis for DO, it is not possible to know the relative importance of each cause of variation. If the simultaneous records of temperature, sunlight intensity, biochemical oxygen demand are known, cross spectral analysis can be applied to each pair of records and certain characteristics such as coherence, phase, transfer function etc. can be calculated to give some information about the relative importance of each source of variation.

One of the important properties of stationary time series that enables this technique to be applied is that not only the estimation of spectrum at one frequency is independent of other frequencies of the same process but it is also independent of other frequencies of the other processes. It may or may not be correlated with the same frequency component of the other processes [34]. This correlation of a particular frequency component of one series with the same frequency component of the other series is measured by a characteristic called coherency.

Another important advantage of using cross spectral analysis is that it permits to retain the phase relationship between the two series. Thus the time lag after which a particular frequency component of one series will follow the same frequency component of other series could be obtained. This statistics is very useful for the analysis of water pollution data. It is known that temperature and DO for an estuary follow a diurnal fluctuation. In this case, it would be desirable to have an estimate of time difference between the peaks of two fluctuations. The phase difference between the two fluctuations is found to be about 180^{o} i.e. when the temperature rises, concentration of DO falls and vice versa.

This chapter deals with the application of cross-spectral analysis to temperature, specific conductance and flow rate data at the Ontario River station.

3.1 Cross spectral analysis:

As in simple spectrum analysis, the auto-covariance function measures the correlation of observations of a time series at different time spans, the cross-covariance function measures the correlation of two different series. The cross-covariance function at lag k is defined as

$$\gamma_{12}(k) = E\,[(x_1(t) - \mu_1)(x_2(t+k) - \mu_2)]$$

$$\gamma_{21}(k) = E[(x_2(t) - \mu_2)(x_1(t+k) - \mu_1)]$$

where $\gamma_{12}(k)$ = cross-covariance function at lag k with series 2 leading series 1.

$\gamma_{21}(k)$ = cross covariance function at lag k with series 1 leading series 2.

The corresponding estimates of $\gamma_{12}(k)$ and $\gamma_{21}(k)$ for a sampled discrete time series are given by

$$C_{12}(k) = \frac{1}{N} \sum_{t=1}^{N-1} (X_{1t} - \bar{X}_1)(X_{2t+k} - \bar{X}_2), \qquad k \geq 0 \tag{3.1}$$

$$C_{21}(k) = \frac{1}{N} \sum_{t=1}^{N-k} (X_{1t+k} - \bar{X}_1)(X_{2t} - \bar{X}_2), \quad k \geq 0 \tag{3.2}$$

where, $\bar{X}_1 = \frac{\sum_{i=1}^{N} X_{1i}}{N}$, $\bar{X}_2 = \frac{\sum_{i=1}^{N} X_{2i}}{N}$

It has the properties that

$$C_{12}(k) \neq C_{12}(-k) \tag{3.3}$$

$$\left.\begin{aligned} C_{12}(k) &= C_{21}(-k) \\ \text{or} \quad C_{21}(k) &= C_{12}(-k) \end{aligned}\right\} \tag{3.4}$$

(3.4) shows that in general the cross covariance function is not an even function of lag.

The corresponding cross correlation function is defined as

$$\rho_{12}(k) = \frac{c_{12}(k)}{c_{11}(0)\ c_{22}(0)} \tag{3.5}$$

with the properties that

$$|\rho_{12}(k)| \leq 1$$

$$\rho_{12}(u) = \rho_{21}(-u).$$

It is interpreted in about the same manner as an auto-correlation function.

A positive cross-correlation indicating that a high (low) observation in one series tends to follow a high (low) observation in the other series. A zero correlation at all lags indicates that the two processes are completely uncorrelated.

As before, main use of the cross covariance function is as an intermediate step in the calculation of its Fourier transform, the cross spectrum. The sample cross spectrum estimator is defined as

$$S_{12}(\omega) = \int_{-N}^{N} c_{12}(k)\, e^{-i\omega k}\, dk \qquad (3.6)$$

This sample cross spectrum is a complex quantity and can be written as

$$S_{12}(\omega) = A_{12}(\omega)\, e^{-i\phi\omega} \qquad (3.7)$$

An alternative expression for (3.7) is

$$S_{12}(\omega) = c0_{12}(\omega) + jQ_{12}(\omega) \qquad (3.8)$$

where $c0_{12}(\omega)$ is the real part called cospectrum

and $Q_{12}(\omega)$ is the imaginary part called Quadrative spectrum.

$$A_{12}^2(\omega) = c0_{12}^2(\omega) + Q_{12}^2(\omega) \qquad (3.9)$$

$$\text{and } \phi_{12}(\omega) = \arctan \frac{Q_{12}(\omega)}{c0_{12}(\omega)} \qquad (3.10)$$

Writing $c_{12}(k)$ as the sum of an even part and odd part gives,

$$\ell_{12}(k) = \frac{1}{2}\,(c_{12}(k) + c_{12}(-k)), \qquad 0 \leq k \leq M-1 \qquad (3.11)$$

$$q_{12}(k) = \frac{1}{2}\,(c_{12}(k) - c_{12}(-k)), \qquad 0 \leq k \leq M-1 \qquad (3.12)$$

the estimation of the sample cospectrum and quadrative spectrum is obtained as

$$c0_{12}(\omega_j) = \frac{1}{2\pi}\left\{\frac{\ell_{12}(0)}{2} + \sum_{k=1}^{M} \ell_{12}(k)\cos\omega_j k\right\} \tag{3.13}$$

$$Q_{12}(\omega_j) = \frac{1}{2\pi}\sum_{k=1}^{M} q_{12}(k)\sin\omega_j k \tag{3.14}$$

where $\omega_j = \frac{\pi j}{M}$ $\qquad j = 0, \ldots, M$

These new estimates of the co-and quadrative spectra are then smoothed by using a spectral window, as discussed in Chapter II.

In this study, a Hamming window was used. According to this window, the smoothed cospectrum is

$$\bar{c}0_{12}(0) = .54c0_{12}(0) + .46c0_{12}(1)$$

$$\bar{c}0_{12}(\omega_j) = .23\ c0_{12}(\omega_{j-1}) + .54c0_{12}(\omega_j) + .23c0_{12}(\omega_{j+1}) \qquad 0 \le j < m$$

$$\bar{c}0_{12}(\omega_m) = .54c0_{12}(\omega_m) + .46c0_{12}(\omega_{m-1}) \tag{3.15}$$

Similar expressions can be written for the quadrative spectrum.

The smoothed estimate for amplitudes of a cross spectrum is observed as

$$\bar{A}_{12}(\omega_j) = \sqrt{\bar{c}0^2_{12}(\omega_j) + \bar{Q}^2_{12}(\omega_j)}, \qquad 0 \le j \le m \tag{3.16}$$

and the smoothed phase spectral estimate is obtained as

$$\bar{\phi}_{12}(\omega_j) = \arctan \frac{\bar{Q}_{12}(\omega_j)}{\bar{c0}_{12}(\omega_j)} \tag{3.17}$$

As discussed earlier, the coherency function is a very useful in the interpretation of the cross-spectrum and is defined as

$$H(\omega_j) = \frac{\bar{c0}^2_{12}(\omega_j) + Q^2_{12}(\omega_j)}{S_1(\omega_j) \cdot S_2(\omega_j)} \tag{3.18}$$

where $S_1(\omega_j)$ and $S_2(\omega_j)$ are smoothed spectral estimates of series 1 and 2 respectively at frequency ω_j.

$$H(\omega_j) \leq 1$$

The distribution of the coherency function has been studied and the confidence intervals for coherence estimates when the true coherency is zero have been given by Granger and Hatanka in [34] for different levels of N/M and 50%, 90%, 95% significance level. Confidence bands for phase angle for certain levels of N/M and coherency are also given in [34] and were used in this study.

Another important characteristic to be determined is the response function or transfer function. It gives an estimation of the output spectra if the input record were the only parameter dominating it. The relative importance of effect of each causal parameter on a single pollutant may be compared by calculating transfer function of each parameter with the given pollutant separately.

The amplitude of the transfer function is given by

$$\bar{G}(\omega_j) = \frac{\bar{A}_{12}(\omega_j)}{S_1(\omega_j)} \tag{3.19}$$

The confidence interval for the transfer function is given by [33] for 100 (1-α)% confidence interval,

$$\bar{G}(\omega_j) \pm \bar{G}(\omega_j) \cdot \sqrt{\frac{2}{\nu-2} f_{2,\nu-2}(1-\alpha) \left(\frac{1 - H_{12}(c0_j)}{H_{12}(\omega_j)}\right)} \tag{3.20}$$

The phase of the transfer function is the same as derived in (3.17).

The presence of trend in the time series distorts the estimation of autospectra, and cospectra and may produce spurious coherencies. Thus, the data should, initially, be inspected for trend and filtered accordingly. An indication of trend is given by the failure of the autocorrelation function to die quickly. The filtering of the time series does not affect the coherency, phase and transfer function spectra. Also, large spurious cross covariances are generated between two time series as a result of the large autocovariances within the time processes. This may also necessitise a filtering operation on the two series before computing cross-covariances.

Another refinement which may be carried out to improve the coherency estimates is the alignment of the two series. In many cases, it is seen that the cross-covariance function does not have the maximum absolute value at zero lag. Alignment consists of centering the cross-covariance function such that its largest absolute value occurs at zero lag. Let S be the lag at which the maximum cross-covariance occurs, then for alignment,

$$c''_{12}(k) = c_{12}(S + k) \qquad (3.12)$$

Further computation of cross-spectrum should be based on these cross-covariances.

3.3 Analysis of Ontario River data:

Three cross-spectral studies were made for Ontario River data, viz.

(a) Temperature and specific conductance

(b) Flow rate and temperature

(c) Flow rate and specific conductance.

(a) Temperature and specific conductance:

The autocorrelation and power spectral plots of temperature and specific conductance are shown in Figures 2.5, 2.7, 2.9 and 2.11 respectively. Failure of the auto-correlation function to damp quickly along with high peaks at zero frequency in the power spectral estimates of the two pollutants indicate the presence of trend in the data. Figures 3.1 and 3.2 show the cross-correlation function for the original and differenced data. The plot of the original data does not damp quickly whereas the cross-correlation of the differenced data oscillates about zero line. Maximum cross-correlation is obtained at a lag of 67 days. This implies that there is a time lag of 67 days between the responses of the two pollutants and that the net direction of casuality is from temperature to specific conductance which may be expected under natural circumstances. As the presence of a trend is indicated by the above plots, it was decided to perform further analysis using differenced data. The cross-correlation plot of

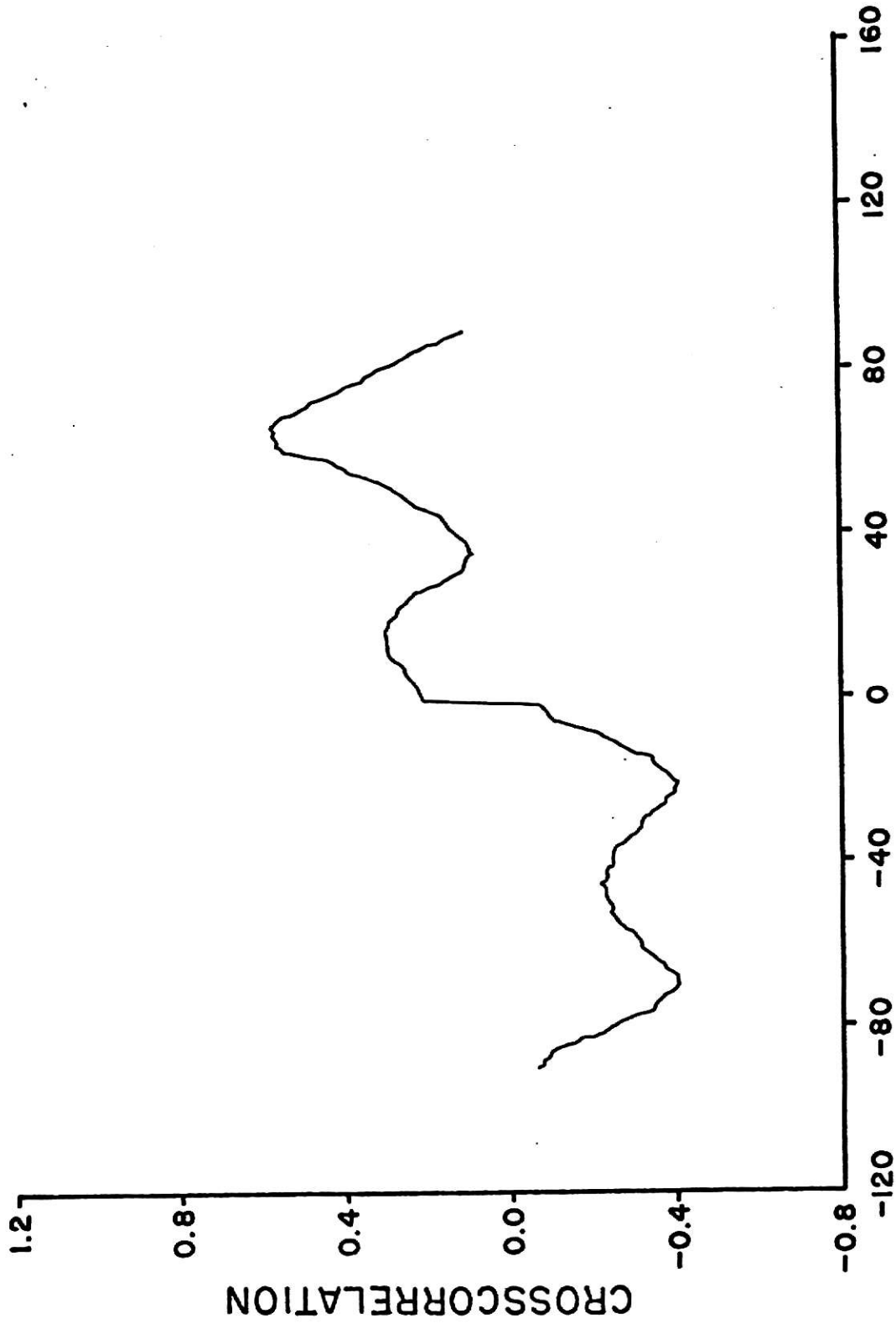

Fig. 3.1 Crosscorrelation of original data - temperature and sp. conductance.

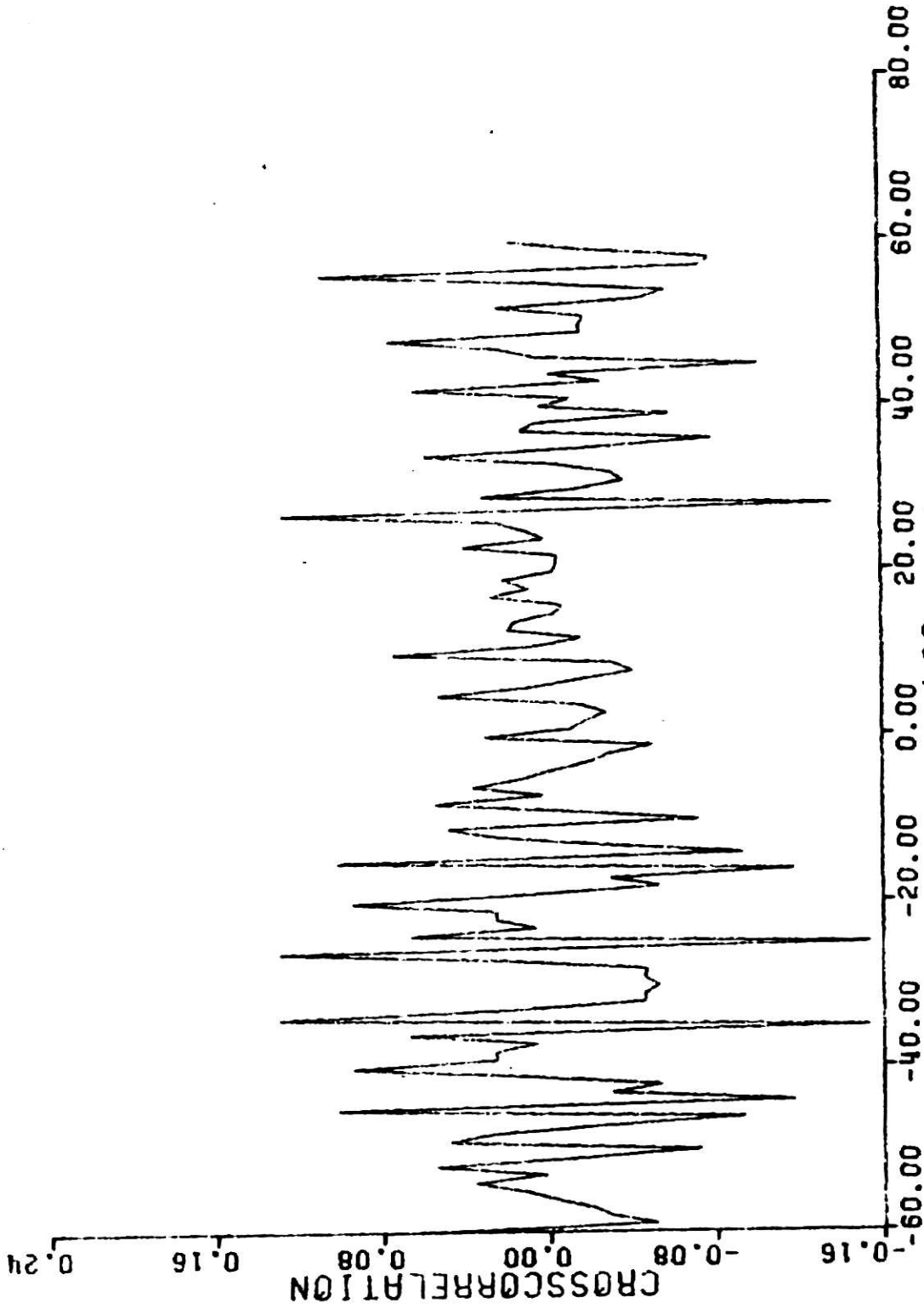

Fig. 3.2 Crosscorrelation of differenced data - temperature and sp. conductance.

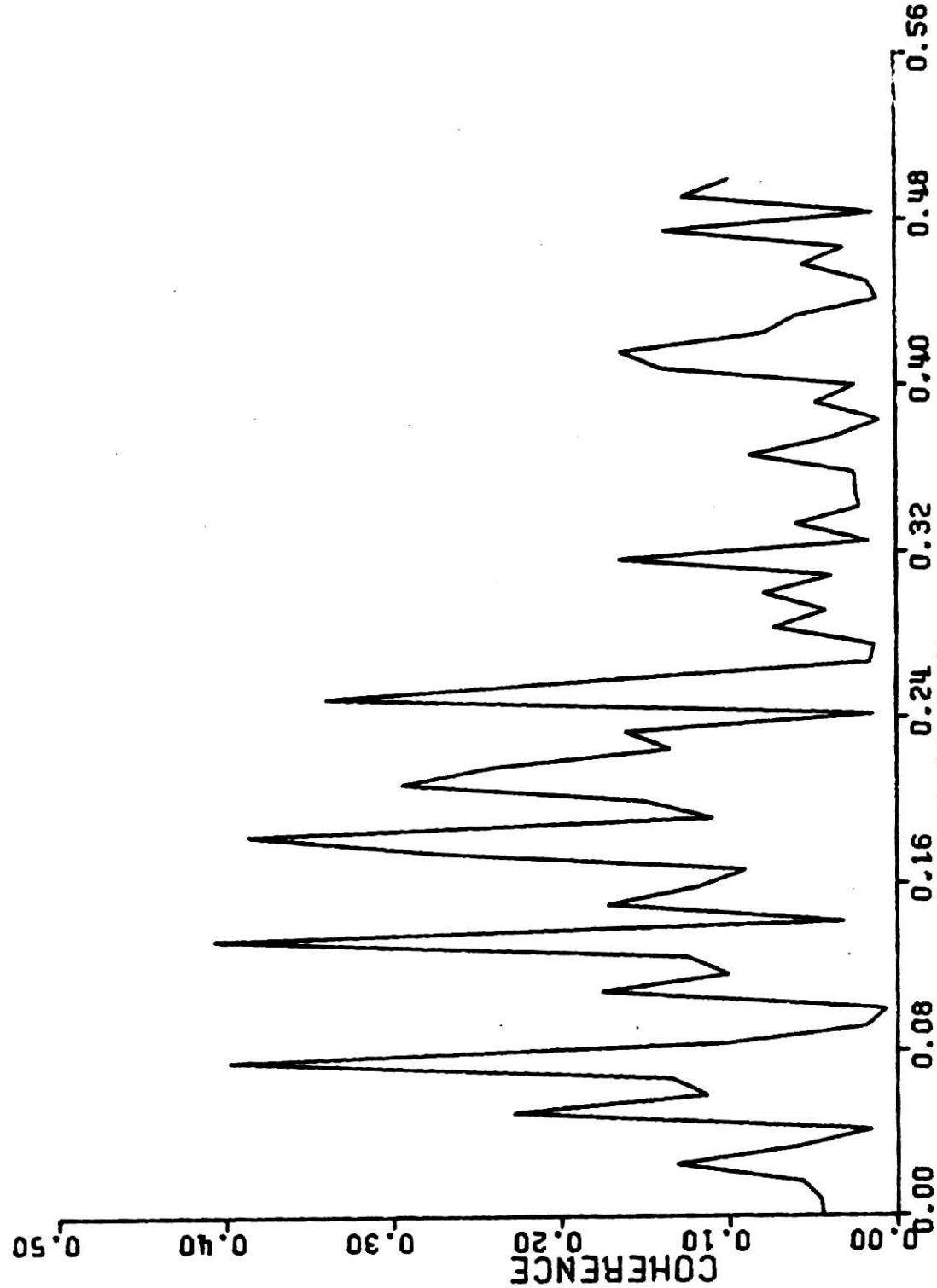

Fig. 3.3 Coherence spectra - temp. and sp. conductance (before alignment).

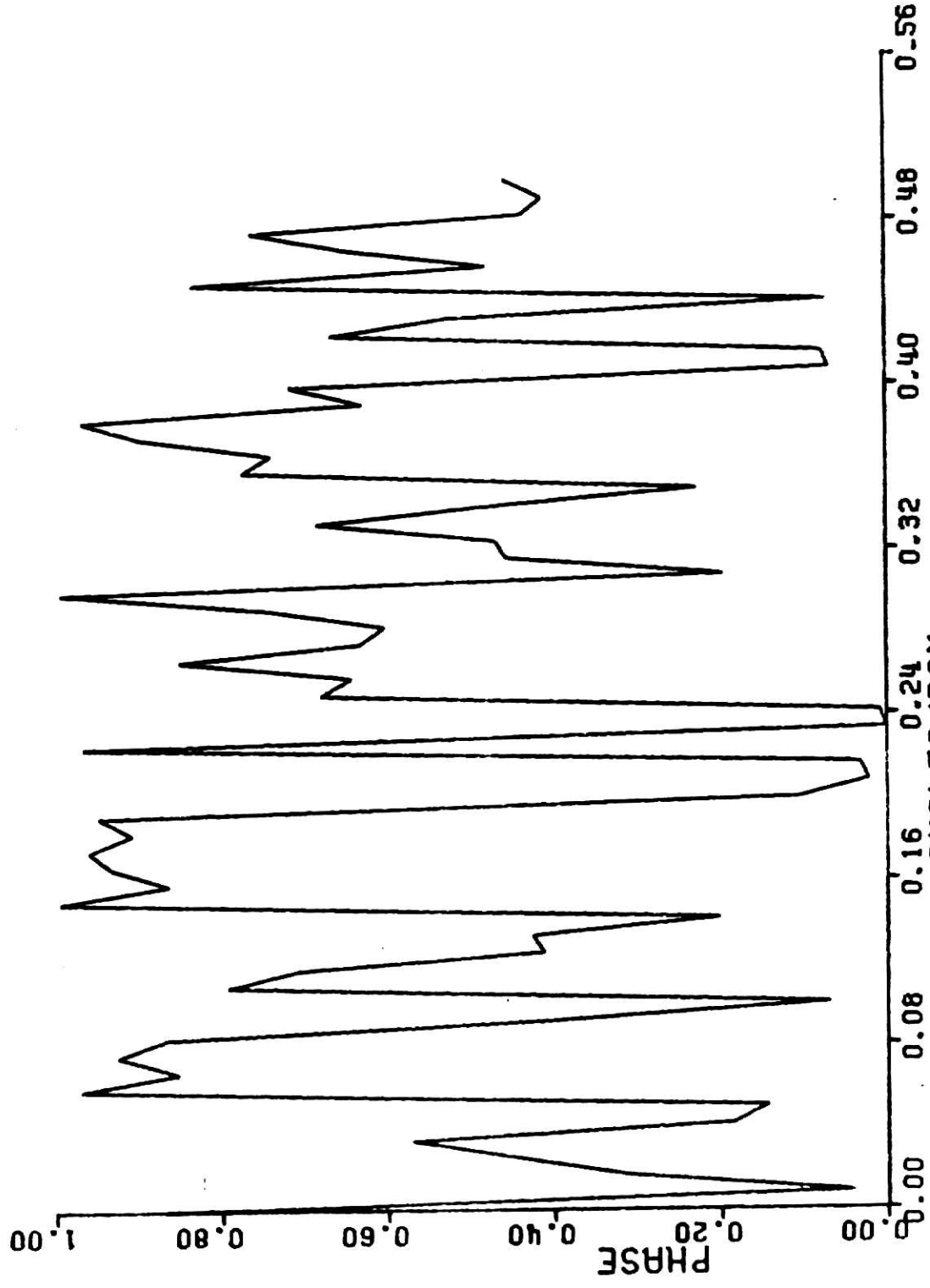

Figure 3.4 Phase spectra - temp. and sp. conductance (before alignment).

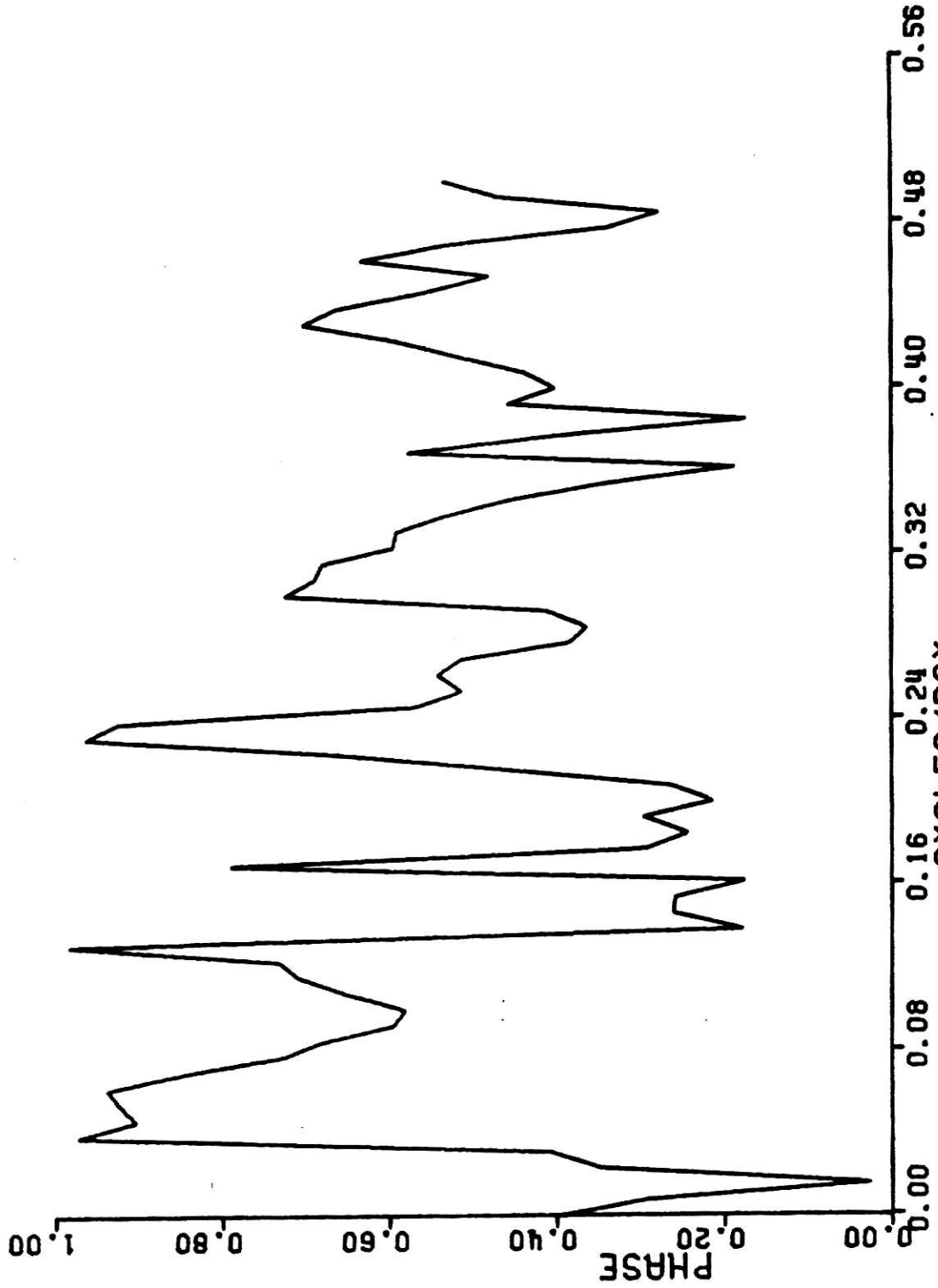

Fig. 3.5 Phase spectra - temp. and sp. conductance (after alignment).

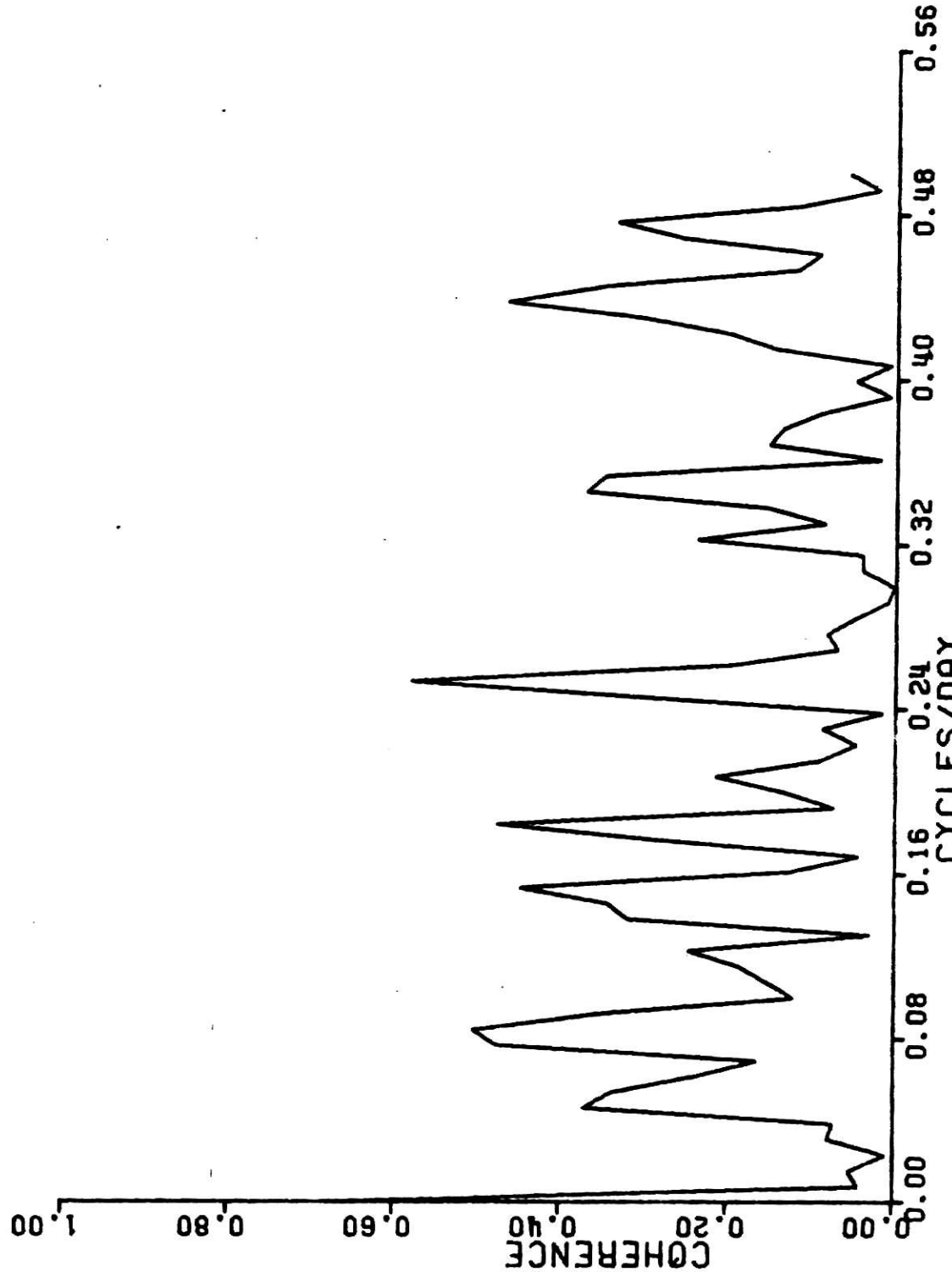

Fig. 3.6 Coherence spectra - temp. and sp. conductance (after alignment).

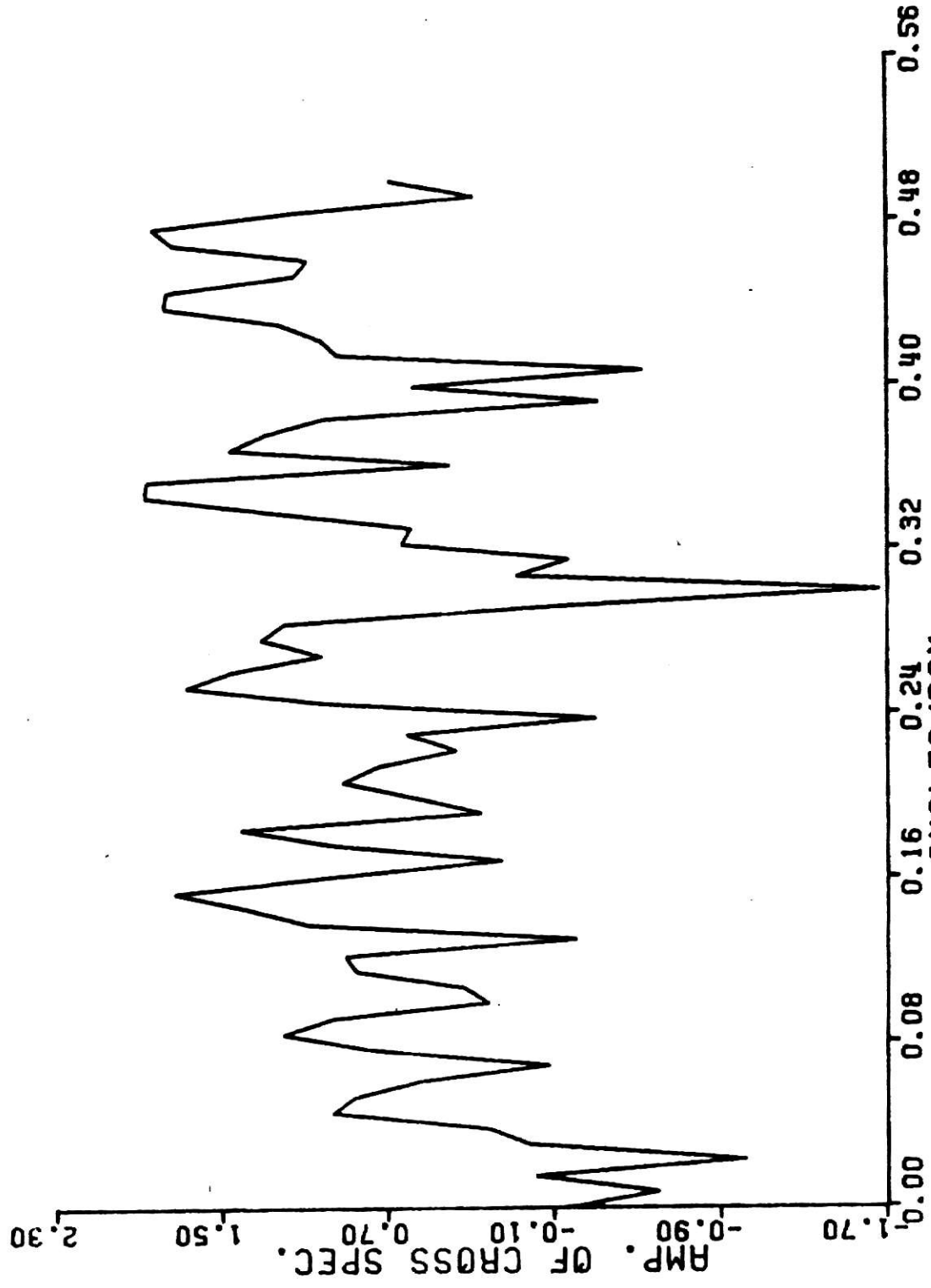

Fig. 5.7 Amplitude of cross-spectrum - temp. and sp. conductance.

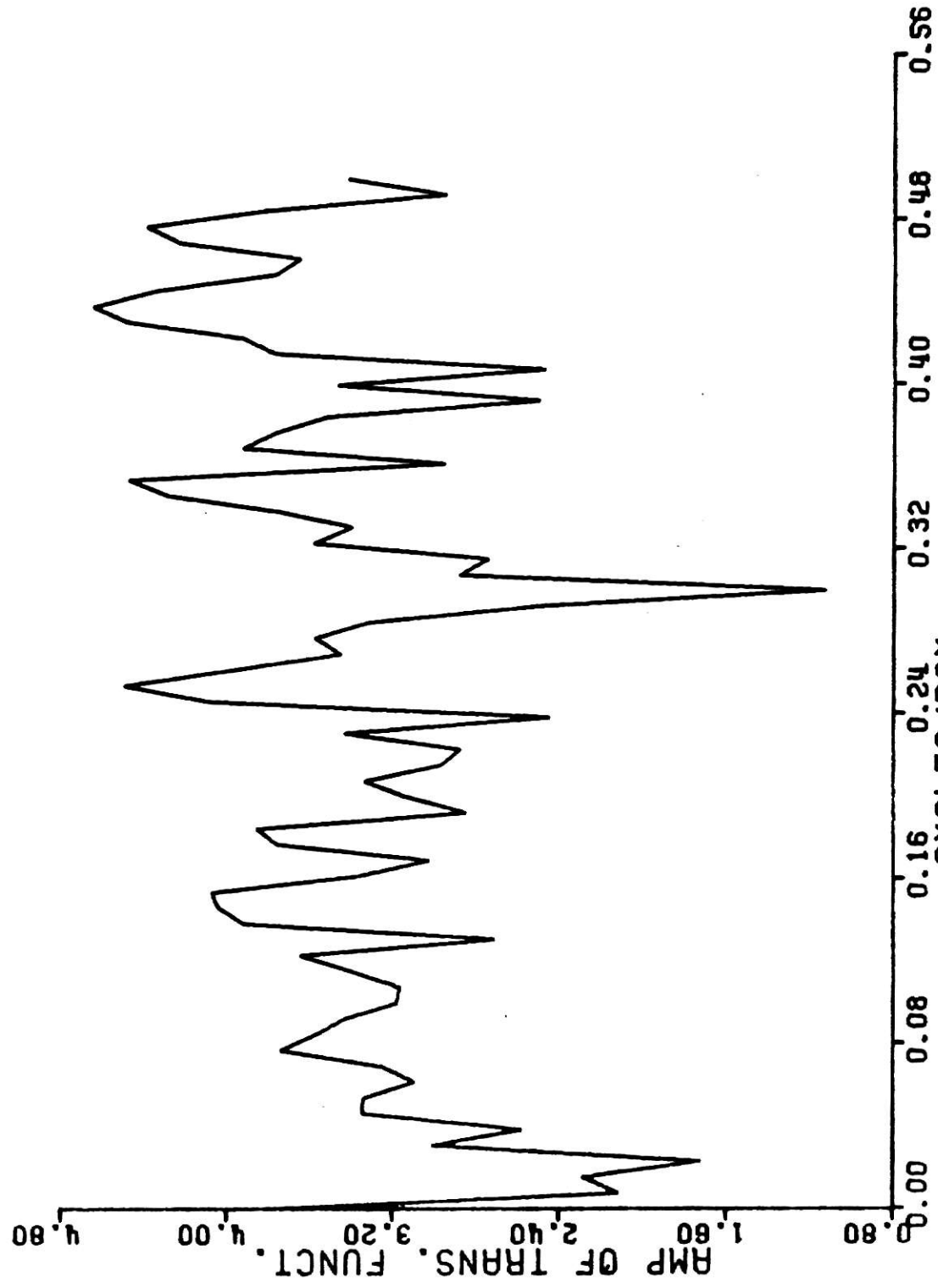

Fig. 3.8 Amplitude of transfer function from temp. to sp. conductance.

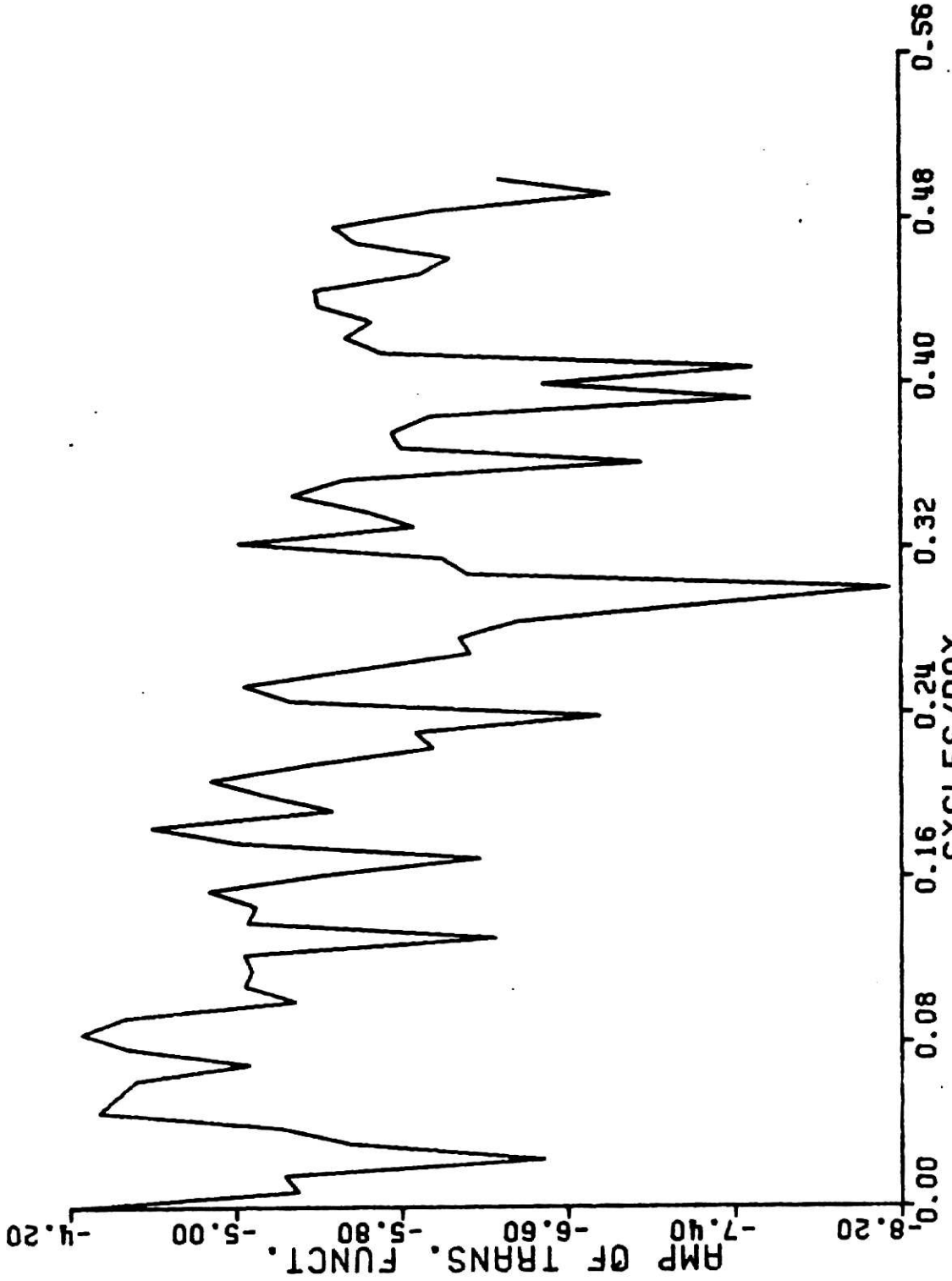

Fig. 3.9 Amplitude of transfer function from sp. conductance to temp.

differenced data shows maximum absolute value at -35 days lag. This suggests that the cross-correlation be aligned for calculation of phase, coherency, and transfer function spectra. Figures 3.3 and 3.4 show the coherency and the phase spectra before alignment and Figures 3.5 thru 3.9 show the various spectra after alignment. It is seen that the coherency spectrum is improved by alignment and the phase spectrum does not show any linear trend, and hence further alignment was not considered necessary. The confidence interval for coherency was read off from tables given by Granger and Hatanka in [34].

In general, the coherency appears to be low at all frequencies, especially at high ones, indicating that the variations in temperature and specific conductance at high frequencies are independent of each other. Some notable peaks in coherency are observed at zero frequency and at 13 day , 4 day periods. High coherency at zero frequency is important in the sense that the individual spectra for both pollutants have high peaks at this frequency. But the exact nature of correlation at zero frequency is difficult to be determined by this analysis due to lack of data. Data for about 3 to 4 years should be available to obtain more meaningful information at low frequencies. A high transfer function from temperature to specific conductance is shown at all frequencies. The transfer function gives a relationship between the input and the output spectra of a process as

$$\text{output spectra} = |\bar{G}(\omega_j)|^2 \text{ Input spectra}$$

A high value of transfer function indicates that a high variance at a particular frequency in input data will cause a high variation in output data. It may be noted here that two records may have low

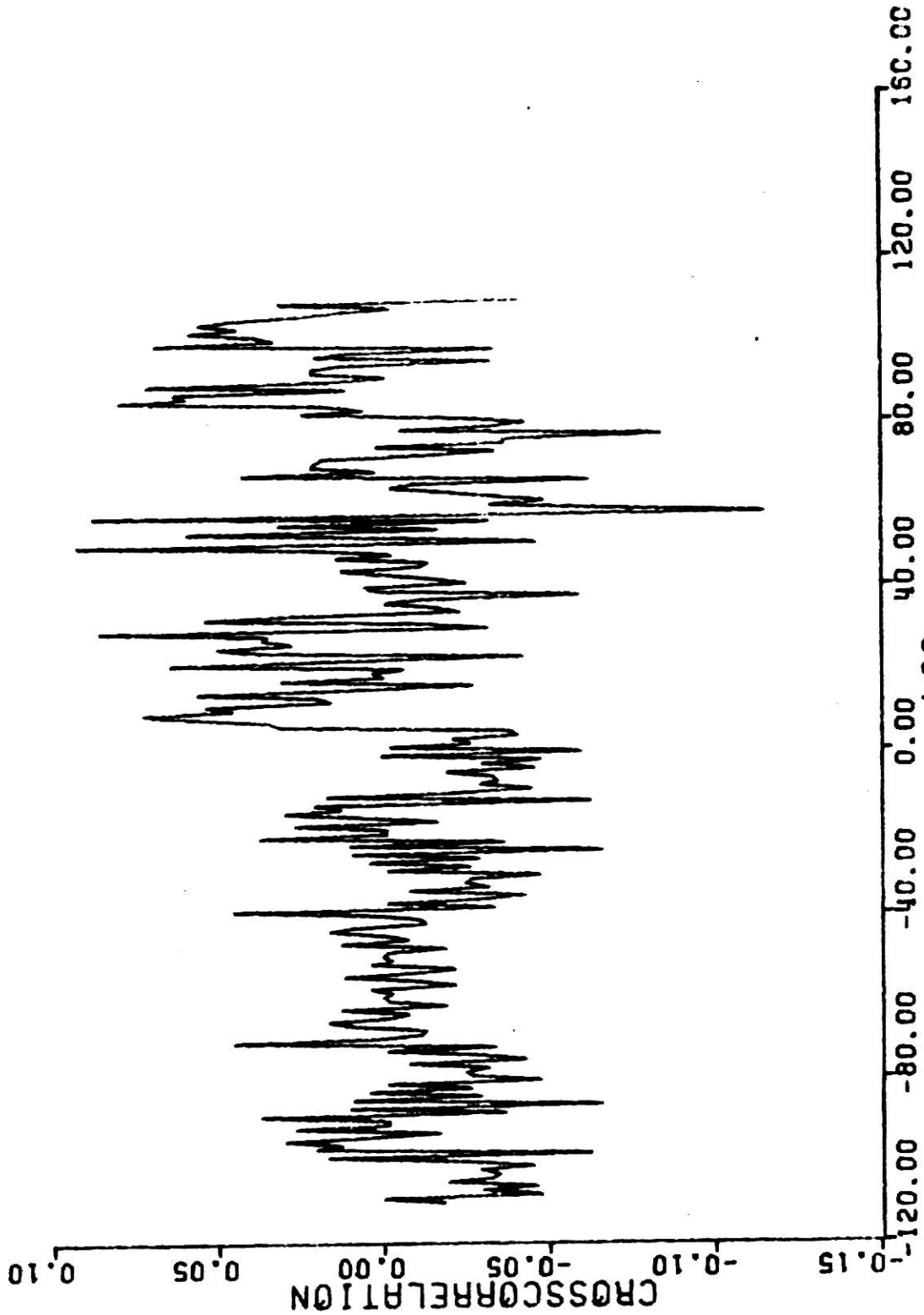

Fig. 5.10 Cross correlation of differenced data - flow rate and sp. conductance.

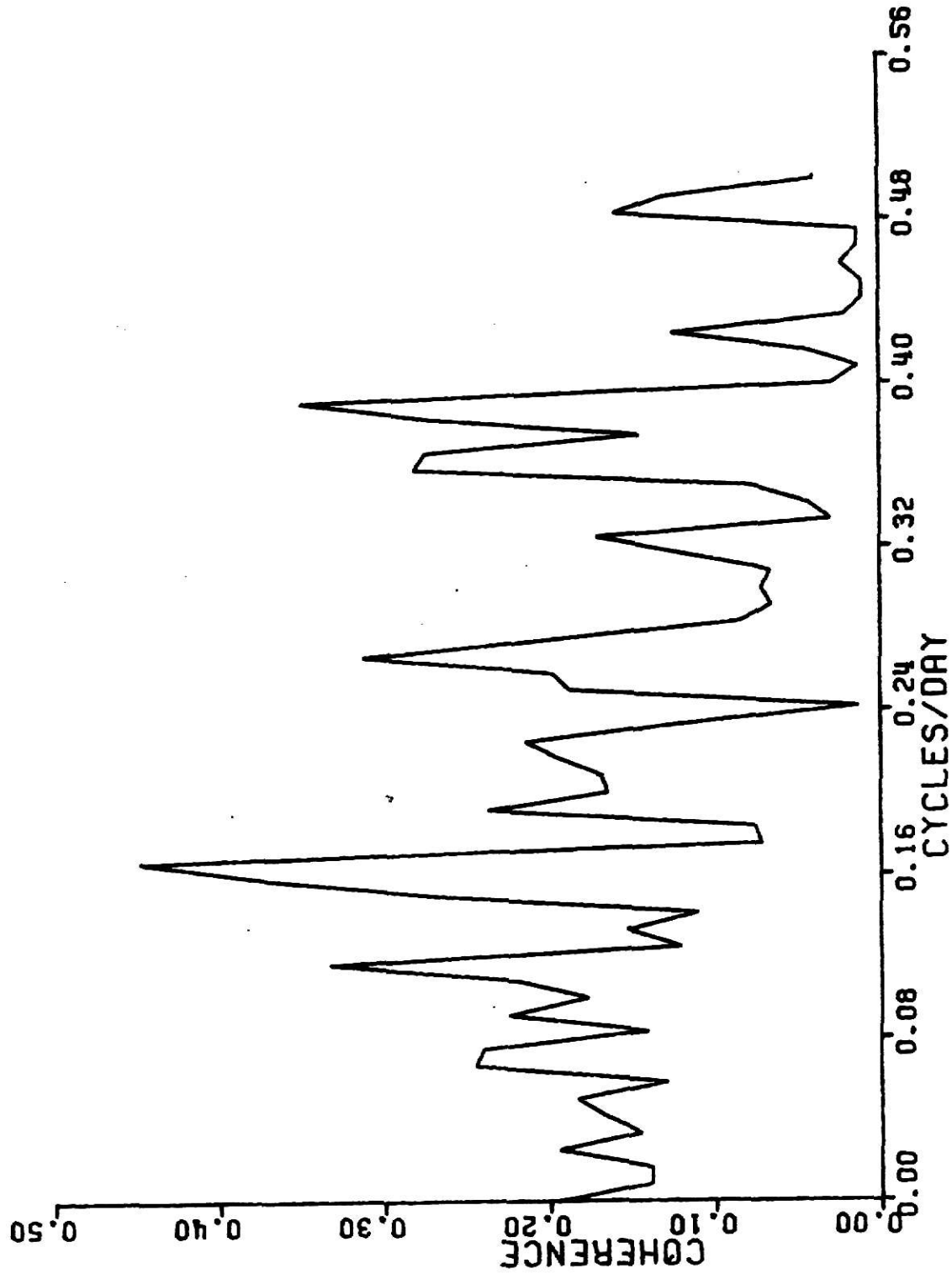

Fig. 3.11 Coherency spectra - flow rate and sp. conductance (before alignment).

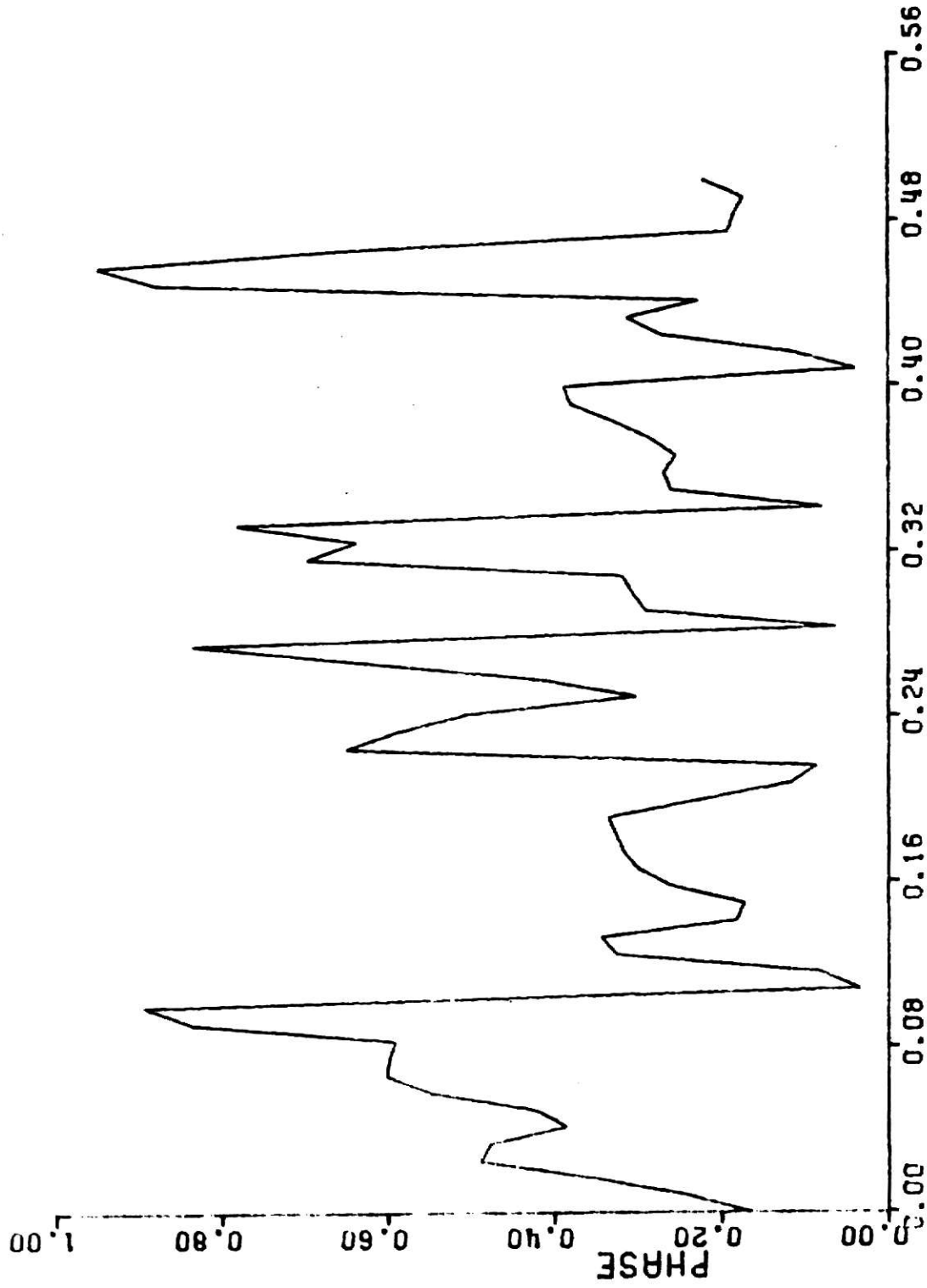

Fig. 3.12 Phase spectra - flow rate and sp. conductance (before alignment).

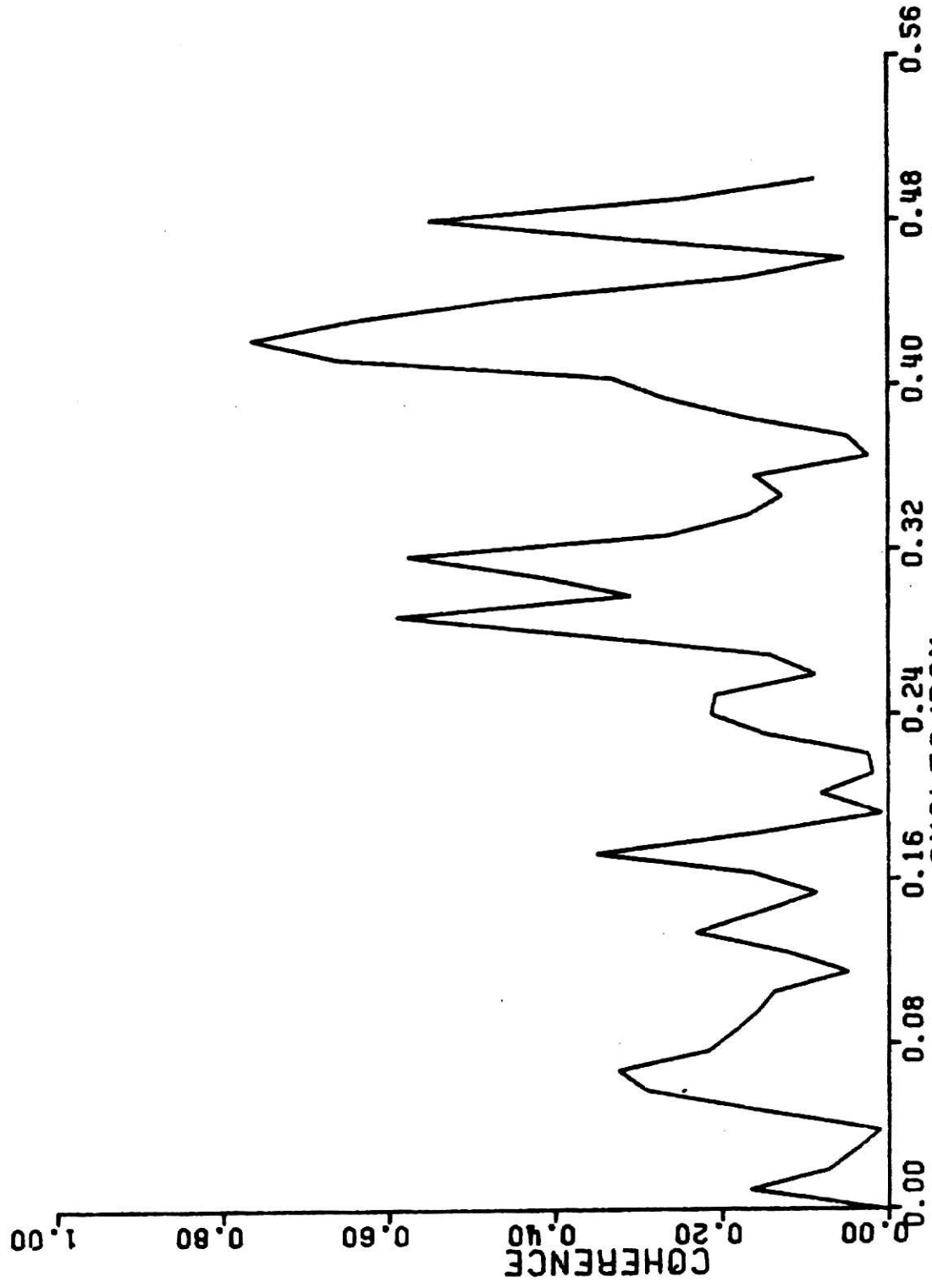

Fig. 3.15 Coherency spectra - flow rate and sp. conductance (after alignment).

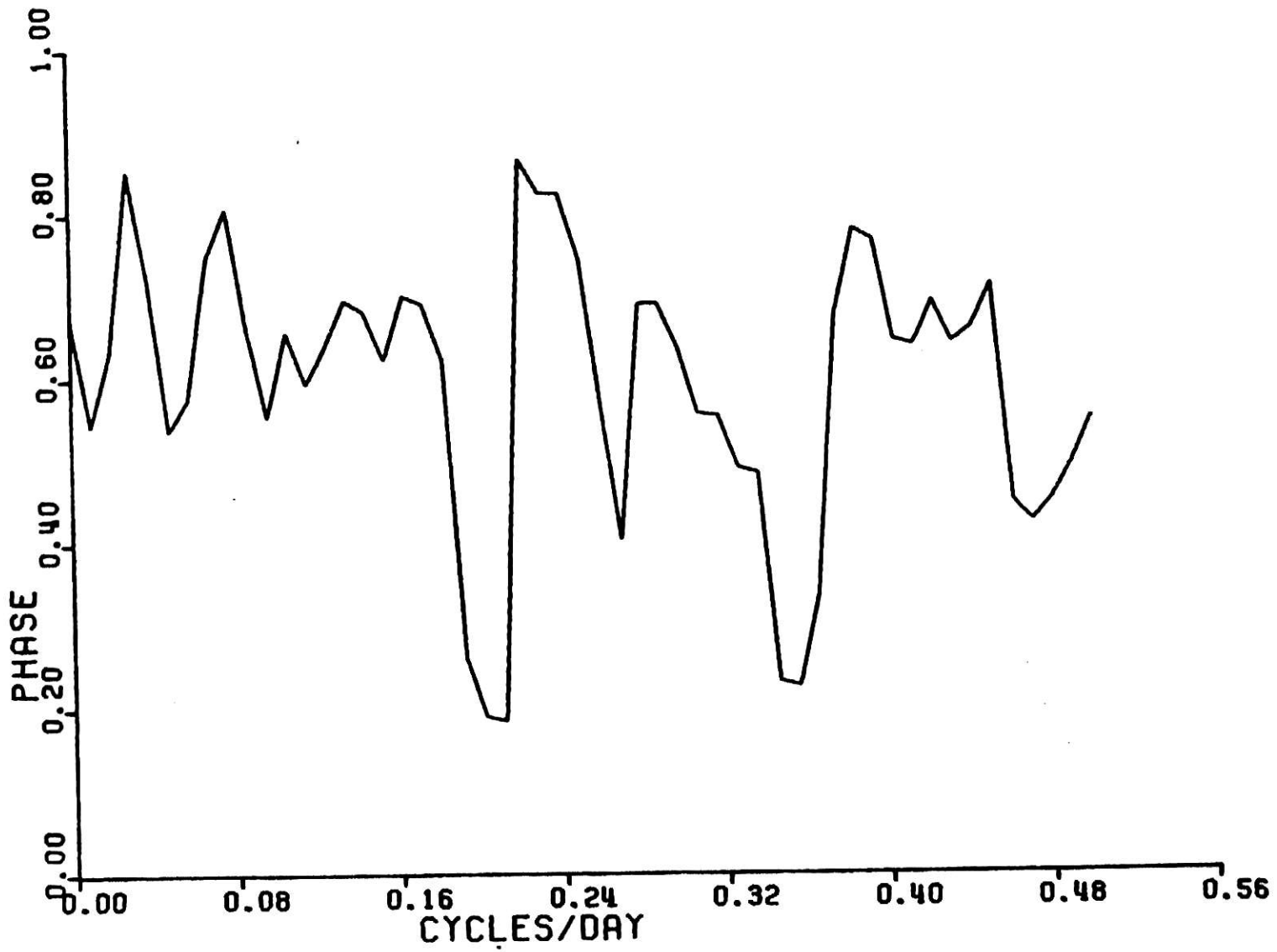

Fig. 3.14 Phase spectra - flow rate and sp. conductance (after alignment).

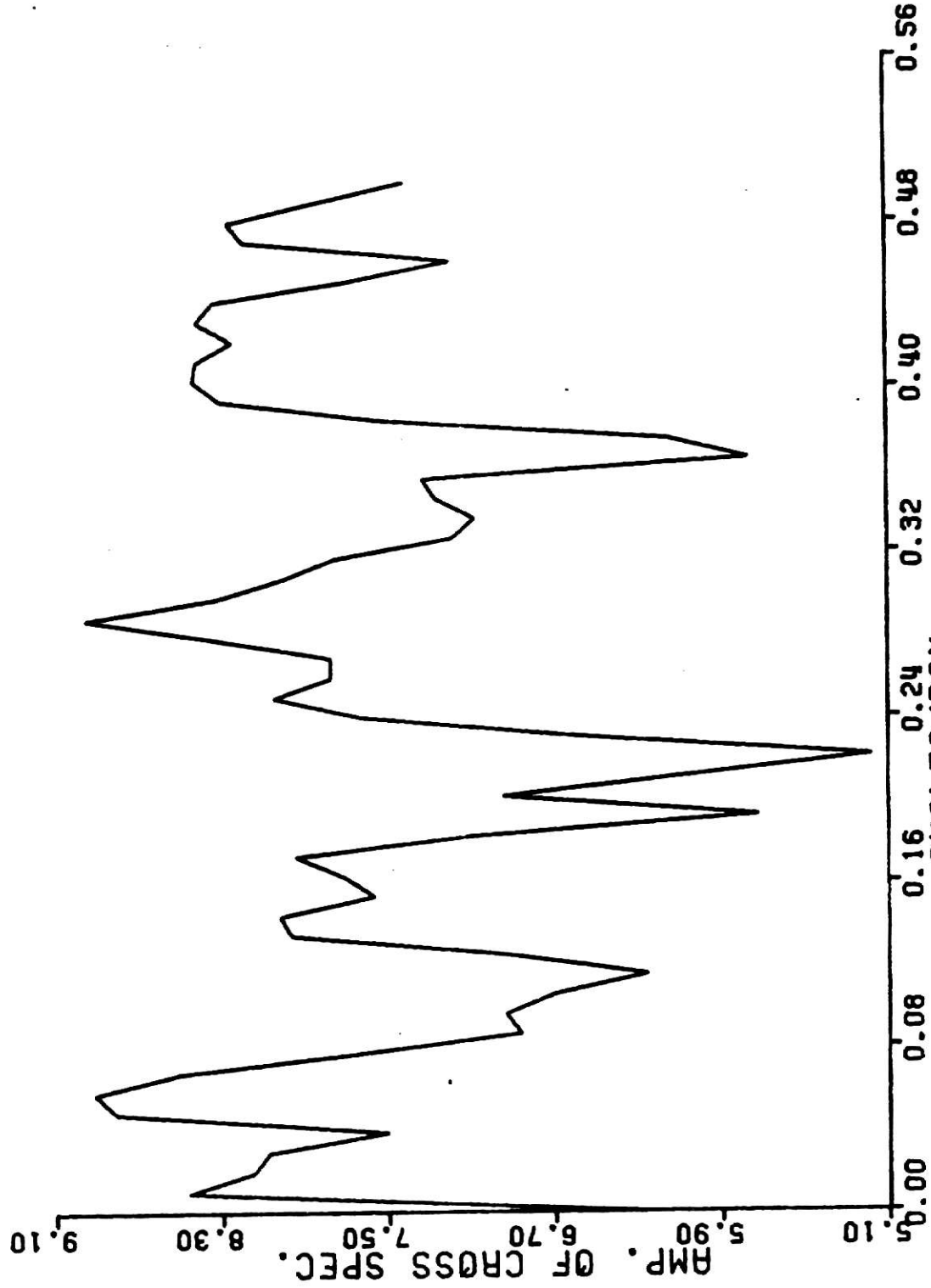

Fig. 3.15 Amplitude of cross-spectrum - flow rate and sp. conductance.

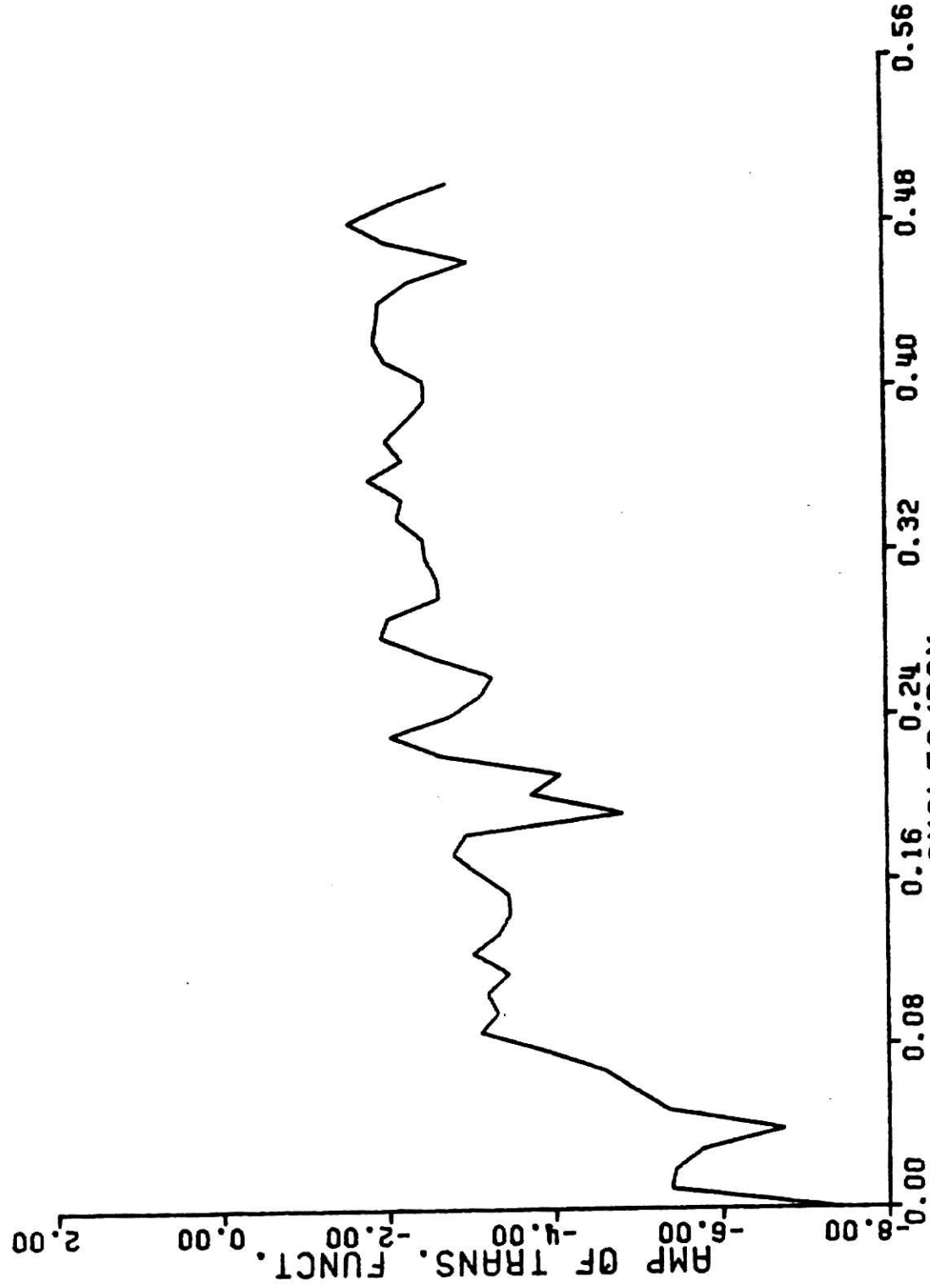

Fig. 3.16 Amplitude of transfer function from flow rate to sp. conductance.

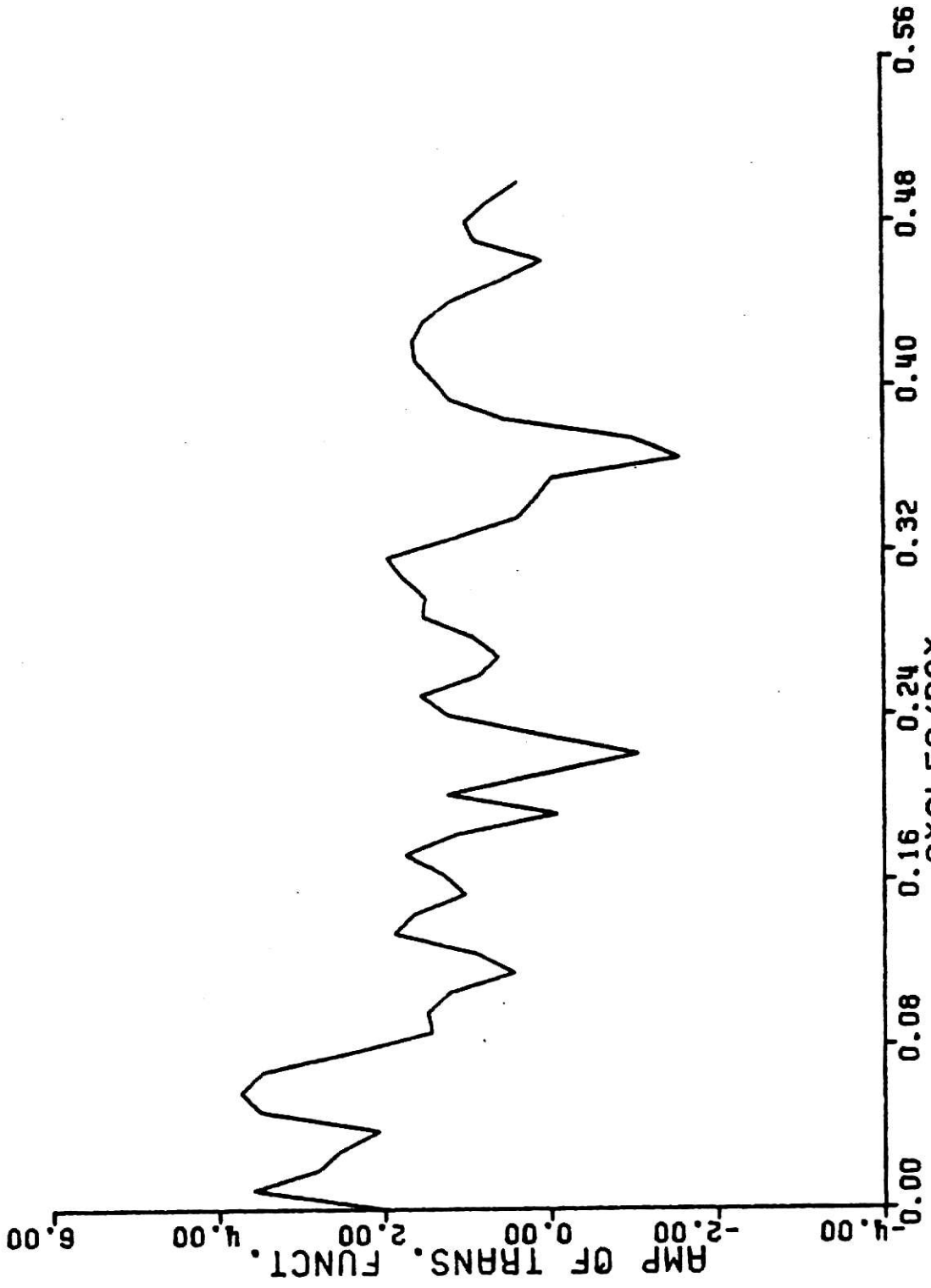

Fig. 3.17 Amplitude of transfer function from sp. conductance to flow rate.

coherency at a given frequency even though the corresponding trasfer function is high.

(b) Flow rate and specific conductance:

Figure 2.9 and 2.13 show the autocorrelation plot of the flow rate and the specific conductance data respectively. Their failure to damp out quickly indicates the presence of trend in the data and hence all further analysis was done using the filtered data. As before, a simple difference filter was used for both the series. Figure 3.10 shows the cross-correlation function for the differenced data for 110 lags. It oscillates about zero and has a maximum absolute value at a lag of 58 days. This indicates a lag of 58 days between the responses of the two pollutants and the direction of causality from flow rate to specific conductance. It also suggests that the two series be aligned before computation of coherency and phase spectra. Figures 3.11 and 3.12 show the coherency and phase spectra before alignment and Figures 3.13 thru 3.17 show the various spectra after alignment. The confidence limit on the coherency spectra have been drawn for 90% significance level and N/M = 7. Low coherency is observed at low frequencies and high correlation is obtained at frequencies of .4057 cycles/day, .4151 cycles/day, .4245 cycles/day. It implies that the long period variations in specific conductance and flow rate are uncorrelated. No physical meaning can be attached to high coherencies at high frequencies as these fluctuations are not dominant in the individual power spectrum of each pollutant. The phase spectrum (Figure 3.14) shows that the two responses have 180^{o} phase difference at low frequencies. This may be expected if the water flow in the river is mainly fresh water

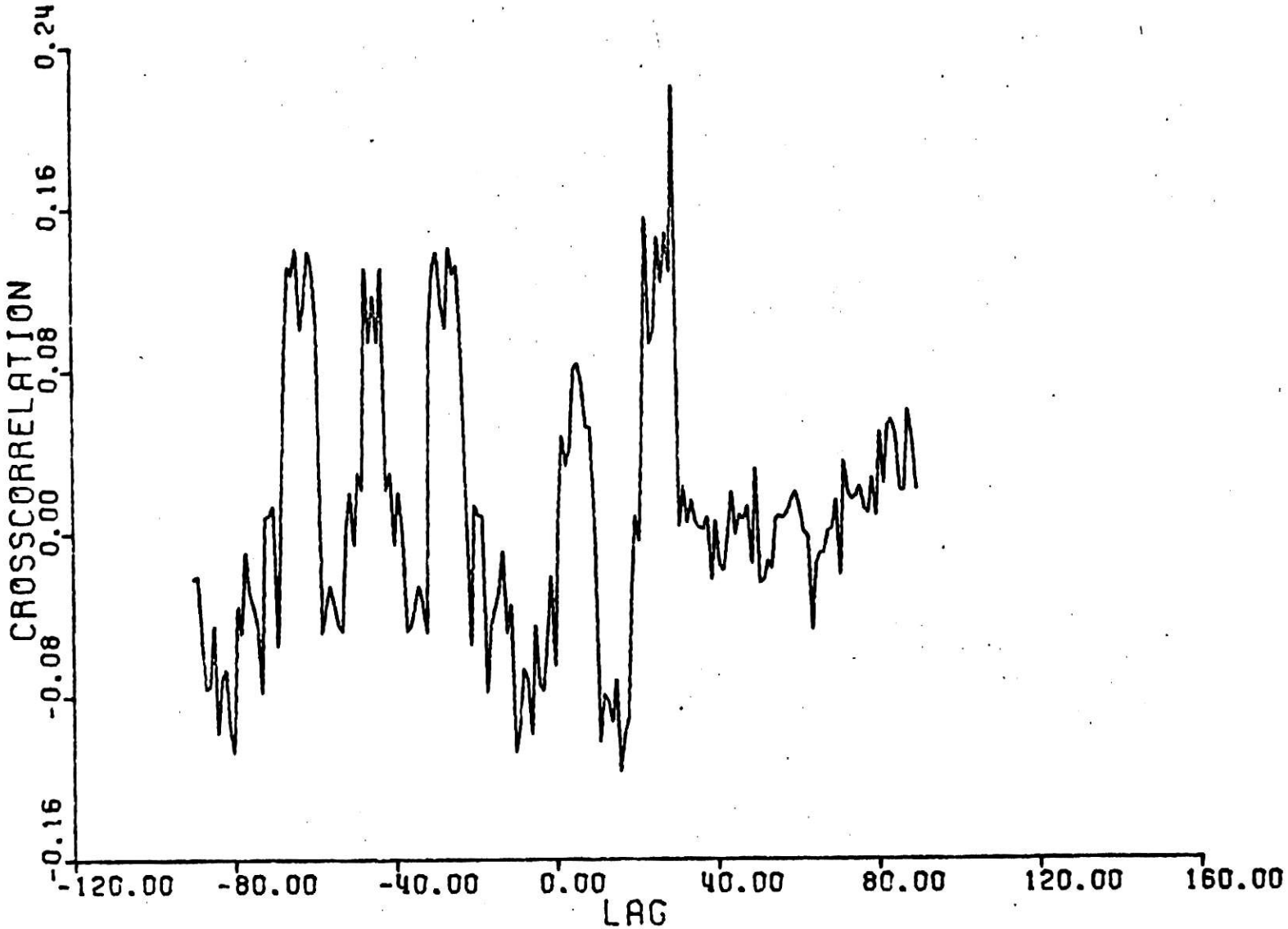

Fig. 3.18 Cross correlation of differenced data - flow rate and temp.

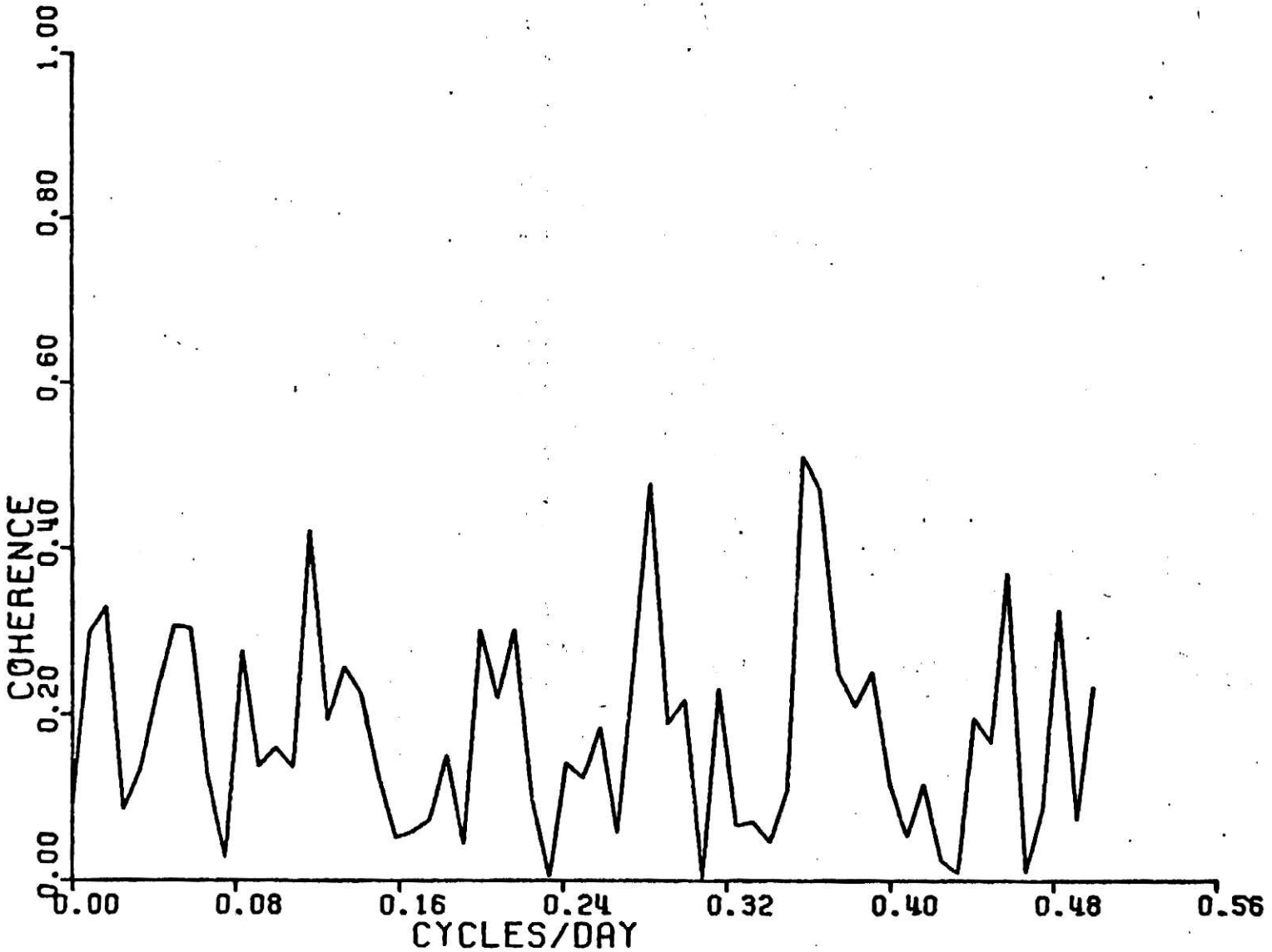

Fig. 3.19 Coherency spectra - flow rate and temp. (before alignment).

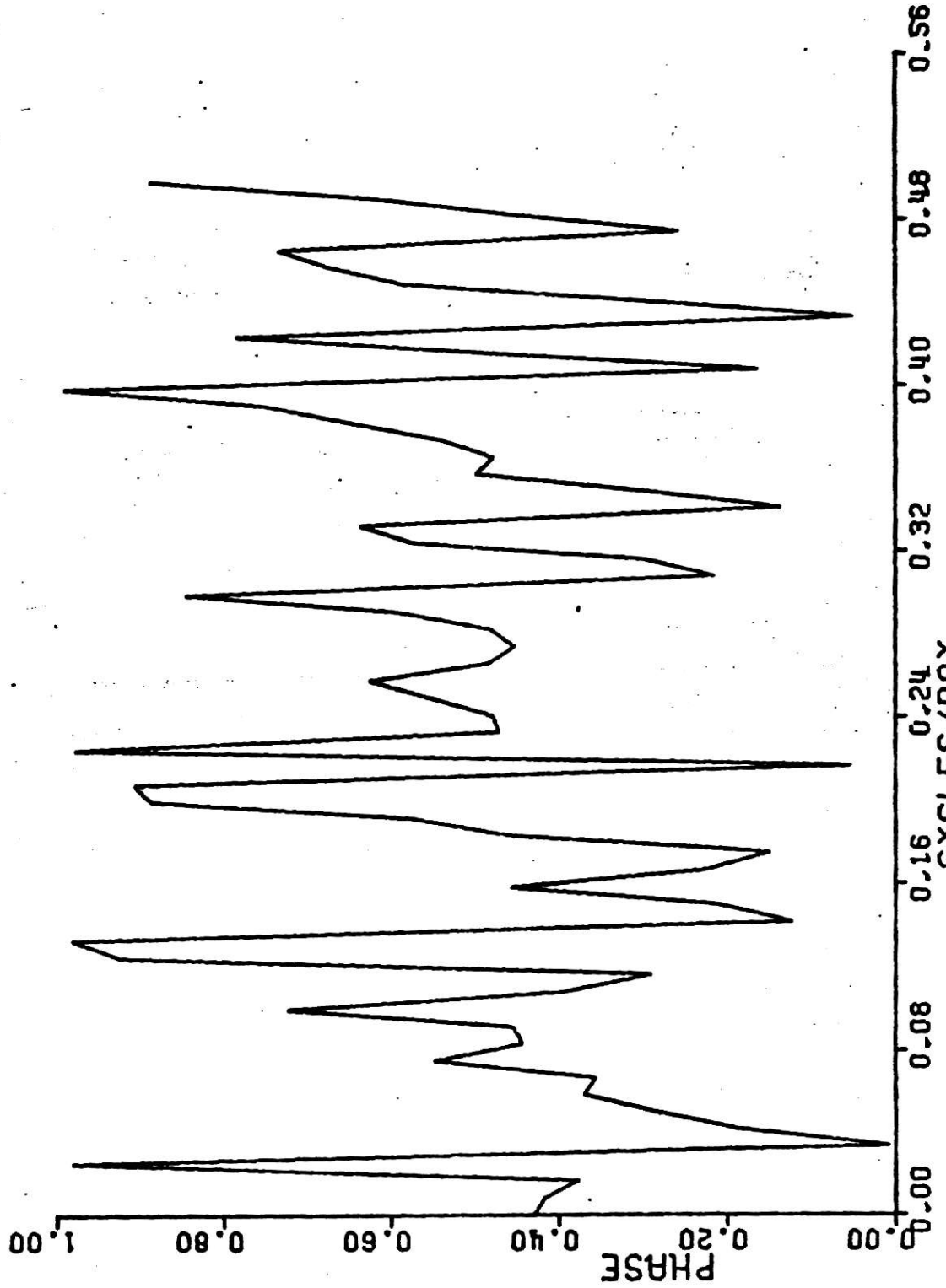

Fig. 3.20 Phase spectra - flow rate and temp. (before alignment).

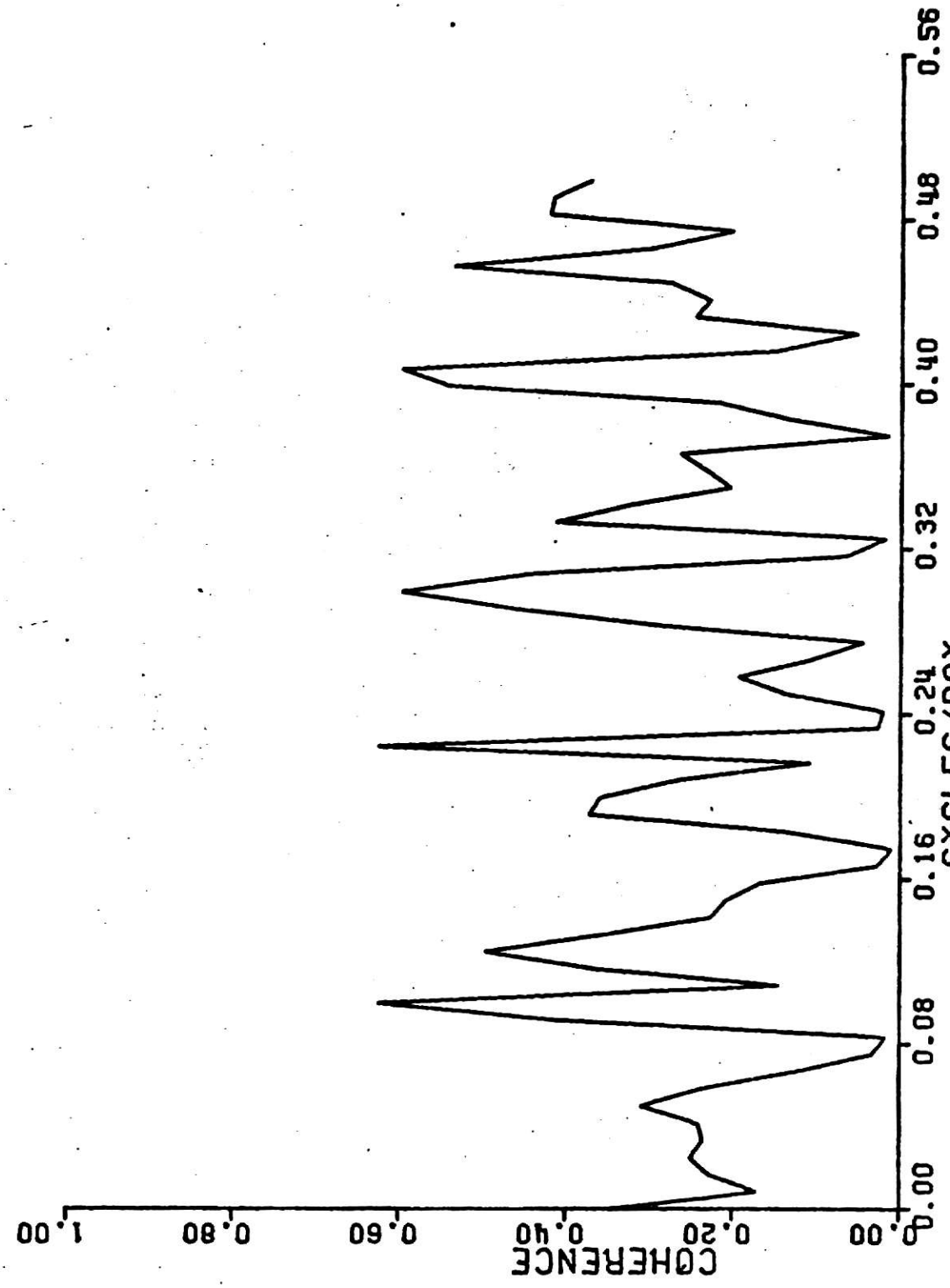

Fig. 3.21 Coherency spectra - flow rate and temp. (after alignment).

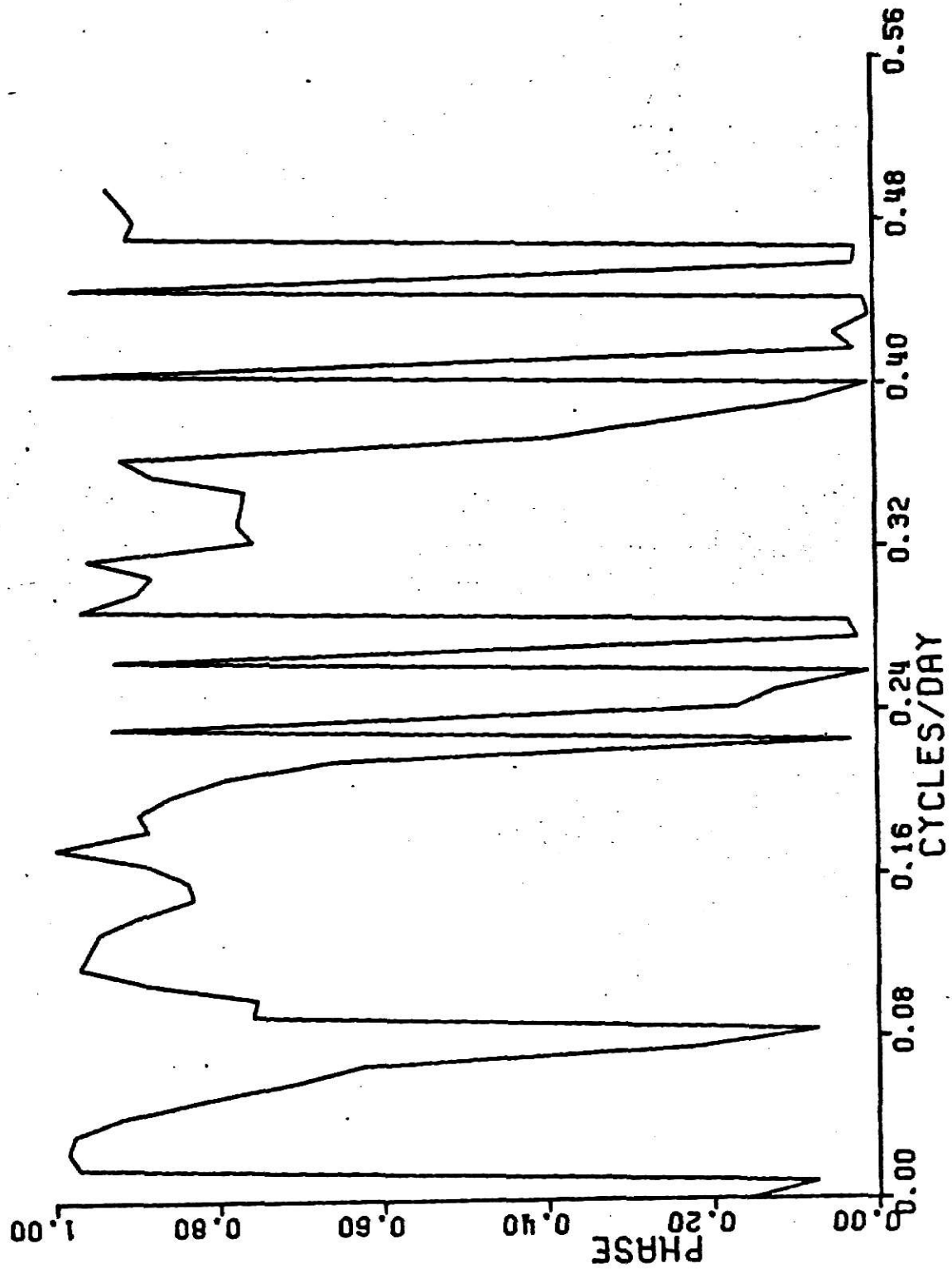

Fig. 3.22 Phase spectra - flow rate and temp. (after alignment).

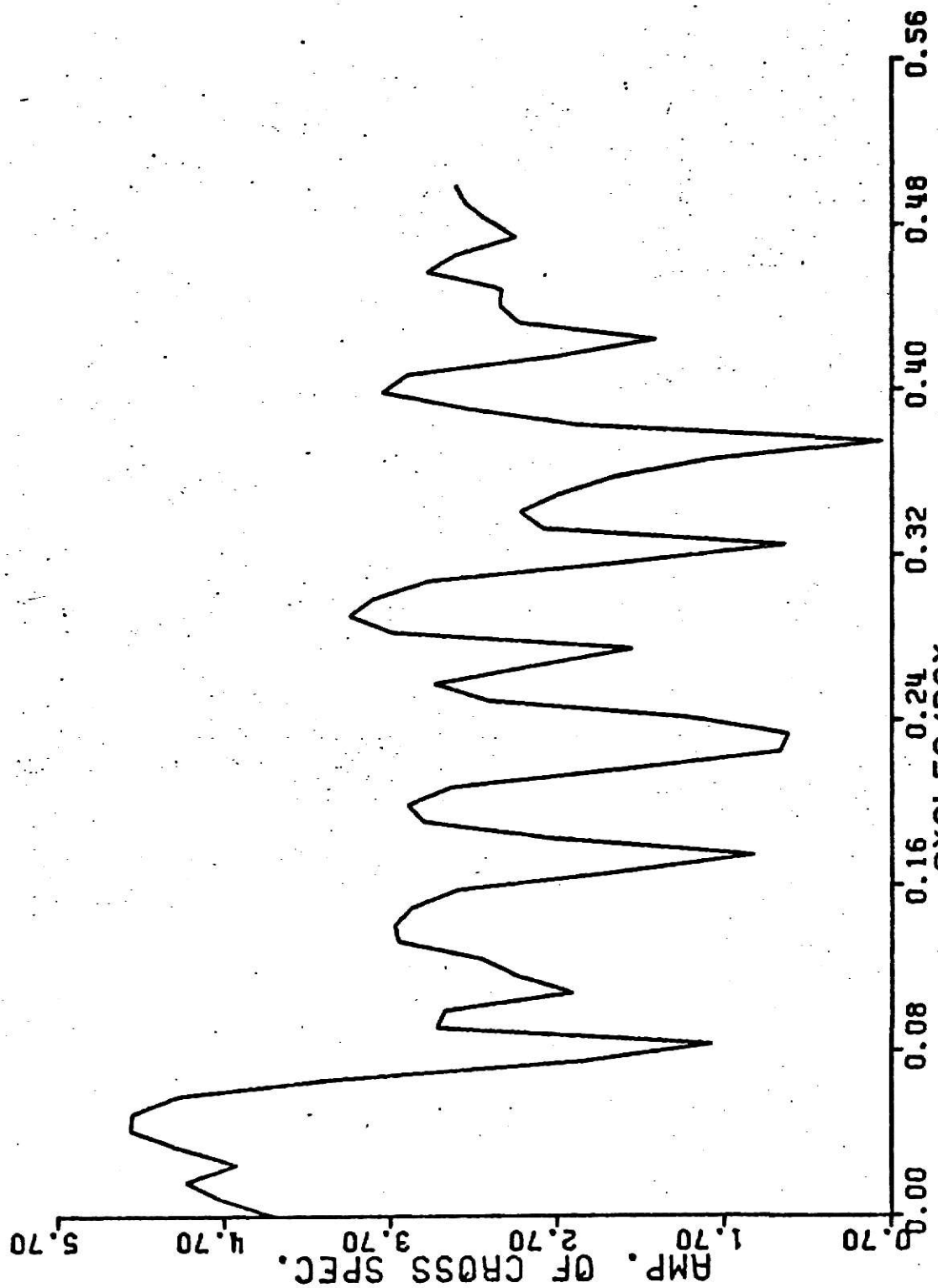

Fig. 3.23 Amplitude of cross-spectra - flow rate and temp.

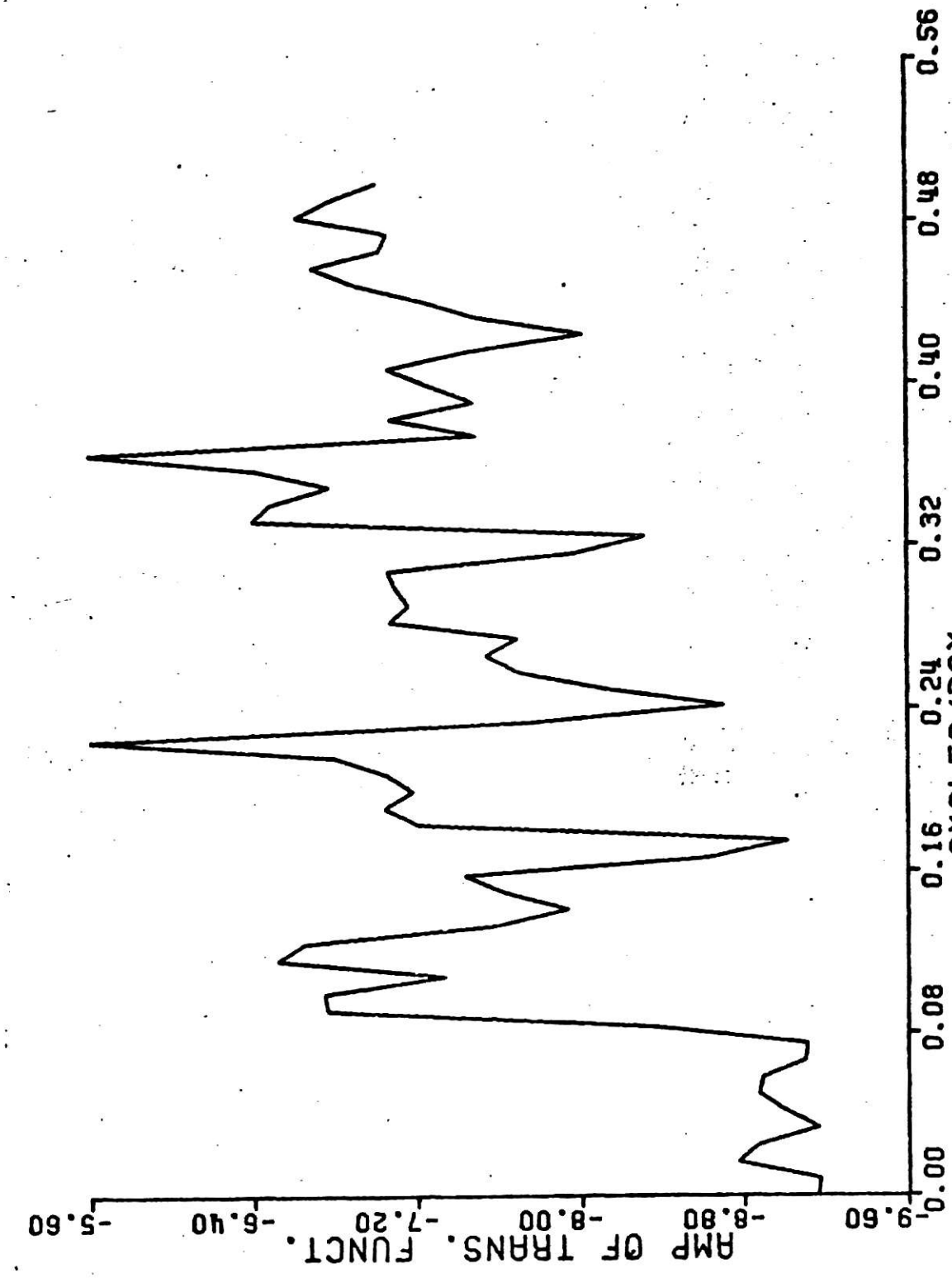

Fig. 3.24 Amplitude of transfer function from flow rate to temp.

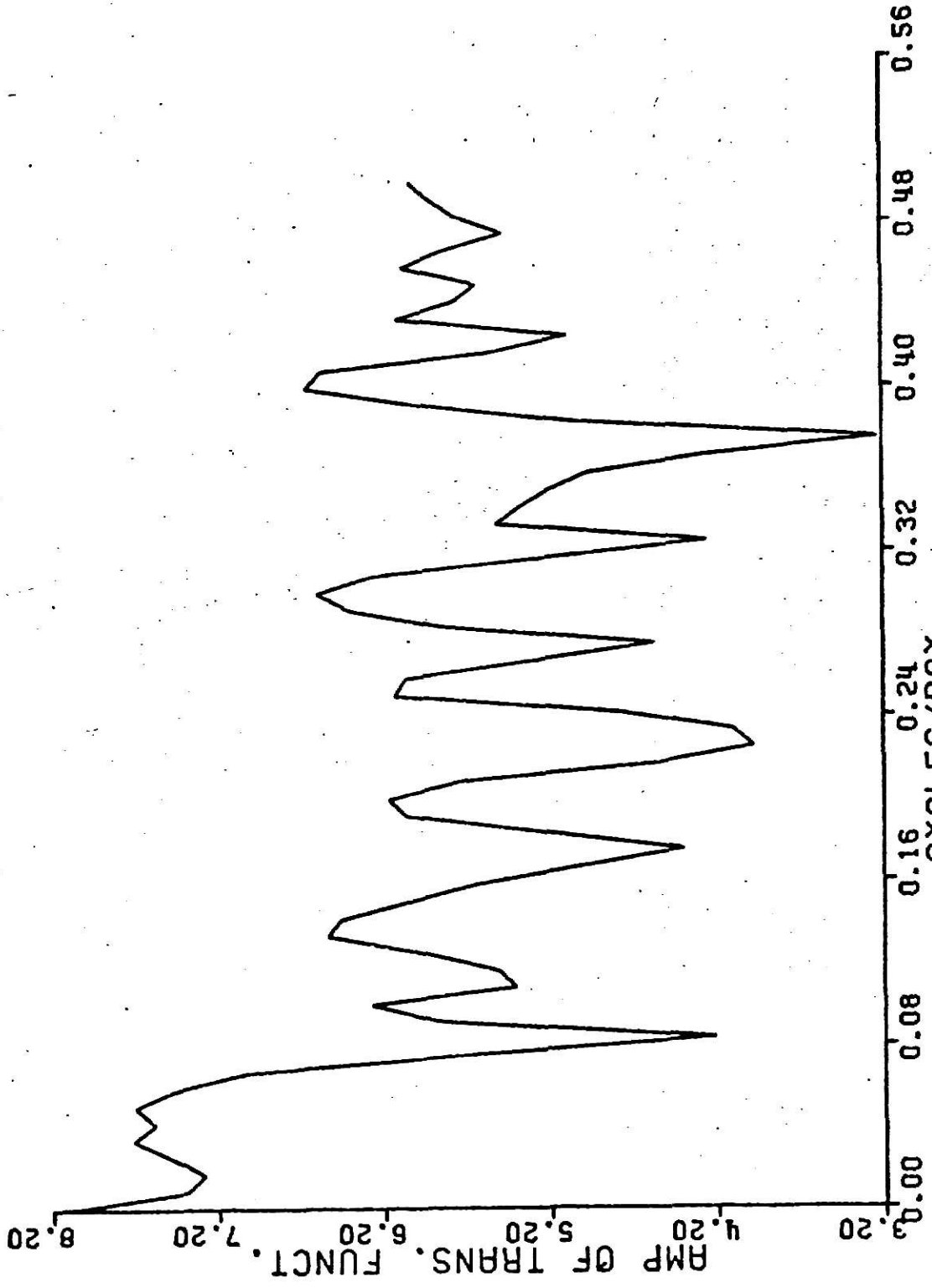

Fig. 3.25 Amplitude of transfer function from temp. to flow rate.

and there is a growth in the concentration of salts to account for increase in specific conductance. A low transfer function from flow rate to specific conductance is observed at all frequencies.

(c) Flow rate and temperature:

The auto-correlation function of the two pollutants shown in Figures 2.5 and 2.13 suggest the use of filtered data for a cross spectral study. As before, a difference filter was used to remove the low frequency components. The cross-correlation plot of the differenced data is shown in Figure 3.18. It has a maximum absolute value at a lag of 30 days, indicating that there is a delay of 30 days between the two processes. Again, alignment was necessary before computing the cross spectral estimates. Figures 3.19 and 3.20 show the coherency and phase spectra respectively before alignment and the various spectra after alignment are shown in Figures 3.21 thru 3.25. Low coherencies are observed at all the frequencies suggesting that there is no significant correlation between temperature and flow rate. Fluctuations in both the pollutants are in phase with each other at zero frequency and 0.0082 cycles/day (120 days period).

The cross spectral analysis of Ontario data brings forth the following points about the behaviour of the three pollutants:

(1) There is a strong correlation between the long range fluctuations of temperature and specific conductance. This included all variations with period more than 120 days. The exact nature of this relationship could not be found due to lack of availability of longer data.

(2) There is a low correlation between flow rate and specific conductance at low frequencies. It means that long range fluctuations in specific conductance are either due to their correlation with temperature variations or due to the growth in the concentration of salts in the water. As before, a longer data record is necessary to arrive at a definite conclusion.

CHAPTER IV

PARAMETRIC TIME SERIES MODELING

Chapter 2 dealt with the analysis of time series using spectral analysis. Another important approach for building stochastic models for discrete time series in the time domain is parametric modeling. One useful characteristic of this technique is that good models can be built using only a small number of parameters and efficient forecasting of future values can be made using them. This chapter provides a brief survey of parametric time series modeling and its application to the Ontario River data.

4.1 Classification of Models: The various models can be broadly classified into the following catagories [39]

(a) Autoregressive Model, AR(p):

In this model, the current value of the data is expressed as a weighted sum of the previous observations plus a random shock.

Let $x_1, x_2, \ldots, x_1 \ldots x_N$ be the N observations of a process and $\tilde{x}_1 = x_1 - \mu$, $\tilde{x}_2 = x_2 - \mu$ etc. be the deviations of the observations from their mean. Then the process at time 't' is defined as

$$x_t = \phi_1 \tilde{x}_{t-1} + \phi_2 \tilde{x}_{t-2} + \ldots + \phi_p \tilde{x}_{t-p} + a_t \qquad (4.1)$$

This is called an autoregressive process of order p. a_t is the random shock assumed to be independently normally distributed variable with zero mean and variance σ_a^2.

This model can be abbreviated in the form

$$\phi(B)\ \tilde{x}_t = a_t \tag{4.2}$$

where $\phi(B) = (1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p)$

and B is the backward shift operator such that

$$Bx_t = x_{t-1}$$

$$B^2 x_t = x_{t-2}$$

and $B^p x_t = x_{t-p}$

As given in (4.1), this model contains p+1 parameters which can be estimated from the available data.

In practice, models of first or second order are sufficient to explain most of the time series.

(b) Moving Average Model, MA(q):

This model expresses the current observation as the finite weighted sum of the previous random shocks. A moving average model of order q can be written as

$$\tilde{x}_t = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \theta_q a_{t-q} \tag{4.3}$$

or $$\tilde{x}_t = \theta(B) a_t \tag{4.4}$$

where $\theta(B) = (1 - \theta_1 B - \theta_2 B^2 - \ldots \theta_q B^q)$

It has q+1 unknown parameters to be estimated from the available data. Of particular importance are the models of first and second order which

generally suffice for all practical problems.

(c) Mixed Autoregressive Moving Average Models, ARMA (p,q):

Sometimes it becomes necessary to combine both autoregressive and moving average models into one mixed model to obtain certain desired characteristics. A general mixed model of order (p,q) may be written as

$$\tilde{x}_t = \phi_1 \tilde{x}_{t-1} + \ldots + \phi_p \tilde{x}_{t-p} + a_t - \theta_1 a_{t-1} \ldots - \theta_q a_{t-q} \qquad (4.5)$$

or

$$\phi(B)\ \tilde{x}_t = \theta(B)\ a_t. \qquad (4.6)$$

The models described above are applicable only to stationary time series. In practice, however, a non-stationary series is not uncommon. This is particularly true for water pollution problems. For use with non-stationary series, a general Autoregressive Integrated Moving Average model is used. It essentially consists of transforming a non-stationary series into a stationary series using a difference filter.

(d) Autoregressive Integrated Moving Average Model, ARIMA (p,d,q)!

The general form of ARIMA (b,d,q) is given as

$$\tilde{\omega}_t = \phi_1 \tilde{\omega}_t + \ldots + \phi_p \tilde{\omega}_{t-p} + a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \theta_q a_{t-q} \qquad (4.7)$$

where $\omega_t = \nabla^d x_t$

d = order of differencing necessary to produce a stationary series. The indication of non-stationarity in the time series is given by the

auto-correlation function. This property will be described later in Section 4.3.

In general, first or second order differencing is sufficient to produce stationarity in the time series.

Another important class of models is the seasonal models which takes into account any seasonal fluctuation in the time series.

(e) Seasonal Models:

Let S be the period of cyclic fluctuation and $B^s x_t = x_{t-s}$. Then the seasonal model can be defined as

$$\left.\begin{aligned} \phi_P(B^s)\ \nabla_s^D x_t &= \theta_Q(B^s)\alpha_t \\ \phi_p(B)\ \nabla^d \alpha_t &= \theta_q(B)a_t \end{aligned}\right\} \qquad (4.8)$$

The first part of the model links the terms 's' time intervals apart and the second part links the consecutive terms.

This model can also be written in multiplicative form as

$$\phi_p(B)\ \phi_P(B^s)\nabla^d\ \nabla_s^D x_t = \theta_q(B)\ \theta_Q(B^s)a_t \qquad (4.9)$$

4.2 Stationarity and Invertibility Conditions

Certain limitations are imposed on the θ and ϕ weights of a moving average and autoregressive processes respectively to ensure stationarity and invertibility conditions. It can be seen that an auto-regressive model can be written as a moving average type and vice versa. e.g. AR(1) model is written as

$$\tilde{x}_t = \phi_1 \tilde{x}_{t-1} + a_t$$

or $\phi(B)\, \tilde{x}_t = a_t$

$$\tilde{x}_t = \frac{1}{\phi(B)}\, a_b = \phi^{-1}(B) a_t$$

or $$\tilde{x}_t = \sum_{j=0}^{\infty} \phi_1^j\, a_{t-j}$$

$$\text{Let } \psi(B) = (1 - \phi_1 B)^{-1} = \sum_{j=0}^{\infty} \phi_1^j\, B^j \tag{4.10}$$

ψ's are called the pure moving average weights.

The variance of this process can be obtained as

$$\gamma_0 = \sigma_a^2 \sum_{j=0}^{\infty} \psi_j^2$$

Hence for the variance to be finite, ψ's must converge fast enough. This can be achieved by taking $|\phi_1| \leq 1$. This is the limitation on ϕ for a AR(1) process to be stationary.

Considering a MA(1) process,

$$\tilde{x}_t = (1 - \theta_1 B) a_t$$

$$a_t = \frac{\tilde{x}_t}{(1 - \theta_1 B)} = (1 - \theta_1 B)^{-1}\, \tilde{x}_t$$

$$\text{Let } \pi(B) = (1 - \theta_1 B)^{-1} = \sum_{j=0}^{\infty} \theta_1^j\, B^j \tag{4.11}$$

π's are called the pure auto-regressive weights.

It is desirable that π's form a convergent series in (4.11) otherwise it would imply that current observation $\tilde{x}_t$ depends on previous observations $x_{t-1}, x_{t-2}, \ldots, x_{t-j}$, with the weights increasing as j increases. To ensure convergence of π's, it is necessary for MA(1) process to have $|\theta_1| \leq 1$. This is called the invertibility condition for a moving average process.

Similar limitations on the weights of higher order models can be obtained.

The model building procedure consists essentially of three steps:

(i) Identification: It is the stage where the data is analyzed to obtain information about the kind of model (e.g. AR(1), MA(1) or ARMA(1,1) etc.) to be selected for further investigation. A rough estimate of parameters is also obtained.

(ii) Estimation of Parameters: The parameters of the candidate model are determined by least square methods using the rough estimate obtained in the identification stage as starting points.

(iii) Diagnostic Checking: The residuals obtained by using the candidate model are subjected to statistical testing to check if the model should be accepted as satisfactory. In case of inadequate model, the procedure returns to step 1 and a new candidate model is entertained for acceptance.

Next three sections deal with a brief discussion about each of the three steps.

4.3 Identification:

It involves the selection of a particular model to be entertained as

a candidate for acceptance. Two functions that are useful for identification purposes are the auto-correlation and (acf) and the partial auto-correlation (pacf). The auto-correlation function has been defined in Chapter 2. For an auto-regressive process, the auto-correlation function satisfies the difference equation,

$$\rho_k = \phi_1\rho_{k-1} + \phi_2\rho_{k-2} + \ldots + \phi_p\rho_{k-p}, \qquad k > 0 \tag{4.12}$$

i.e. for first order auto-regressive process,

$$\rho_k = \phi_1\rho_{k-1} \qquad k > 0$$

which has a solution,

$$\rho_k = \phi_1^k\rho_0 = \phi_1^k \tag{4.13}$$

It has already been shown that for an auto-regressive process to be stationary, $|\phi_1| \leq 0$. This together with (4.13) implies that the auto-correlation function of an auto-regressive process decays exponentially. Similar results can be derived for higher order auto-regressive processes. Unlike an auto-correlation function, the partial auto-correlation function of an AR(p) process has a cut off after a lag of p.

A general recursive relationship to obtain partial auto-correlation can be given as [40]

$$\rho'_{k+1} = \frac{\rho_{k+1} - \sum_{j=1}^{k} \rho'_{k,j}\,\rho_{k+1-j}}{1 - \sum_{j=1}^{k} \rho'_{k,j}\,\rho_j} \tag{4.14}$$

$$\rho'_{k+1,j} = \rho'_{k,j} - \rho'_{k+1}\rho'_{k,k-j+1}, \quad (j = 1, 2, \ldots, k) \tag{4.15}$$

where $\rho'_0 = 1$

and $\rho'_1 = \rho_1$

Intuitively, it can be seen that an AR(p) process means that the current observation depends on p previous observations only, hence once these are known, the partial auto-correlations with the earlier observations should be zero. But for a moving average process, the partial auto-correlations tail off due to the invertibility condition.

The auto-correlation function of a moving average process satisfies the relationship,

$$\rho_k = \frac{-\theta_k + \theta_1\theta_{k+1} + \ldots + \theta_{q-k}\theta_q}{1 + \theta_1^2 + \theta_2^2 + \ldots + \theta_q^2}, \qquad k = 1,2, \ldots, q \tag{4.16}$$

$$= 0 \qquad k > q$$

It implies that the auto-correlation function of an MA(q) process has a cut off after lag q whereas its partial auto-correlation tails off.

These properties of AR(p) and MA(q) processes are extremely useful for identification of a model. Table 4.1 summarizes these properties.

The relationships given by expressions (4.12) and (4.16) are also used for obtaining initial estimates of the parameters of the candidate model. Charts given in [39] and [40] provide good estimates for MA(1), MA(2), AR(1), AR(2), AR(3), ARMA(1,1), ARMA(2,1) and ARMA(1,2) processes.

Table 4.1 Properties of MA(q), AR(p) and ARMA(p,q) processes.

Model	Autocorrelation function	Partial auto-correlation function
MA(q)	cuts off after q lags	tails off
AR(p)	tails off	cuts off after p-lags
ARMA(p,q)	tails off	tails off

An indication of non stationarity in a time series is given by the failure of the auto-correlation function to die off quickly. Differencing of data is carried out to achieve stationarity and it is assumed that the non-stationarity has been removed from the series when the auto-correlation function of the differenced data dies off quickly. After stationarity has been achieved, results shown in Table (1) are applicable for identification of a tentative model.

The identification of a seasonal model is done in a similar manner. The presence of a seasonal component is indicated by a high auto-correlation at lags corresponding to the seasonal period and its integral multiples.

In the earlier expressions, use of estimated auto-correlations has been made whenever needed. But the theoretical auto-correlations differ from estimated auto-correlations and an expression for estimation of the variance of auto-correlations is given by [40],

$$\text{Var } (\hat{\rho}_k) = \frac{1}{n} \left(1 + 2 \sum_{j=1}^{k} \hat{\rho}_j^2\right) \tag{4.17}$$

This can be used to find whether $\hat{\rho}_k$ is effectively zero.
The estimation of variance of partial auto-correlation function on the hypothesis that the process is AR(k-1) is given by [40].

$$\text{Var } (\rho_k') = \frac{1}{n-k}$$

After the candidate model has been identified using the procedure described above, more accurate estimation of the model parameters is needed.

4.4 Estimation

In the case of a moving average of ARMA (p,q) processes, the parameters occur in non linear forms. In this study, Marquardt's algorithm for non linear least squares estimation of parameters was used.

A stationary ARMA(p,q) model can be written in the form,

$$a_t = \tilde{\omega}_t - \phi_1 \tilde{\omega}_{t-1} - \ldots - \phi_p \tilde{\omega}_{t-p} + \theta_1 a_{t-1}$$

$$+ \theta_2 a_{t-2} + \ldots + \theta_q a_{t-q} \qquad (4.18)$$

The sum of squares function of errors a_t is given by

$$S(\phi,\theta) = \sum_{t=1-Q}^{n} [a_t/\phi,\theta,\omega]^2 \qquad (4.19)$$

Marquardt's algorithm is based on a compromise between the linearization method and the steepest descent method. The linearization method uses the initial guess values to start the process and then expands the function in the vicinity of guess values as

$$[a_t] = [a_{t,0}] - \sum_{i=1}^{k} (\beta_i - \beta_{i,0})\, x_{i,t}$$

where $x_{i,t} = \dfrac{-\partial [a_t]}{\partial \beta_i} \Big|_{\beta=\beta_0}$

or in Matrix form

$$[a_0] = X\,(\beta - \beta_0) + [a] \qquad (4.20)$$

where [a] and $[a_0]$ are column vectors with (n+Q) rows. The adjustments $\beta - \beta_0$, which minimize $S(\phi,\theta) = [a]'[a]$ are obtained by least square methods. The adjusted values of the parameters 'β' are then obtained which are again used as new guess values and the whole process is repeated.

The steepest descent method uses an iterative approach to find minimum error sum of squares by moving from an initial point $[\phi_1,\phi_p,\theta_1,\theta_q]$ along the vector with components

$$\frac{-\partial S(\phi\theta)}{\partial \phi_1} , \frac{\partial S(\phi,\theta)}{\partial \phi_p} \quad . \quad . \quad . \quad \frac{-\partial S(\phi,\theta)}{\partial \theta_q}$$

whose value changes continuously as the path is followed.
In Marquard's algorithm, both of these techniques have been combined to provide quick convergence.
It may be pointed out here, that the initial guess values used should be as reliable as possible to obtain quick convergence and reliable solution.

4.5 Diagnostic checking

Having identified the model and the parameters estimated, diagnostic checks are then applied to the model to test its adequacy.
Diagnostic checks are applied to the residuals in the form of an auto-correlation check, lack of fit test, and cumulative periodogram check.
(a) Auto-correlation check: It has been shown by Anderson that the estimated auto-correlations of the residuals would be uncorrelated and distributed approximately normally with zero mean and variance n^{-1} and hence with a standard error $\frac{1}{\sqrt{n}}$.

It was shown later on by Box and Piere, that this estimate of standard error would be satisfactory for high lags but would be unsatisfactory at low lags [39].

(b) A portmanteau lack of fit test:

This test considers the effect of auto-correlations as a whole to indicate the inadequacy of the model. Let r_k be auto-correlations of the residual, k = 1,2, ..., M. It can be shown that if the model is adequate then

$$Q = n \sum_{k=1}^{M} r_k^2(\hat{a}) \tag{4.20}$$

follows a chi-square distribution with $\nu = (M - p - q - P' - Q')$ degrees of freedom, where $n = (N - d - D * IS)$. The Q-value can be calculated from (4.20) and tested against the tabulated value of χ^2_ν.

A cumulative periodogram test may be made for checking the presence of some cyclic fluctuation in the residuals. In this study, spectral analysis was used for testing the presence of cyclic fluctuation. Theoretically, the spectrum of the residuals should be a horizontal straight line.

4.6 Forecasting

After a model has been found statistically adequate, it is desirable that it can be used for forecasting future values. Consider an ARMA(p,q) model, observation x_{t+k} can be written as

$$x_{t+k} = \phi_1 x_{t+k-1} + \ldots + \phi_p x_{t+k-p}$$

$$+ a_{t+k} - \theta_1 a_{t+k-1} - \theta_2 a_{t+k-2} \ldots - \theta_q a_{t+k-q}$$

Let $x_t(k)$ denote the expected value of x_{t+k}, given observations through time t, then we have

$$\hat{x}_t(k) = \phi_1 \hat{x}_t(k-1) + \phi_2 \hat{x}_t(k-2) + \ldots + \phi_p \hat{x}_t(k-p)$$

$$- \theta_1 a_{t+L-1} - \ldots - \theta_q a_{t+L-q} + a_{t+\ell}$$

To calculate the conditional expectations which occur in the above expression, the following observations are noted. The x_{t-j} which have already happened at time t are left unchanged, the x_{t+j} which have not yet happened one replaced by their forecast values at origin t'. The a_{t-j} which have happened are available from $x_{t-j} - \hat{x}_{t-j-1}(1)$

The a_{t+j} which have not yet happened are replaced by zeros.

Thus this equation can be used to forecast values one step ahead.

The 95% confidence interval for $\hat{x}_t(k)$ can be given as

$$\hat{x}_t(k) \pm 1.96 \sqrt{\sigma^2 [1 + \psi_1^2 + \psi_2^2 + \ldots + \psi_{k-1}^2]}$$

where ψ's are the pure moving average weights for ATMA(p,q) process. This forecasting procedure may not be useful for predicting values in the distant future as the future unknown values in the forecast equation are replaced by their expected values instead of the actual observed values which are unknown at the origin. This may be improved by updating the forecast at every step in the future.

4.7 Analysis of Ontario River data

(a) Temperature: Figures 4.1, 4.2 show the autocorrelation and partial

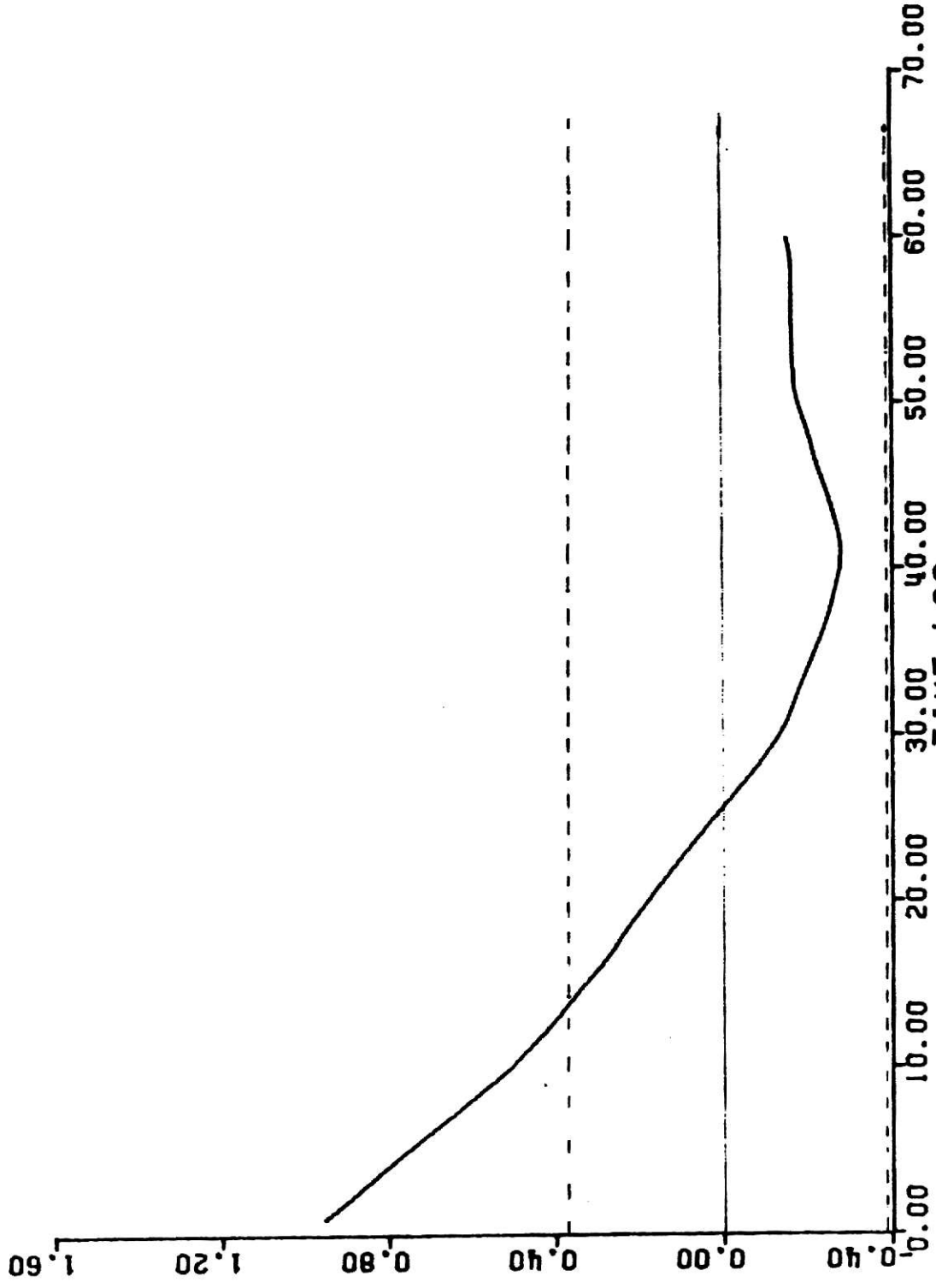

Fig. 4.1 Autocorrelation of original temperature data.

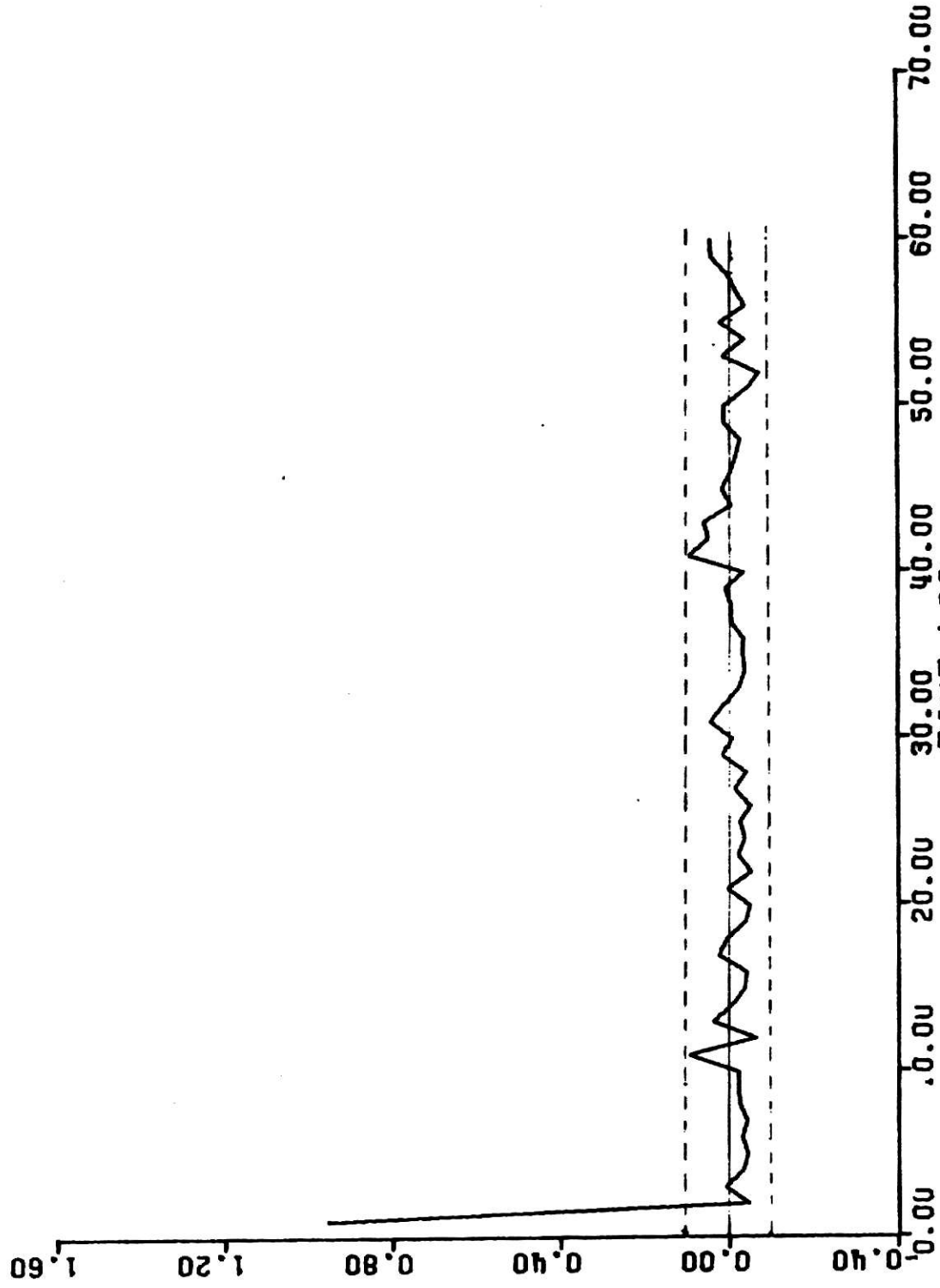

Fig. 4.2 Partial autocorrelation of original temperature data.

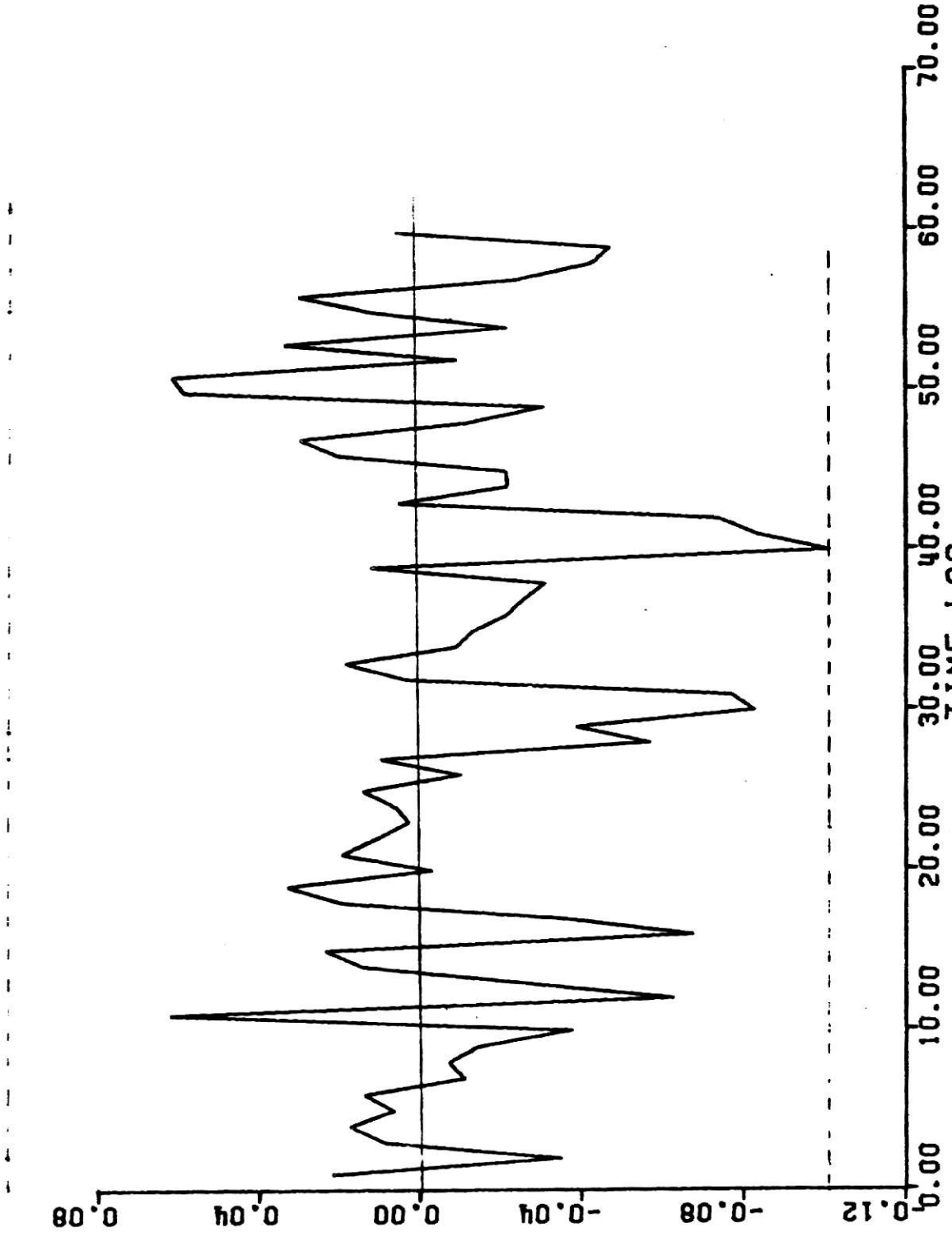

Fig. 4.3 Autocorrelation of temperature data (∇x).

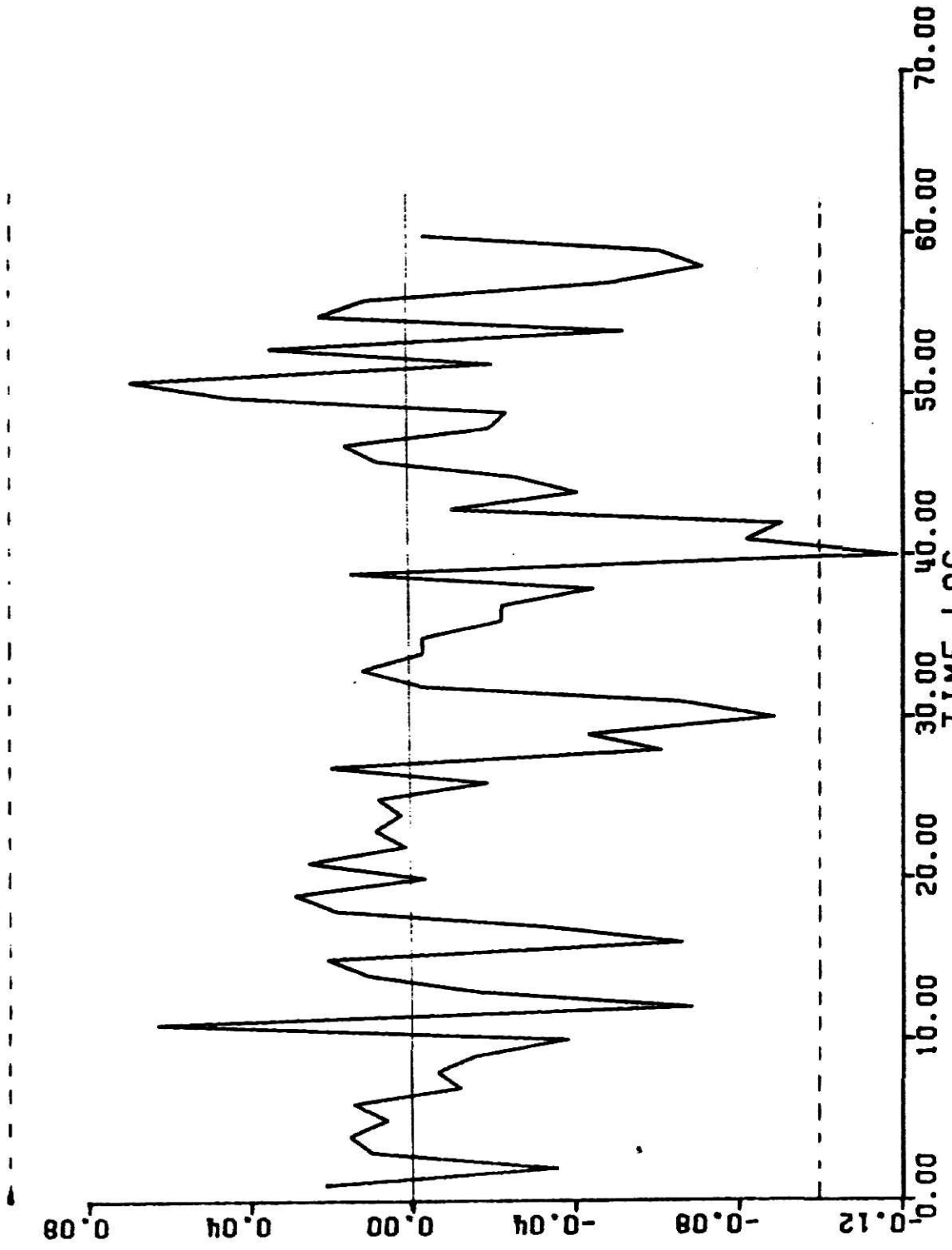

Fig. 4.4 Partial autocorrelation of temperature data (Vx).

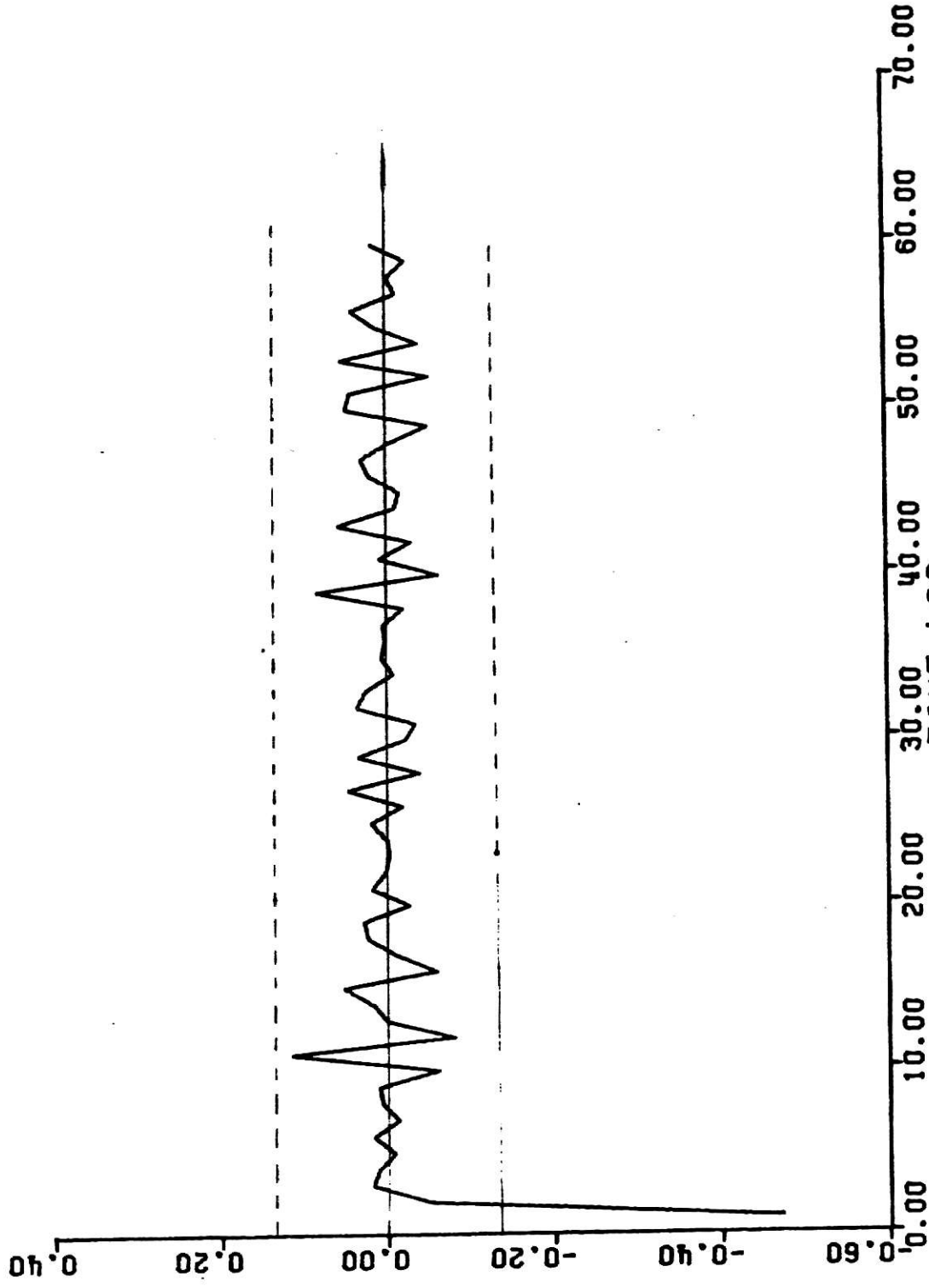

Fig. 4.5 Autocorrelation of temperature data (v^2x)

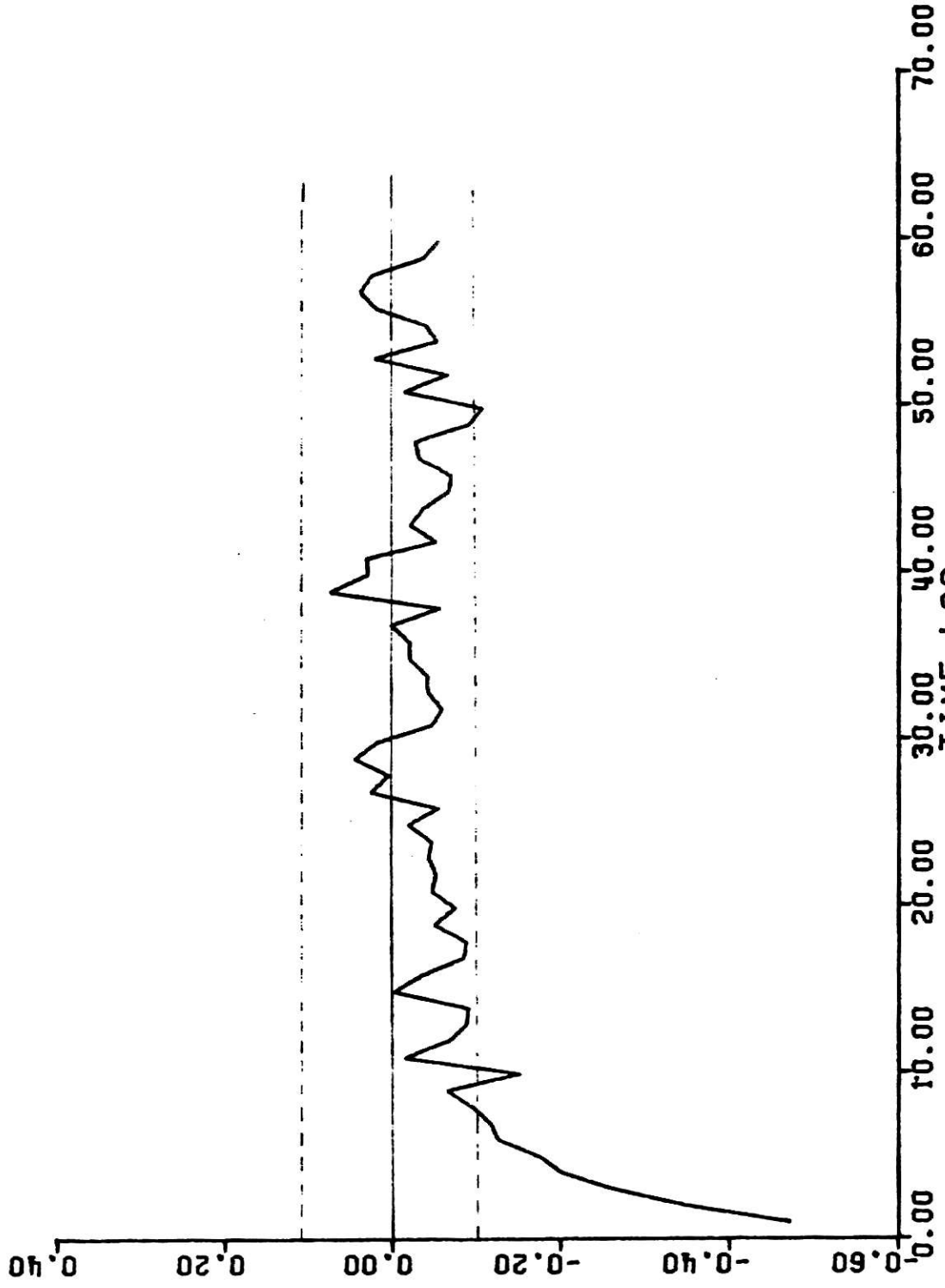

Fig. 4.6. Partial autocorrelation of temperature data ($\nabla^2 x$).

autocorrelation plot of raw temperature data. It is seen that the auto-correlation function tails off gradually but the partial auto-correlation function has a cut off afterone lag suggesting a tentative ARMA(1,0,0) model. Corresponding plots for first differenced data are shown in Figures 4.3 and 4.4. Here both the plots have an immediate cut off at lag one implying that the first differenced data behaves as white noise. A tentative ARMA(0,1,1) model may be tried for this data. The plots for second differenced data are shown in Figures 4.5 and 4.6. In this case the auto-correlation plot has a cut off after one lag whereas the partial auto-correlation function tails off rapidly. This suggests a tentative ARMA(0,2,1) model.

Hence the plots of the auto-correlation and partial auto-correlation function suggest three tentative models

(1) ARMA(1,0,0) (2) ARMA(0,1,1) (3) ARMA(0,2,1).

Initial estimates of parameters for these models may be obtained using either (4.12) and (4.16) or the charts given in [40].

In this case, estimates were obtained from charts given in [40] and are given as

ARMA(1,0,0) $\phi_1 = 0.95$

ARMA(0,1,1) $\theta_1 = 0.02$

ARMA(0,2,1) $\theta_1 = 0.70$

These values were then used as starting points for the nonlinear least squares estimation procedure. The value of the parmeters obtained by this procedure are

ARMA(1,0,0) $\phi_1 = 0.9522$

ARMA(0,1,1) $\theta_1 = -0.023$

ARMA(0,2,1) $\theta_1 = 0.988$

Diagnostic checks were then applied to the residuals of the sample data for all these models.

Tables 4.2, 4.3, 4.4 show the autocorrelations for residuals together with the ratio

$$\rho_k / \text{S.D.}(\rho_k)$$

It is seen that in all cases the estimated autocorrelations fall within 2σ limits of true value (i.e. 0). Hence all the auto-correlations are effectively zero.

Q values as given by (4.20) were calculated for chi square tests

Model	Q	d.f.	Tabulated $\chi^2_{0.90}$
ARMA(1,0,0)	32.76	59	63.2
ARMA(0,1,1)	29.12	59	63.2
ARMA(0,2,1)	7.54	24	33.2

It is seen that in all cases the tabulated value is well above the calculated Q value, hence there seems to be no reason to doubt the adequacy of models.

Thus the models for temperature for the Ontario River can be given as

(1) $\tilde{x}_t = 0.95^2 \tilde{x}_{t-1} + a_t$

(2) $\nabla x_t = (1 + 0.023B) a_t$

Table 4.2 Autocorrelation of residuals for temperature. Model ARMA(1,0,0)

Lag k	Autocorrelation ρ_k	ρ_k/S.D.(ρ_k)
1	0.03276	0.62596
2	-0.02390	-0.45614
3	0.01842	0.35136
4	0.02616	0.49885
5	0.01434	0.27327
6	0.02100	0.40001
7	-0.00491	-0.09344
8	-0.00161	-0.03071
9	-0.00991	-0.18875
10	-0.02979	-0.56723
11	0.06805	1.29455
12	-0.05617	-1.06358
13	-0.01959	-0.36977
14	0.01846	0.34836
15	0.02732	0.51546
16	-0.06339	-1.19485
17	-0.03288	-0.61745
18	0.02150	0.40326
19	0.03442	0.64540
20	-0.00123	-0.02295
21	0.02060	0.38578
22	0.01159	0.21695
23	0.00288	0.05386
24	0.00631	0.11802
25	0.01324	0.24778
26	-0.01151	-0.21540
27	0.00785	0.14695
28	-0.05854	-1.09527
29	-0.04144	-0.77284
30	-0.08557	-1.59326
31	-0.07945	-1.46911
32	-0.00074	-0.01361
33	0.01423	0.26149
34	-0.01333	-0.24491
35	-0.01791	-0.32907
36	-0.02648	-0.48650
37	-0.03060	-0.56174
38	-0.03646	-0.66884
39	0.00602	0.11025
40	-0.10626	-1.94673
41	-0.08865	-1.60735
42	-0.07970	-1.43506
43	-0.00132	-0.02370
44	-0.02777	-0.49720
45	-0.02738	-0.48989
46	0.01422	0.25434
47	0.02411	0.43103
48	-0.01619	-0.28929
49	-0.03526	-0.62986

Table 4.2 Autocorrelation of residuals for temperature. Model ARMA(1,0,0) (continued)

Lag k	Autocorrelation ρ_k	ρ_k/S.D.(ρ_k)
50	0.05280	0.94216
51	0.05552	0.98830
52	-0.01434	-0.25452
53	0.02833	0.50297
54	-0.02657	-0.47131
55	0.00593	0.10515
56	0.02498	0.44275
57	-0.02860	-0.50667
58	-0.04745	-0.84019
59	-0.05151	-0.91028
60	0.00090	0.01580

Table 4.3 Autocorrelation of residuals for temperature Model ARMA (0,1,1)

Lag k	Autocorrelation ρ_k	ρ_k/S.D.(ρ_k)
1	-0.00138	-0.02627
2	-0.03544	-0.67608
3	0.00880	0.16777
4	0.01715	0.32674
5	0.00542	0.10320
6	0.01381	0.26307
7	-0.01141	-0.21728
8	-0.00666	-0.12687
9	-0.01063	-0.20249
10	-0.03889	-0.74049
11	0.06428	1.22184
12	-0.06391	-1.20986
13	-0.02396	-0.45182
14	0.01415	0.26672
15	0.02457	0.46293
16	-0.06743	-1.26986
17	-0.03511	-0.65823
18	0.01930	0.36132
19	0.03305	0.61872
20	-0.00440	-0.08221
21	0.01903	0.35579
22	0.00991	0.18518
23	0.00241	0.04502
24	0.00537	0.10034
25	0.01388	0.25940
26	-0.01089	-0.20357
27	0.01118	0.20881
28	-0.05677	-1.06064
29	-0.03600	-0.67054
30	-0.08037	-1.49506
31	-0.07524	-1.39116
32	0.00445	0.08184
33	0.01814	0.33356
34	-0.00936	-0.17206
35	-0.01249	-0.22962
36	-0.02045	-0.37587
37	-0.02489	-0.45728
38	-0.03107	-0.57054
39	0.01506	0.27638
40	-0.10021	-1.83817
41	-0.07954	-1.44571
42	-0.07188	-1.29893
43	0.00730	0.13136
44	-0.02183	-0.39274
45	-0.02115	-0.38028
46	0.01996	0.35881
47	0.02945	0.52912
48	-0.01109	-0.19918
49	-0.03246	-0.58265

Table 4.3 Autocorrelation of residuals for temperature Model ARMA (0,1,1) (continued)

Lag k	Autocorrelation ρ_k	ρ_k/S.D.(ρ_k)
50	0.05733	1.02824
51	0.06034	1.07903
52	-0.01142	-0.20356
53	0.03405	0.60686
54	-0.02309	-0.41105
55	0.01149	0.20449
56	0.02974	0.52914
57	-0.02383	-0.42374
58	-0.04094	-0.72758
59	-0.04672	-0.82910
60	0.00767	0.13580

Table 4.4 Autocorrelation of residuals for temperature

Model ARMA (0,2,1)

Lag k	Autocorrelation ρ_k	ρ_k/S.D.(ρ_k)
1	0.02017	0.38431
2	-0.03661	-0.69721
3	0.00708	0.13458
4	0.01624	0.30882
5	0.00488	0.09271
6	0.01237	0.23527
7	-0.01254	-0.23830
8	-0.00844	-0.16042
9	-0.01232	-0.23414
10	-0.03843	-0.73035
11	0.06154	1.16754
12	-0.06337	-1.19802
13	-0.02541	-0.47853
14	0.01403	0.26394
15	0.02312	0.43504
16	-0.06798	-1.27842
17	-0.03650	-0.68332
18	0.01903	0.35578
19	0.03337	0.62372
20	-0.00327	-0.06112
21	0.01907	0.35610
22	0.01027	0.19167
23	0.00269	0.05024
24	0.00560	0.10449
25	0.01360	0.25373

(3) $\nabla^2\tilde{x}_t = (1 - 0.988B)a_t$.

b) Specific conductance:

Figures 4.7 and 4.8 show the auto-correlation and partial auto-correlation plots of the raw data. The auto-correlation does not tail off even upto 30 lags indicating the presence of some kind of non-stationarity in the series. The partial auto-correlation function tails off after 3 days implying that a ARMA(3,0,0) model may be entertained. The data was differenced once and Figures 4.9 and 4.10 show the corresponding auto-correlation and partial autocorrelation functions plots. It is seen that the auto-correlation cuts off immediately after the first lag as shown by the confidence limit drawn at 2σ distance. The partial auto-correlations tail off quickly. This indicates that a ARMA(0,1,1) process should be entertained for this data. Another differencing was tried and Figures 4.11 and 4.12 show the corresponding auto-correlation and partial auto-correlation plots. Figure 4.11 shows that the autocorrelation function cuts off immediately after the first lag whereas the partial correlation tails off smoothly. Again an ARMA(0,2,1) model is suggested.

Hence two tentative models are suggested by these plots

(1) ARMA(0,1,1), $\nabla\tilde{x}_t = (1 - \theta_1 B)a_t$

(2) ARMA(0,2,1), $\nabla^2\tilde{x}_t = (1 - \theta_1 B)a_t$.

The initial estimates of the parameters for model 1 and 2 are obtained either using (4.16) or the charts in [40].

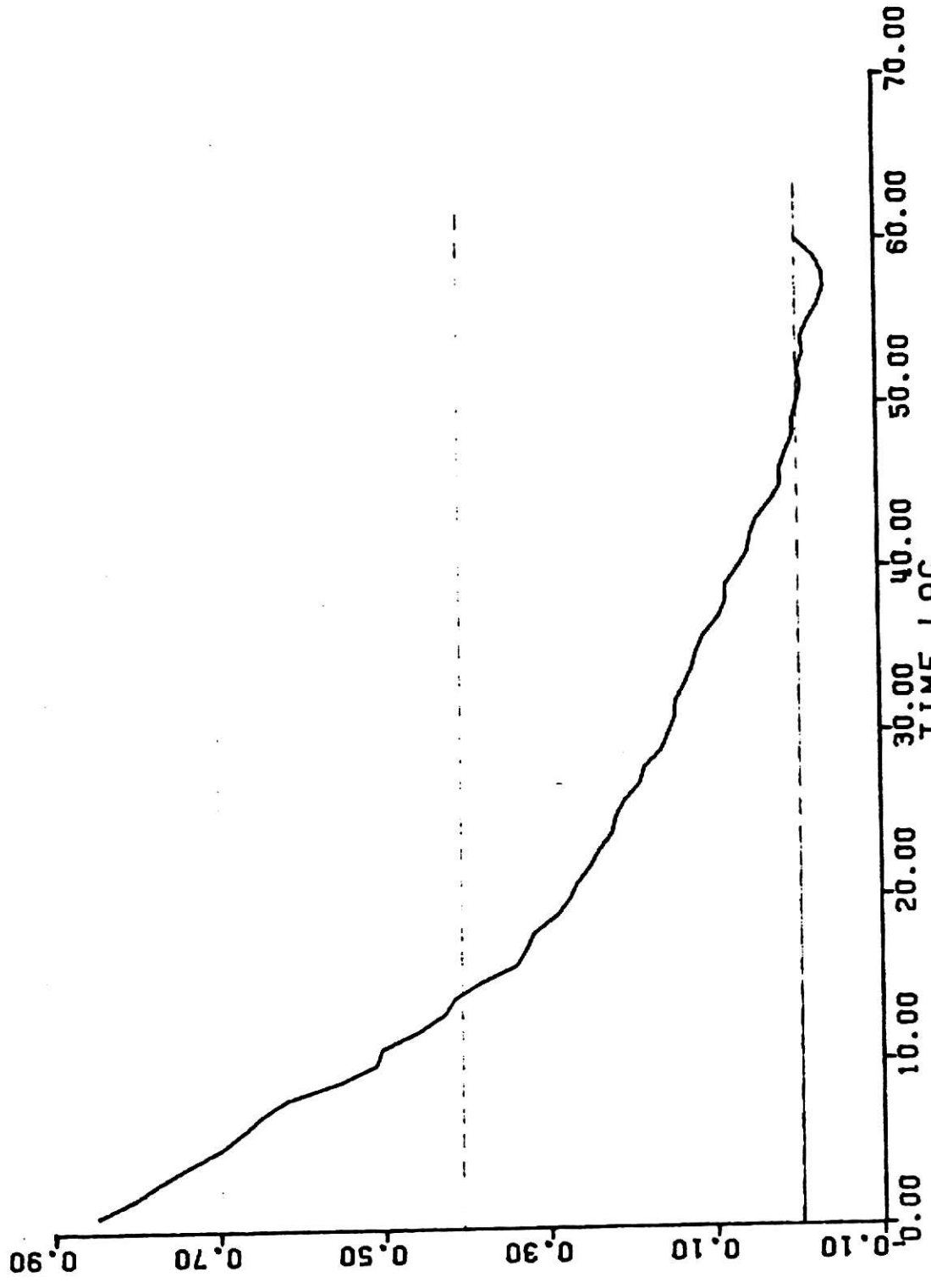

Fig. 4.7 Autocorrelation of original specific conductance data.

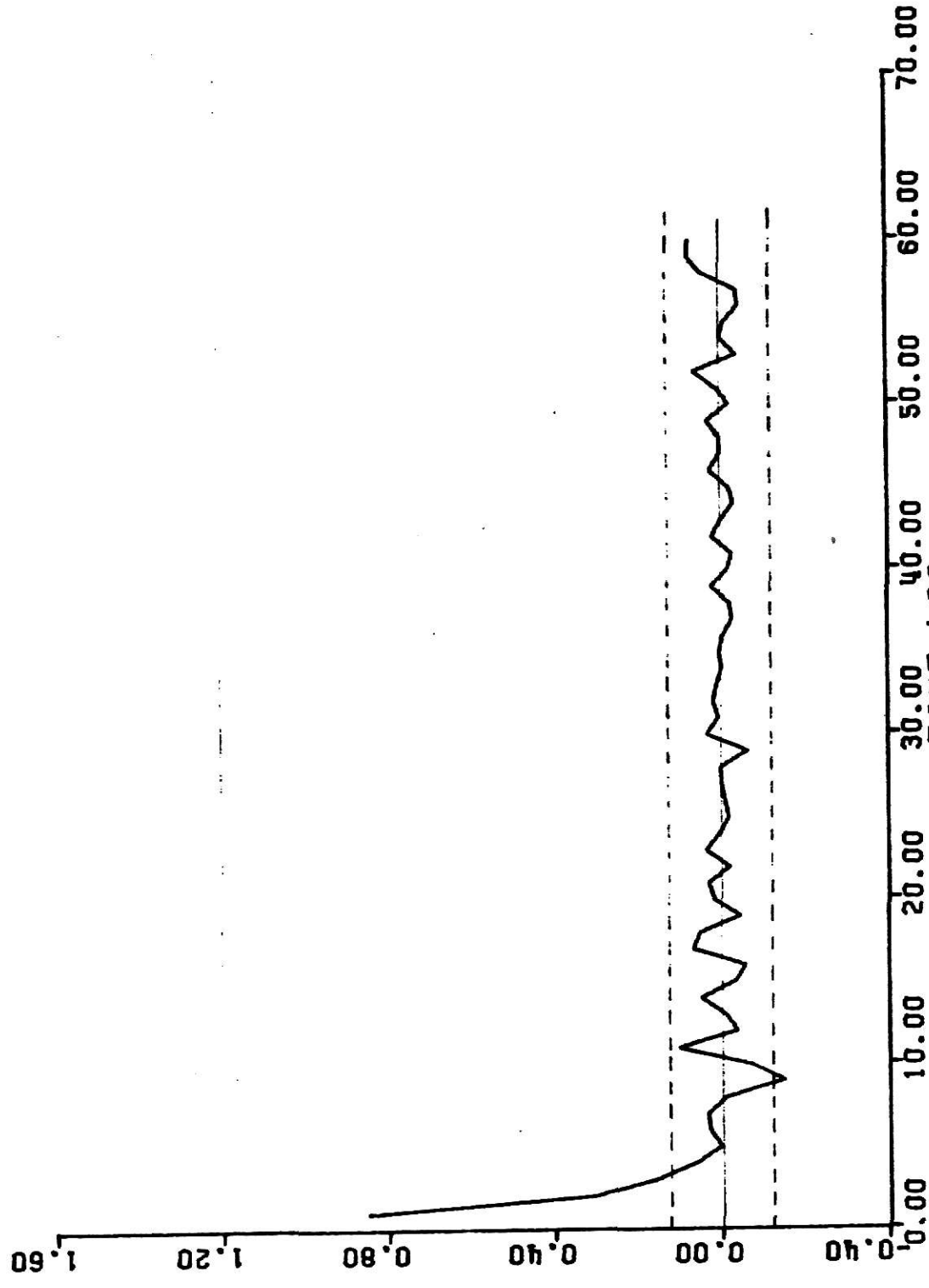

Fig. 4.8 Partial autocorrelation of original specific conductance data.

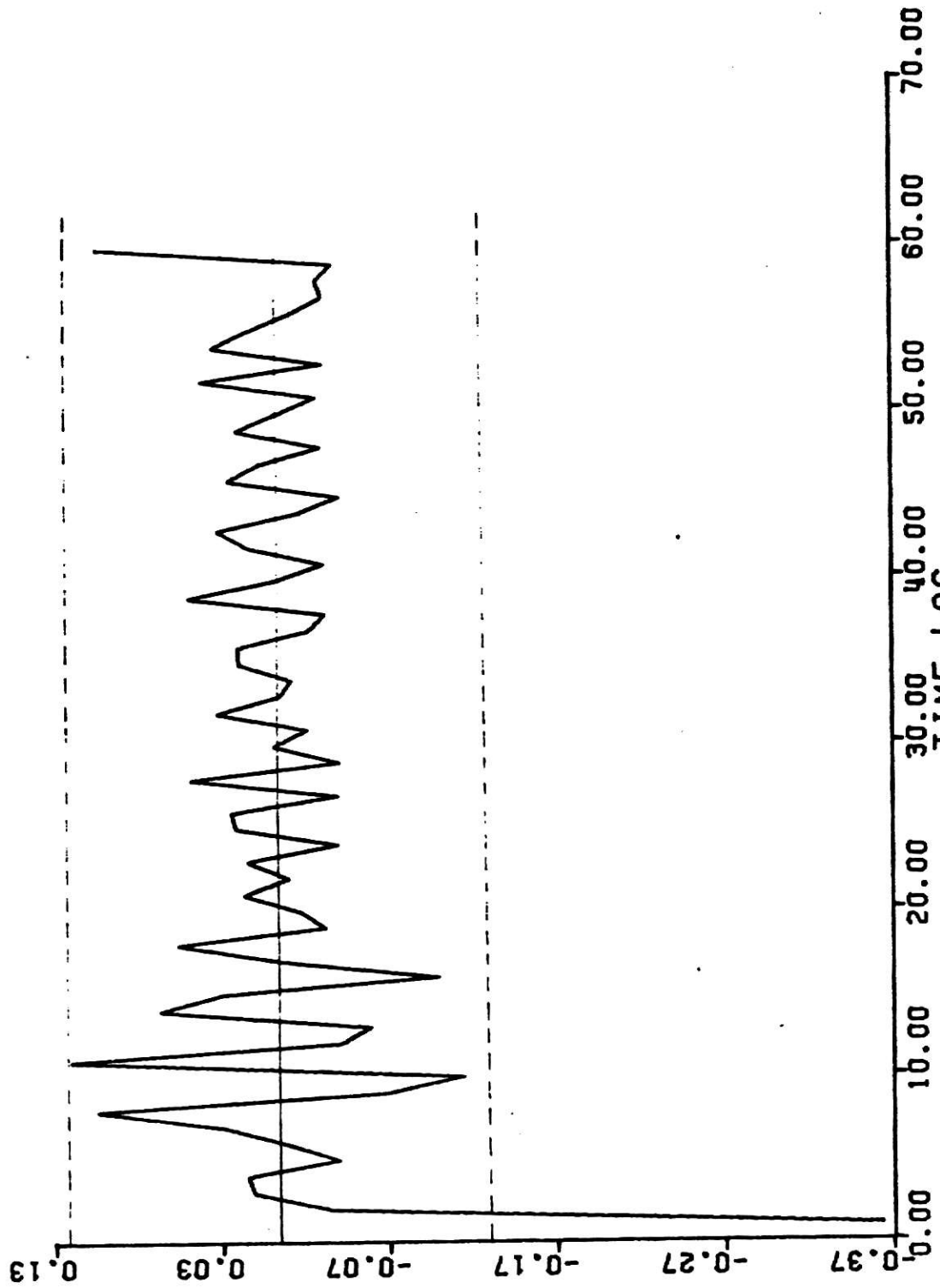

Fig. 4.9 Autocorrelation of specific conductance data (∇x).

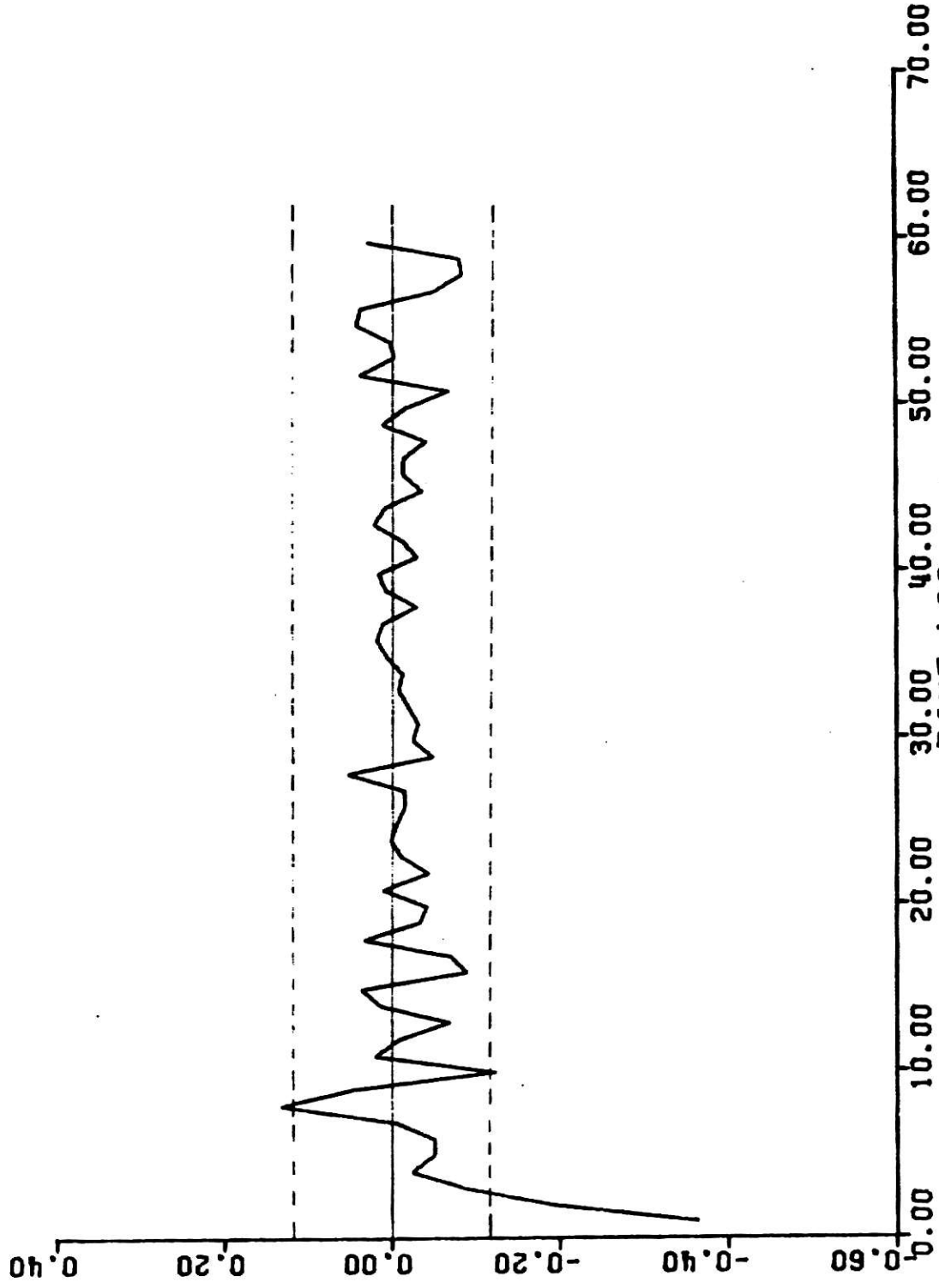

Fig. 4.10 Partial autocorrelation of specific conductance data (Vx).

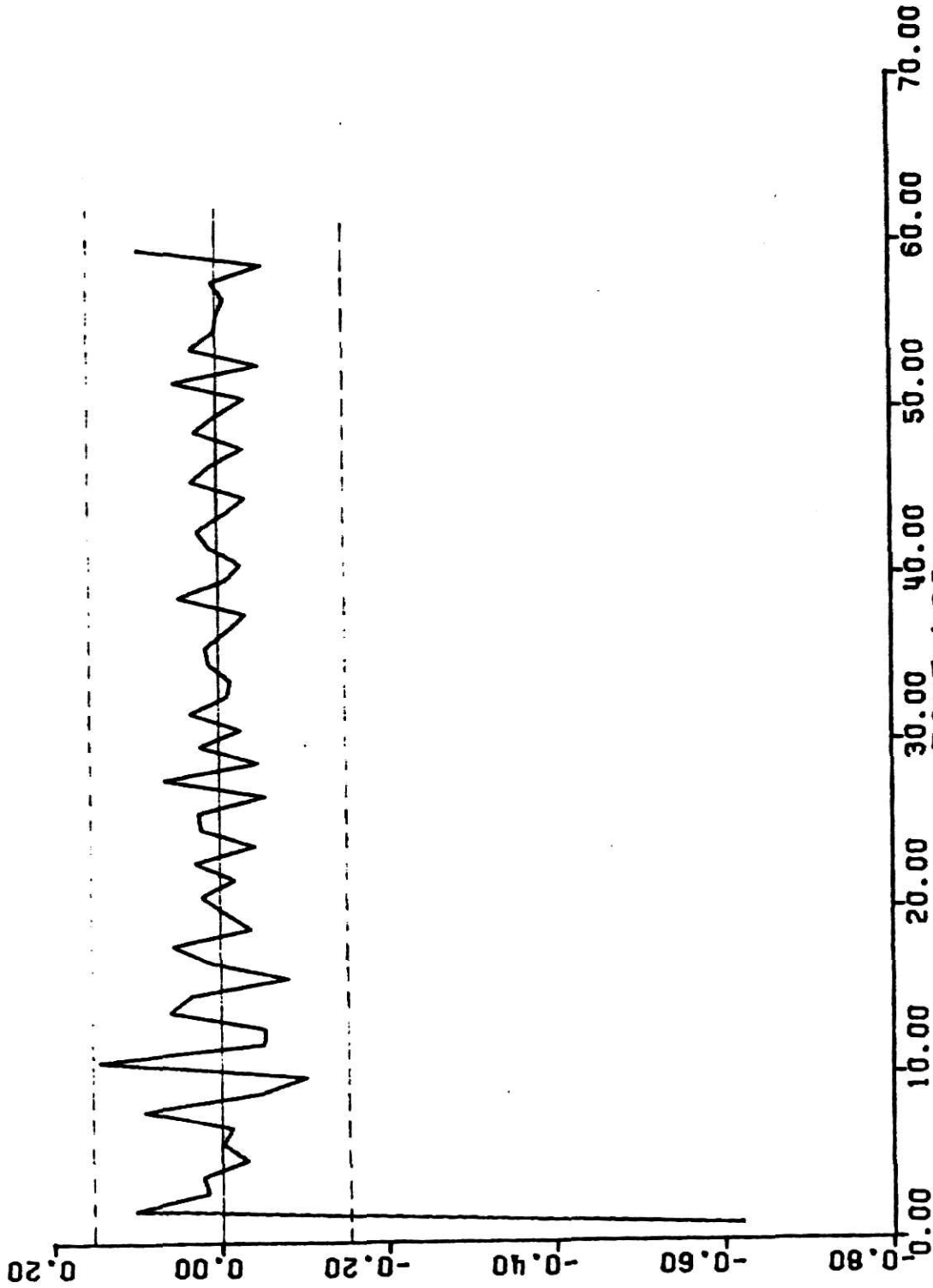

Fig. 4.11 Autocorrelation of specific conductance data ($\nabla^2 x$).

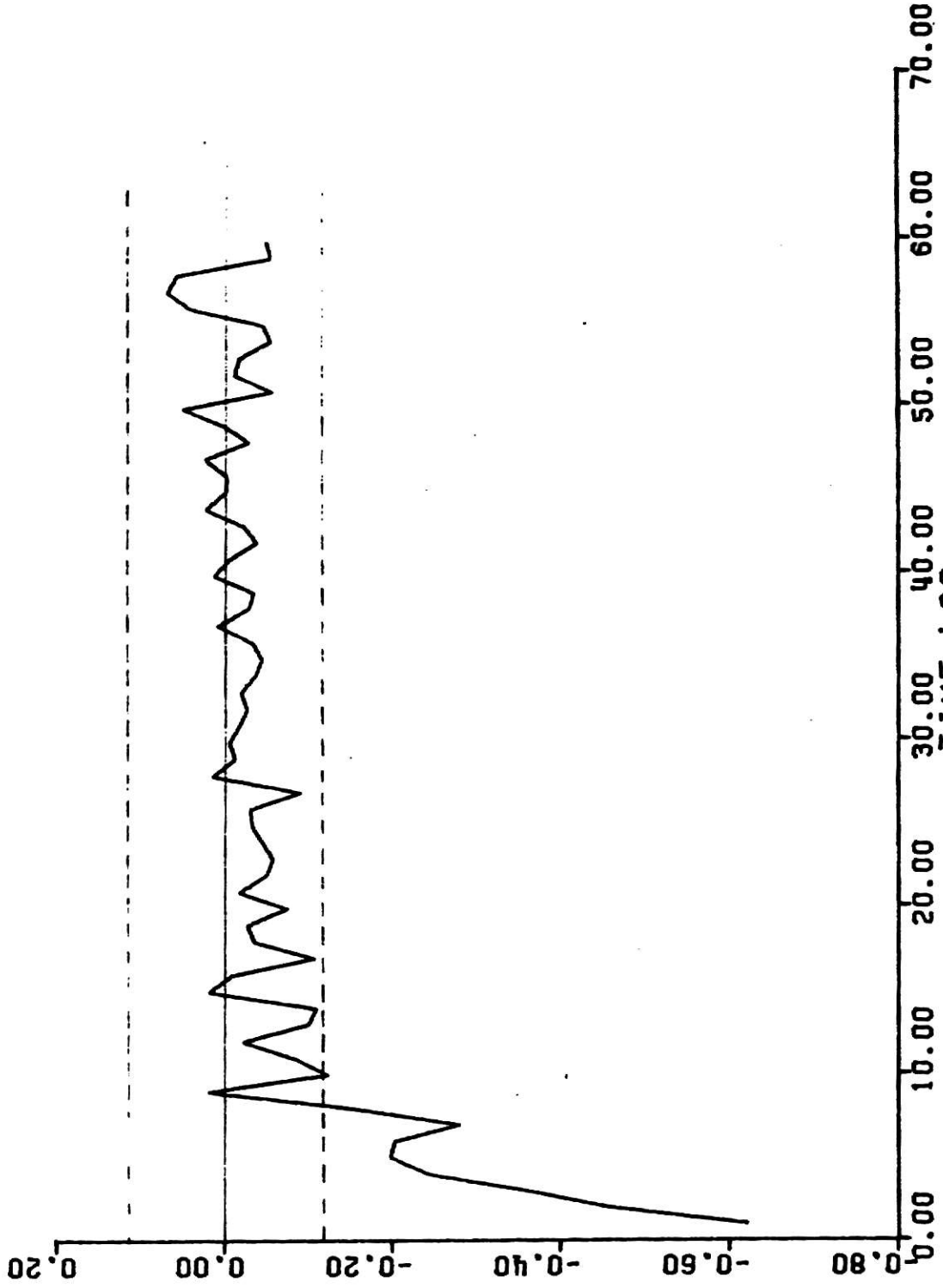

Fig. 4.12 Partial autocorrelation of specific conductance data ($\nabla^2 x$).

Table 4.5 Autocorrelation of residuals for specific conductance.

Model ARMA (0,1,1)

Lag k	Autocorrelation ρ_k	ρ_k/S.D.(ρ_k)
1	0.01917	0.36579
2	-0.02683	-0.51166
3	0.00095	0.01813
4	0.00150	0.02867
5	-0.03914	-0.74599
6	0.00232	0.04406
7	0.06876	1.30850
8	0.10102	1.91332
9	-0.07487	-1.40405
10	-0.12124	-2.26134
11	0.05545	1.02000
12	-0.04225	-0.77506
13	-0.05854	1.07203
14	0.04344	0.79300
15	0.00017	0.00302
16	-0.11508	-2.09721
17	-0.04078	-0.73439
18	0.02700	0.48550
19	-0.03405	-0.61183
20	-0.03094	-0.55538
21	-0.00161	-0.02894
22	-0.01325	-0.23762
23	-0.00512	-0.09176
24	-0.03635	-0.65177
25	0.01408	0.25225

Using (4.16),

$$\rho_1 = \frac{-\theta_1}{1+\theta_1^2}$$

estimates of ρ_1 is obtained from the auto-correlation plot. Putting the value of ρ_1 in above expression,

$$-0.36 = \frac{-\theta_1}{1+\theta_1^2} \quad \text{(for model 1)}$$

and $$-0.62 = \frac{-\theta_1}{1+\theta_1^2} \quad \text{(for model 2)}$$

First relation gives $\theta_1 = 0.312$

Second relation doesn't yield any real roots, showing thereby that the model is not feasible for this data. Hence only model ARMA (0,1,1) was entertained for further investigation.

This initial estimate was used as starting value for least squares estimation procedure. Thus a more accurate value of parameter θ_1 was obtained

$$\theta_1 = 0.4715.$$

Diagnostic checks:

Table 4.5 contains the correlation coefficients for the sample residuals to 25 lags. Also the ratio,

$$\rho_k / \text{S.D.}(\rho_k)$$

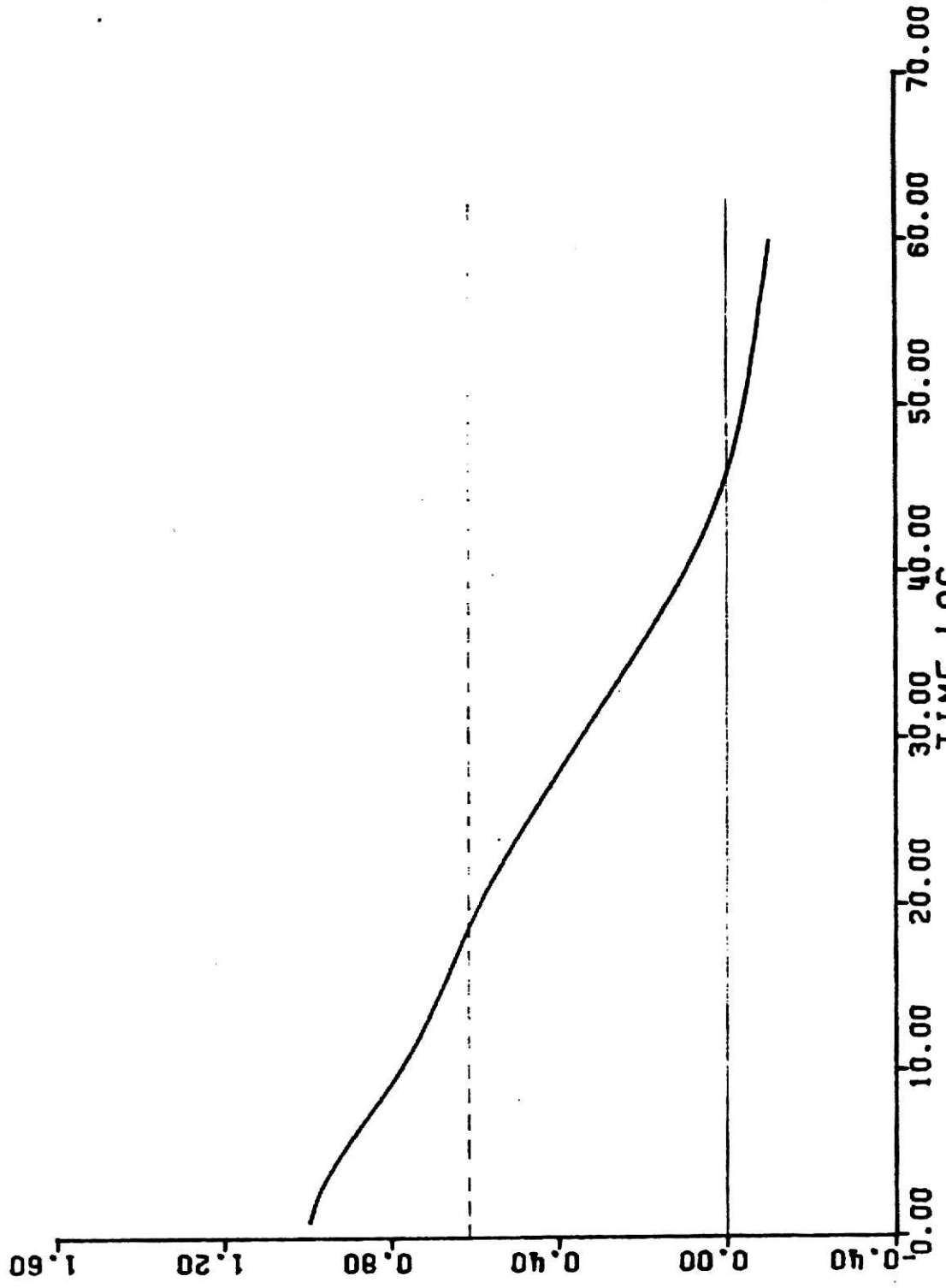

Fig. 4.13 Autocorrelation of original flow rate data.

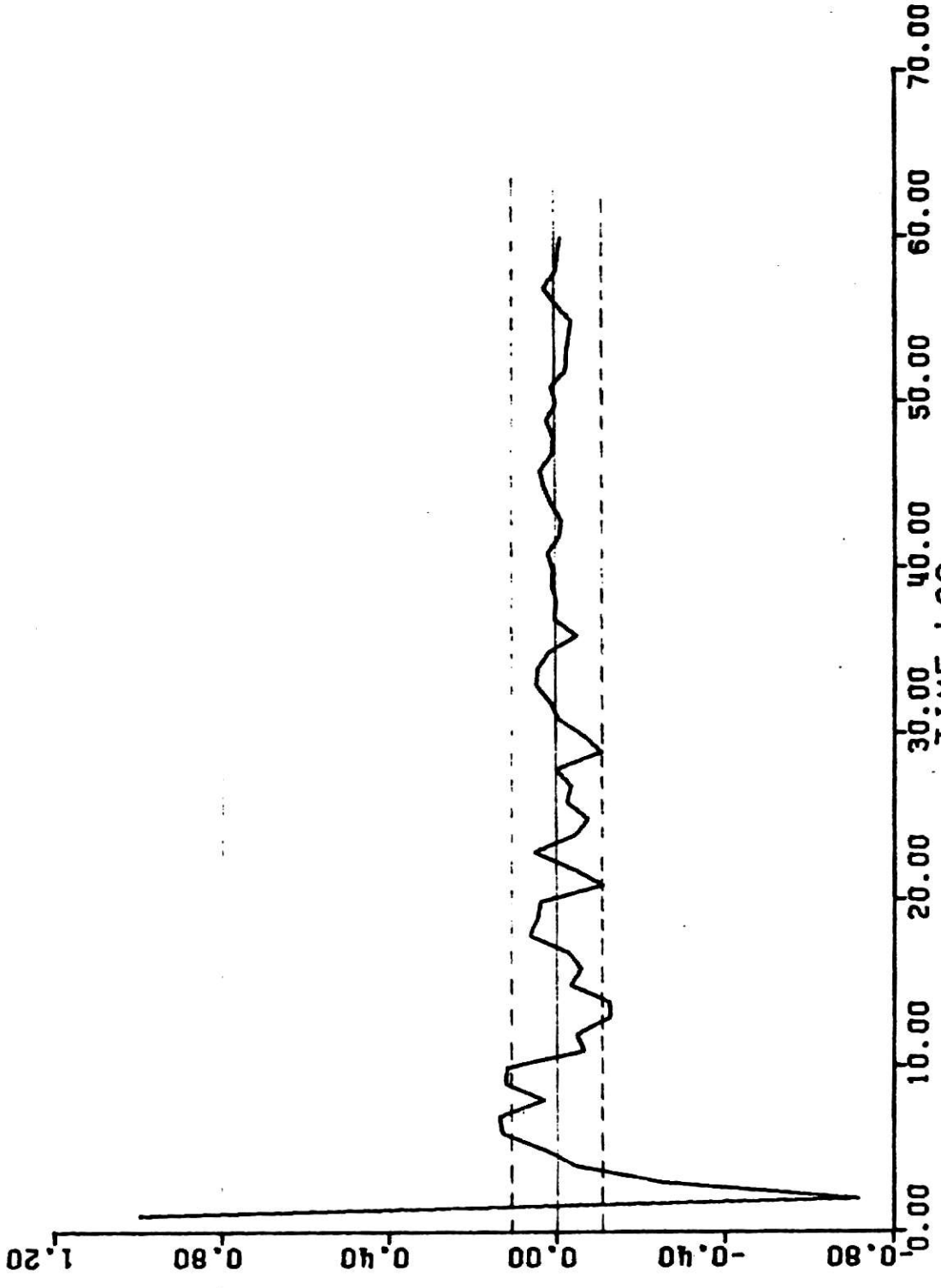

Fig. 4.14 Partial autocorrelation of original flow rate data.

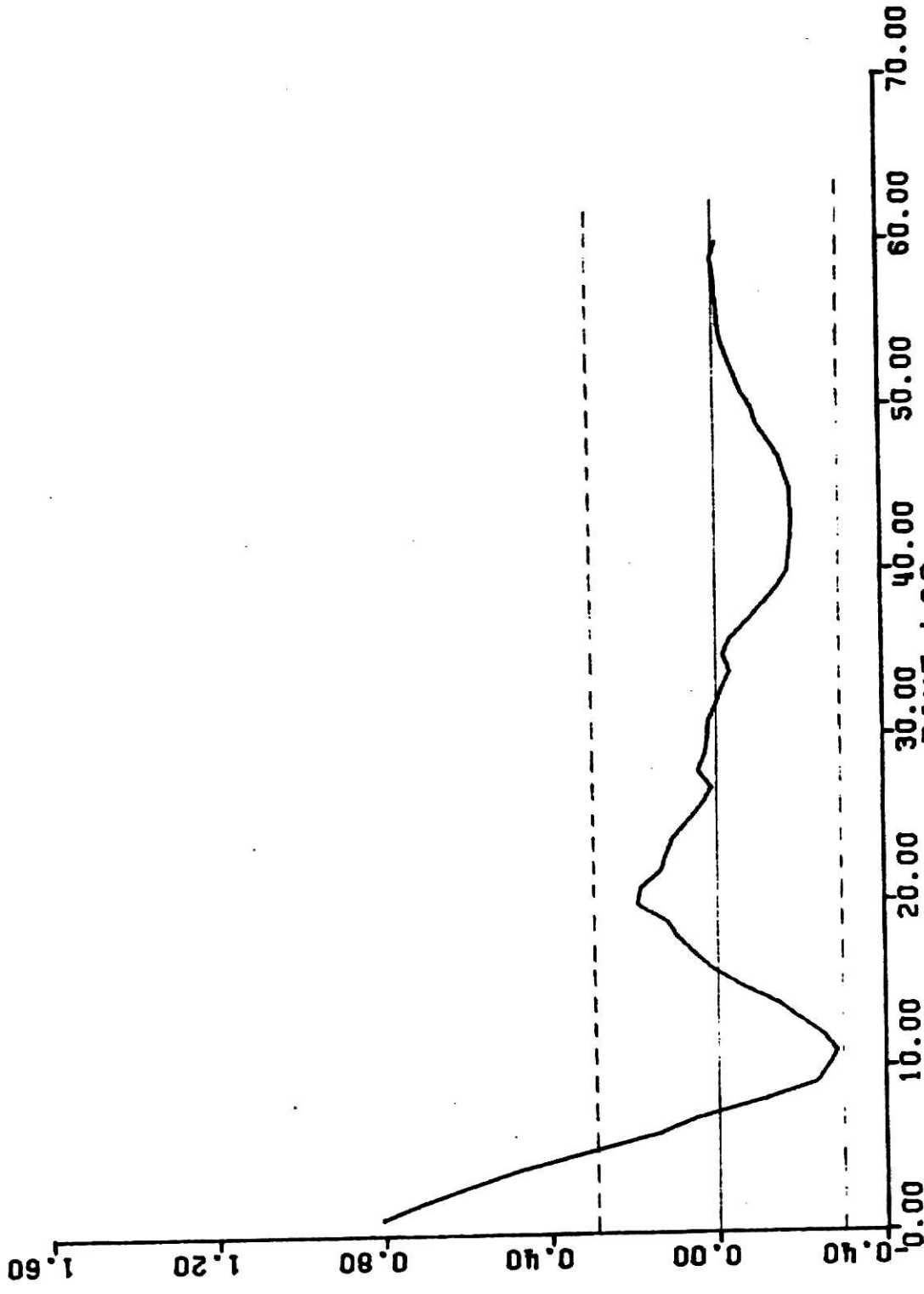

Fig. 4.15 Autocorrelation of flow rate data (Vx).

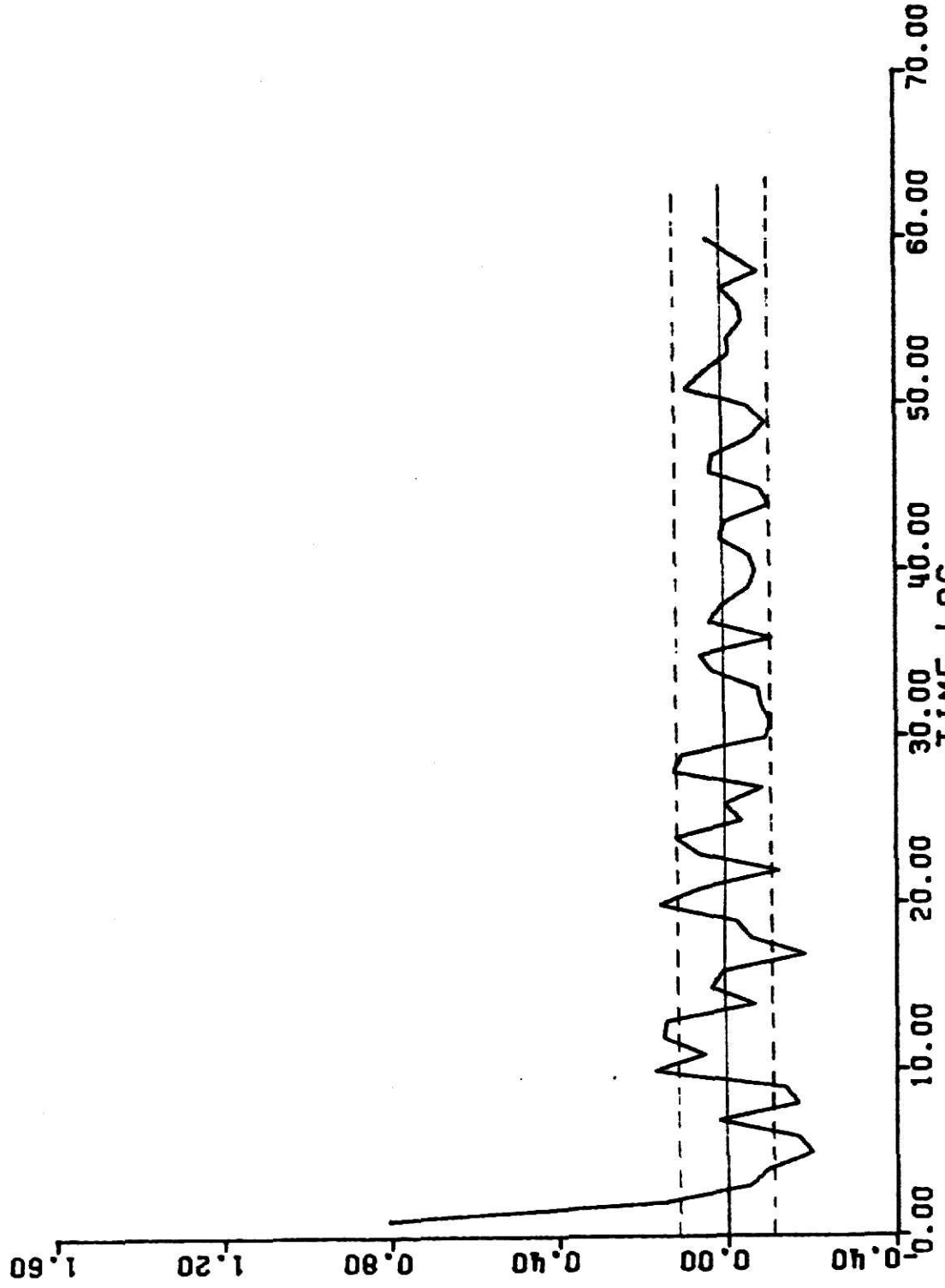

Fig. 4.16 Partial autocorrelation of flow rate data (Vx).

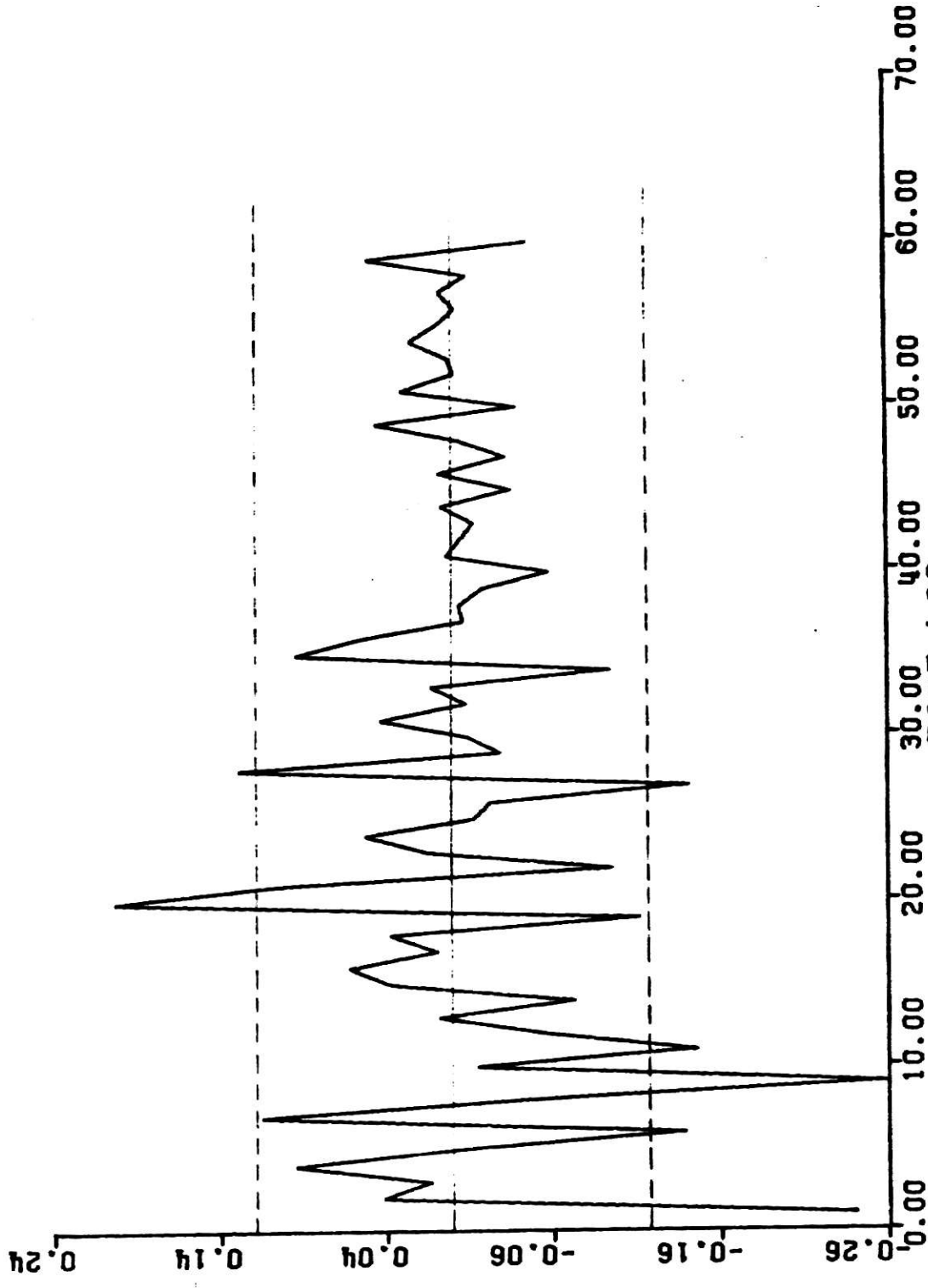

Fig. 4.17 Autocorrelation of flow rate data ($\nabla^2 x$).

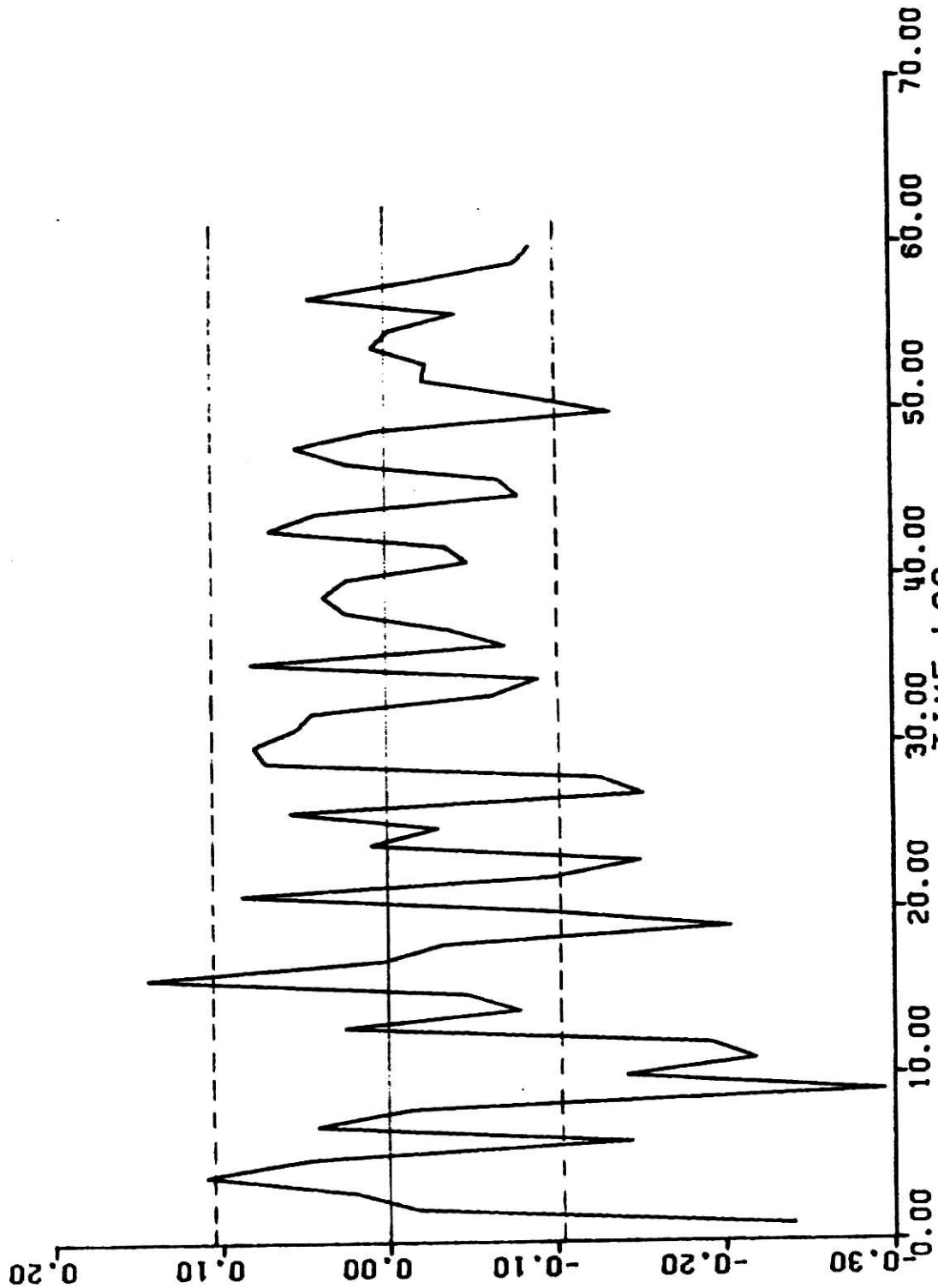

Fig. 4.18 Partial autocorrelation of flow rate data ($\nabla^2 x$).

indicates that all residual correlations are effectively zero.

A value of Q was calculated to use chi square test

$Q = 23.8$ $\quad$ d.f. = 25-1 = 24.

Tabulated chi square value for 24 d.f. and 90% confidence level is 34.382 which is well above the Q value.

Hence there seems to be no reason to doubt the adequacy of the model. Thus the model could be written as

$$\nabla\tilde{x}_t = (1 - 0.4715B)a_t.$$

(c) Flow: Figures 4.13 thru 4.18 show the auto-correlation and partial auto-correlation functions of the raw flow data, first differenced data and twice differenced data. These plots were analyzed and the following tentative models were suggested

(1) ARMA(1,20) (2) ARMA(1,1,1)

(3) ARMA(1,2,1).

Following table lists the results obtained from these models.

Model ARMA	Initial Estimate of Parameters	Final Estimate of Parameters	Q	$x^2_\alpha(0.90)$	d.f.
(1,2,0)	$\theta_1 = 0.24$	-0.24	134.60	63.2	59
(1,1,1)	$\theta_1 = 0.85$	0.86			
	$\theta_1 = 0.19$	0.157	126.40	63.2	59
(1,2,1)	$\theta_1 = 0.15$	-0.189			
	$\theta_1 = -0.15$	0.538	130.7	63.2	59

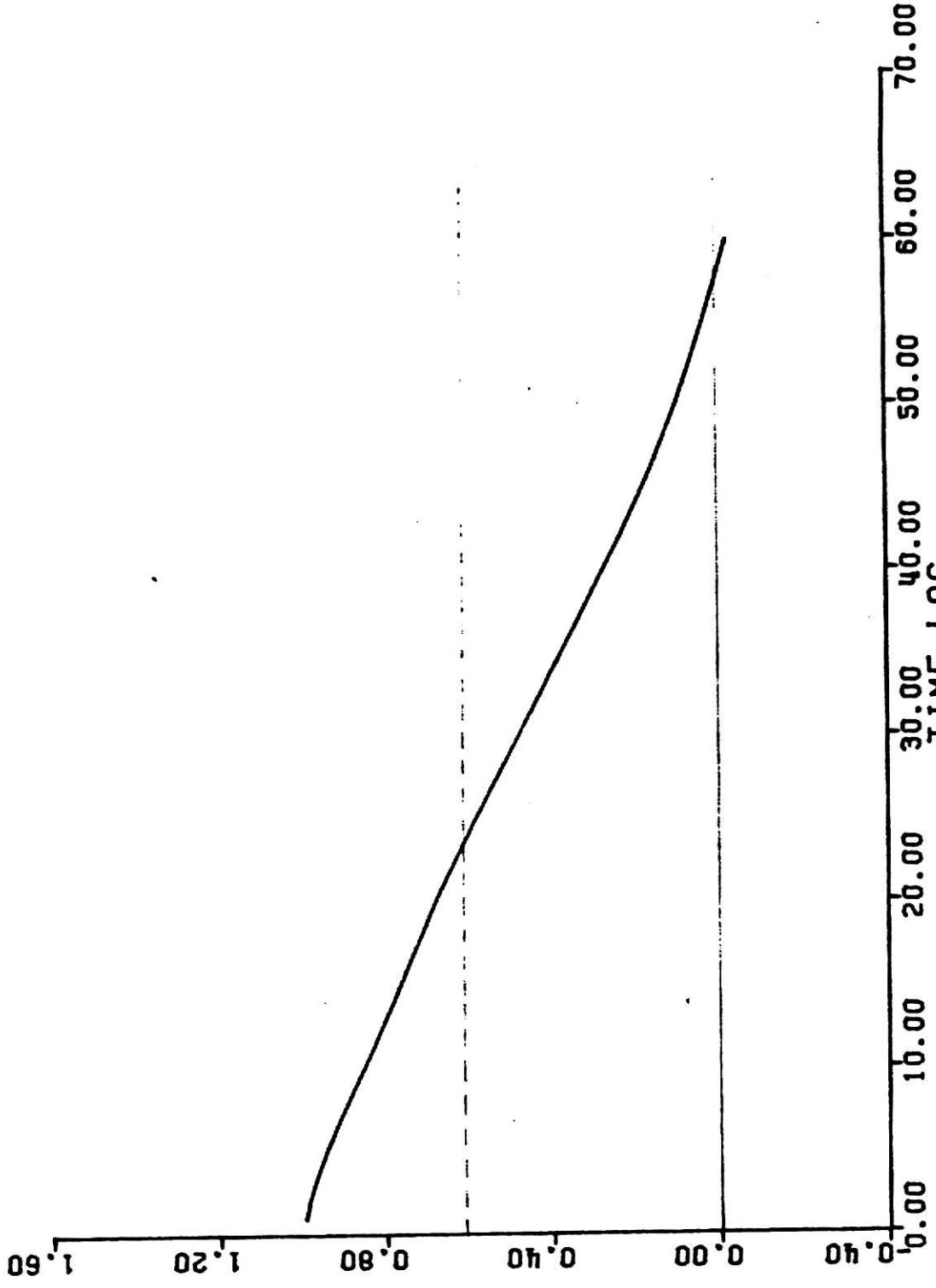

Fig. 4.19 Autocorrelation of lag flow rate data.

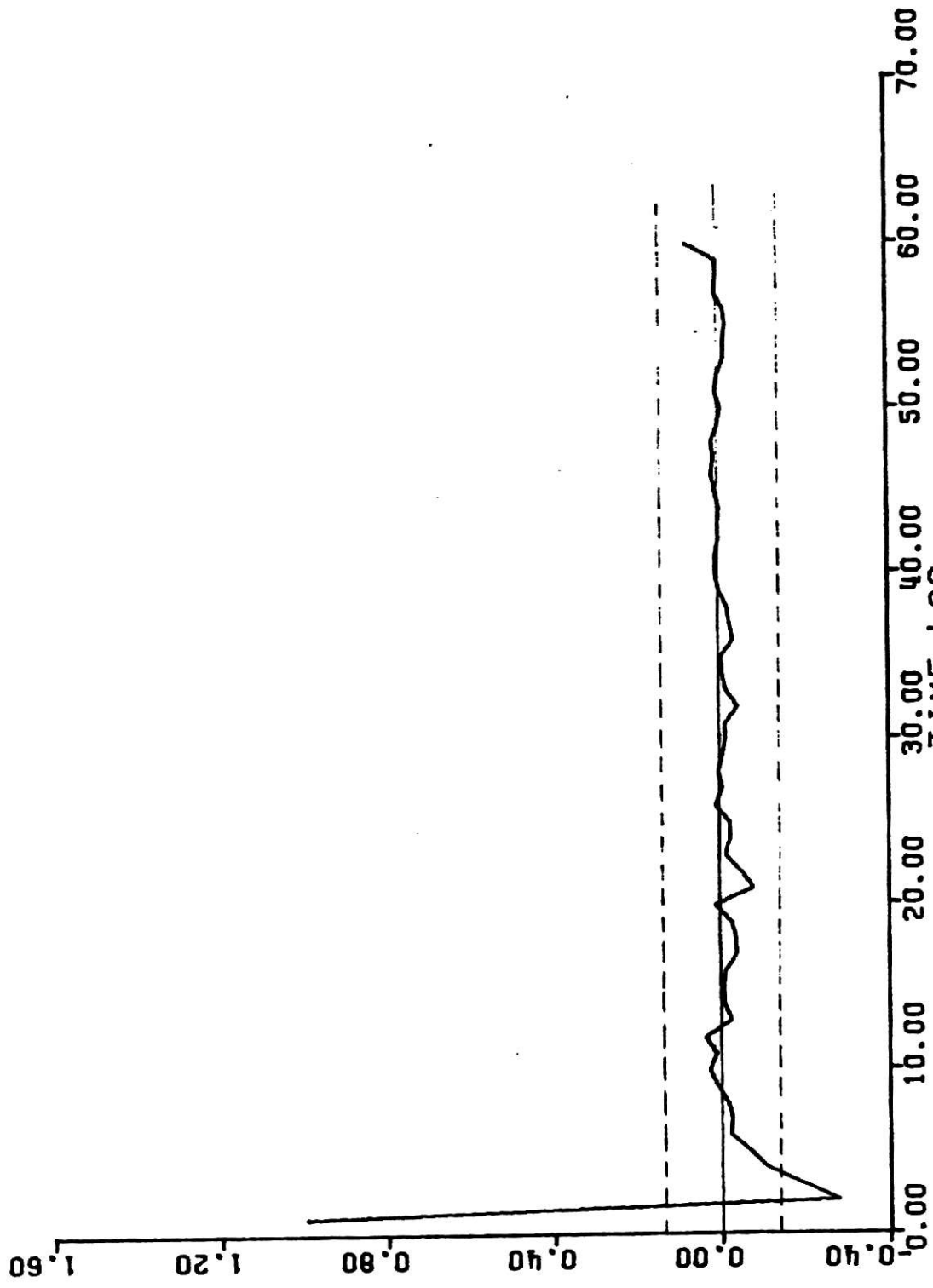

Fig. 4.20 Partial autocorrelation of log flow rate data.

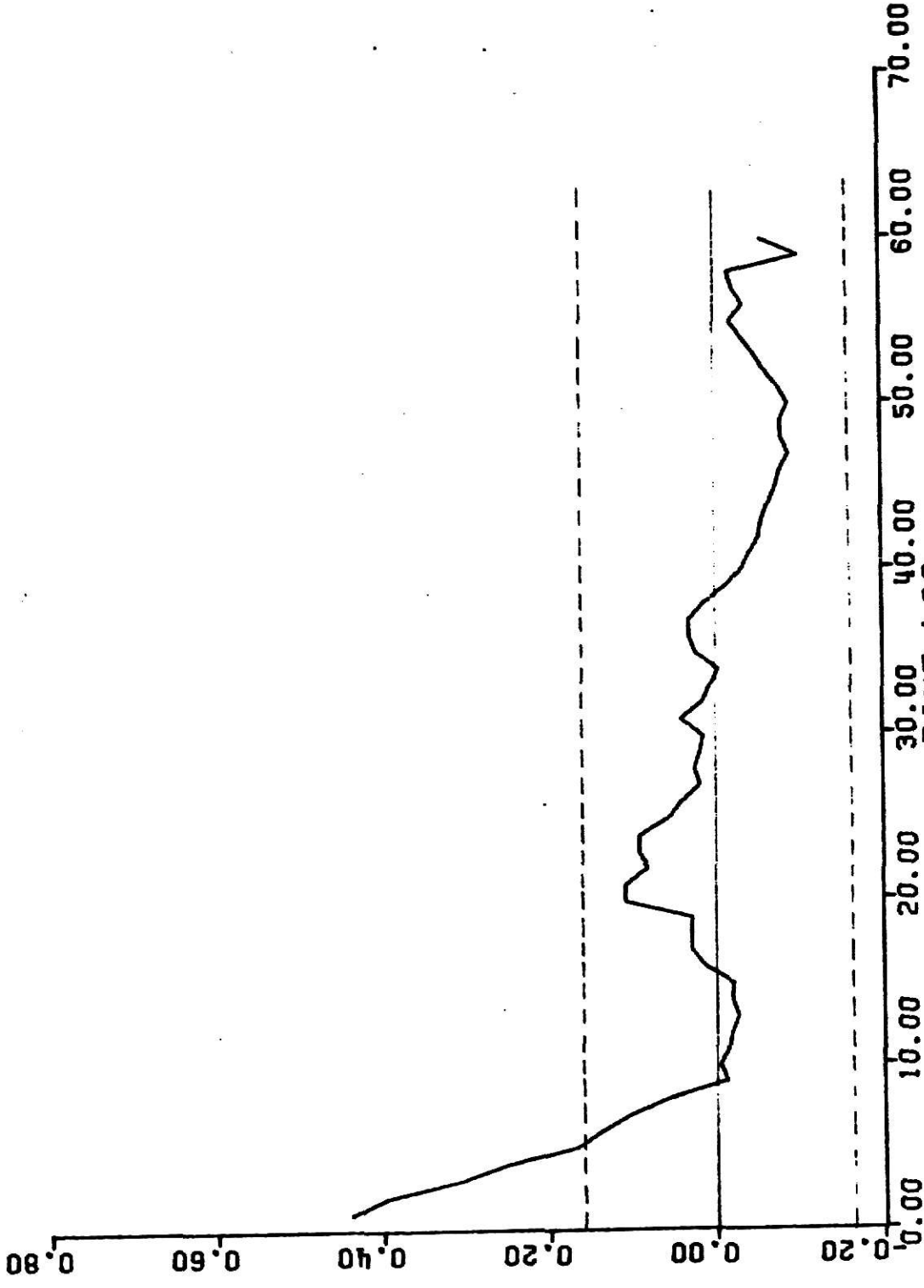

Fig. 4.21 Autocorrelation of log flow rate data (V_x).

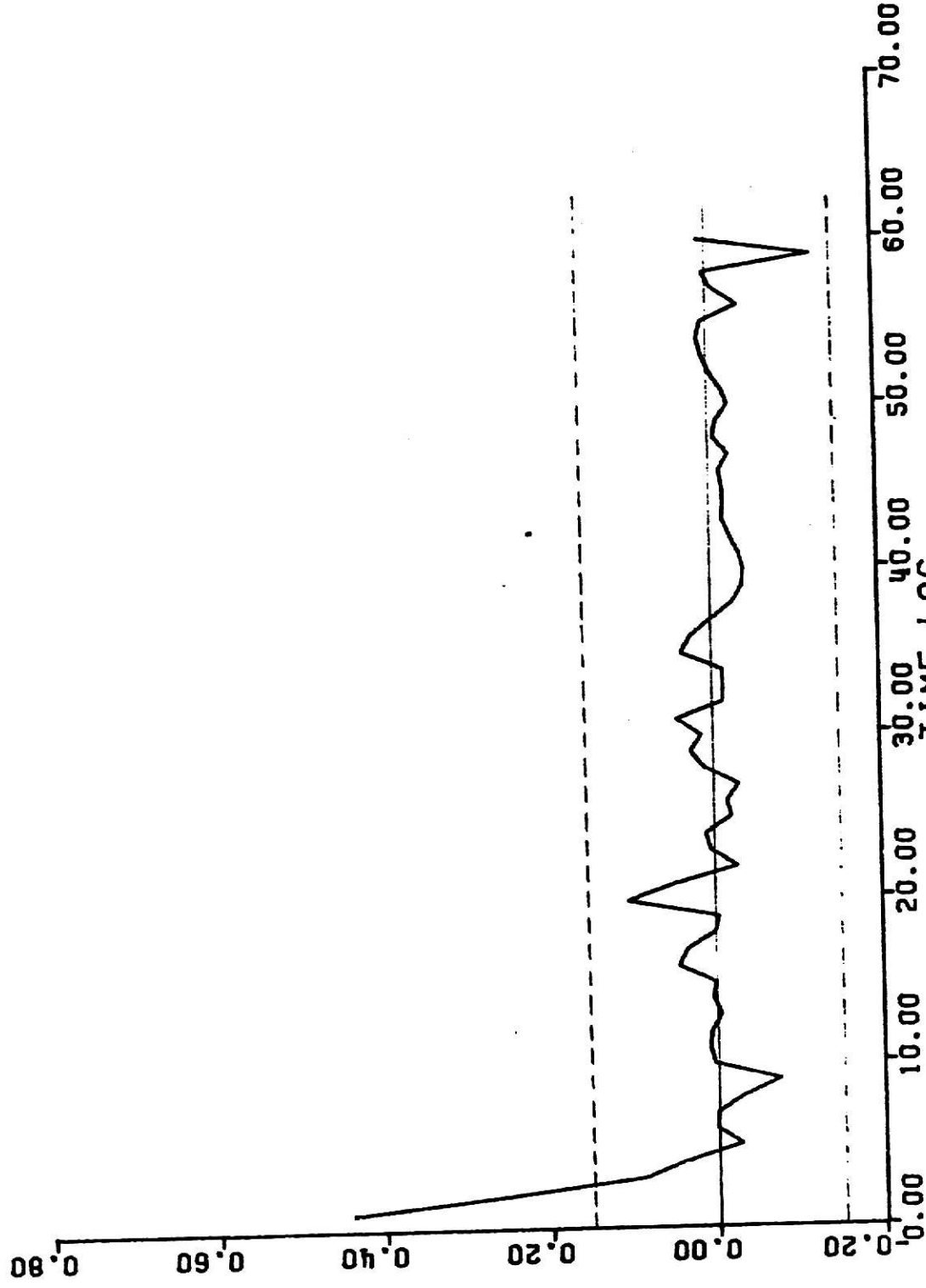

Fig. 4.22 Partial autocorrelation of log flow rate data (Yx).

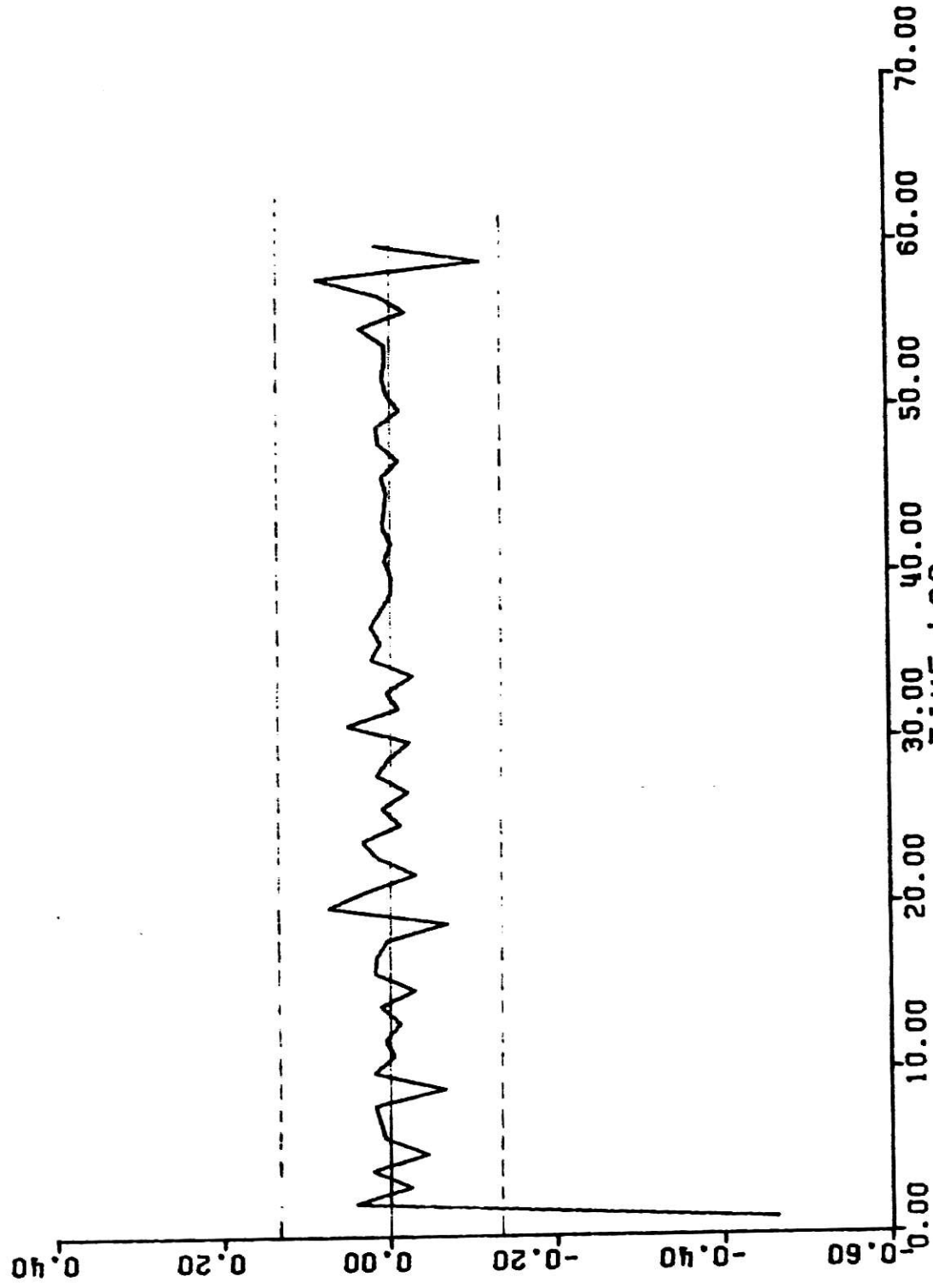

Fig. 4.25 Autocorrelation of log flow rate data ($\nabla^2 x$).

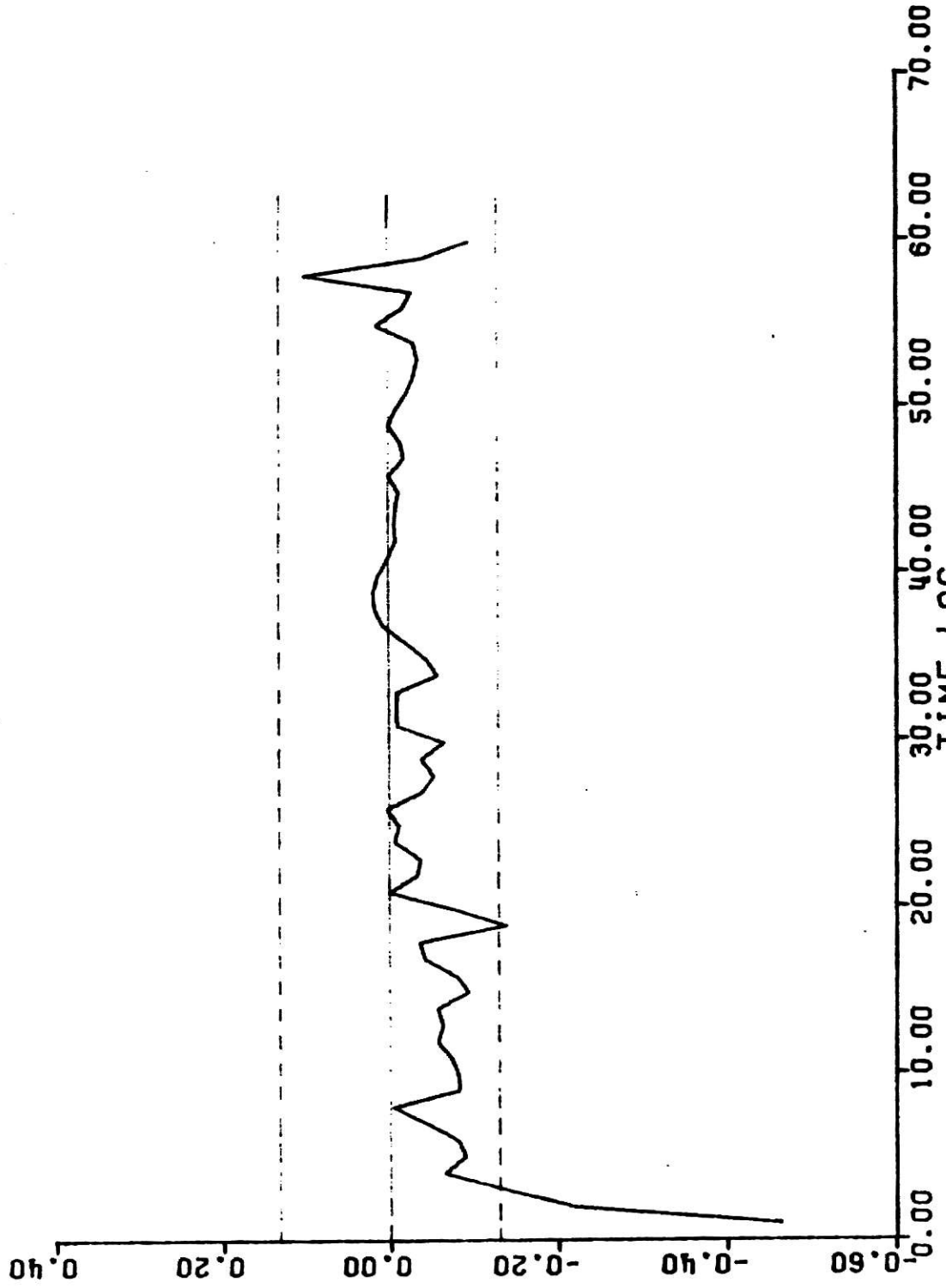

Fig. 4.24 Partial autocorrelation of log flow rate data ($\nabla^2 x$).

The tabulated Chi square value for 59d.f. and 0.90 significance level is 63.2, which is less than the Q values obtained above. This indicates the inadequacy of these models in explaining the variation in this data. Another approach for modeling flow data was made by transforming the original data to a natural log scale. Figures 4.19 to 4.24 show the auto-correlation and partion auto-correlation plots for the transformed data. The following tentative models were suggested by these plots

(1) ARMA(2,1,0)

(2) ARMA(0,2,1)

(3) ARMA(1,1,1).

The following table lists the results obtained for these models:

Model	Estimated Parameters	Q-value	d.f.	$\chi^2_\nu(0.90)$
(2,1,0)	$\phi_1 = 0.328$	25.48	59	63.2
	$\phi_2 = 0.253$			
(0,2,1)	$\theta_1 = 0.639$	32.67	59	63.2
(0,1,1)	$\phi_1 = 0.816$	21.82	59	63.2
	$\theta_1 = 0.484$			

It is seen that for all 'the above three models, the tabulated χ^2 value is greater than the Q - value. Hence there does not seem to be any reason to doubt the adequacy of these models.

Thus, it can be concluded that the Onterior River flow rate can be easily modeled by transforming it to a natural logarithmic scale.

Table 4.6 Comparison of variance of raw series and residuals.

(A) Pollutant: Temperature

Model	Variance $(^{o}C)^2$
Raw series	2.624
ARMA (1,0,0)	0.216
ARMA (0,1,1)	0.223
ARMA (0,2,1)	0.226

(B) Pollutant: Specific Conductance

Model	Variance $(MHO/Cm \times 10^{-2})^2$
Raw series	7658.33
ARMA (0,1,1)	1951.0

(C) Pollutant: Flow rate

Model	Variance $(cfs)^2$
(a) Original Data	
Raw series	68831760.0
ARMA (1,2,0)	204500.0
ARMA (1,2,1)	205100.0
ARMA (1,1,1)	191900.0
(b) After transforming to natural logarithmic scale.	
Raw series	0.87166
ARMA (0,1,1)	0.0047
ARMA (2,1,0)	0.0047
ARMA (0,2,1)	0.0051

As seen above, several models seem to fit for each pollutant. One way to decide the best model would be to see the amount of variance reduced by each model. Table 4.6 shows the variance of the raw data and the variance of the residuals after fitting the particular model. It is observed that each model reduces the variance considerably but the difference in variance of residuals for each model seems to be insignificant. The particular choice of use of any model is arbitrary for these cases.

It may be noted here that though no model could be fitted adequately for original flow data, yet the tentative models significantly reduced the variance in the data.

CHAPTER V

ANALYSIS OF POTOMAC RIVER DATA

This chapter deals with the analysis of temperature, dissolved oxygen, biochemical oxygen demand and chloride data for Potomac river. Spectral analysis was conducted for each individual pollutant and cross-spectral analysis was performed to study the behaviour of a pollutant at different points of a stream and also the relationship among different pollutants at a station. Prediction models were developed using parametric time series modeling.

5.1 Data Acquisition:

The data were obtained from two different sources. The first set of data consists of four stations which cover the upper 30 miles of the Potomac Estuary. The approximate location of these stations is shown in Figure 5.1. Sampling was done at half hour interval for a period of one month (from 6/20/67 to 7/20/1967) for the first station and for 15 days (from 7/6/67 to 7/20/67) for all other stations.

The principal sources of water pollution for the Potamac River in this region are the effulents from the sewage treatment plants. The points of effulent discharge and the relative organic loads in the various discharges are shown in Figure 5.1 and Table 5.1 respectively. [41] Industrial wastes do not contribute much to the water pollution in this area. Cooling water from electric generating plants constitute the major thermal pollution source.

ILLEGIBLE DOCUMENT

THE FOLLOWING DOCUMENT(S) IS OF POOR LEGIBILITY IN THE ORIGINAL

THIS IS THE BEST COPY AVAILABLE

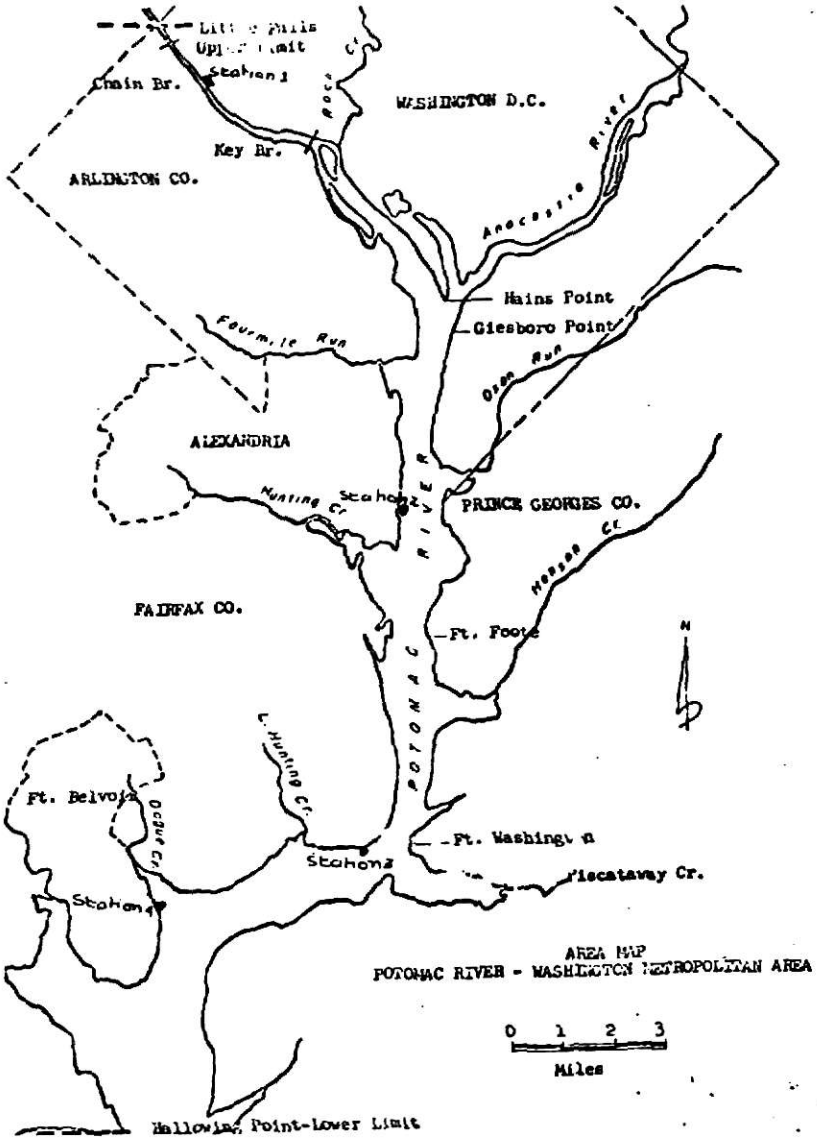

Fig. 5.1 Map showing the location of four stations on Potomac river.

Table 5.1 Sources of Sewage Discharged to the Potomac River

Source	Discharge Location	BOD lbs/day
Alexandria, Va.	Hunting Creek	9500
Arlington Co., Va.	Four Mile Run	3600
Fairfax Co., Va. Westgate	Hunting Creek	12600
Little Hunting Creek	Little Hunting Creek	1000
Dogue Creek	Douge Creek	600
District of Columbia	Potomac River	92000

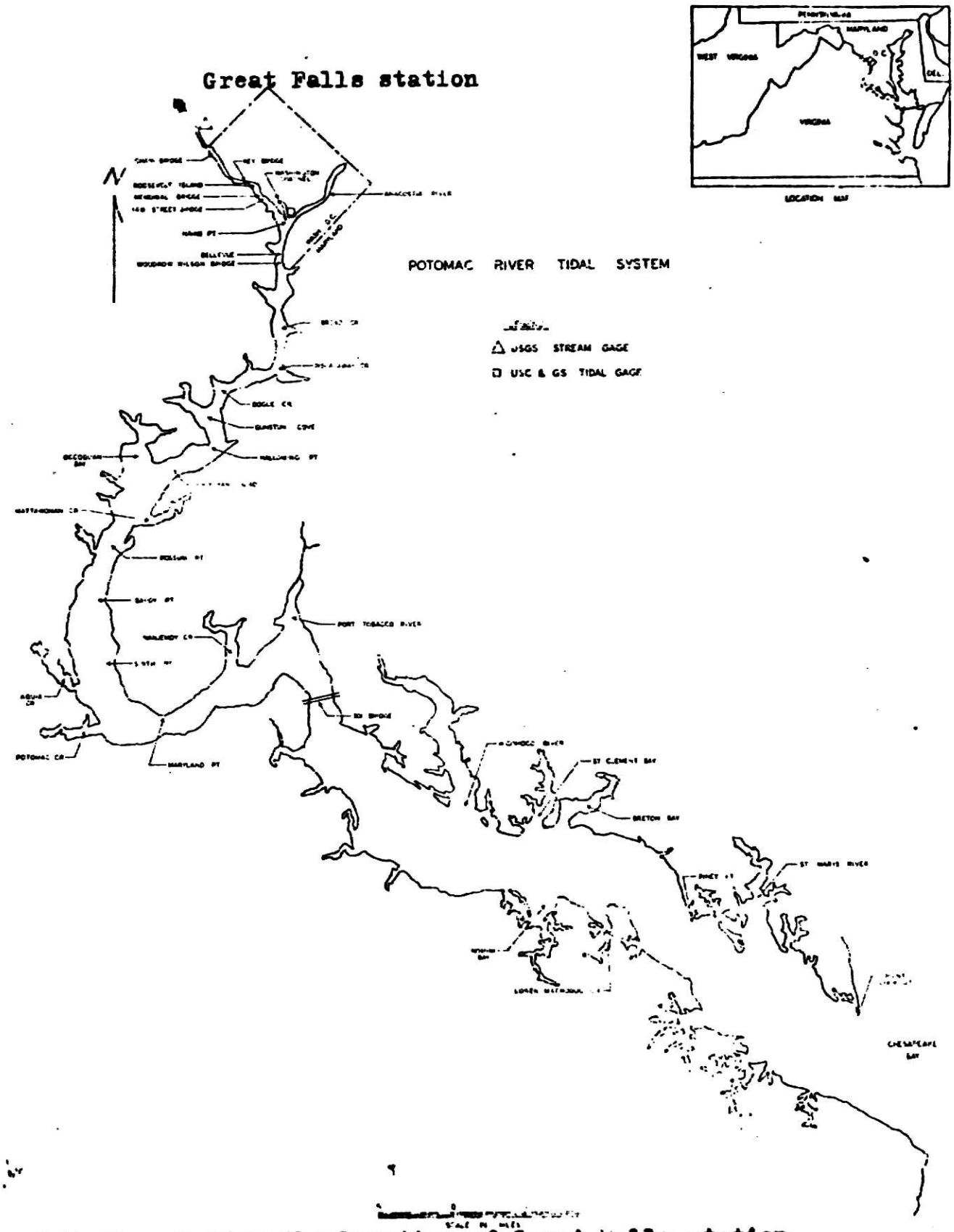

Fig. 5.2 Map showing the location of Great Falls station.

The second set of data consist of temperature, dissolved oxygen, biochemical oxygen demand and chloride records for Great Falls station on Potomac River. The approximate location of this station is shown in Figure 5.2. Sampling was done on a weekly basis for the period from 1/4/65 to 12/28/71.

For this study, the first set of data was obtained from a publication by the Environment Protection Agency [43], and the second set of data was provided in a separate communication by the same agency [42].

5.2 Analysis of Data for Stations 1,2,3, and 4

5.2.1 Introduction: The Potomac Estuary is a highly complex tidal system. The waste discharges in the river remain in the vicinity of the discharge point for some time before they are passed on to the sea. The major consequences of the sewage discharges are

(i) high bacterial density

(ii) low dissolved oxygen.

(iii) excessive algal growth.

The low dissolved oxygen concentration is caused mainly by the oxidation of the organic wastes in the water; the oxygen for this purpose being extracted from the water itself.

Figures 5.3 thru 5.10 show the temperature and dissolved oxygen plots for stations 1,2,3, and 4 respectively. Visual inspection of these plots does not indicate the presence of a definite periodicity, though there is indication of small (12 hrs. and 24 hrs.) periodic fluctuations in dissolved oxygen and 24 hrs. periodicity in temperature data. Table 5.2 below shows the mean DO and temperature levels at each station.

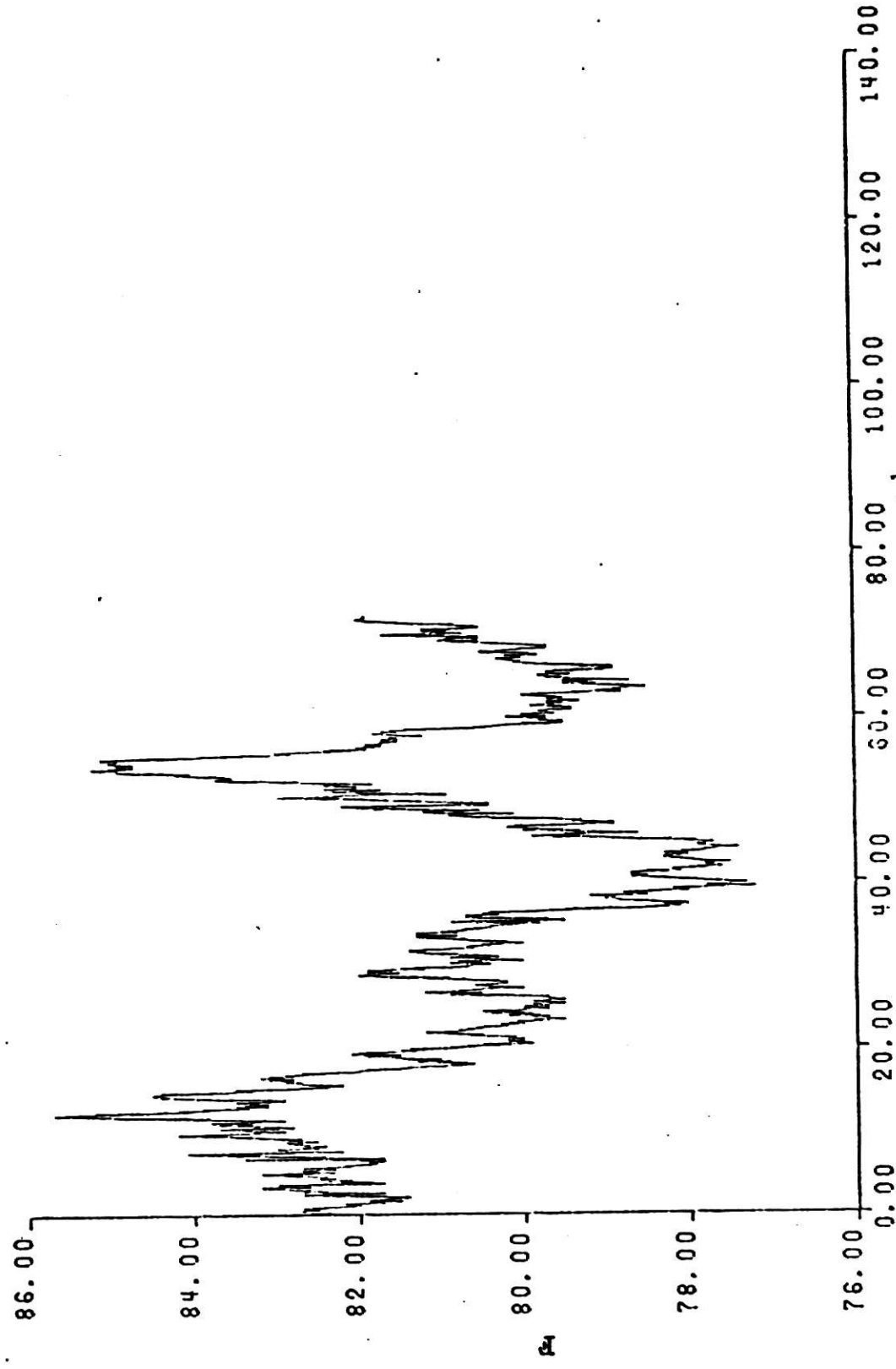

Fig. 5.3 Temperature record - station 1.

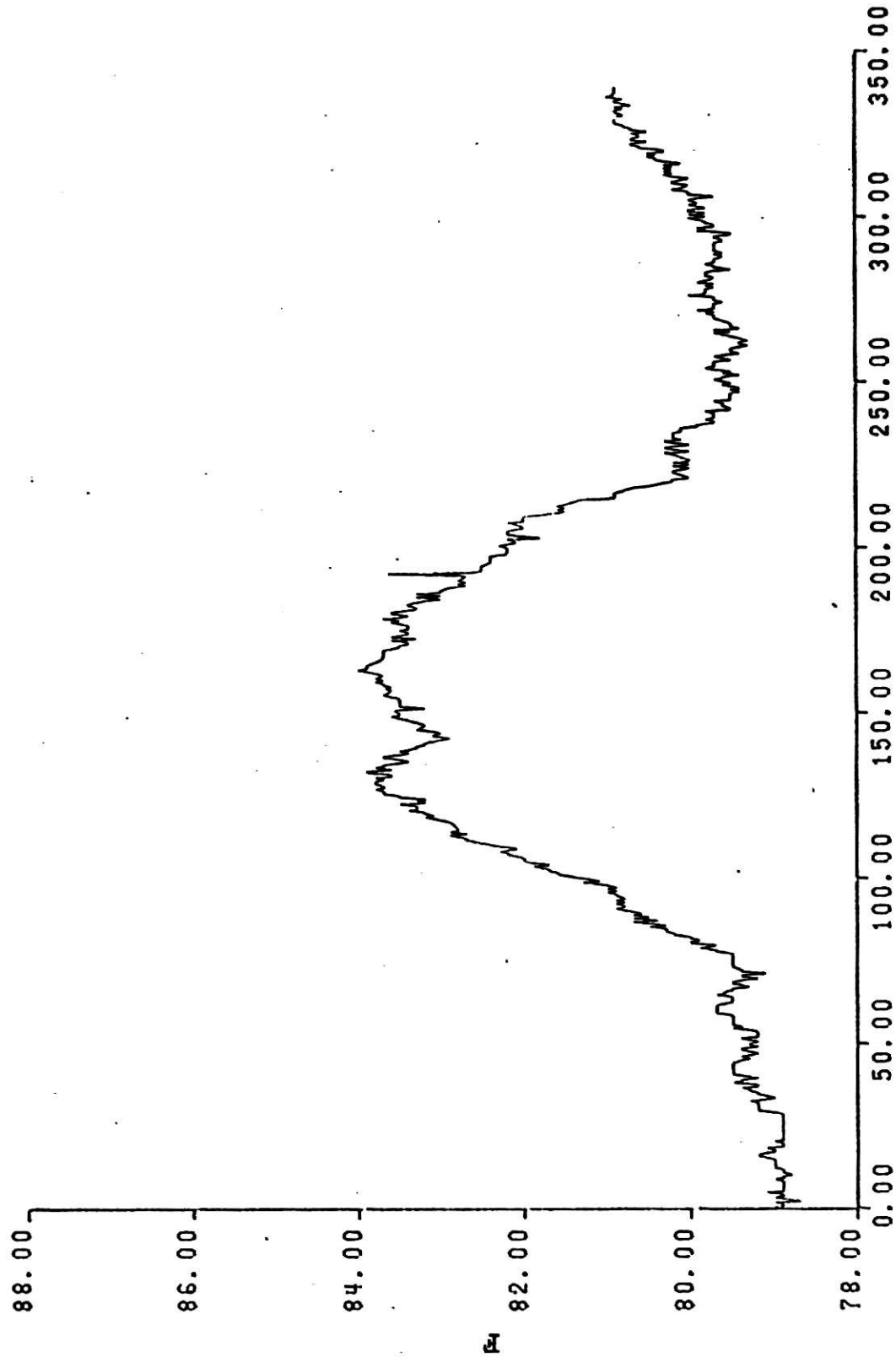

Fig. 5.4 Temperature record - station 2.

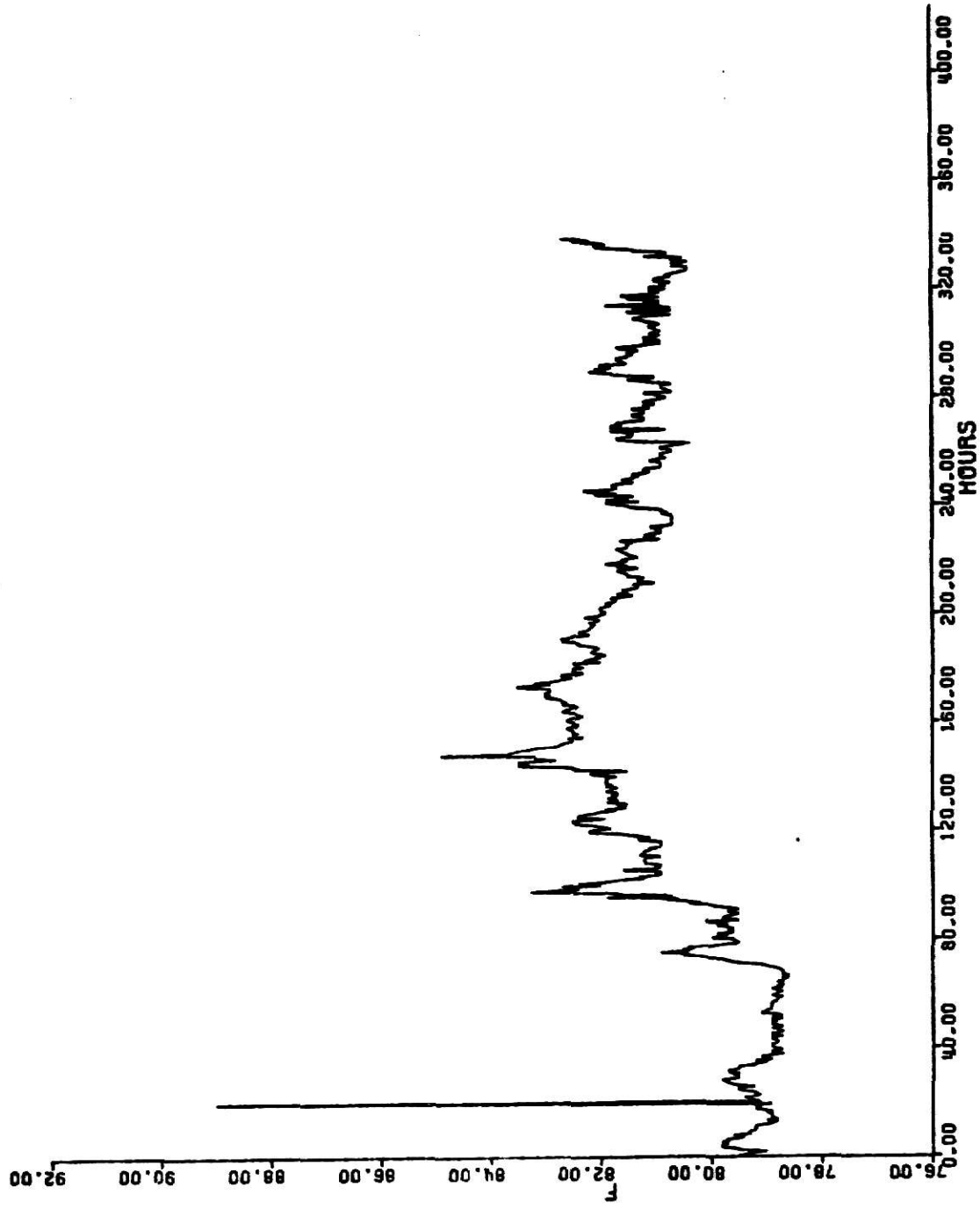

Fig. 5.5 Temperature record -station 3.

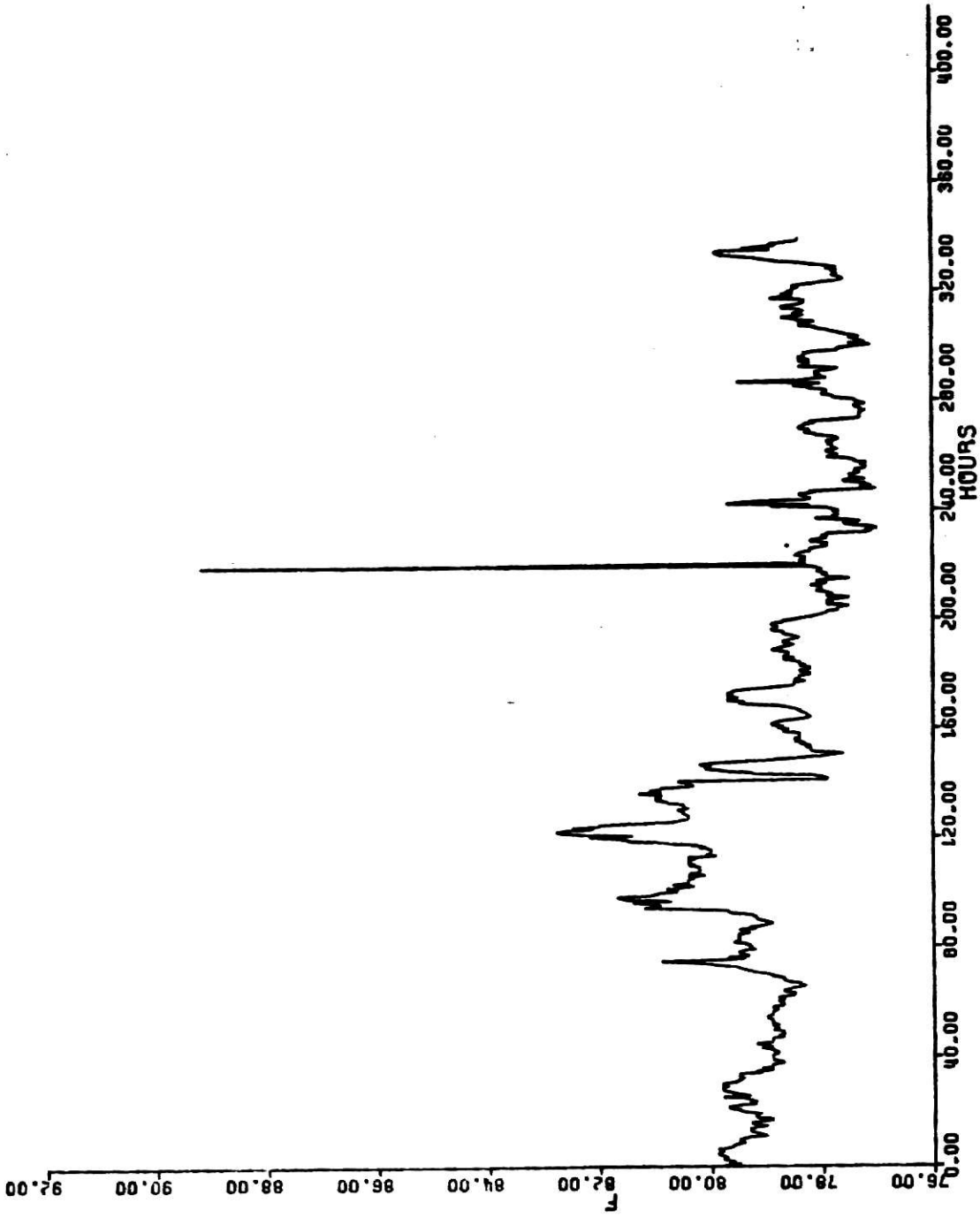

Fig. 5.6 Temperature record - station 4.

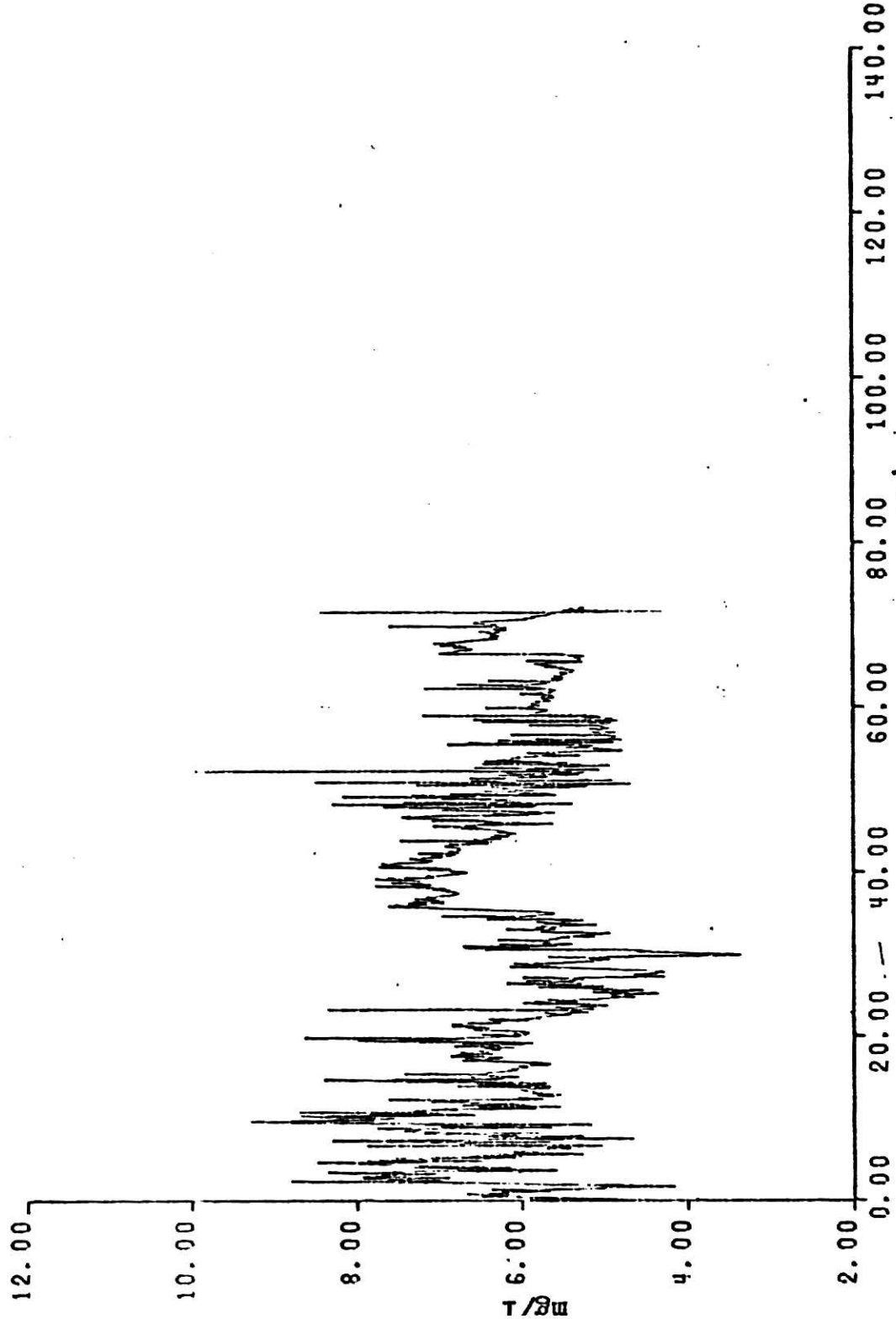

Fig. 5.7 DO record - station 1.

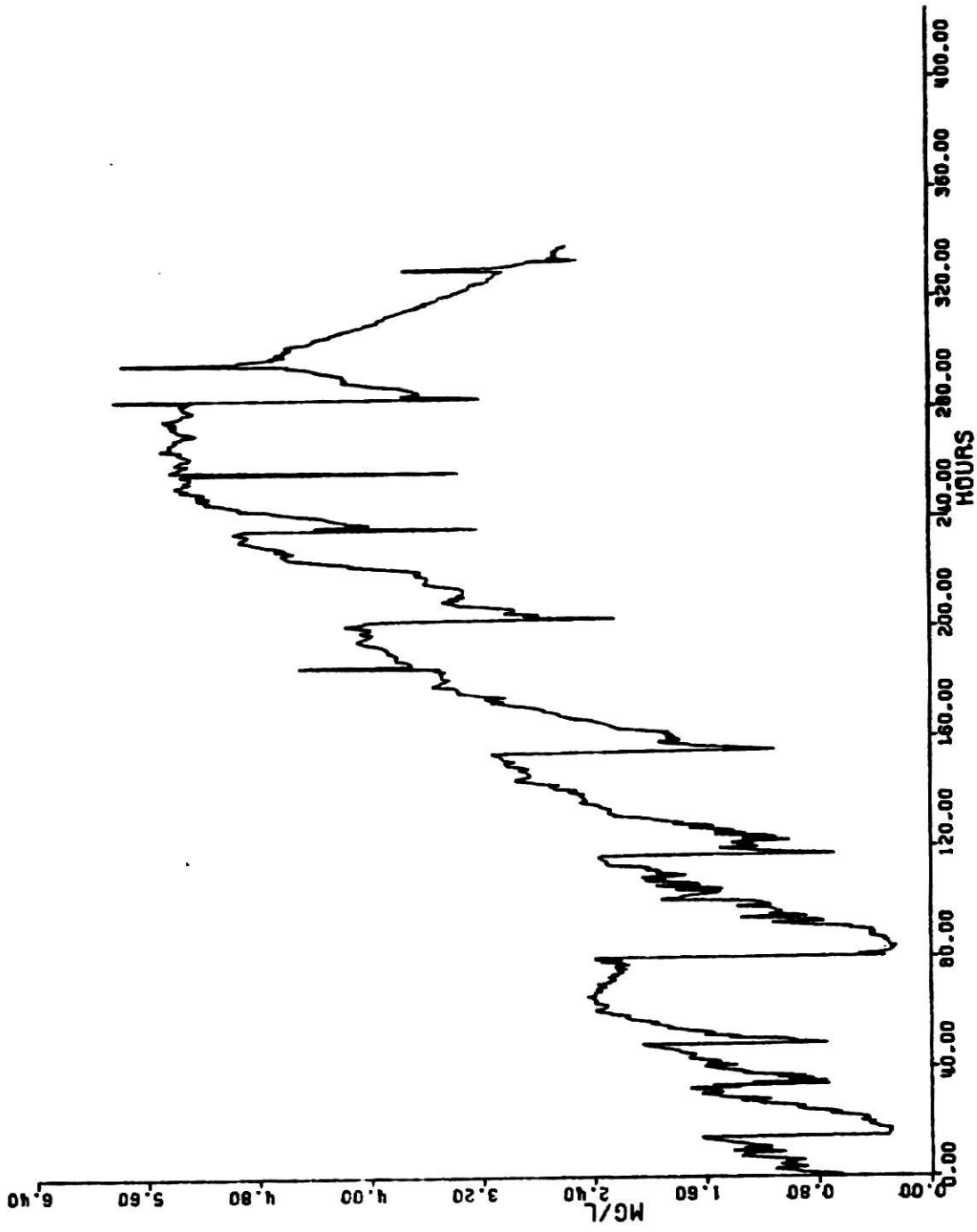

Fig. 5.8 DO record - station 2.

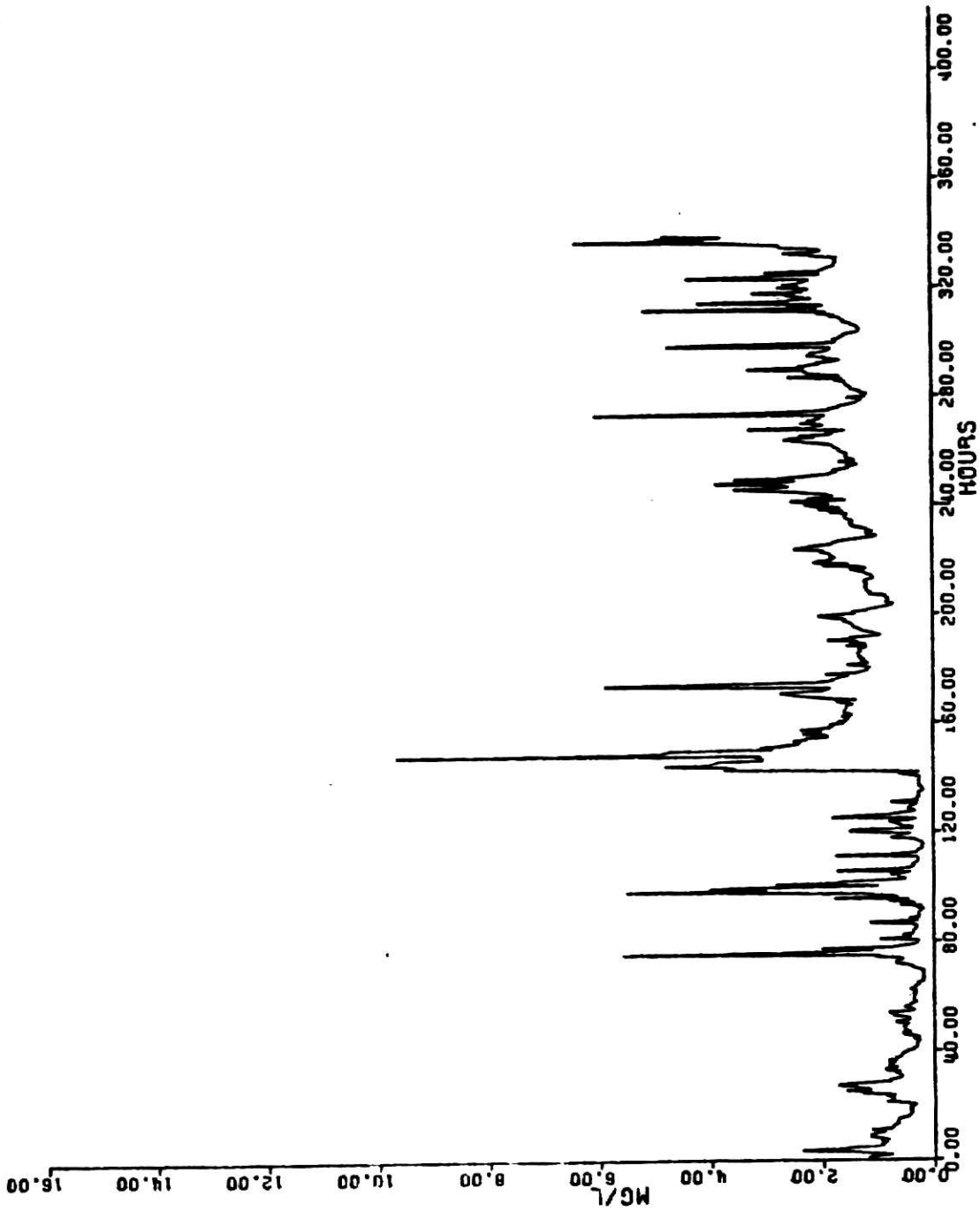

Fig. 5.9 DO record - station 3.

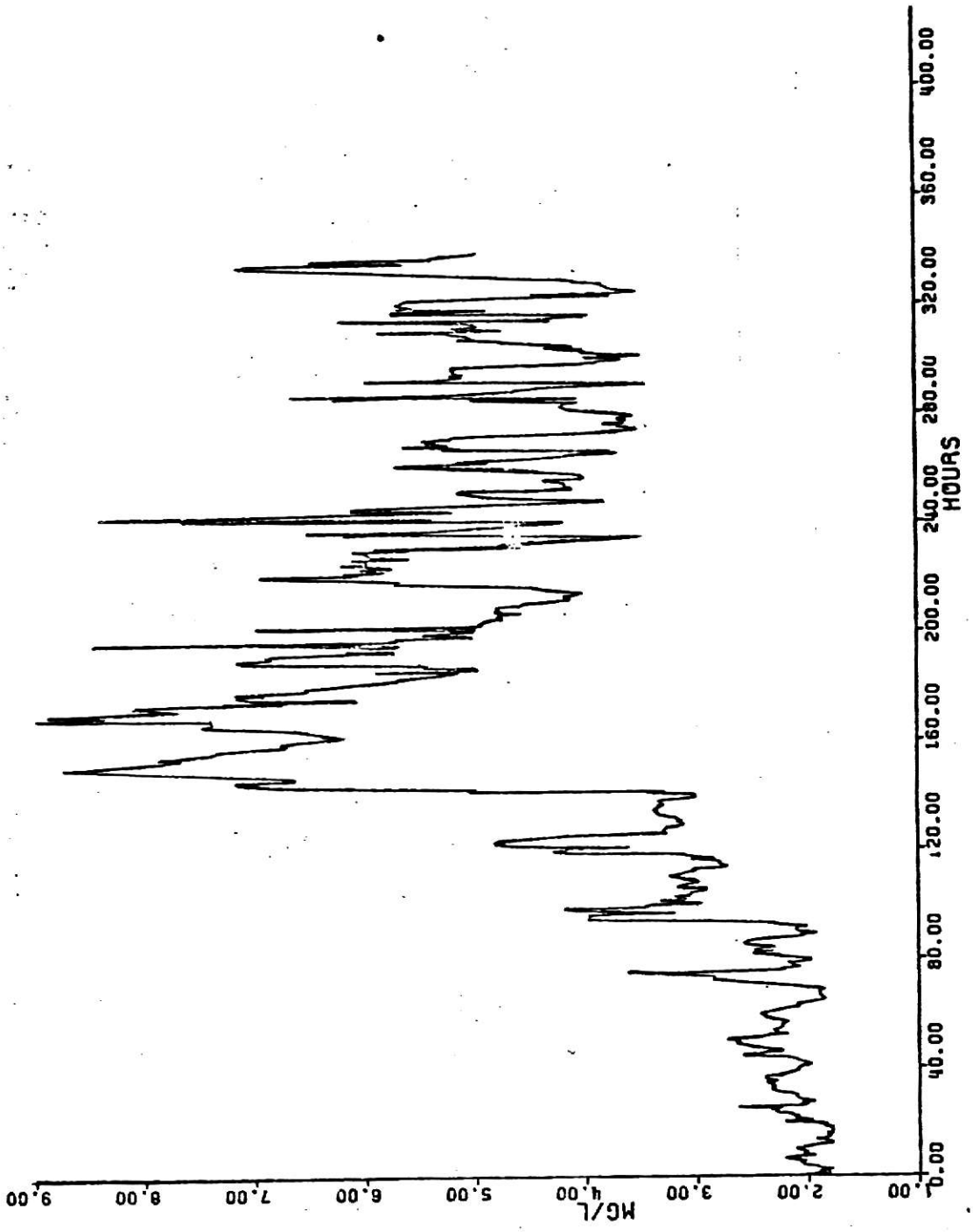

Fig. 5.10 DO record - station 4.

Table 5.2

Mean DO and temperature levels

Station	DO mg/l	Temperature $^{o}_{F}$
1	6.19	80.31
2	2.90	80.87
3	1.45	81.00
4	4.25	78.87

The dissolved oxygen level falls sharply after the first station and shows a rise at the fourth station. The sharp decline in DO level from station 1 to station 3 may be attributed to the oxidation of the sewage discharges from Arlington County, Washington D.C., Alexandria County and Fairfax County. The oxidation of these organic discharges nearly reaches completion at station 4, thereby increasing the DO level. The temperature level does not show a wide variation over the stations.

5.2.2 Harmonic Analysis: Tables 5.3 thru 5.10 summarize the results for harmonic analysis of temperature and dissolved oxygen data for stations 1, 2, 3 and 4. The results for station 1 are based on one month of data and for other stations are based on 14 days data. It is observed that mean alone accounts for about 70-90% of the total mean square value for DO and about 98% of the total value for temperature. The decomposition of contributions to the mean square due to each harmonic was done by taking

Table 5.3 Harmonic Analysis - Temperature Station 1

Mean = 80.31°F Total Variance = 2.63(°C)^2

Source	Amplitude °C	Phase (in degrees)	Percentage contribution to total variance
Fundamental	0.678	31.6	29.68
2nd harmonic	0.522	-30.3	17.55
3rd harmonic	0.672	-79.3	29.14
4th harmonic	0.495	-33.2	15.83
5th harmonic	0.164	-36.79	1.75
30th harmonic	0.221	48.30	3.15

Table 5.4 Harmonic Analysis - Dissolved Oxygen Station 1

Mean = 6.19 mg/ℓ Total Variance = 0.656 $(mg/\ell)^2$

Source	Amplitude (mg/ℓ)	Phase (in degrees)	Percentage contribution to total variance
fundamental	0.061	51.1	1.14
2nd harmonic	0.361	73.8	40.46
3rd harmonic	0.147	73.1	6.72
4th harmonic	0.076	60.7	1.83
5th harmonic	0.119	14.6	4.39
9th harmonic	0.085	-3.02	2.26
10th harmonic	0.110	-4.96	3.78
11th harmonic	0.108	-7.7	3.67
13th harmonic	0.089	78.46	2.49
30th harmonic	0.122	62.85	4.65
58th harmonic	0.0847	-44.2	2.23

Table 5.5 Harmonic Analysis - Temperature Station 2

Mean = 80.87°F Total Variance = 3.60 $(^{o}F)^{2}$

Source	Amplitude $^{o}_{F}$	Phase (in degrees)	Percentage contribution to total variance
fundamental	0.931	-86.6	67.39
2nd harmonic	0.613	-6.1	29.21
5th harmonic	0.168	-81.0	2.21
13th harmonic	0.054	-81.7	0.22

Table 5.6 Harmonic Analysis - Temperature Station 3

Mean = 81.0°F Total Variance = 2.13 $(^{o}F)^{2}$

Source	Amplitude (^{o}F)	Phase (in degrees)	Percentage contribution to total variance
fundamental	0.643	-43.7	51.65
2nd harmonic	0.4477	-21.4	25.03
9th harmonic	0.094	3.8	1.10
13th harmonic	0.089	72.6	1.01
14th harmonic	0.221	69.6	6.07
16th harmonic	0.103	10.04	1.33

Table 5.7 Harmonic Analysis - Temperature Station 4

Mean = 78.87°F Total Variance = 1.01 $(^{o}F)^{2}$

Source	Amplitude °F	Phase (in degrees)	Percentage contribution to total variance
fundamental	0.543	24.5	48.66
2nd harmonic	0.163	-32.5	4.38
3rd harmonic			
4th harmonic	0.284	-29.0	13.33
6th harmonic	0.143	-11.1	3.37
8th harmonic	0.112	-28.9	2.09
14th harmonic	0.2277	63.6	8.57
27th harmonic	0.1008	57.7	1.68

Table 5.8 Harmonic Analysis - Dissolved Oxygen Station 2.

Mean = 2.90 mg/ℓ Total Variance = 2.15 $(mg/\ell)^2$

Source	Amplitude (mg/ℓ)	Phase (in degrees)	Percentage contribution to total variance
fundamental	0.907	83.0	75.90
2nd harmonic	0.264	33.1	6.43
3rd harmonic	0.184	11.8	3.13
4th harmonic	0.122	78.9	1.39
5th harmonic	0.183	-71.8	3.11
8th harmonic	0.120	-7.6	1.33
9th harmonic	0.149	-38.7	2.07
14th harmonic	0.049	-78.7	0.23

Table 5.9 Harmonic Analysis - Dissolved Oxygen Station 3

Mean = 1.45 mg/ℓ Total Variance = 1.18 $(mg/\ell)^2$

Source	Amplitude mg/ℓ	Phase (in degrees)	Percentage contribution to total variance
fundamental	0.325	-89.5	17.92
2nd harmonic	0.245	-49.2	10.19
4th harmonic	0.233	-38.9	9.16
6th harmonic	0.134	87.9	3.03
7th harmonic	0.092	0.1	1.44
14th harmonic	0.280	32.1	13.26
27th harmonic	0.149	-75.2	3.78

Table 5.10 Harmonic Analysis - Dissolved Oxygen Station 4

Mean = 4.25 mg/ℓ Total Variance = 3.05 $(mg/\ell)^2$

Source	Amplitude (mg/ℓ)	Phase (in degrees)	Percentage contribution to total variance
fundamental	0.911	-30.3	53.83
2nd harmonic	0.446	7.78	12.93
3rd harmonic	0.181	-14.6	2.13
4th harmonic	0.311	12.7	6.31
5th harmonic	0.178	1.17	2.07
6th harmonic	0.183	36.81	2.17
13th harmonic	0.133	20.4	1.15
14th harmonic	0.242	69.9	3.81
15th harmonic	0.172	48.2	1.92
27th harmonic	0.121	30.7	0.96

mean square value of x_i about the mean i.e. variance. As the results for harmonic analysis indicate, the major contribution to variance is made by the first five harmonics. The cause of variability in temperature due to these harmonics could not be determined in the absence of longer data. Another significant variation is caused by the 30th harmonic for the first station and the 13th harmonic for other stations. This 24 hr. cyclic fluctuation is due to variable solar heating during day and night. For dissolved oxygen data at station 1, the first harmonic (corresponding to 30 days period) is relatively non-significant. The variation in dissolved oxygen due to the first five harmonics can be attributed to the non-linear interaction of the solubility of dissolved oxygen with temperature, the variability of photosynthesis with temperature and sunlight etc. The dependence of the saturation value of dissolved oxygen on temperature can be given by the empirical non-linear relation [28]

$$C_S = 14.652 - 0.41022T + 0.0079910T^2 - 0.00077774T^3 \qquad (5.1)$$

A 15 days periodicity (2nd harmonic for 1st station and 1st harmonic for other stations) in dissolved oxygen shows the effect of lunar fortnightly tide. Other harmonics responsible for significant variation in the dissolved oxygen are the 30th and the 58th for first station; the 13th and the 27th for stations 3 and 4. These harmonics correspond to variations due to diurnal variation in temperature, BOD and semidiurnal tides respectively. Station 2 does not show any significant variation due to semidiurnal tide.

After having performed the harmonic analysis to obtain initial information about the behaviour of all pollutants, spectral analysis was carried out

to obtain some more information about the system.

5.2.3 Spectral Analysis: Figures 5.11 thru 5.18 show the autocorrelation plots for temperature and dissolved oxygen for stations 1,2,3 and 4. Their failure to damp off quickly indicates the presence of trend in the data. A pilot study of the spectral estimate of raw data resulted in high peaks at zero frequency. It was decided to use prewhitening. In this analysis, the factor $\alpha = 0.99$ was used for prewhitening. A series of spectral estimates were obtained using different value of lags and it was found that 100 lags give satisfactory results. Figures 5.19 thru 5.26 show the recolored power spectral estimates for the eight series. A high peak at zero frequency is observed in all power spectra. This may be taken as an indication a trend being present, but in view of the short data length, a long periodic fluctuation (annual or seasonal) may also cause this peak. This assumption seems to be more realistic as the temperature and dissolved oxygen are known to follow an annual cycle. Additional data is warranted to obtain more reliable information about the zero frequency component. The temperature power spectra at all stations show a high peak at a 24 hour cycle. This fluctuation may be attributed to variable solar heating. Another peak is observed at a 12 hour cycle. No physical significance could be attributed to this peak. It could be a harmonic of the 24 hour cycle. Dissolved oxygen spectra show two definite peaks at 24 hour and 12 hour cycle corresponding to diurnal variation in temperature, photosynthesis and semidiurnal variation due to tidal phenomena. Table 5.11 summarizes the important components of power spectra for dissolved oxygen records at the four stations. It is seen from the Table 5.11 that the maximum variance at 24 hour and

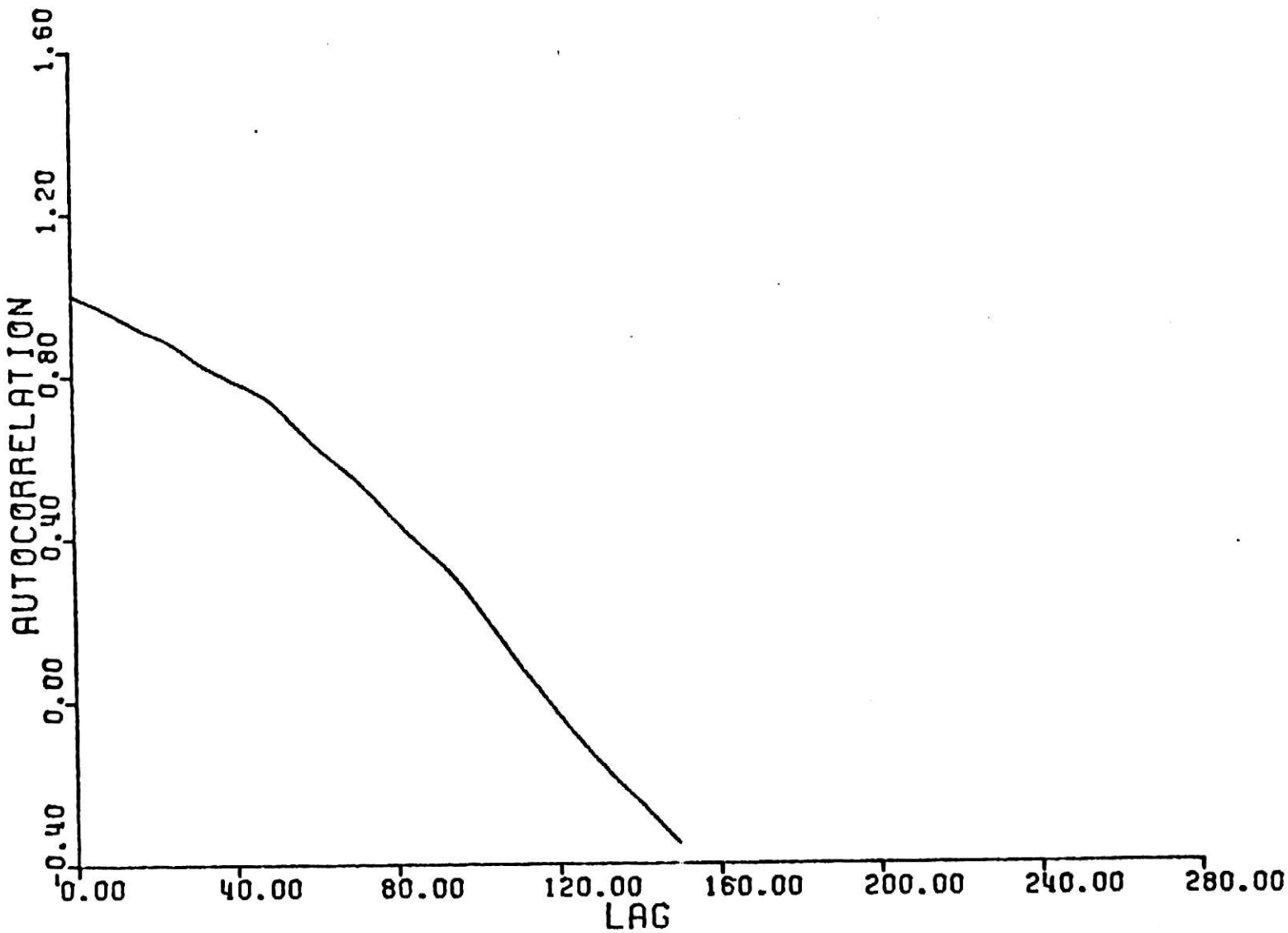

Fig. 5.11 Autocorrelation of original temperature data - station 1.

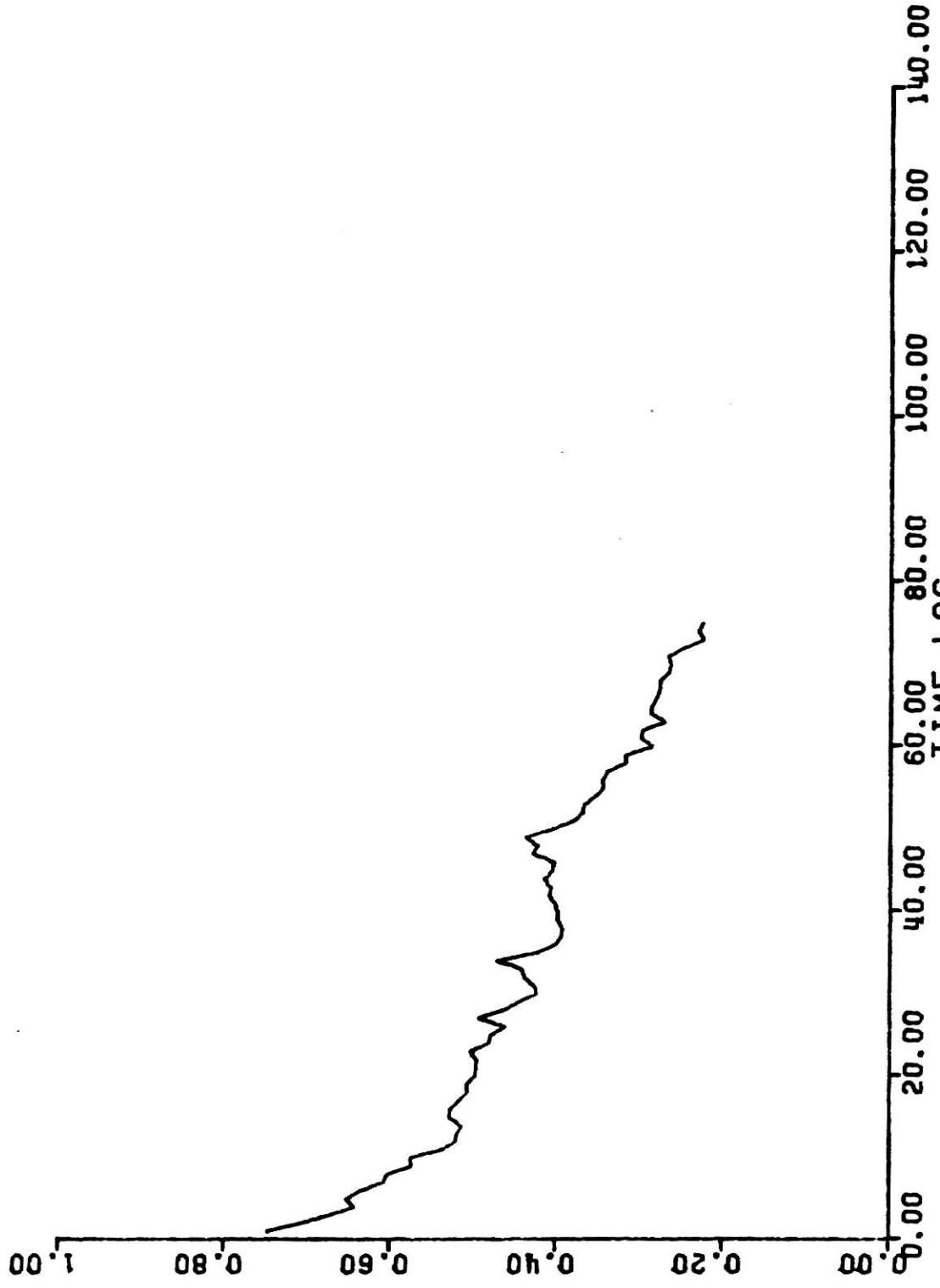

Fig. 5.12 Autocorrelation of original DO data - station 1.

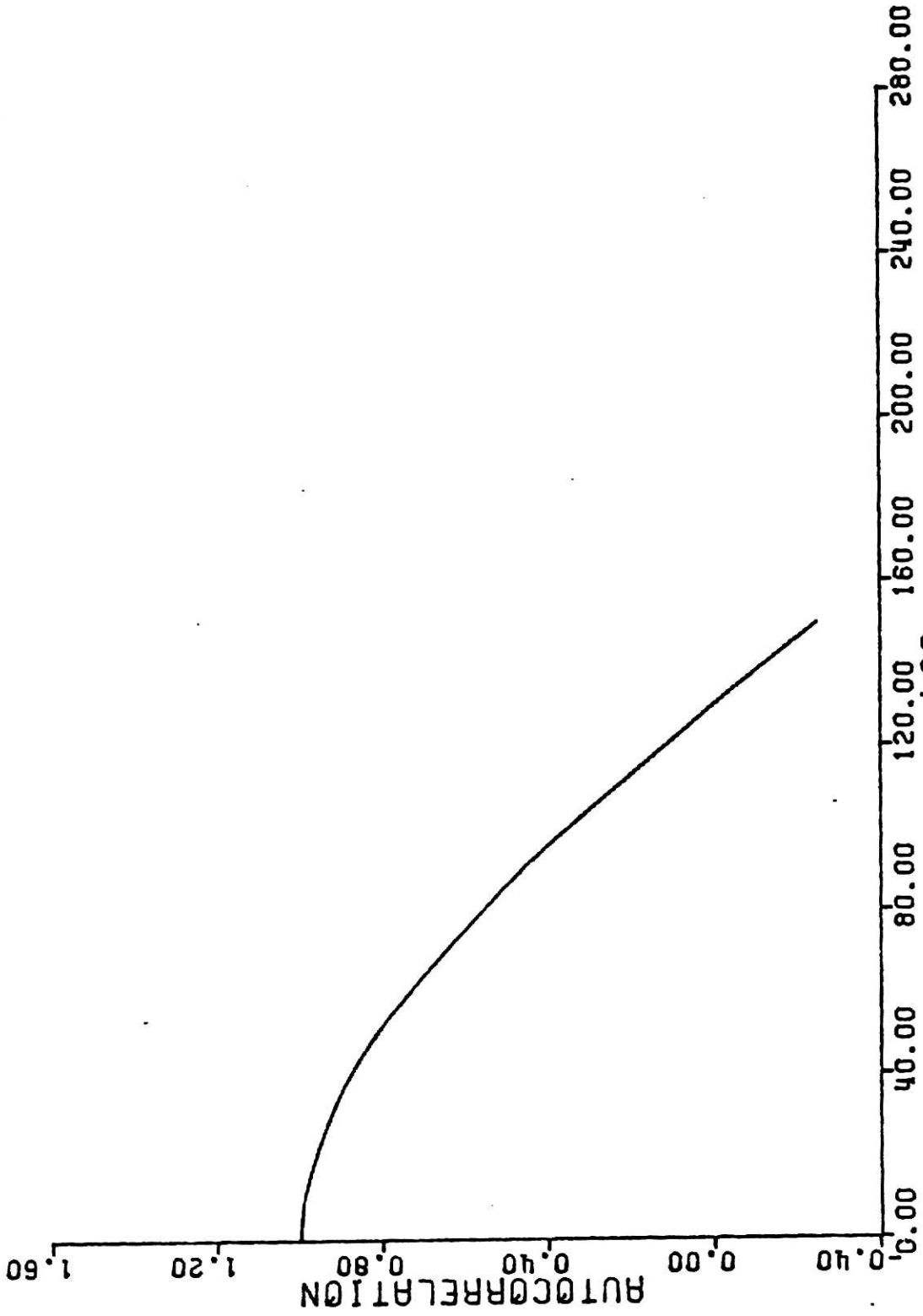

Fig. 5.13 Autocorrelation of original temperature data - station 2.

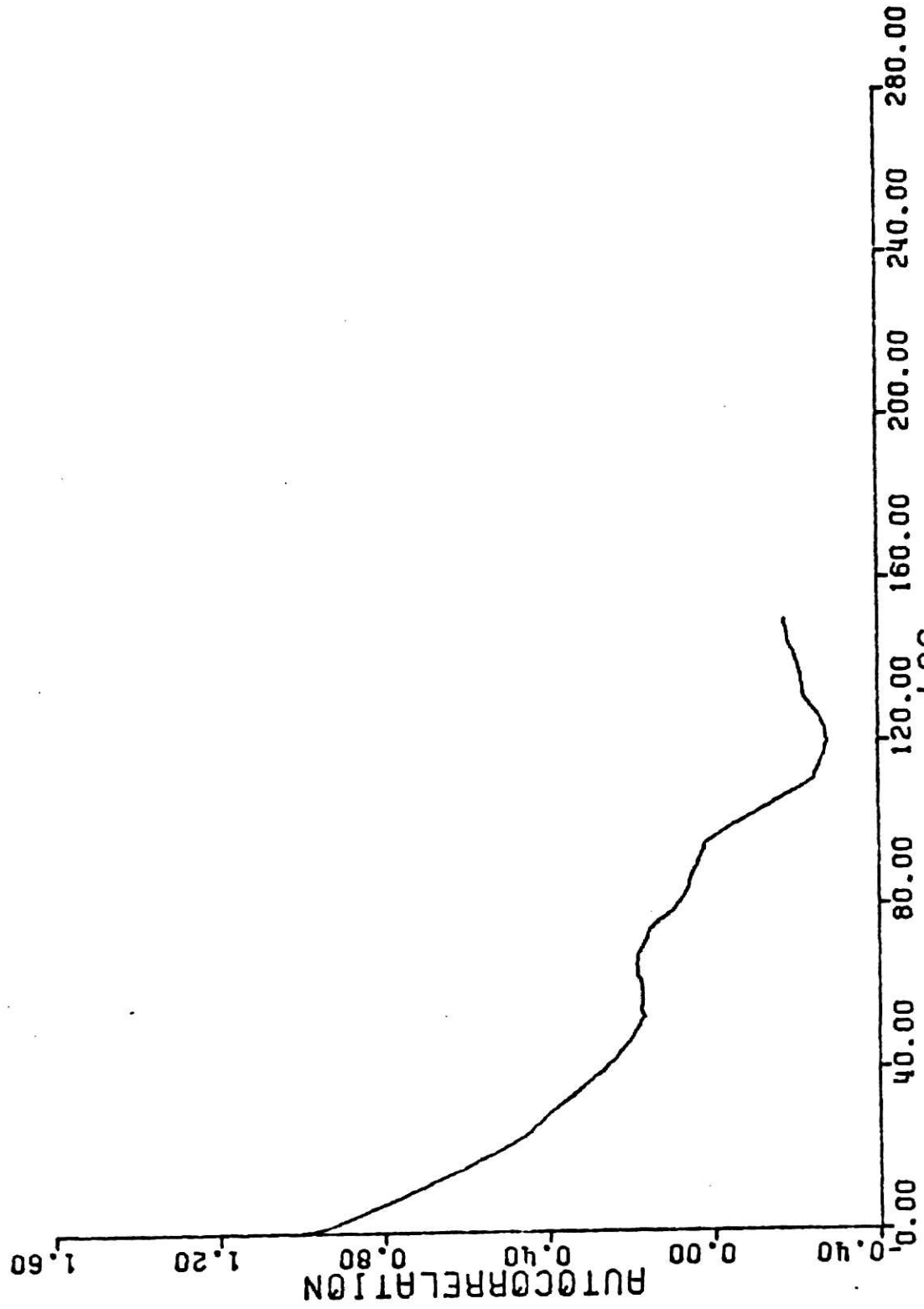

Fig. 5.14 Autocorrelation of original DO data - station 2.

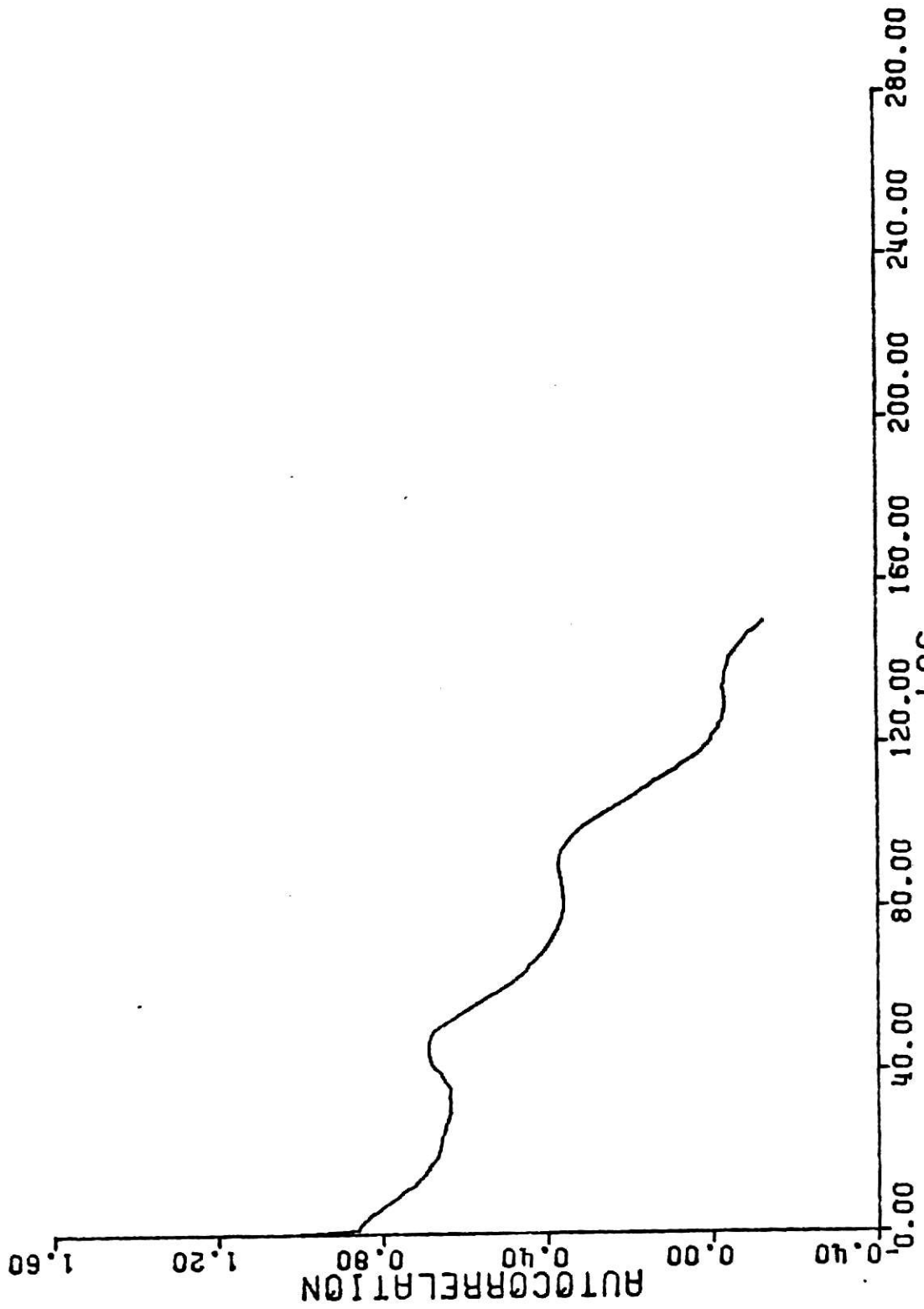

Fig. 5.15 Autocorrelation of original temperature data - station 3.

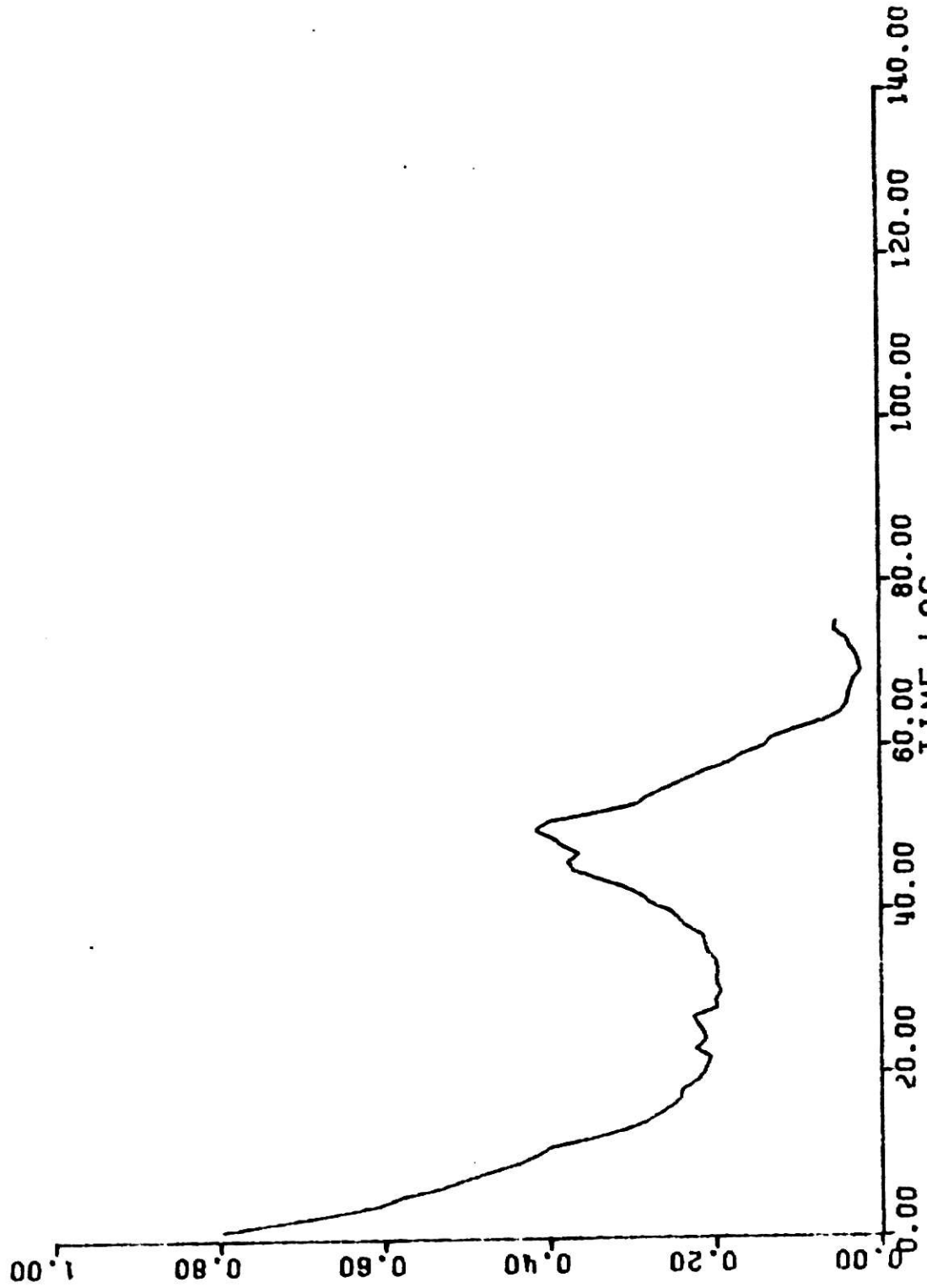

Fig. 5.16 Autocorrelation of original DO data - station 3.

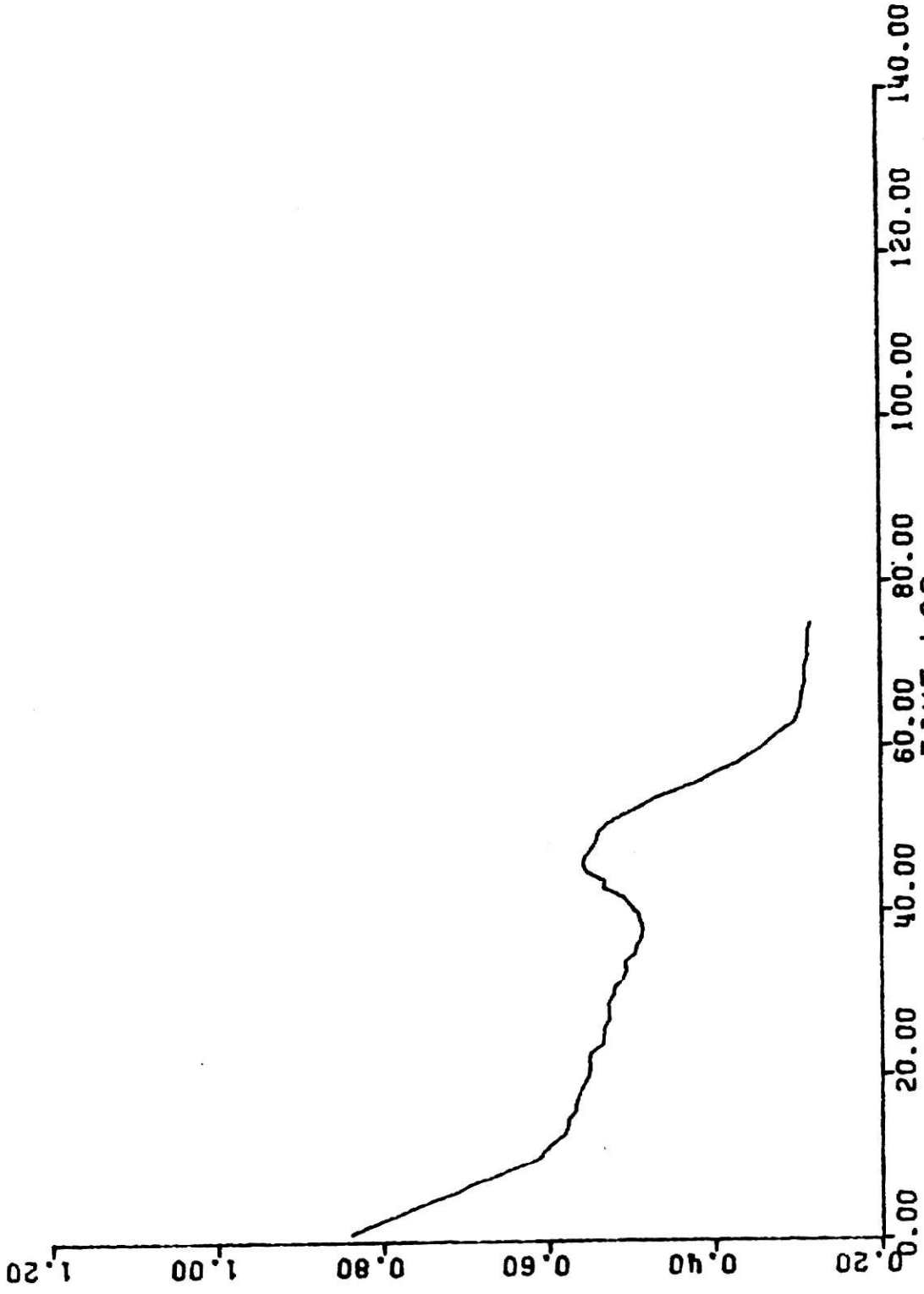

Fig. 5.17 Autocorrelation of original temperature data - station 4.

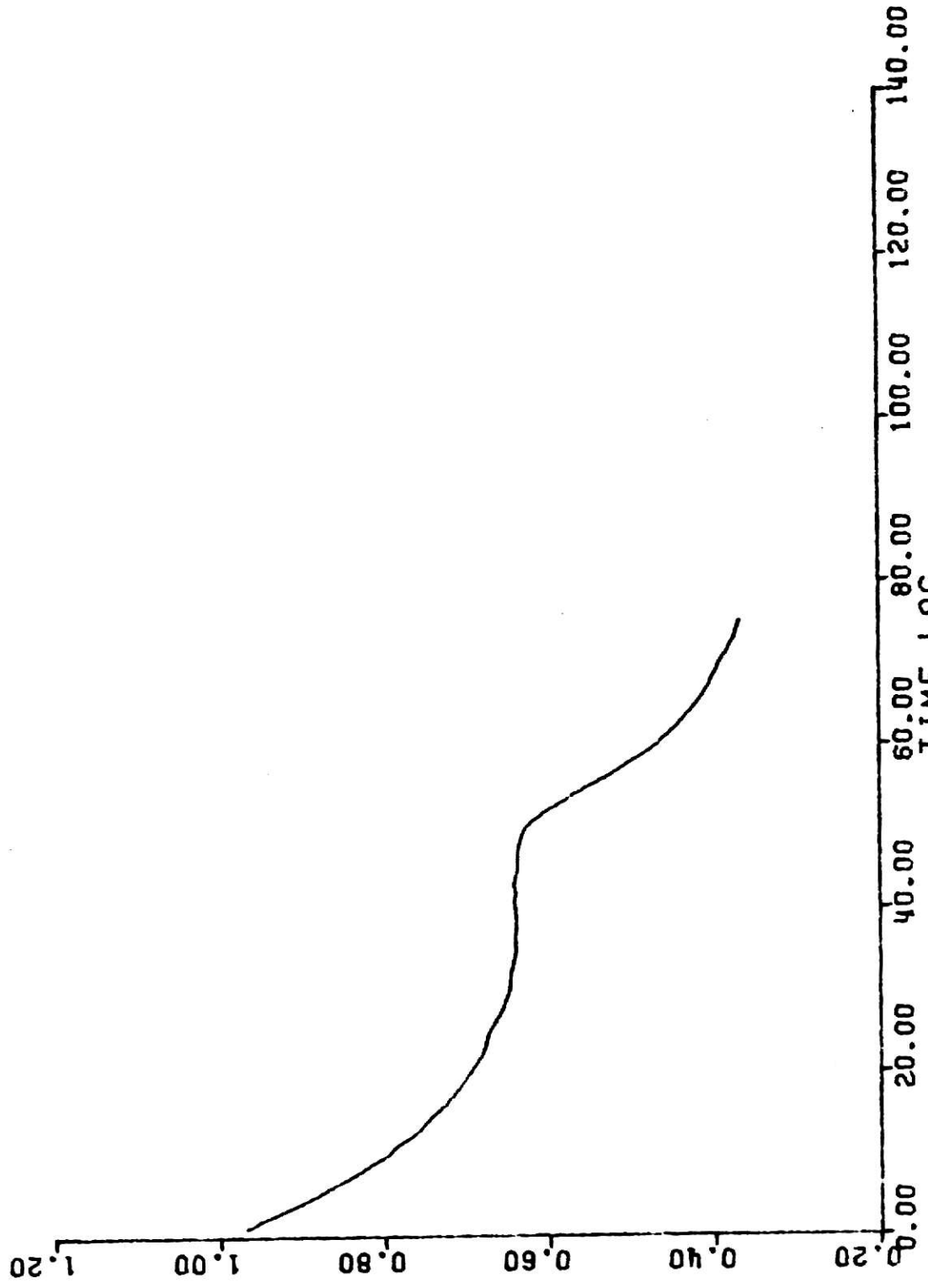

Fig. 5.18 Autocorrelation of original DO data - station 4.

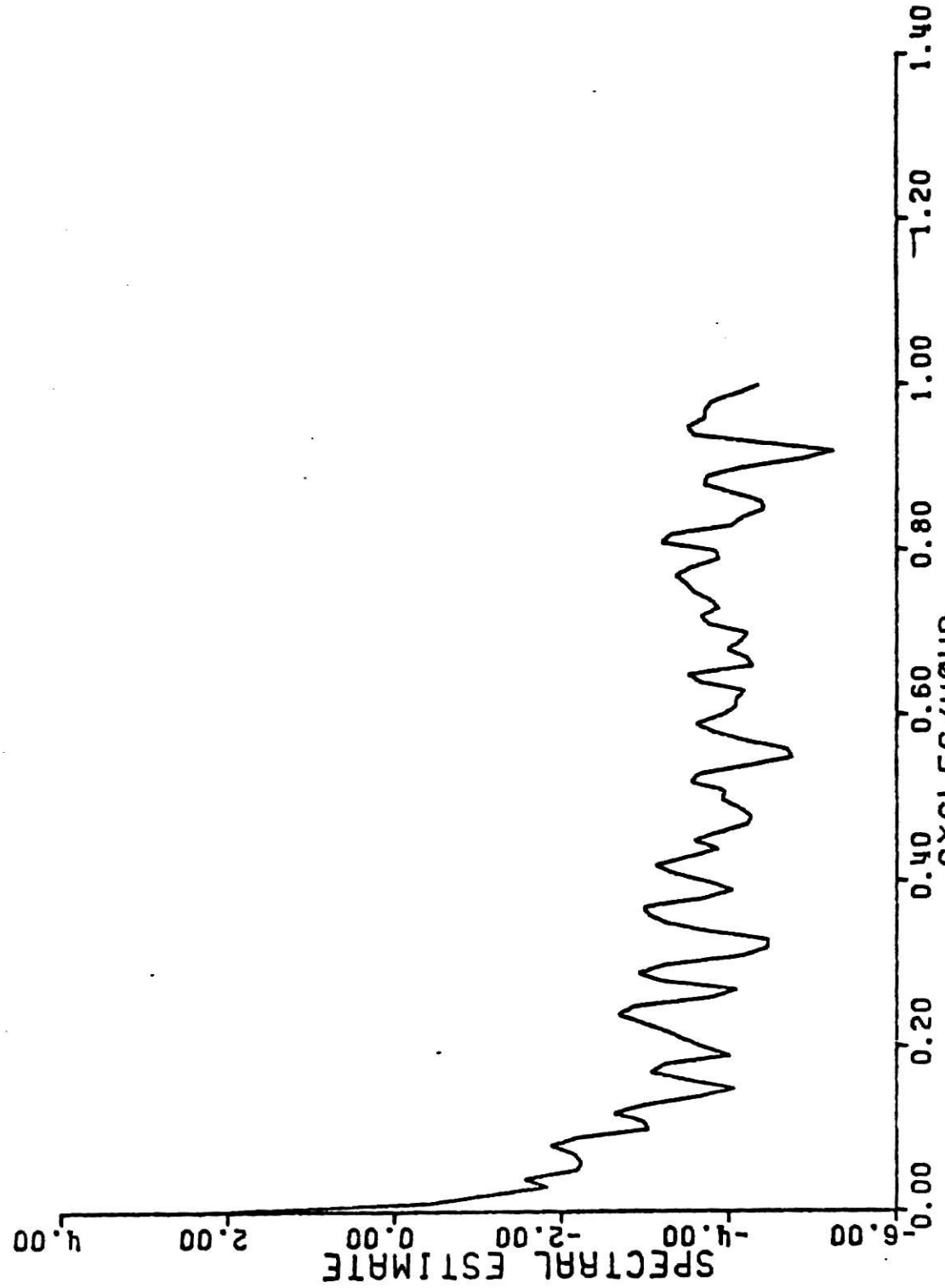

Fig. 5.19 Spectral estimate of DO (recolored) - station 1.

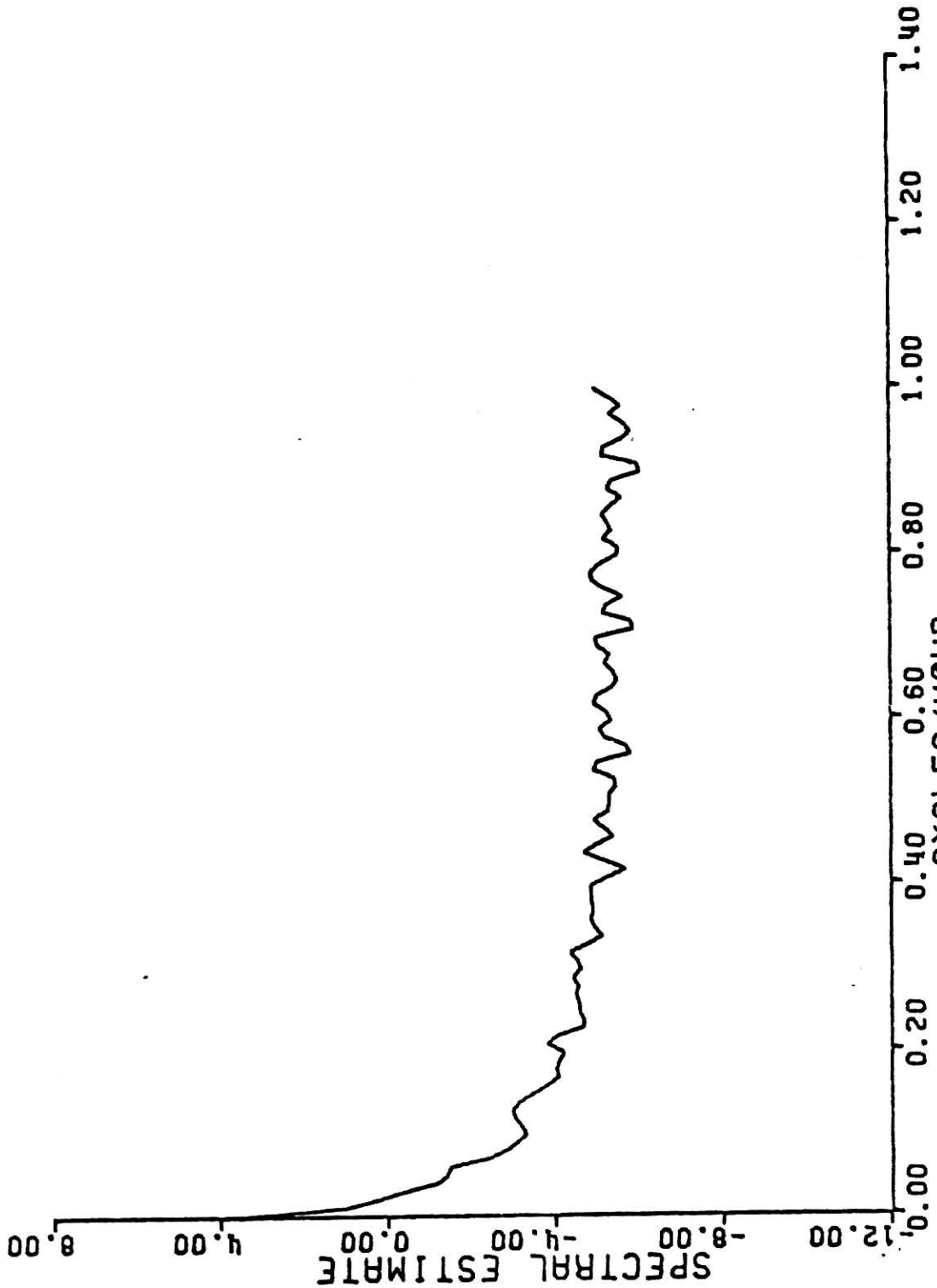

Fig. 5.20 Spectral estimate of DO (recolored) - station 2.

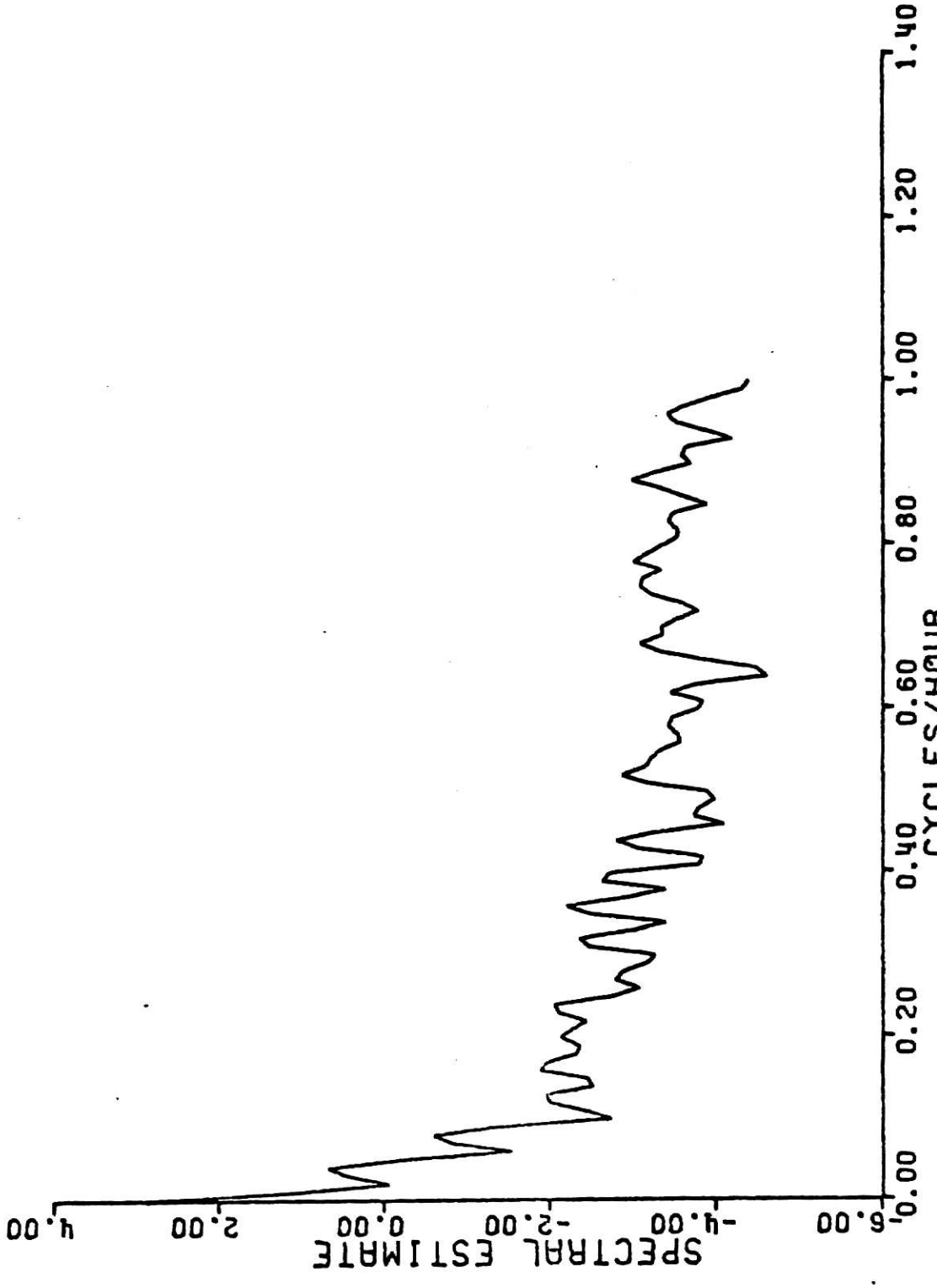

Fig. 5.21 Spectral estimate of DG (recolored) - station 3.

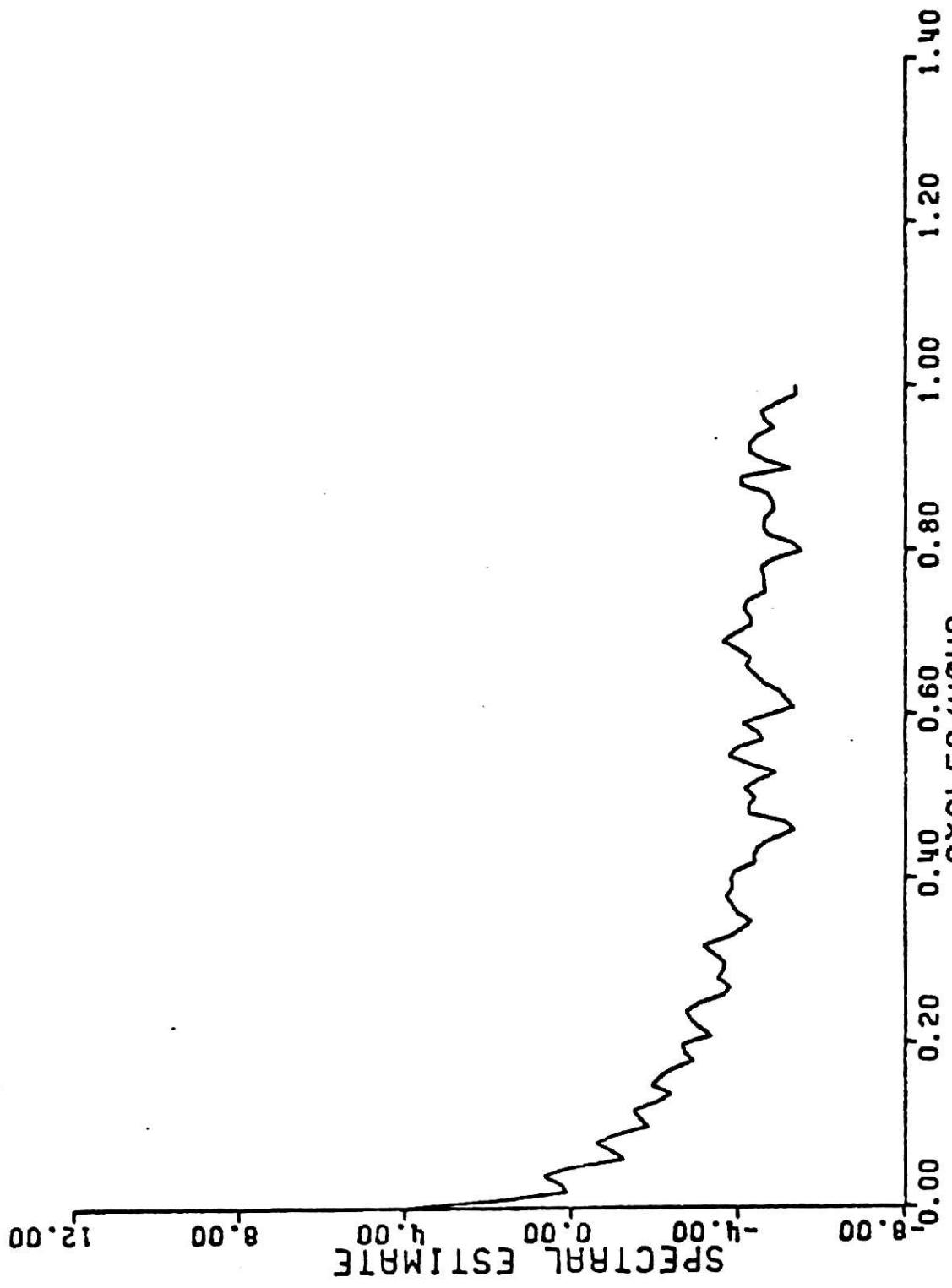

Fig. 5.22 Spectral estimate of DO (recolored) - station 4.

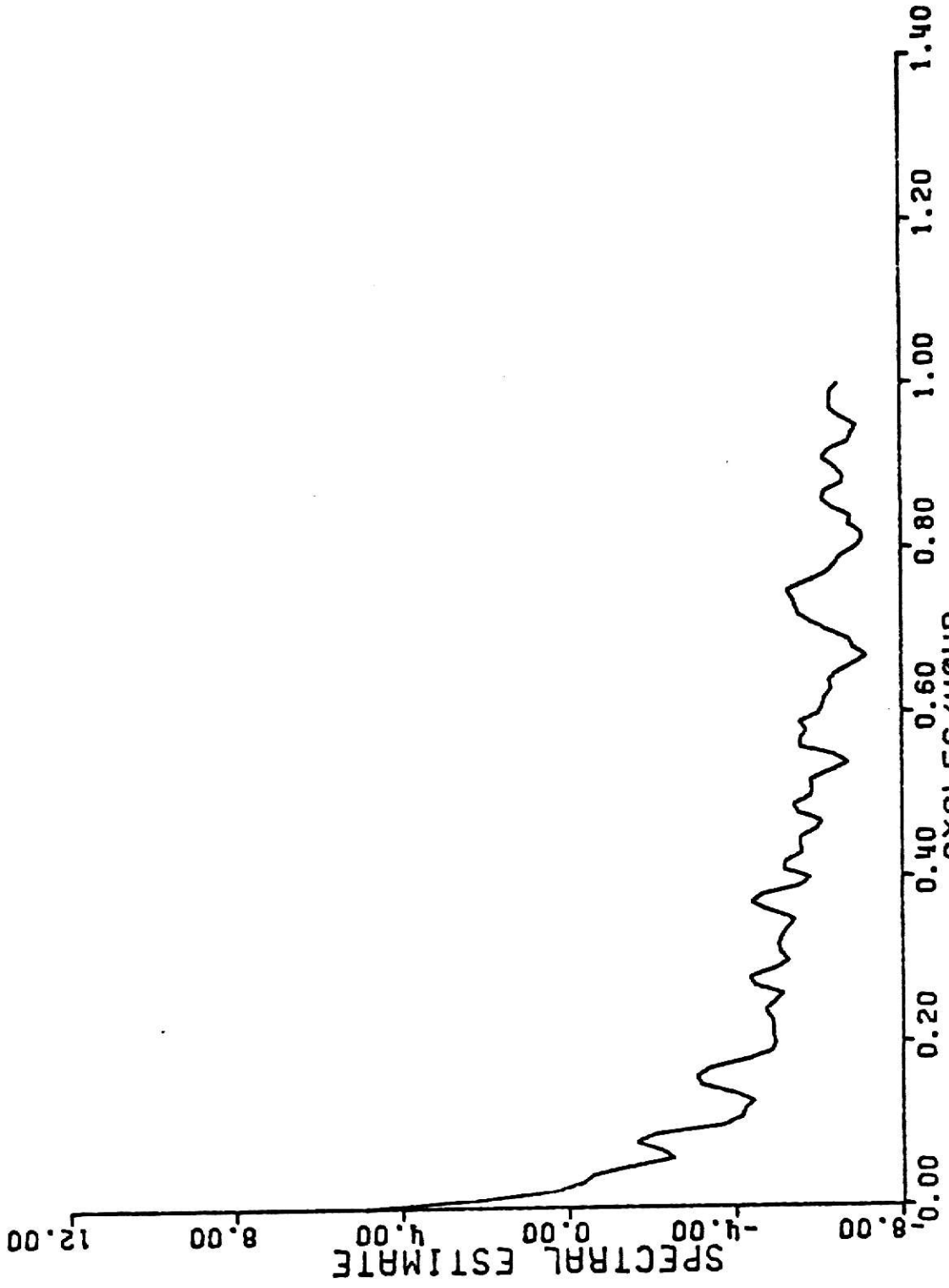

Fig. 5.23 Spectral estimate of temperature (recolored) - station 1.

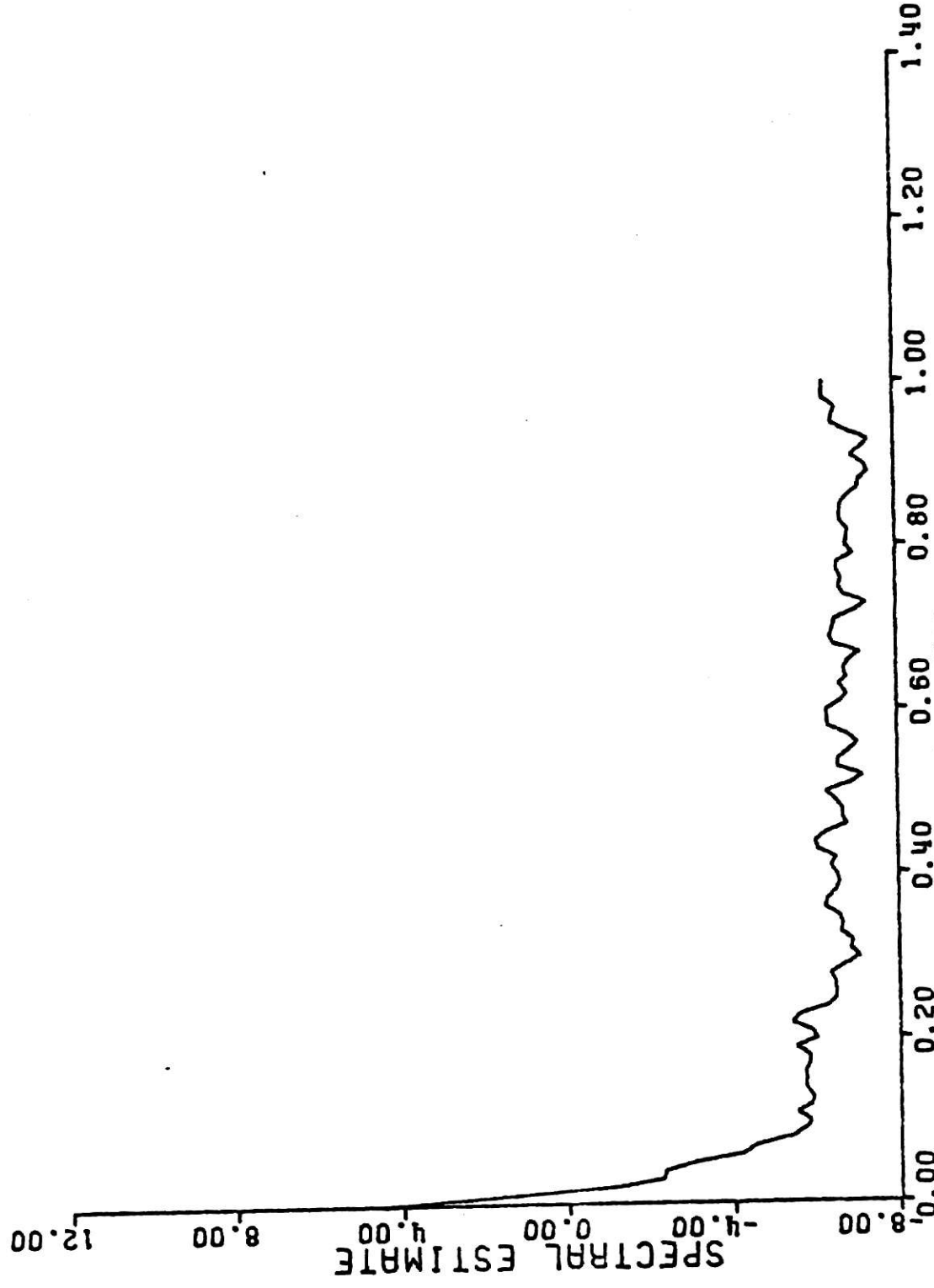

Fig. 5.24 Spectral estimate of temp. (recolorea) - station 2.

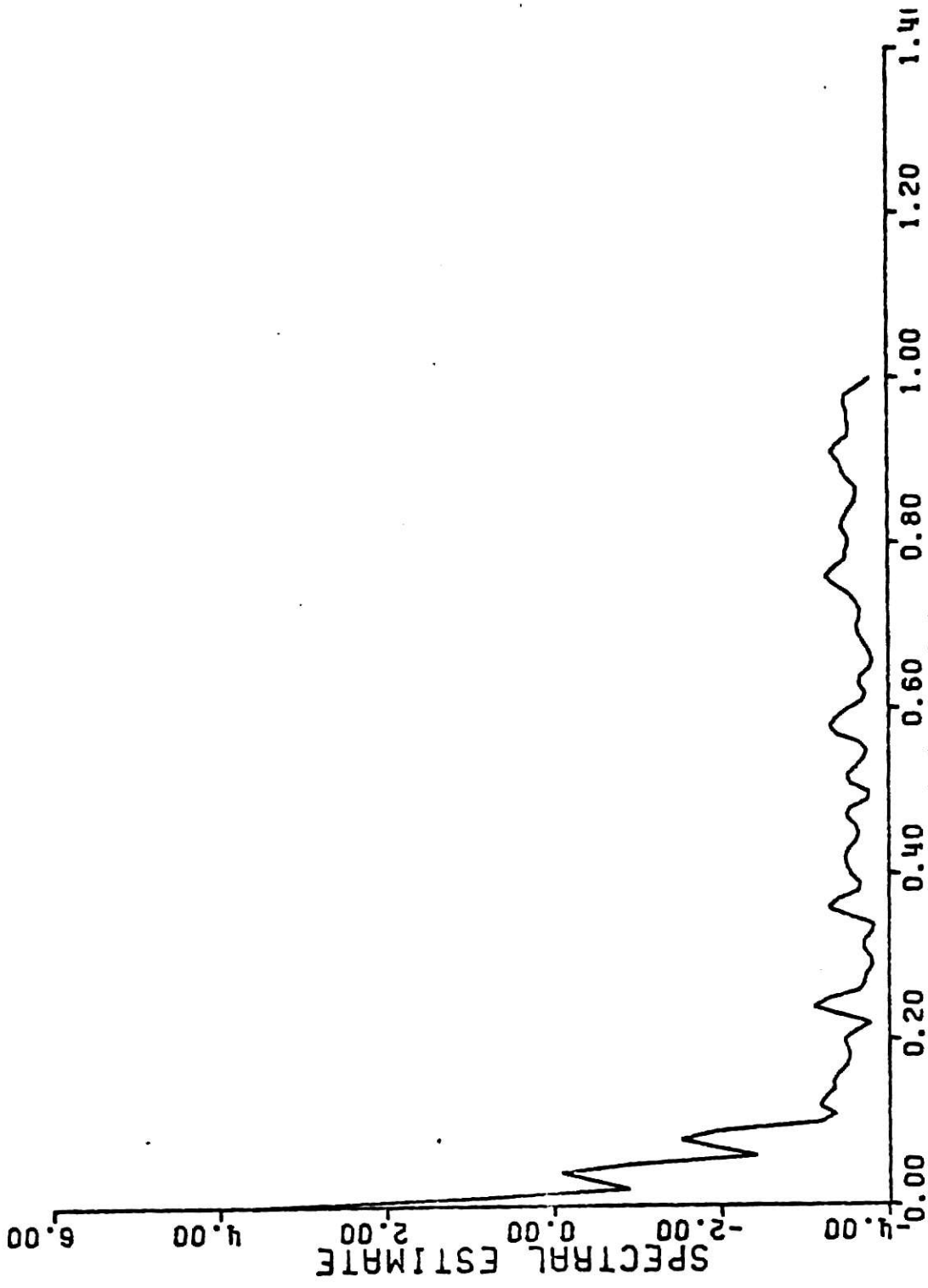

Fig. 5.25 Spectral estimate of temp. (recolored) - station 3.

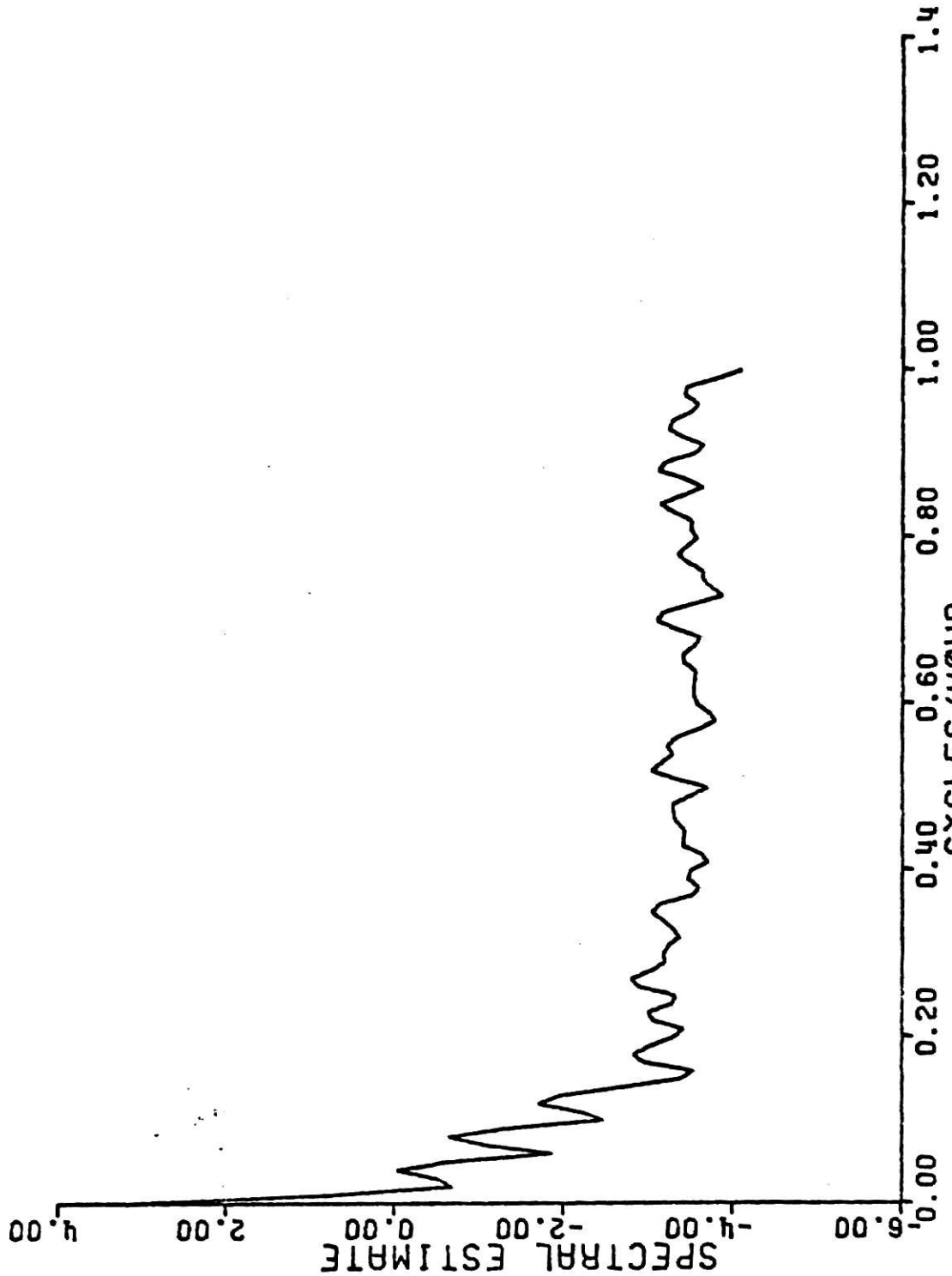

Fig. 5.26 Spectral estimate of temp. (recolored) - station 4.

Table 5.11. Some Components of Power Spectrum of Dissolved Oxygen at Stations 1,2,3 and 4.

Station	Period		
	Long	24 hrs.	12 hrs.
1	10.320	0.212	0.154
2	34.493	0.677	0.058
3	17.457	1.98	0.55
4	59.930	1.99	0.56

12 hour period is associated with stations 3 and 4. Station 3 is the point where the mean DO has a minimum value. Thus the maximum variance in DO occurs where the mean DO is minimum. As station 2 is near to the waste discharge point, hence its diurnal variation may be associated with diurnal variation in stream BOD.

As mentioned earlier, another approach to remove trend is with harmonic regression. This technique was used to remove harmonics from the raw data and thus form a predictive model. The regression equation for temperature at station 1 consisted of regressing 21 independent variables over one dependent variable. For this purpose, the half hourly data for one month was averaged to provide 2 hourly data. Using all 21 variables gave multiple correlation coefficient $R^2 = .974$. It was seen that the addition of the first nine variables resulted in value of $R^2 = .952$. Thus, the addition of 12 more variable does not yield any significant improvement in R^2; though it results in a considerable increase of computational effort. The equation with first nine variables is given as

$$T_t = 79.89 + 0.0051t + 0.5573 \cos\left(\frac{2\pi t}{361}\right) + 1.5231 \sin\left(\frac{2\pi t}{361}\right) + 1.2894 \sin\left(\frac{2\pi t}{120}\right) + 0.8312 \cos\left(\frac{2\pi t}{90}\right) - 0.5344 \sin\left(\frac{2\pi t}{90}\right) + 0.2483 \cos\left(\frac{2\pi t}{12}\right) + 0.3005 \sin\left(\frac{2\pi t}{12}\right) + 1.4961 \cos\left(\frac{2\pi t}{210}\right).$$

with $R^2 = .952$

$F(9,351) = 385.6644$

Similar model was formed for dissolved oxygen for station 1 based on 2 hourly data. The model is

$$DO_t = 6.1970 + 0.1967 \cos\left(\frac{2\pi t}{180}\right) - 0.6960 \sin\left(\frac{2\pi t}{180}\right)$$
$$- 0.2600 \sin\left(\frac{2\pi t}{120}\right) - 0.2340 \cos\left(\frac{2\pi t}{72}\right)$$
$$- 0.1755 \cos\left(\frac{2\pi t}{40}\right) - .2405 \cos\left(\frac{2\pi t}{30}\right)$$
$$- .2192 \cos\left(\frac{2\pi t}{33}\right) + 0.2090 \sin\left(\frac{2\pi t}{12}\right)$$
$$- 0.1503 \sin\left(\frac{2\pi t}{28}\right)$$

with $R^2 = .810$

The power spectra for temperature and dissolved oxygen after removing harmonics is shown in Figures 5.27 and 5.28 respectively.

As these models involve the use of too many parameters, the pollutants at other stations 2,3, and 4 were not modeled using this procedure.

Autoregressive moving average parametric modeling procedure was used to obtain models for temperature and dissolved oxygen at all stations.

5.2.4 Autoregressive Moving Average Models:

Parametric time series models were fitted to all eight series. In the following discussion, temperature and dissolved oxygen models for station 1 will be described in details and the results for other stations will be summarized.

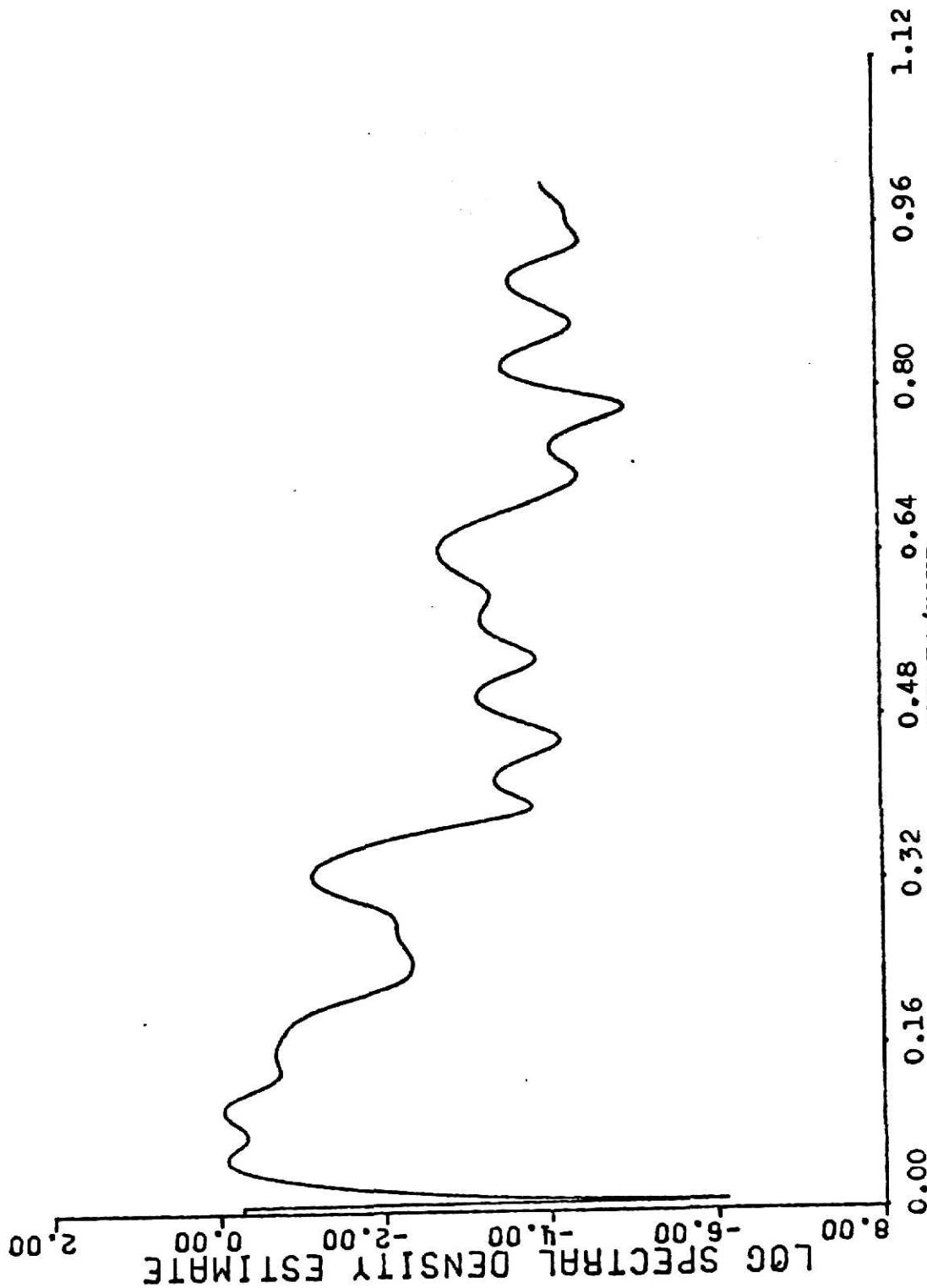

Fig. 5.27 Spectral estimate of temp. (residuals) - station 1.

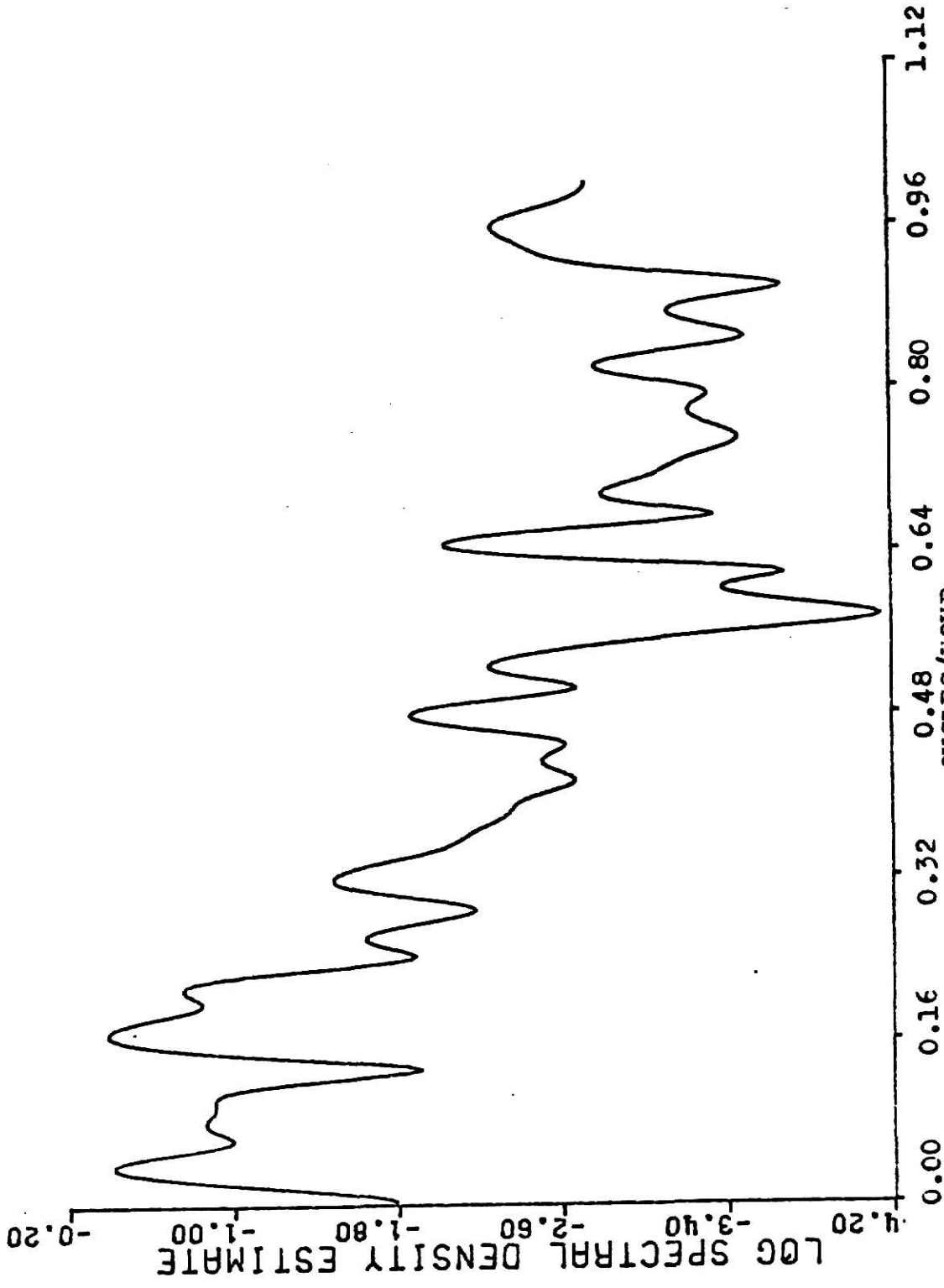

Fig. 5.28 Spectral estimate of DC (residuals) - station 1.

(a) Temperature: The autocorrelation and partial autocorrelation plots for the raw, first differenced (∇x) and second differenced (∇_x^2) data are shown in Figures 5.29 thru 5.34 for 2 hour data interval. The autocorrelations for the raw data are large and fail to die out at higher lags. While simple differencing reduces the correlations in general, a very heavy 12 lag periodic component remains as evidenced by large correlations at lags 12 and 24. Second differencing reduces the correlations isgnificantly but periodic components remain large. Further differencing was carried out to remove the 12 lag periodic component. Figures 5.35 and 5.36 show the autocorrelation and partial autocorrelation obtained by differencing for seasonal component i.e. $\nabla\nabla_{12}x$. This differencing reduces the correlations at all lags considerably. Thus the possible candidate model is

$$w_t = (1 - \theta B)(1 - \theta_{12}B^{12} - \theta_{24}B^{24})\, a_t$$

where, $w_t = \nabla\nabla_{12} x_t$

Using 4.16, the rough estimates of the parameters were obtained as,

$\theta = -0.02$

$\theta_{12} = 0.56$

$\theta_{24} = -0.10$

The least square estimates of these parameters are

$\theta = -0.099$

$\theta_{12} = 0.667$

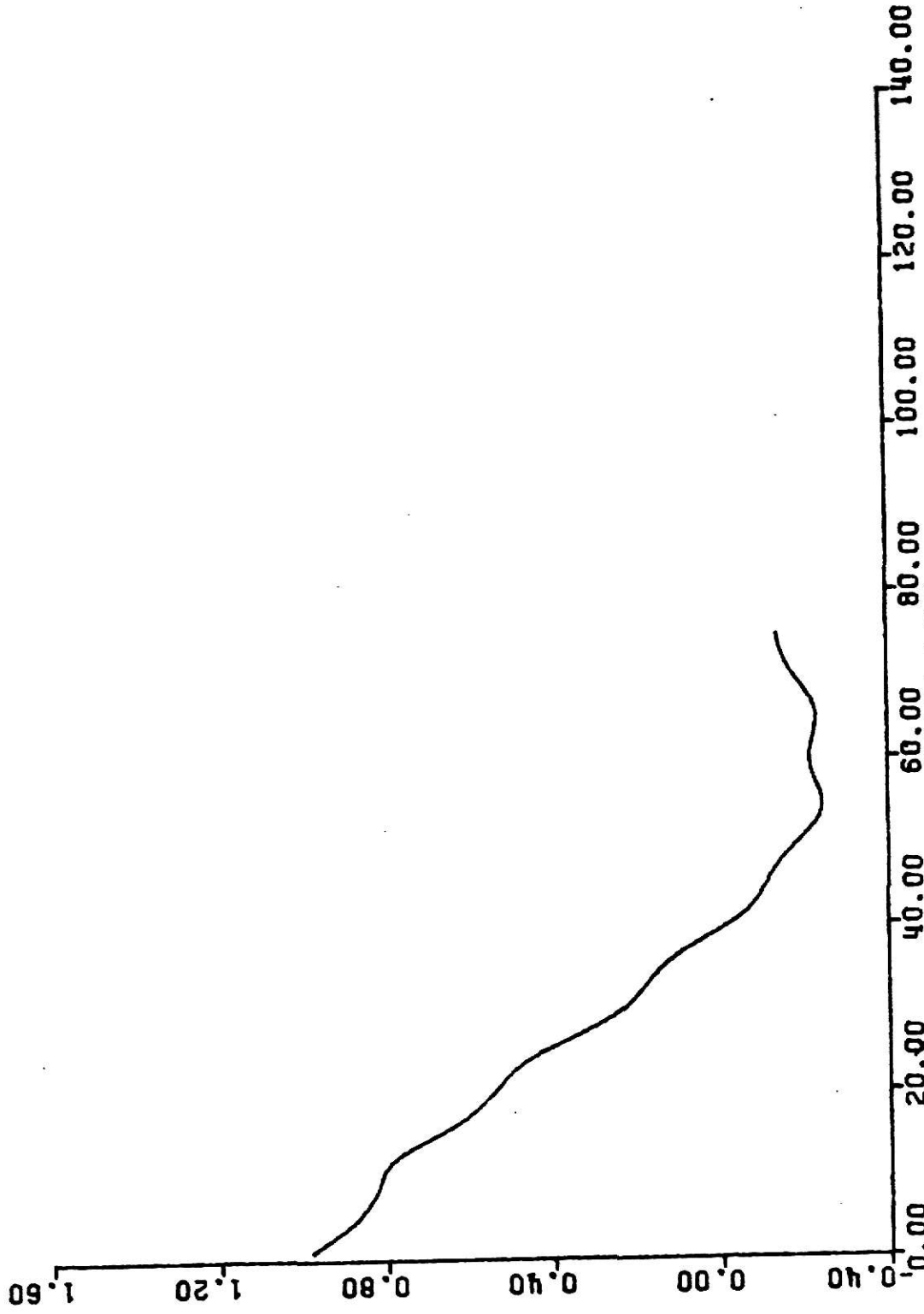

Fig. 5.29 Autocorrelation of temp. (2 hourly data) - station 1.

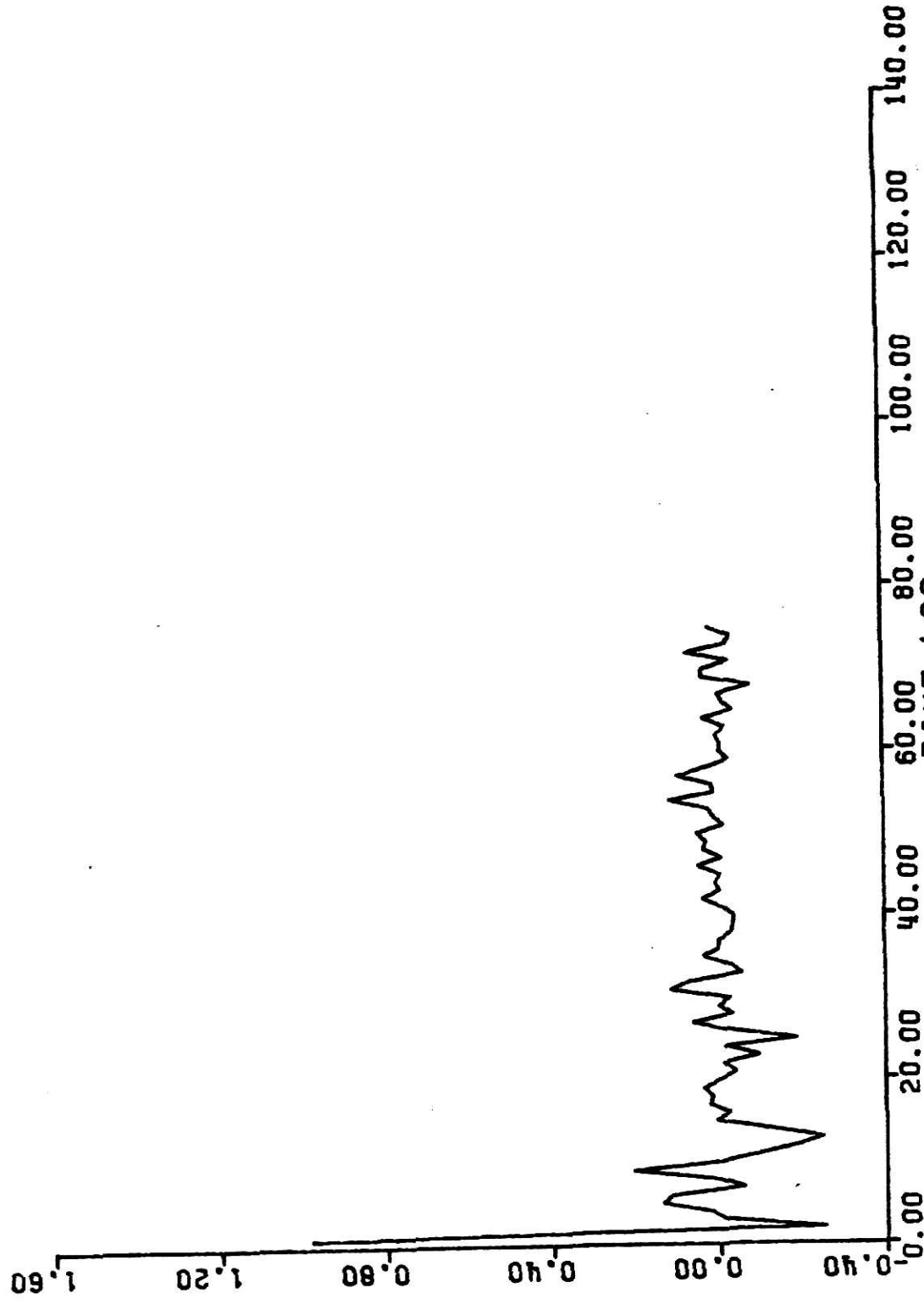

Fig. 5.30 Partial autocorrelation of temp. - station 1.

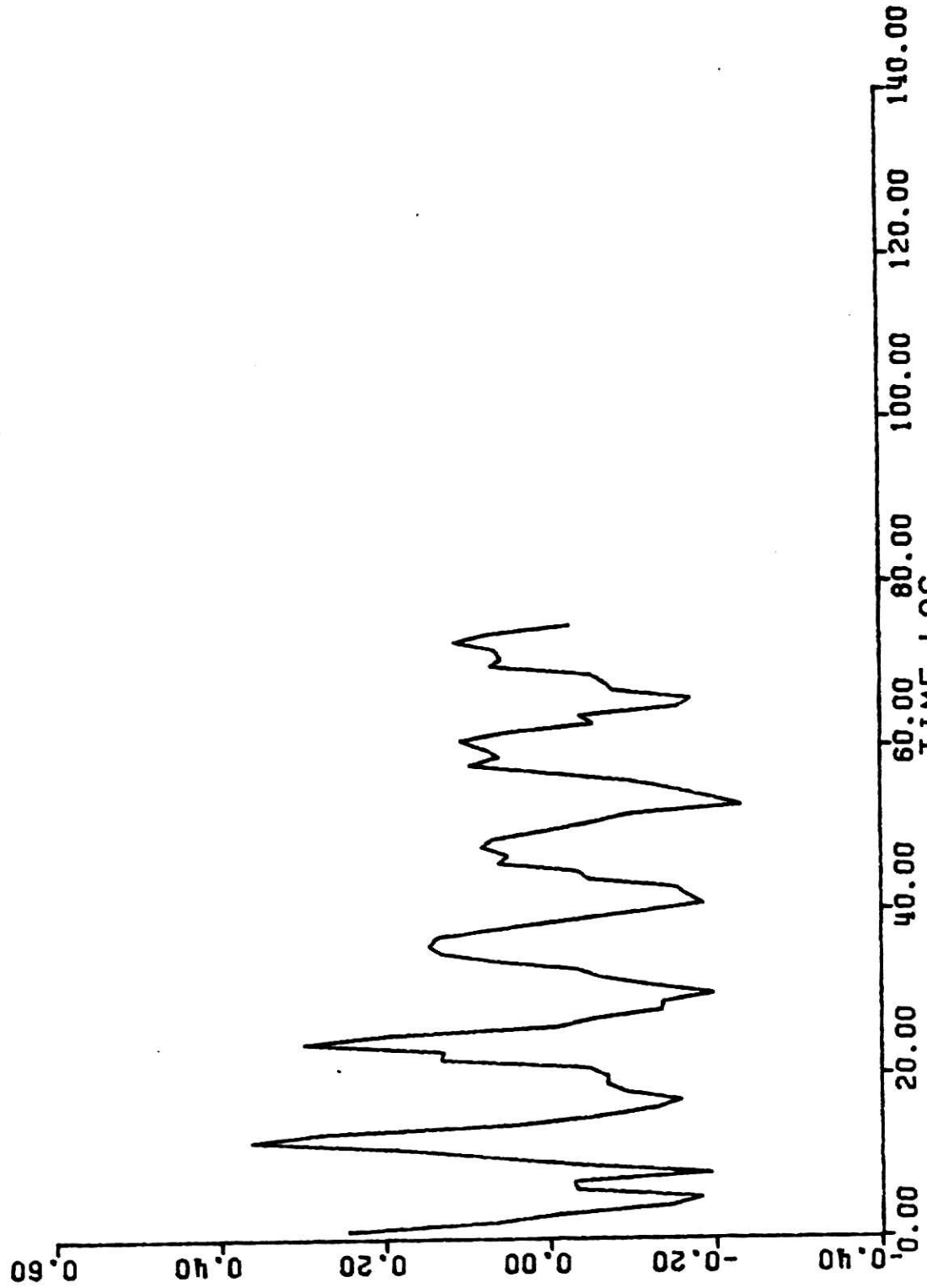

Fig. 5.31 Autocorrelation of temp. (Vx) - station 1.

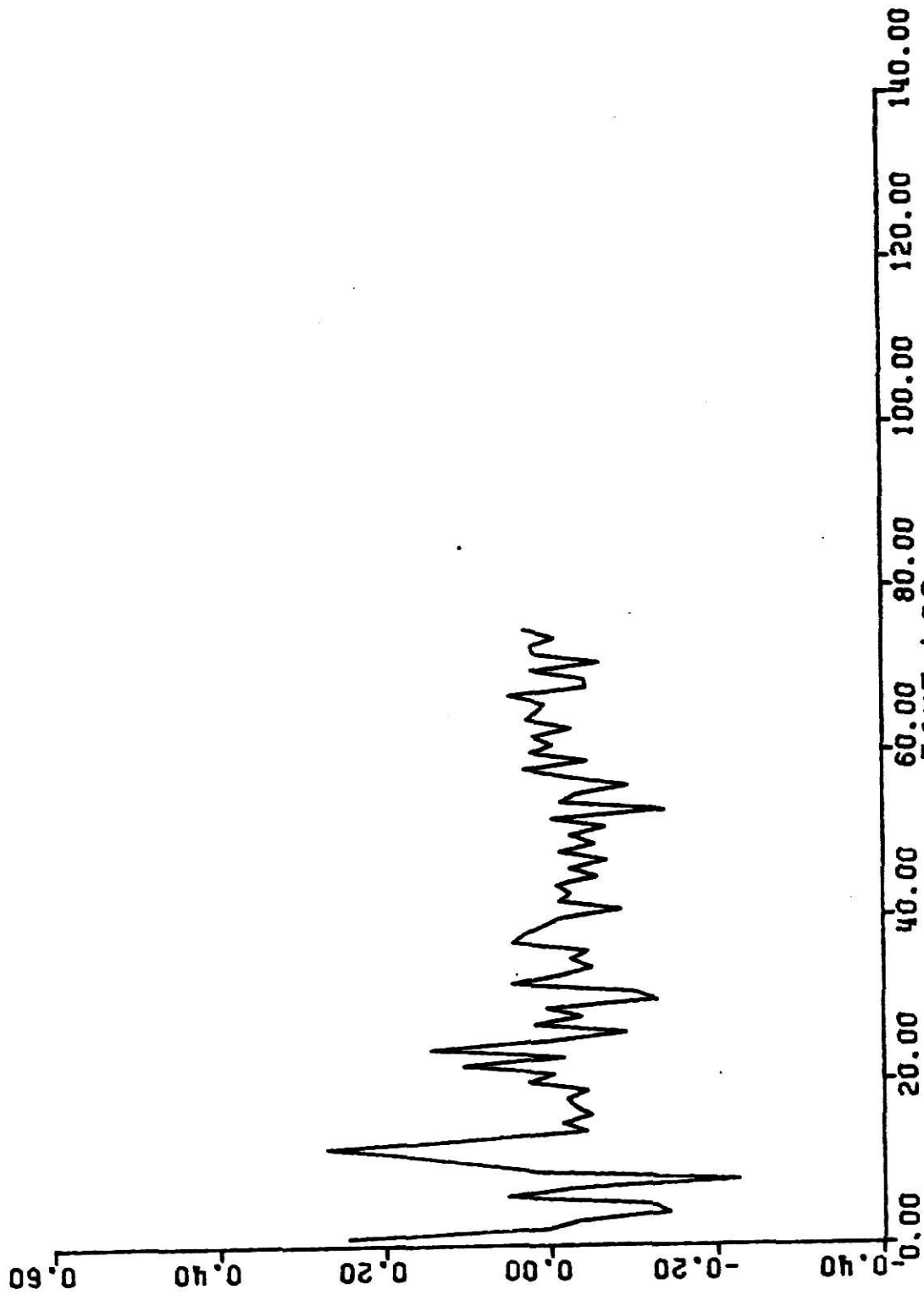

Fig. 5.32 Partial autocorrelation of temp. (Vx) - station 1.

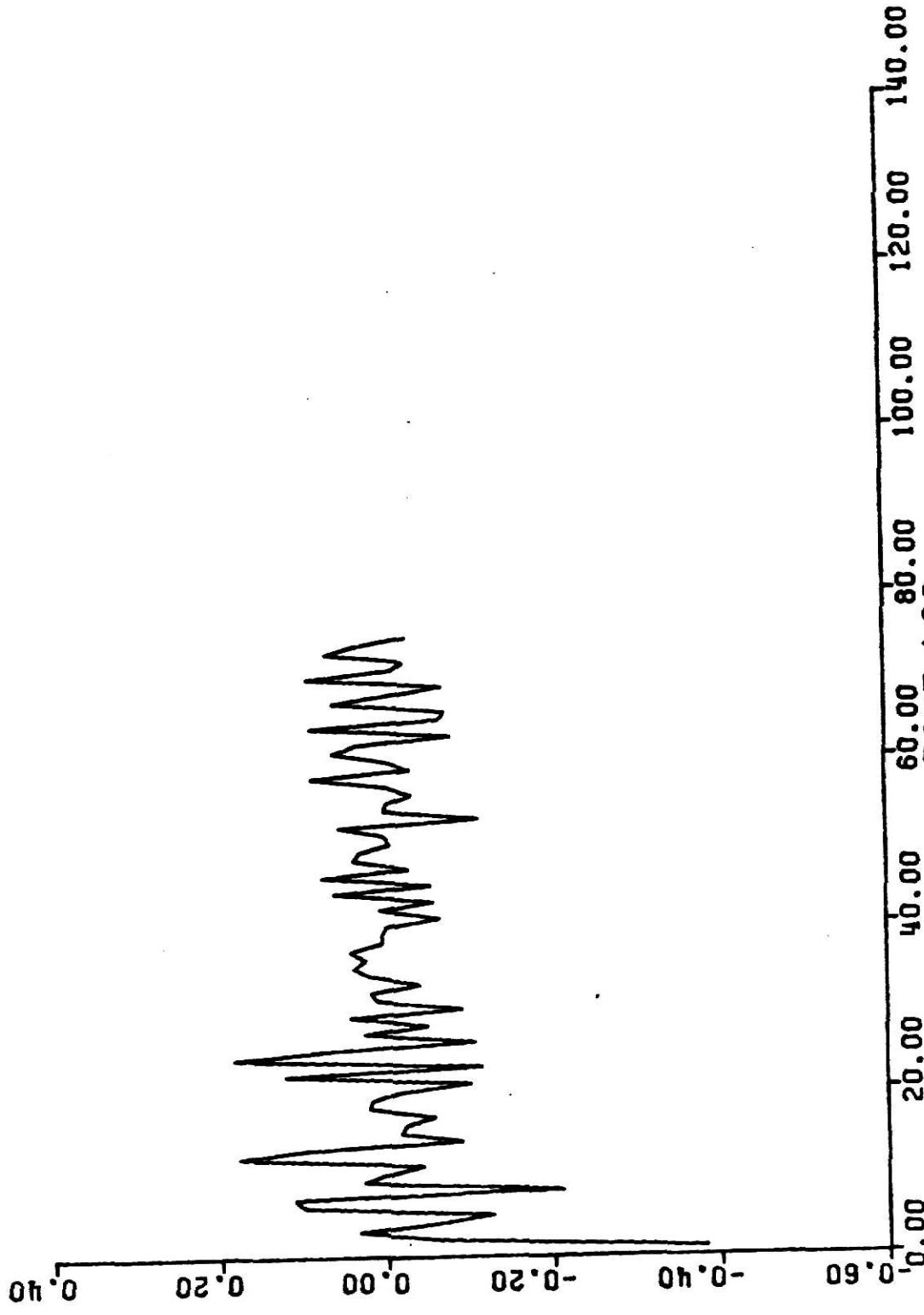

Fig. 5.33 Autocorrelation of temp. (v^2x) - station 1.

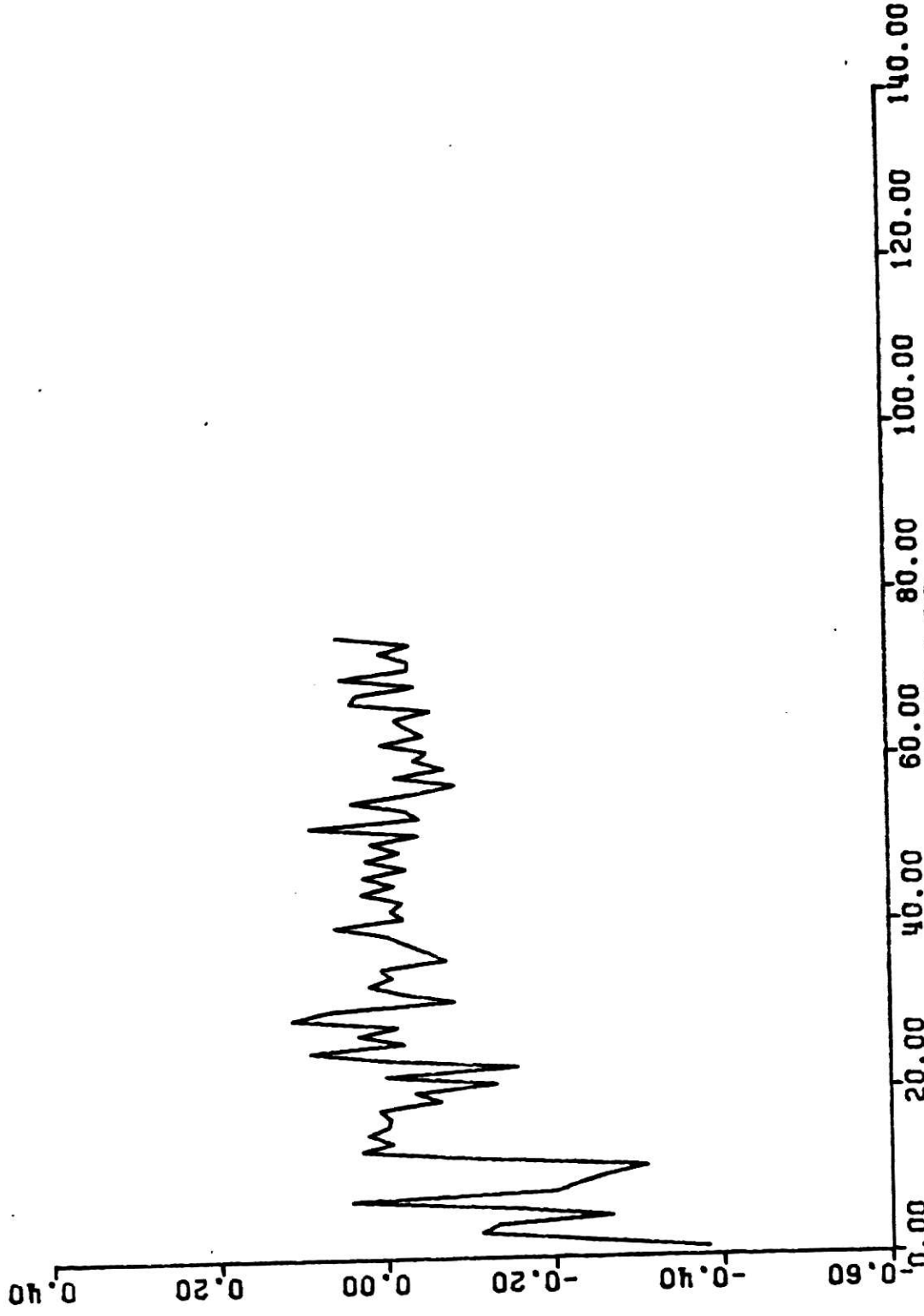

Fig. 5.34 Partial autocorrelation of temp. ($\nabla^2 x$) - station 1.

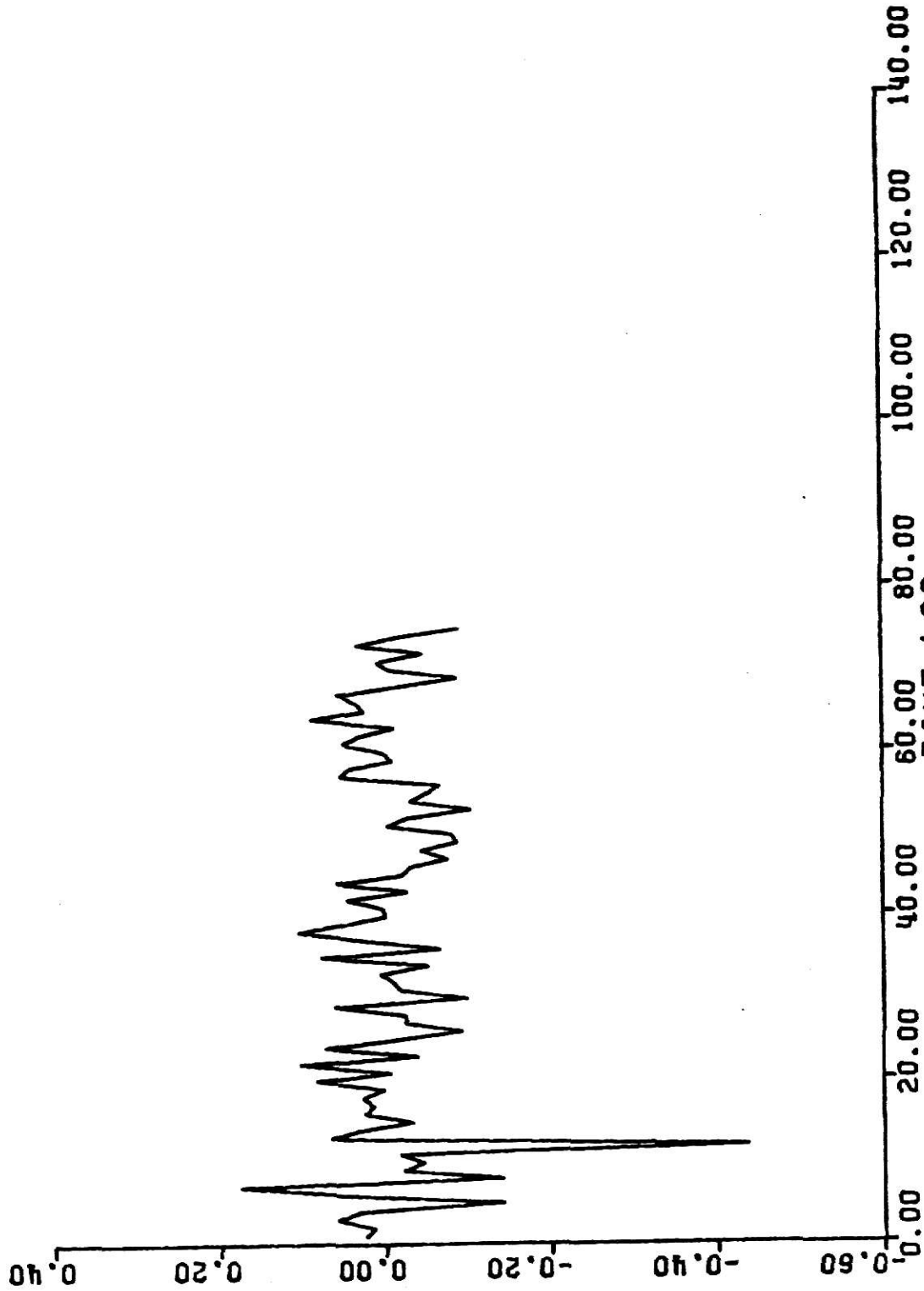

Fig. 5.35 Autocorrelation of temp. ($\nabla\nabla_{12}x$) - station 1.

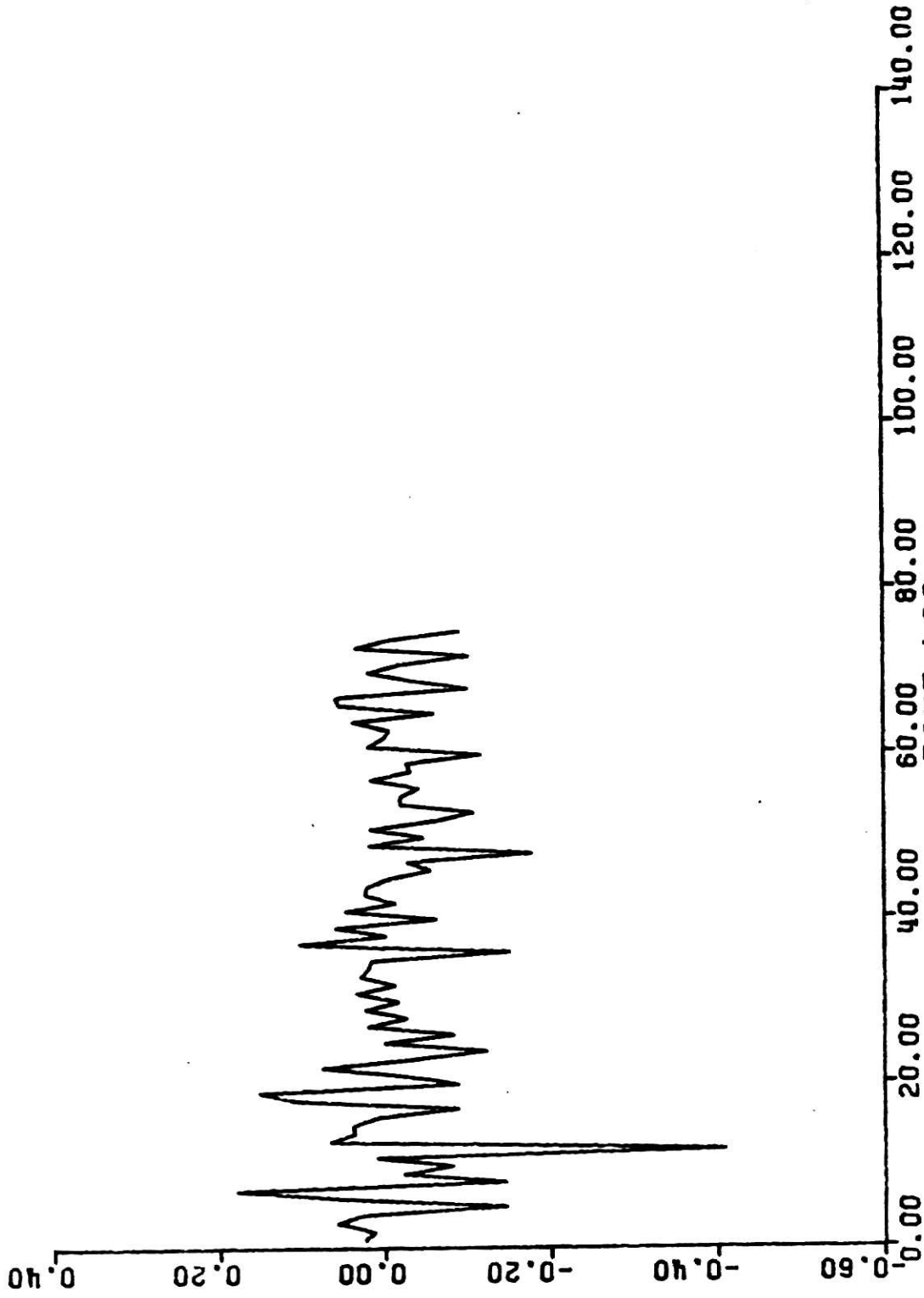

Fig. 5.36 Partial autocorrelation of temp. ($\nabla\nabla_{12}{}^{x}$) - station 1.

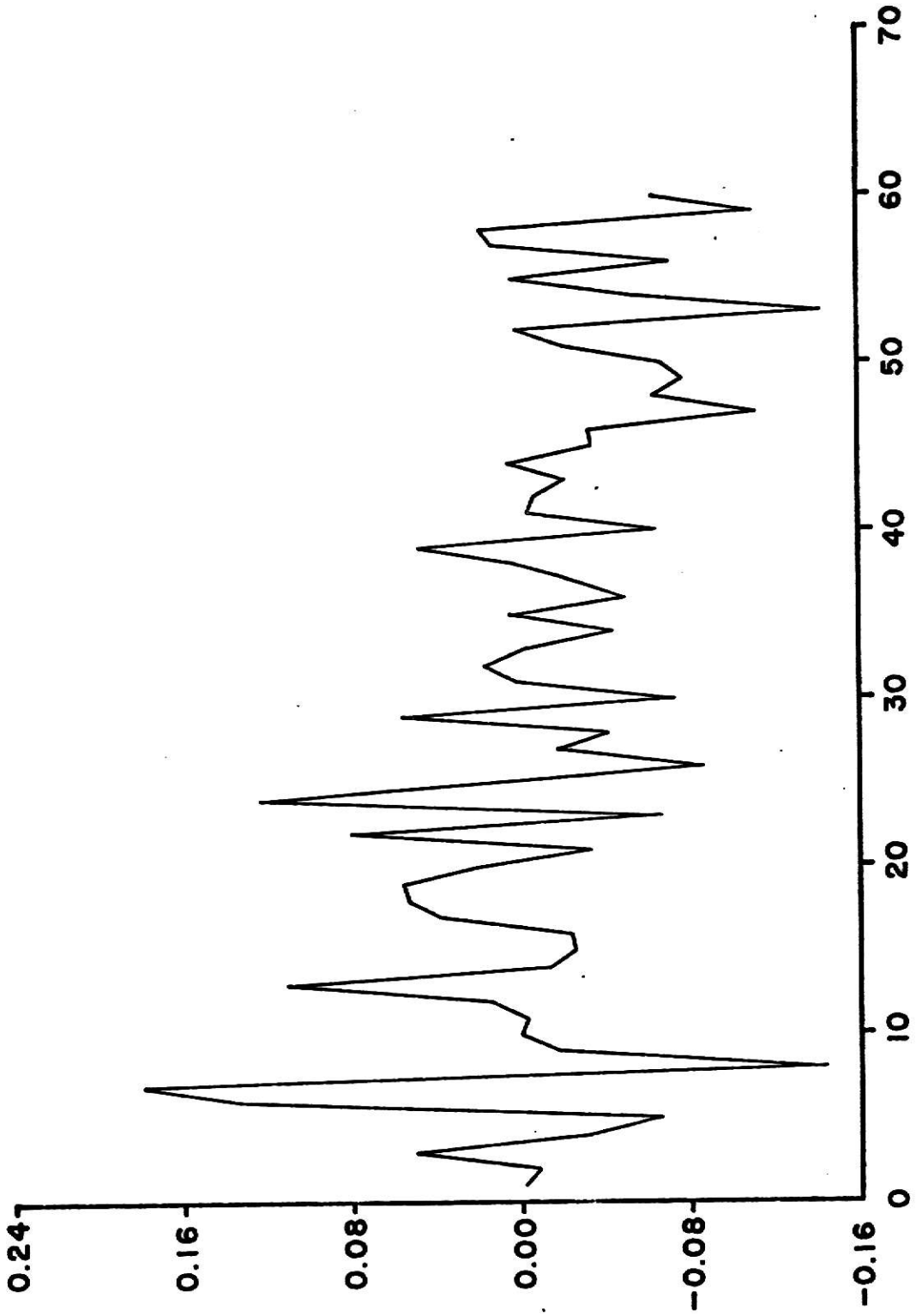

Fig. 5.37 Autocorrelation of temp. residuals. Model ARMA (0,1,1)x(0,1,2)x(12).

and $\theta_{24} = -0.077$

Figure 5.37 shows the autocorrelation plot for the residuals after fitting the above model. All except 3 points are within the $\pm 2\sigma$ limits.

An overall check on autocorrelation function is provided by the quantity $Q = n \sum_{k=1}^{90} r_k^2(\hat{a}) = 111.36$, which is distributed as χ^2 with 87 degrees of freedom. The tabulated $\chi^2_{0.95}$ value (113.14) is greater than Q = 111.36, indicating thereby, that there is no reason to doubt the adequacy of this model.

A spectral analysis was done for the residuals which did not indicate the presence of any dominant cyclic fluctuation.

Thus the model is,

$$w_t = (1 + 0.099B)(1 - 0.667B^{12} + 0.077B^{24})\, a_t$$

or $$w_t = a_t + 0.099a_{t-1} - 0.667a_{t-12} - 0.0660a_{t-13}$$

$$+ 0.077a_{t-24} + 0.0076a_{t-25}$$

A comparison of the original dates variance and the residuals variance is given below which shows that a considerable reduction in variance has been achieved.

Variance of original data = 3.53 $(^oF)^2$

Variance of residuals = 0.092 $(^oF)^2$

$$R^2 = \frac{3.53-.092}{3.53} = .971 = .971$$

(b) Dissolved oxygen: The autocorrelation and partial autocorrelation plots for the raw data. ∇x, $\nabla^2 x$ and $\nabla\nabla_{12}$ are shown in Figures 5.38 thru 5.45. As before, these plots suggest a tentative model

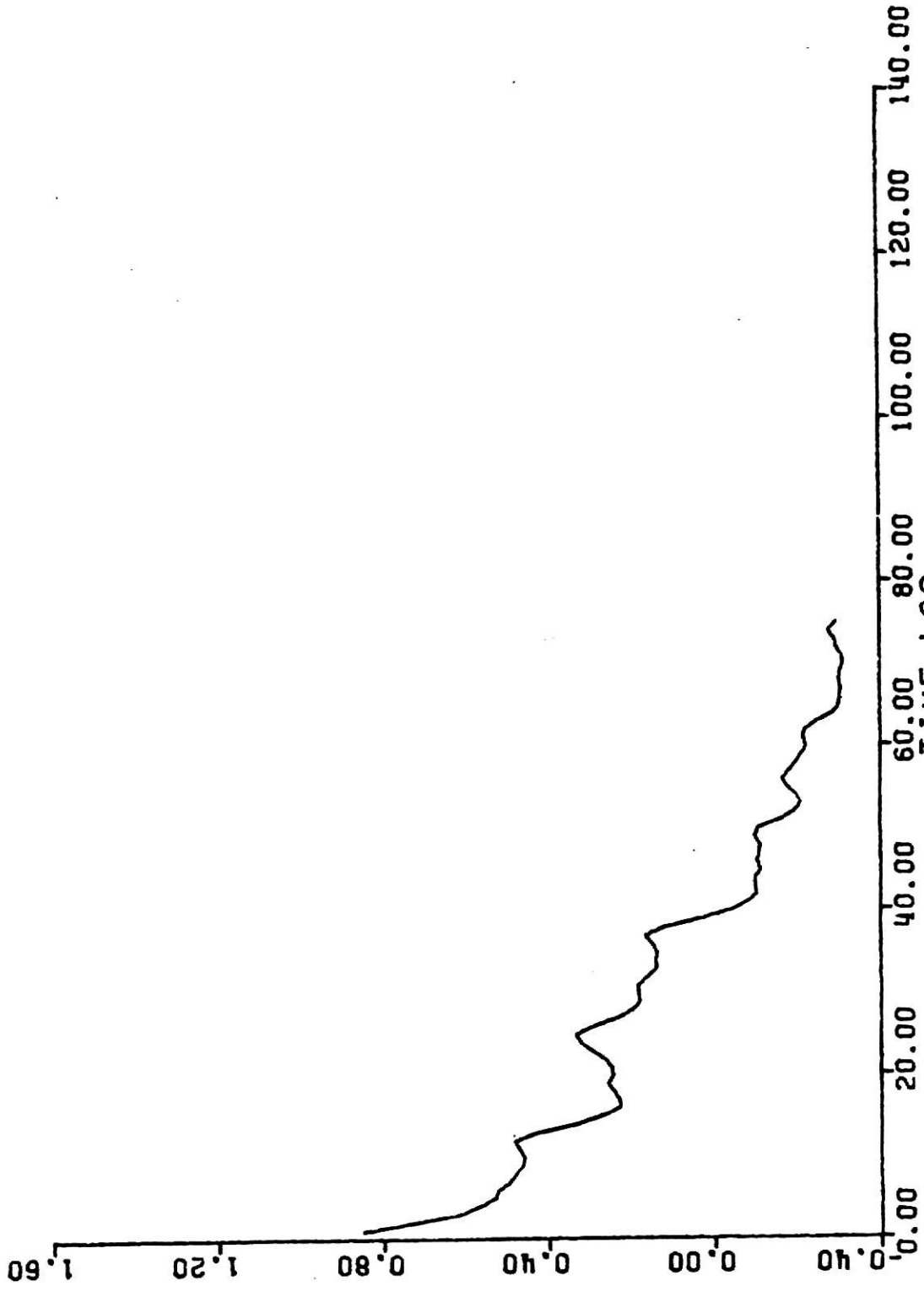

Fig. 5.38 Autocorrelation of original DO (2 hourly data) - station 1.

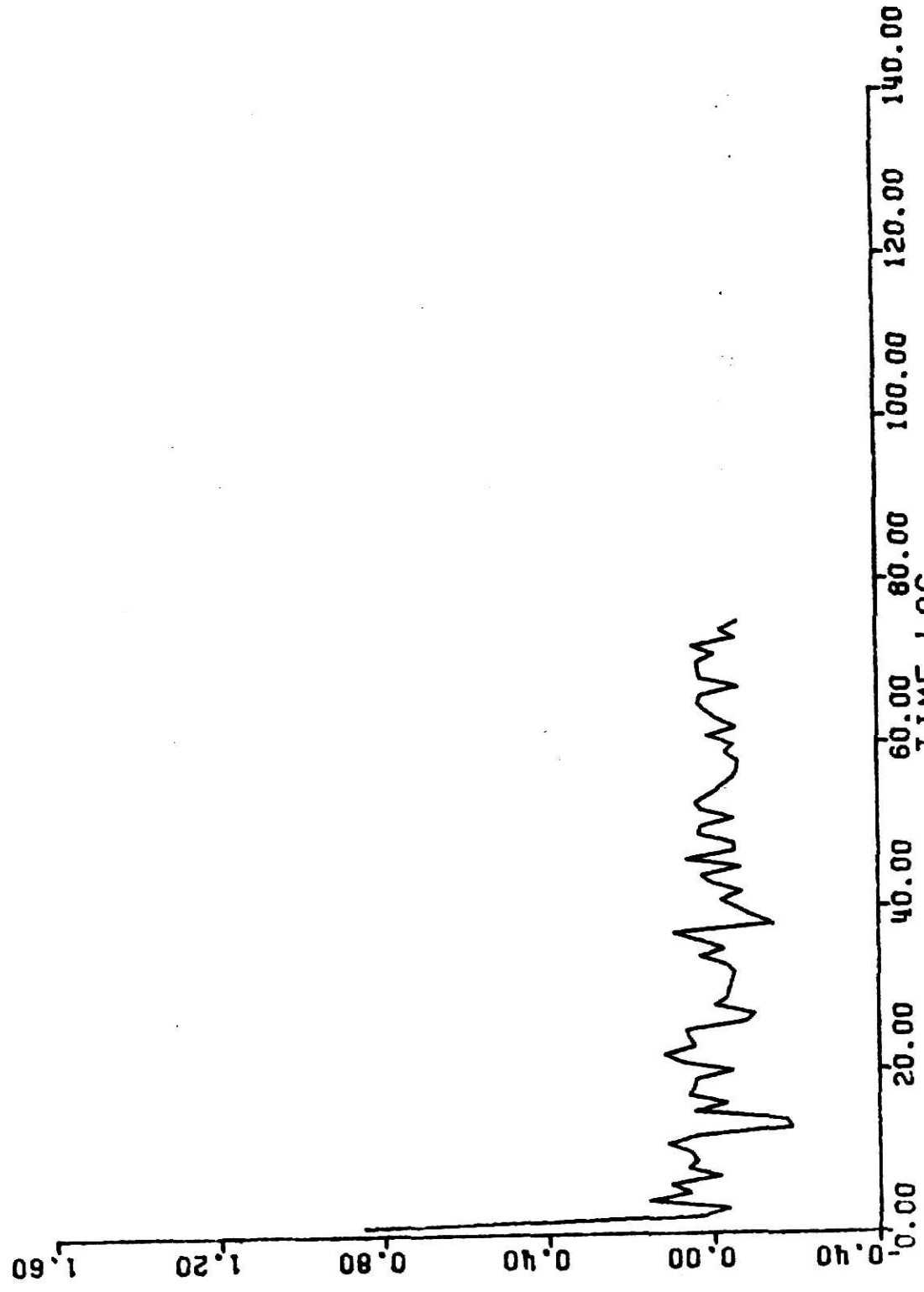

Fig. 5.39 Partial autocorrelation of DO (2 hourly data) - station 1.

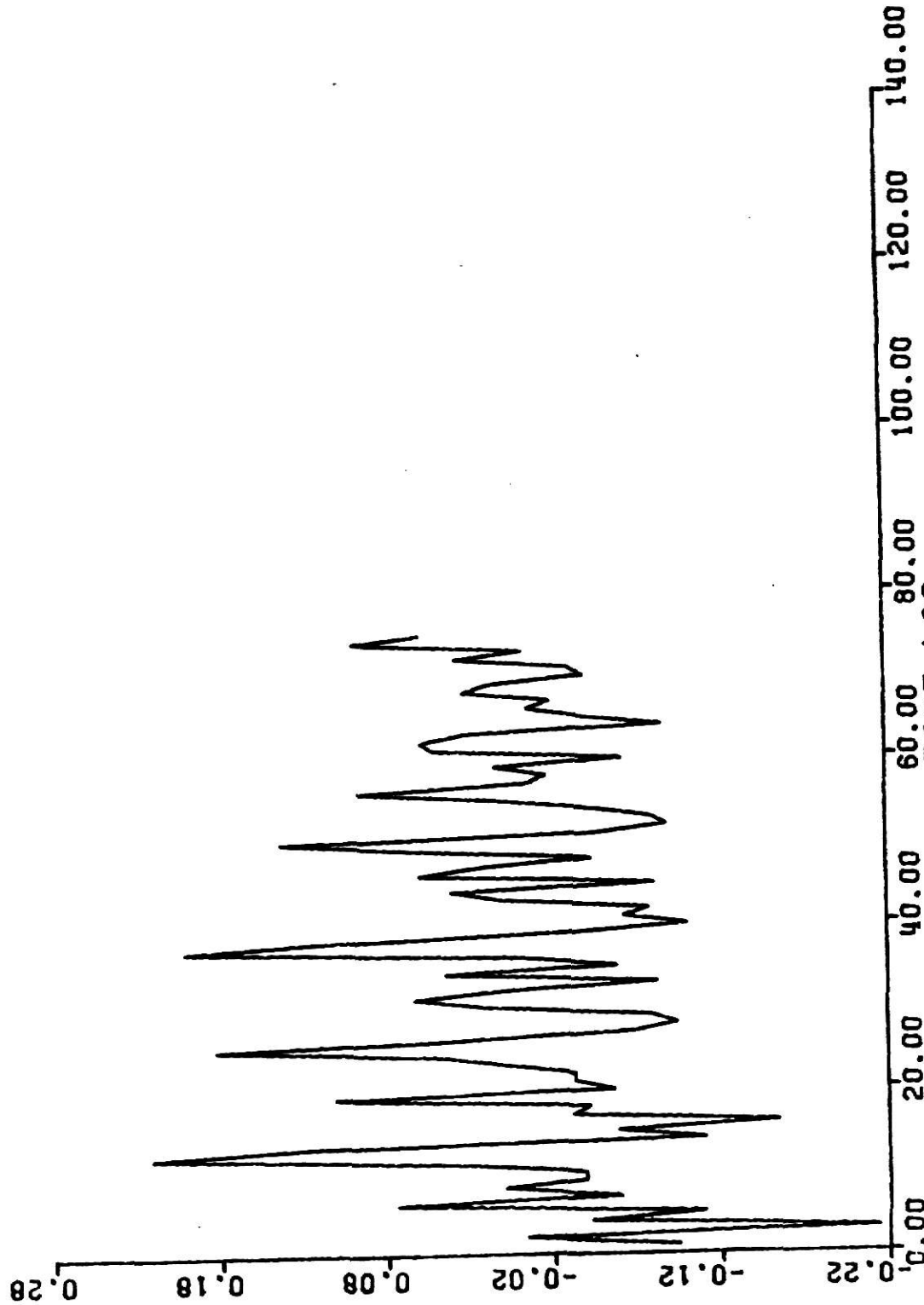

Fig. 5.40 Autocorrelation of DO (Vx) - station 1.

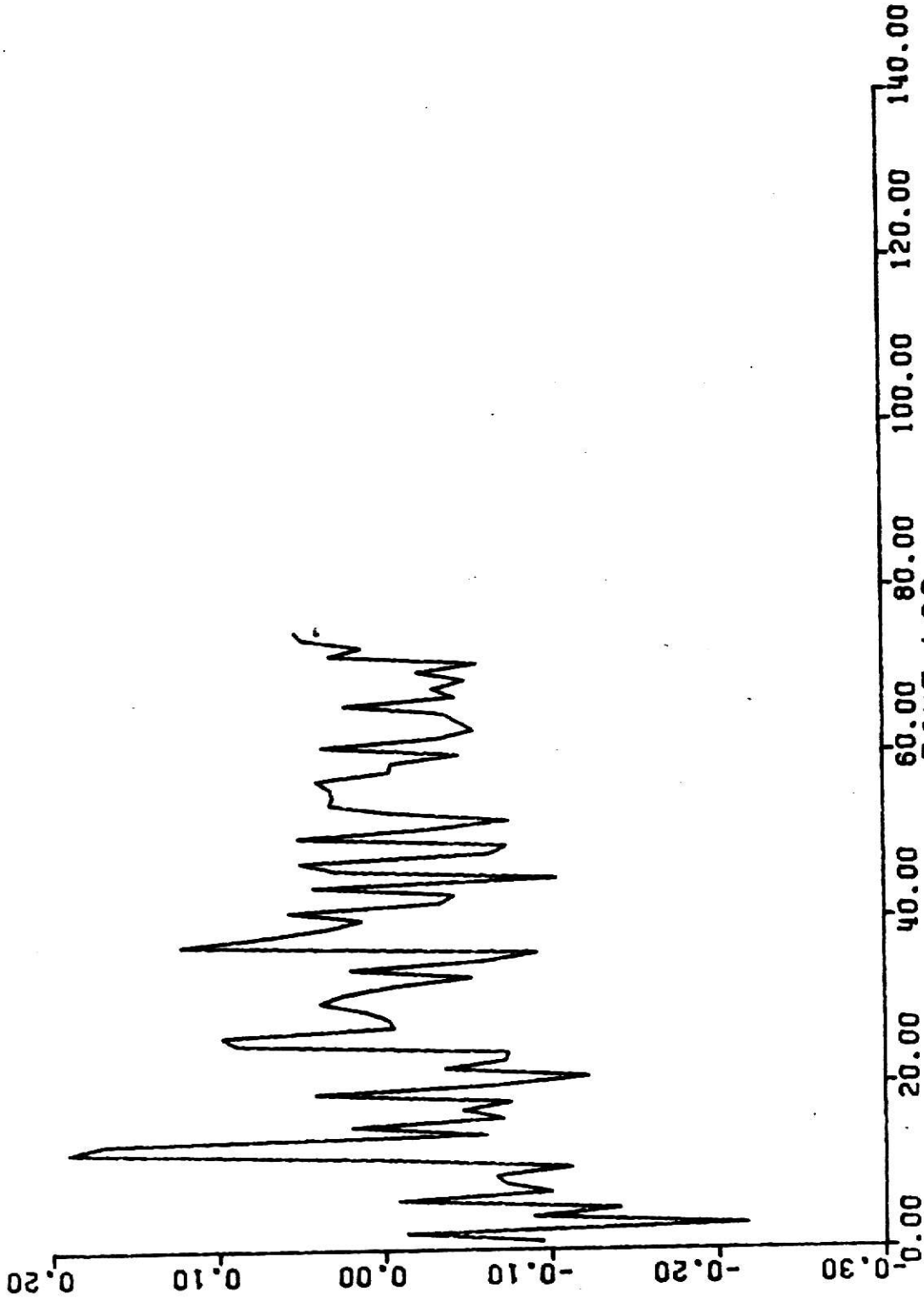

Fig. 5.41 Partial autocorrelation of DO (∇x) - station 1.

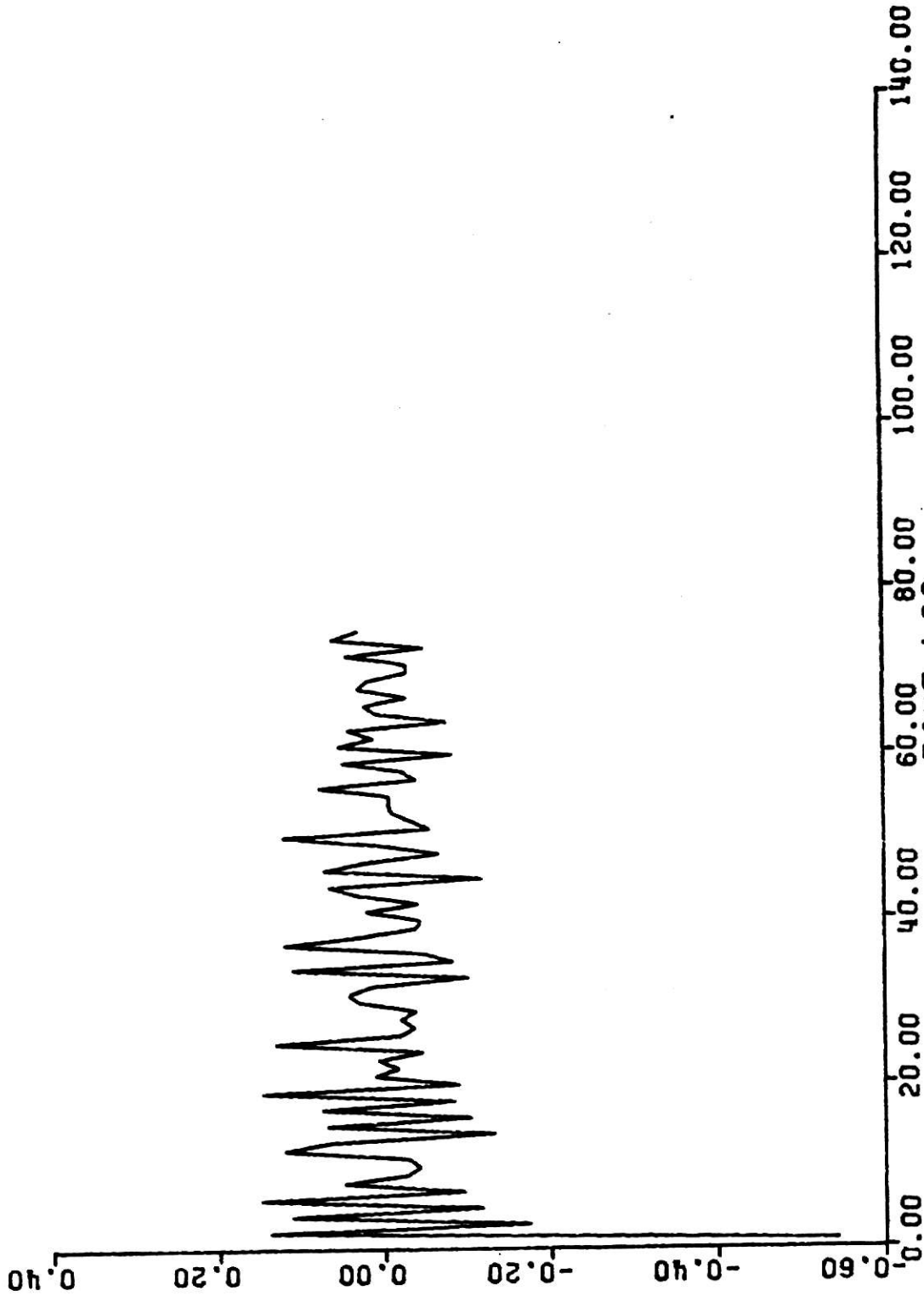

Fig. 5.42 Autocorrelation of DO ($\nabla^2 x$) - station 1.

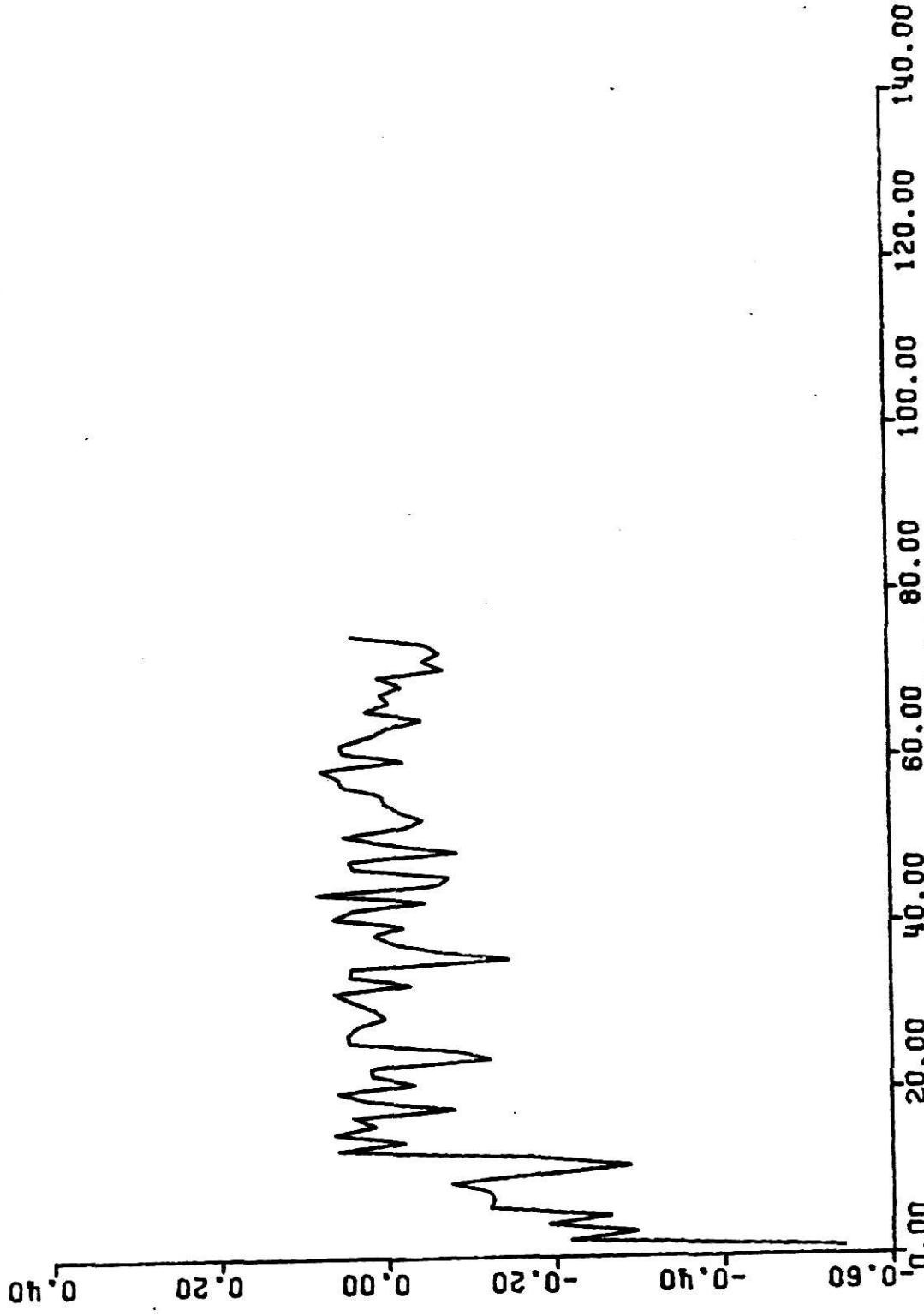

Fig. 5.43 Partial autocorrelation of DO ($\nabla^2 x$) - station 1.

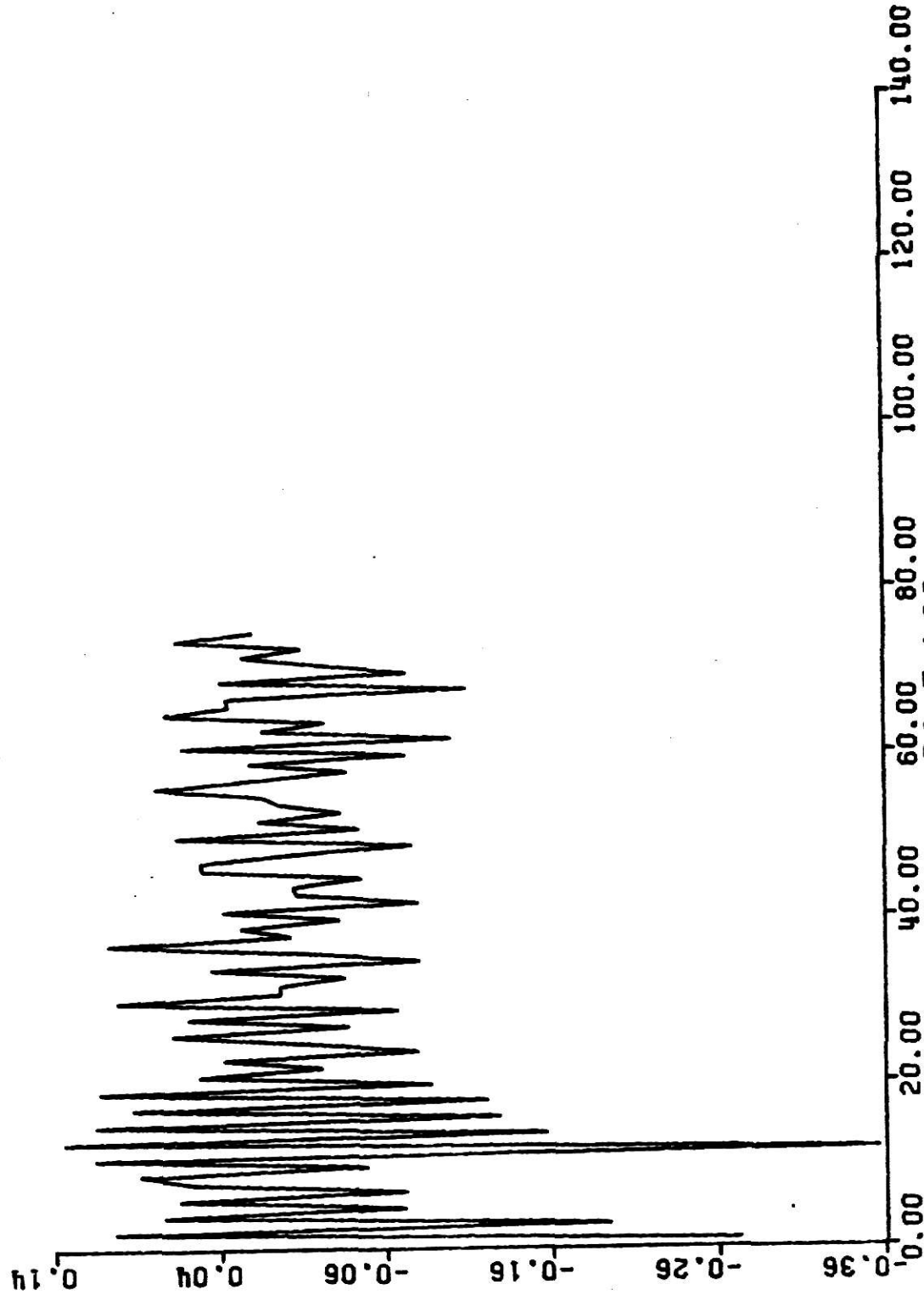

Fig. 5.44 Autocorrelation of DO ($VV_{12}x$) - station 1.

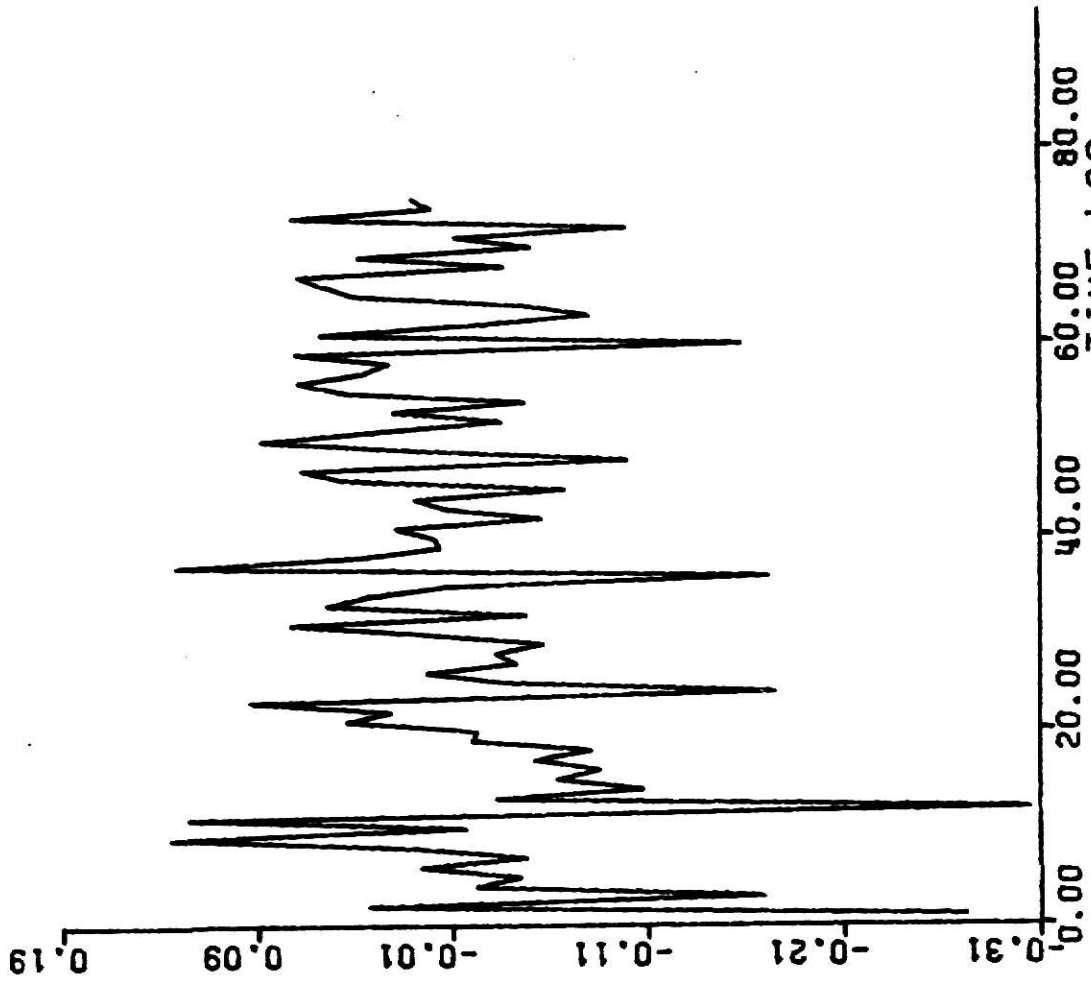

Fig. 5.45 Partial autocorrelation of WO ($\nabla\nabla_{12}x$) - station 1.

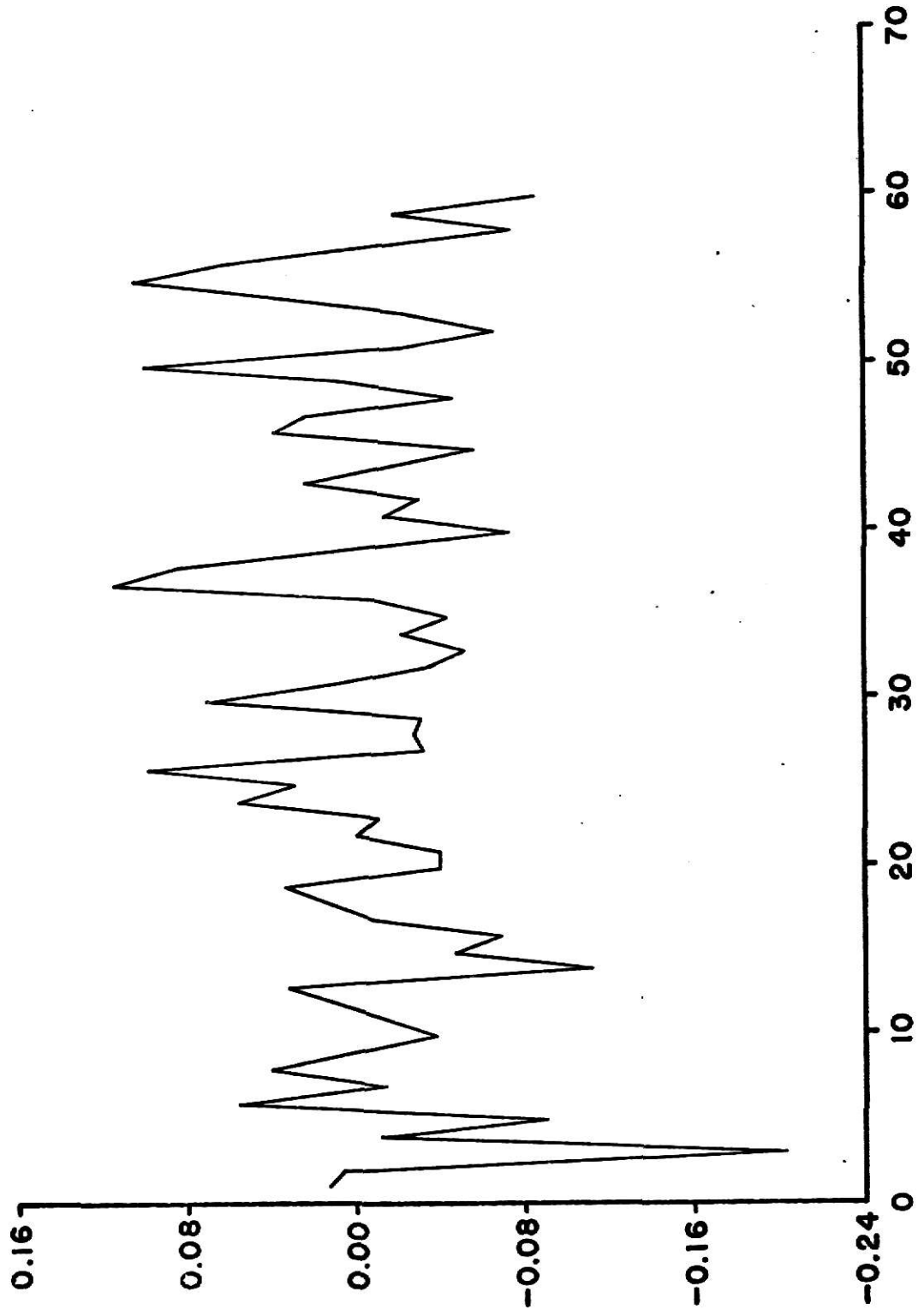

Fig. 5.46 Autocorrelation of DO residuals. Model ARMA (0,1,1)x(0,1,2)x(12).

$$w_t = (1 - \theta_1 B)(1 - \theta_{12} B^{12} - \theta_{24} B^{24}) a_t$$

where, $w_t = \nabla\nabla_{12} x_t$

Initial estimates of the parameters are,

$$\theta_1 = 0.30$$
$$\theta_{12} = 0.50$$
$$\theta_{24} = 0.16$$

Using least squares estimation, final values of the parameters were obtained as

$$\theta_1 = 0.261$$
$$\theta_{12} = 0.536$$
$$\theta_{24} = 0.125$$

Figure 5.46 shows the autocorrelation plot of the residuals. It is seen that all except one value are within $\pm 2\sigma$ limits.

An overall check on autocorrelation gives $Q = n \sum_{i=1}^{60} r_i^2(\hat{a}) = 66.12$, which is distributed as Chi square with 57 df. This value is less than the tabulated value of $\chi^2_{0.90}(74.40)$ which shows that there is no reason to doubt the adequacy of the model.

Thus the model is,

$$w_t = (1 - 0.261B)(1 - 0.536B^{12} - 0.025B^{24}) a_t$$

A comparison of the variance of the original data and the residuals is given below:

Table 5.12 Summary of Autoregressive Moving average Models for Station 2,3,4.

	Series	Model	Parameters	Variance of original series	Variance of residuals
(A)	DO Station 2	1) ARMA(0,1,1)	$\theta_1 = 0.312$	2.15	0.065
		2) ARMA(0,2,2)	$\theta_1^1 = 1.24$ $\theta_2^1 = -0.27$		0.068
(B)	Temperature Station 2	1) ARMA(0,1,1)	$\theta_1 = 0.372$	3.60	0.015
		2) ARMA(0,2,2)	$\theta_1 = 1.144$ $\theta_2^1 = -0.494$		0.014
	DO Station 3	ARMA(0,1,1)	$\theta_1 = 0.443$	1.18	0.384
		ARMA(0,2,2)	$\theta_1^1 = 1.25$ $\theta_2^1 = -0.26$		0.40
(D)	Temperature Station 3	ARMA(0,2,2)	$\theta_1 = 1.40$ $\theta_2^1 = -0.42$	2.13	0.27
	DO Station 4	ARMA(1,2,2)	$\theta_1 = 0.365$ $\theta_1^1 = 1.51$ $\theta_2^1 = -0.53$	3.05	0.186
(F)	Temperature Starting	ARMA(0,2,2)	$\theta_1 = 1.47$ $\theta_2^1 = -0.48$	1.01	0.33

Variance of original data = 0.55

Variance of residuals = 0.17

As discussed earlier in the spectral analysis of the dissolved oxygen data, cyclic fluctuations corresponding to 24 hrs and 12 hrs period were found to be dominant. Thus, an attempt was made to remove these periodicities by harmonic regression and fit a parametric model to the residuals. The autocorrelation plot of the residuals suggested a tentative ARMA(1,1,1) model. The usual procedure of estimation of parameters resulted in the model

$$(1 + 0.092B)w_t = (1 - 0.943B)a_t$$

The residuals, thus obtained still had a high correlation among themselves and hence this model was not considered for further investigation.

Table 5.12 summarizes the models for temperature and dissolved oxygen for stations 2,3 and 4.

5.2.5 Cross-Spectral Analysis. Cross-spectral analysis was conducted to study the behaviour of temperature and dissolved oxygen at different stations in the stream and also the relationship between temperature and dissolved oxygen at each station. This involved the study of sixteen pairs of time series. Some of the results will now be discussed below:

(a) Temperature station 1 and dissolved oxygen station 1: As high auto-correlations were observed for large lags for both the pollutants, differenced data was used for further calculations. Figure 5.47 shows the cross correlation function for the differenced data. It oscillates about zero lag. The maximum value of crosscorrelation is observed at zero lag, hence no alignment is necessary. Figures 5.48 thru 5.51 show the corresponding

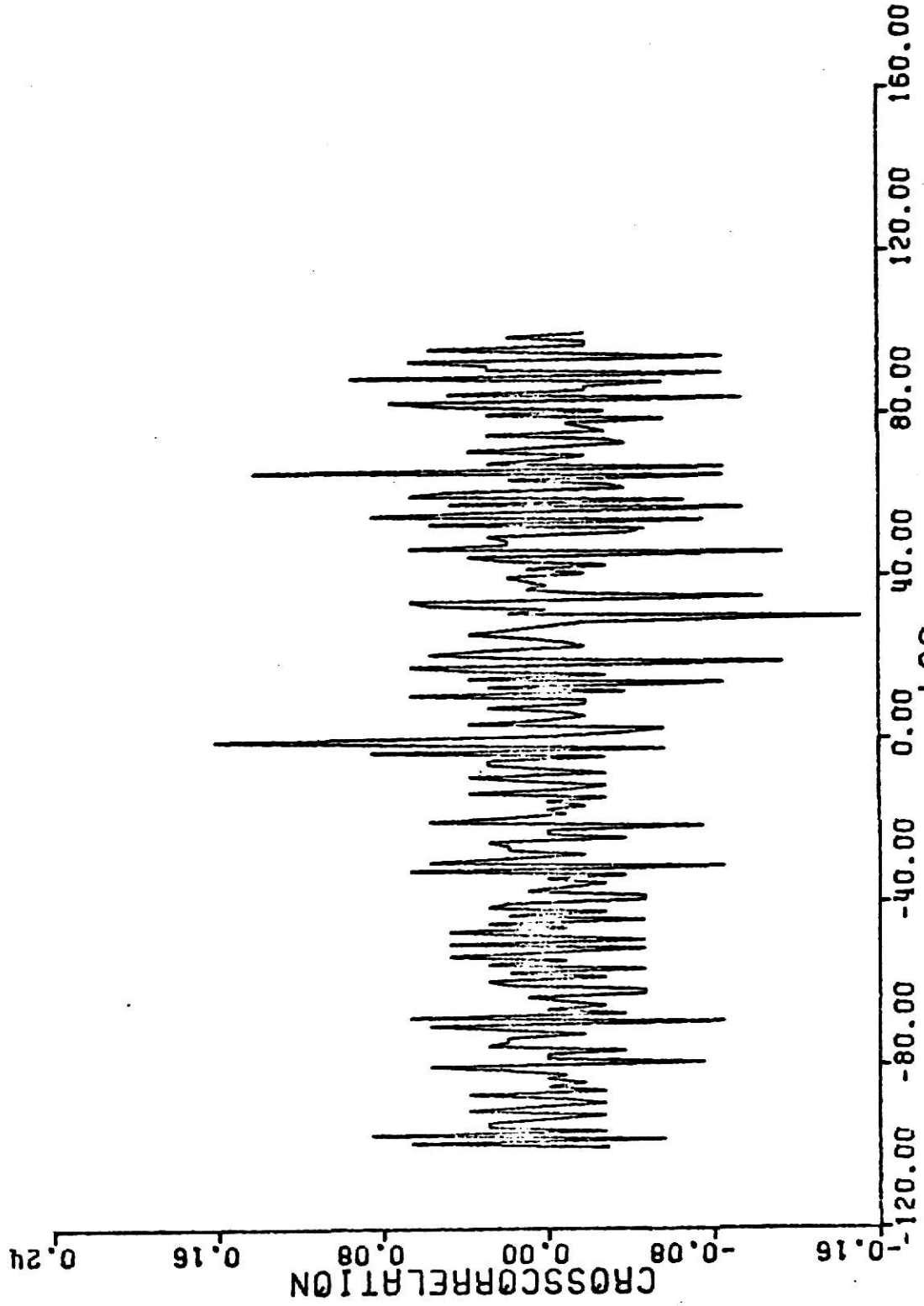

Fig. 5.47 Crosscorrelation temp. - DO , station 1.

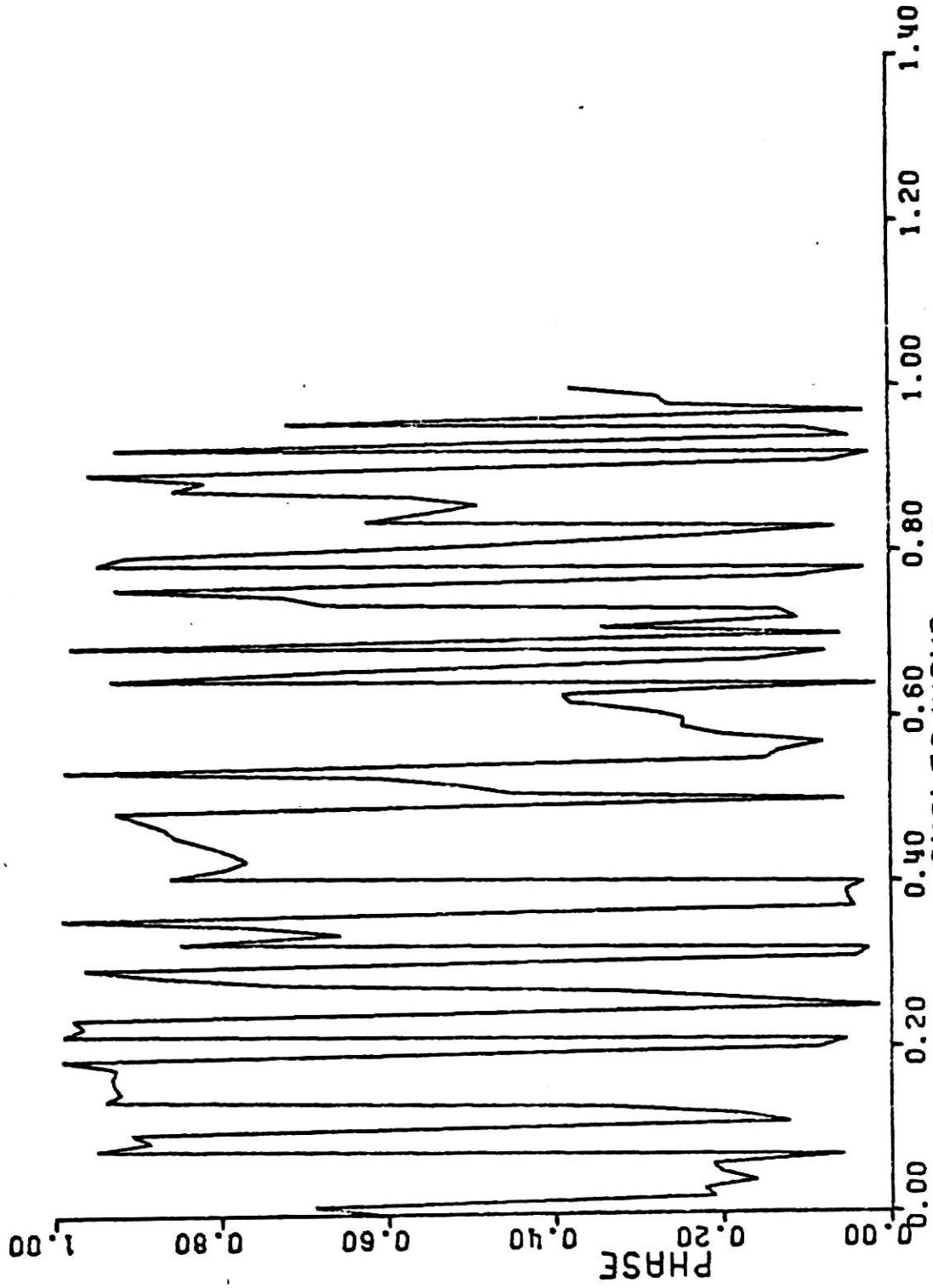

Fig. 5.46 Phase spectra temp. - μ, station 1.

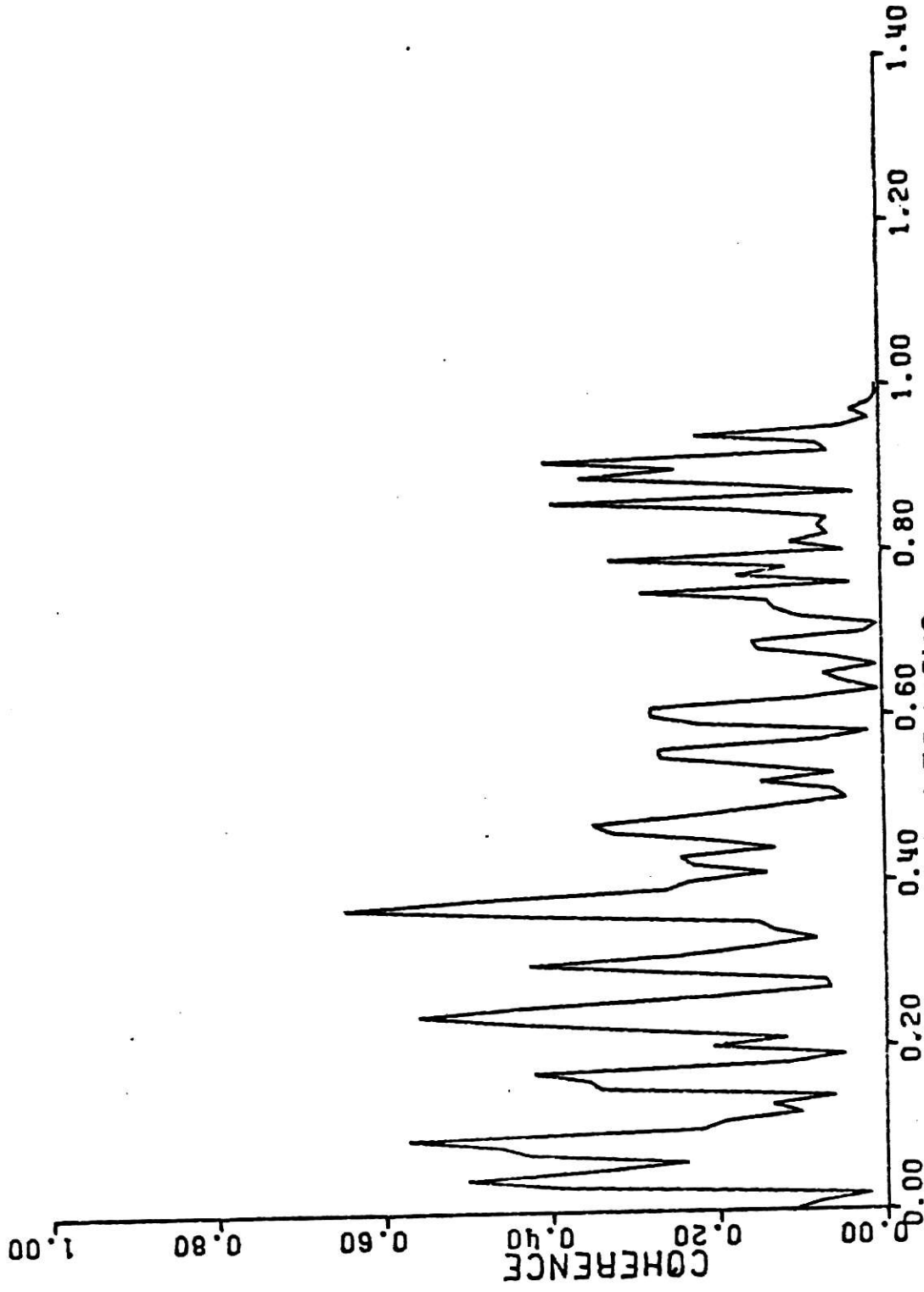

Fig. 5.49 Coherency spectra, temp. - DO , station 1.

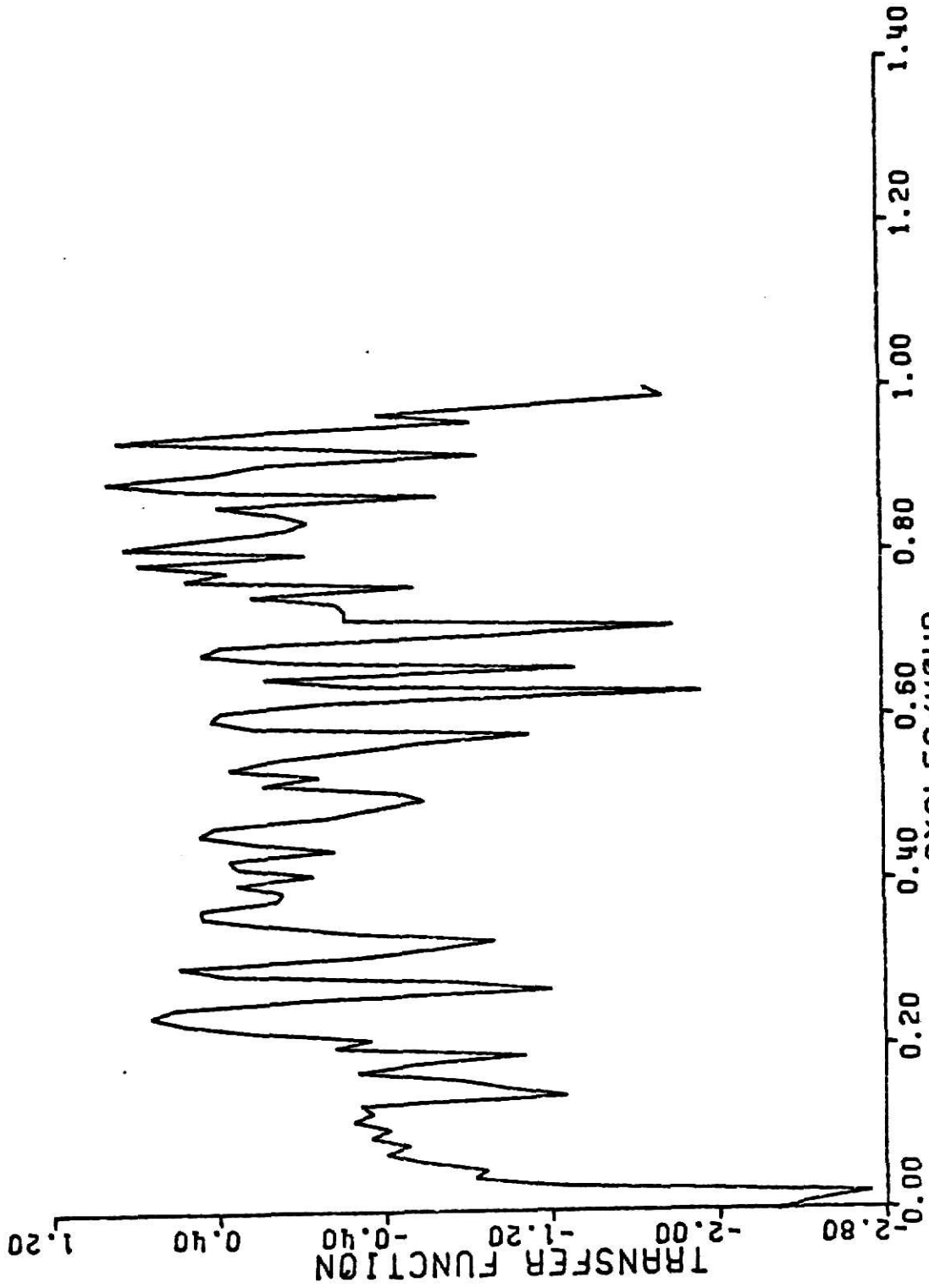

Fig. 5.50 Amplitude of transfer function, temp. - ω , station 1.

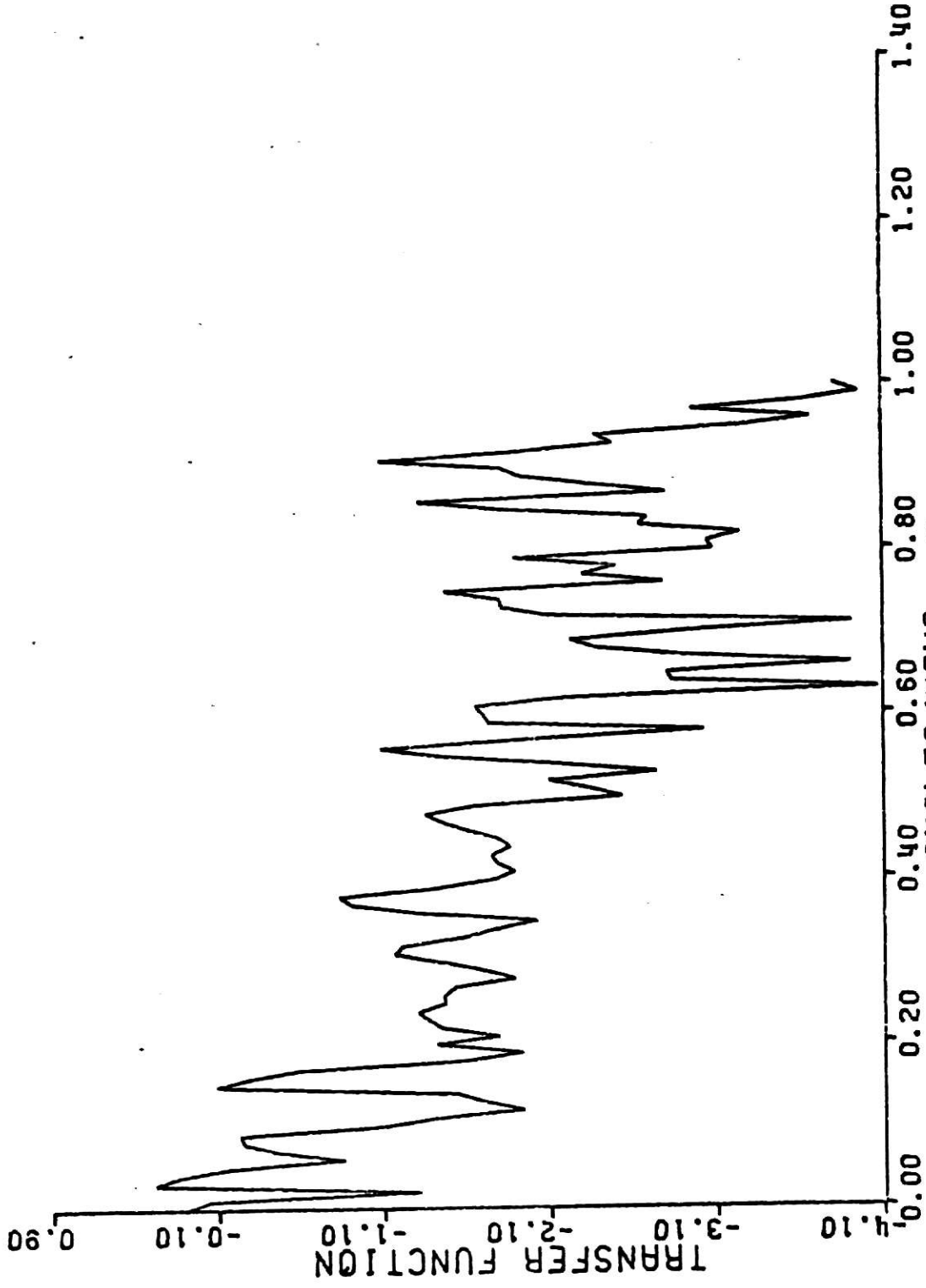

Fig. 5.51 Amplitude of transfer function ΔC - temp., station 1.

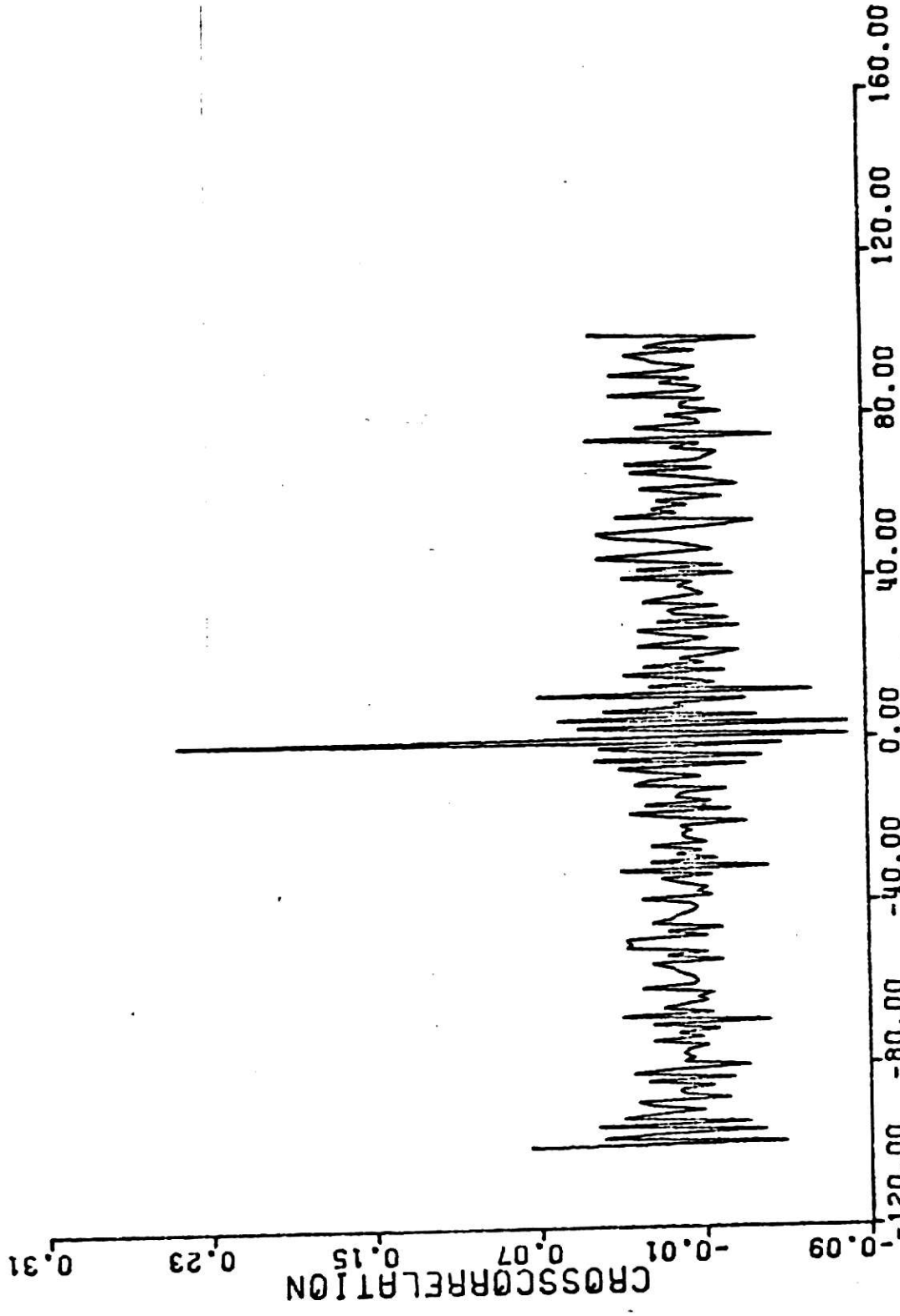

Fig. 5.52 Crosscorrelation of temp. - DO, station 3.

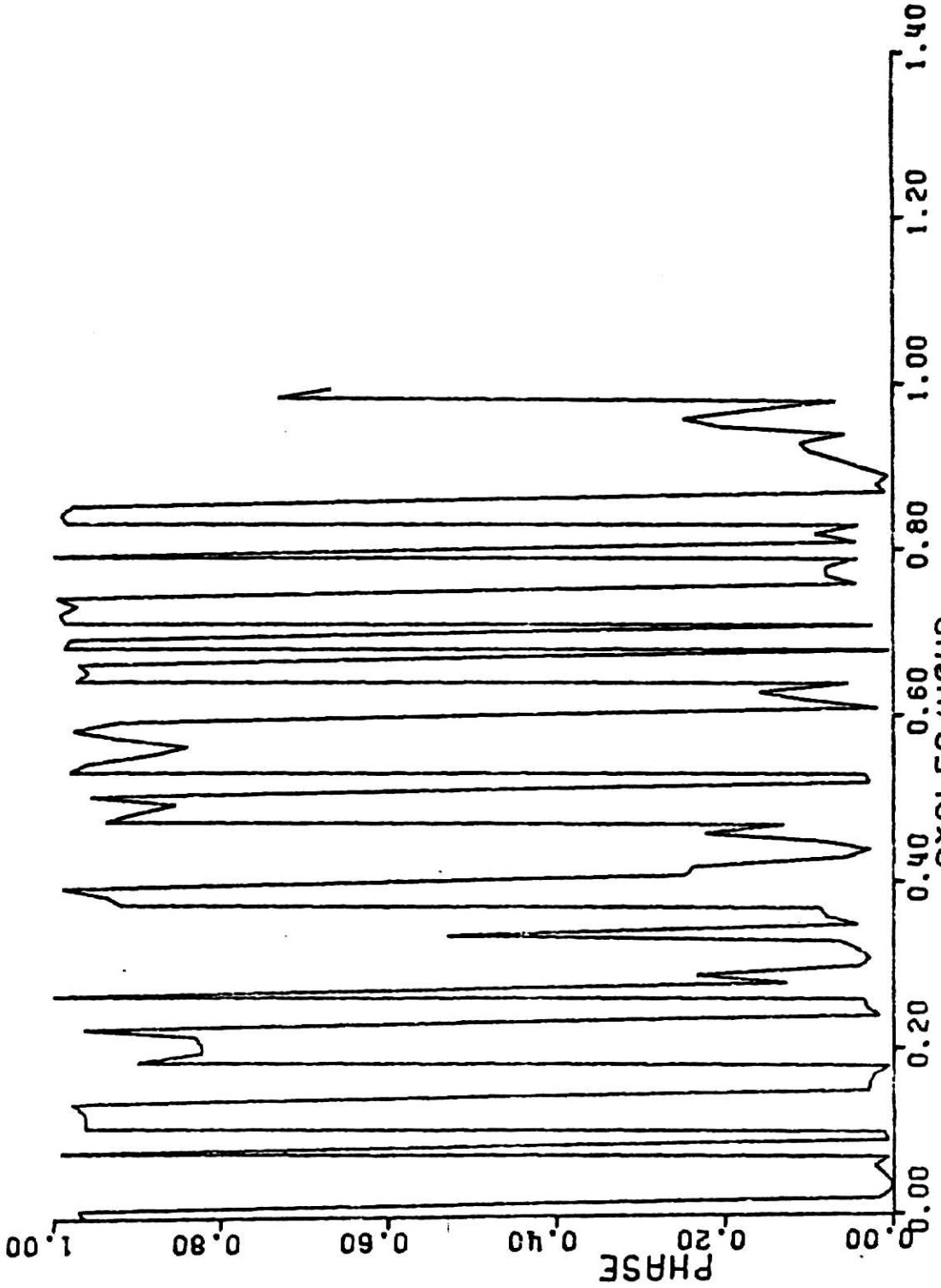

Fig. 5.53 Phase spectra, temp. - DO , station 3.

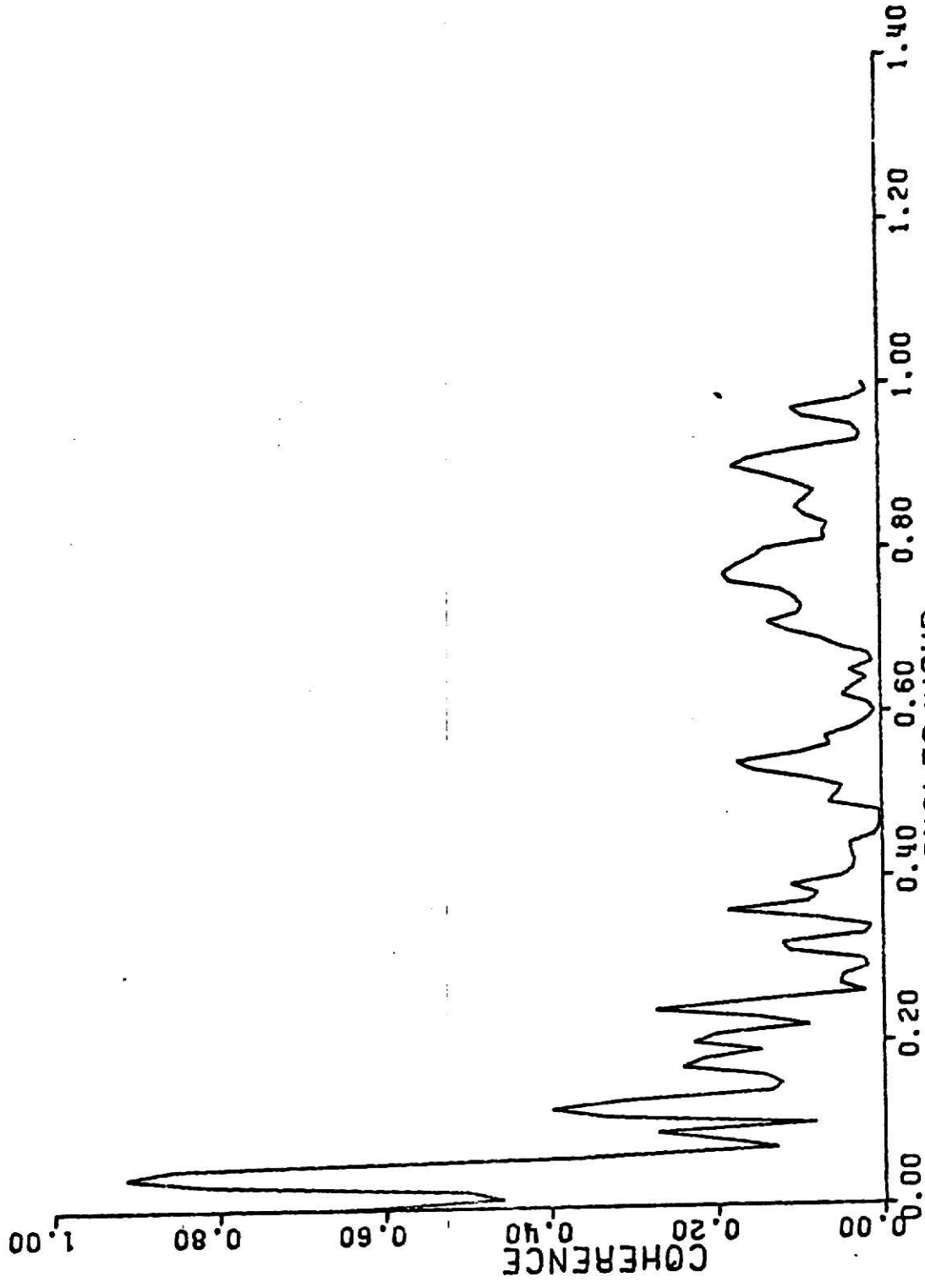

Fig. 5.54 Coherency spectra, temp. -00 , station 3.

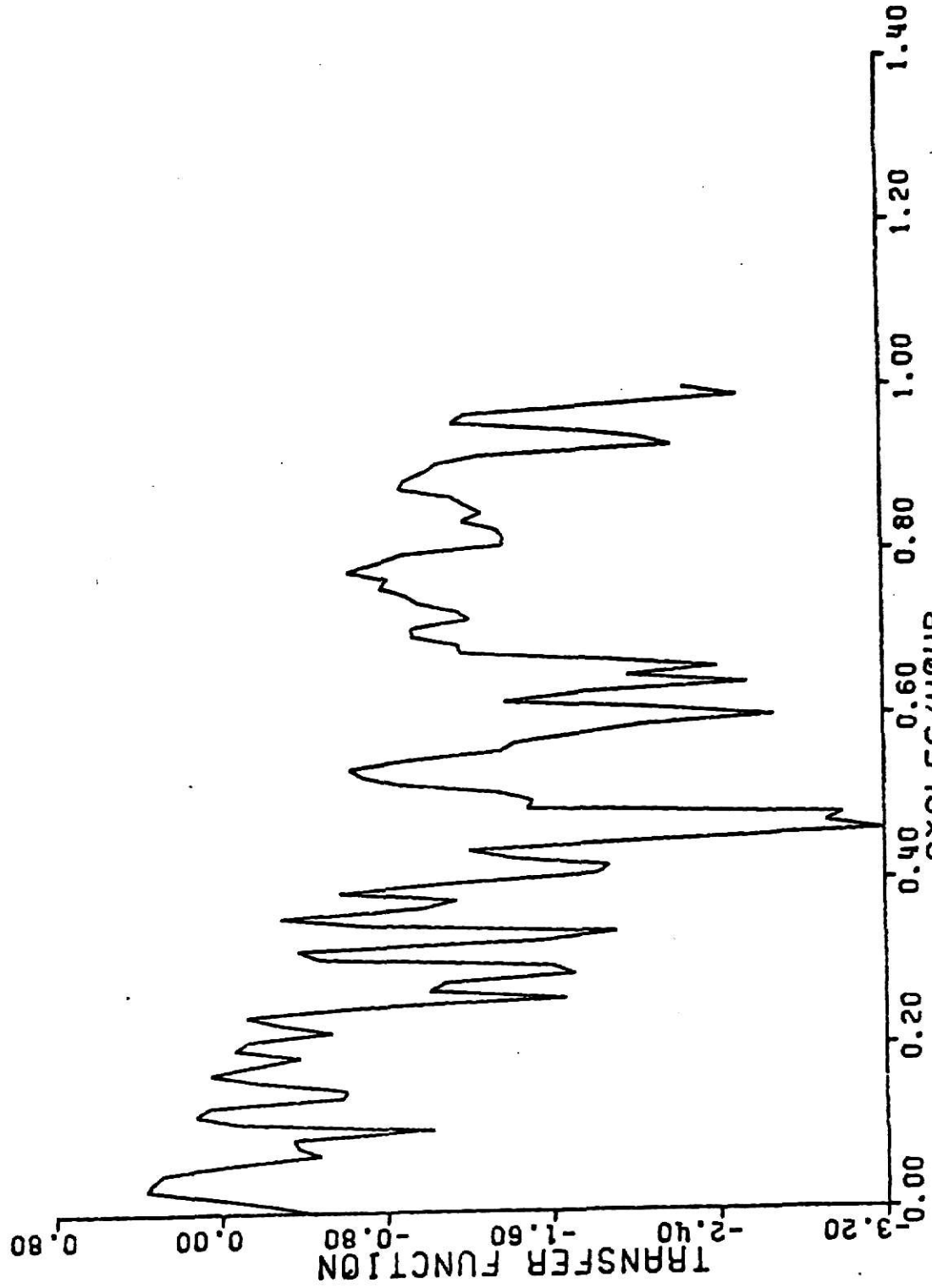

Fig. 5.55 Amplitude of transfer function, temp. - $\mathcal{W}$, station 3.

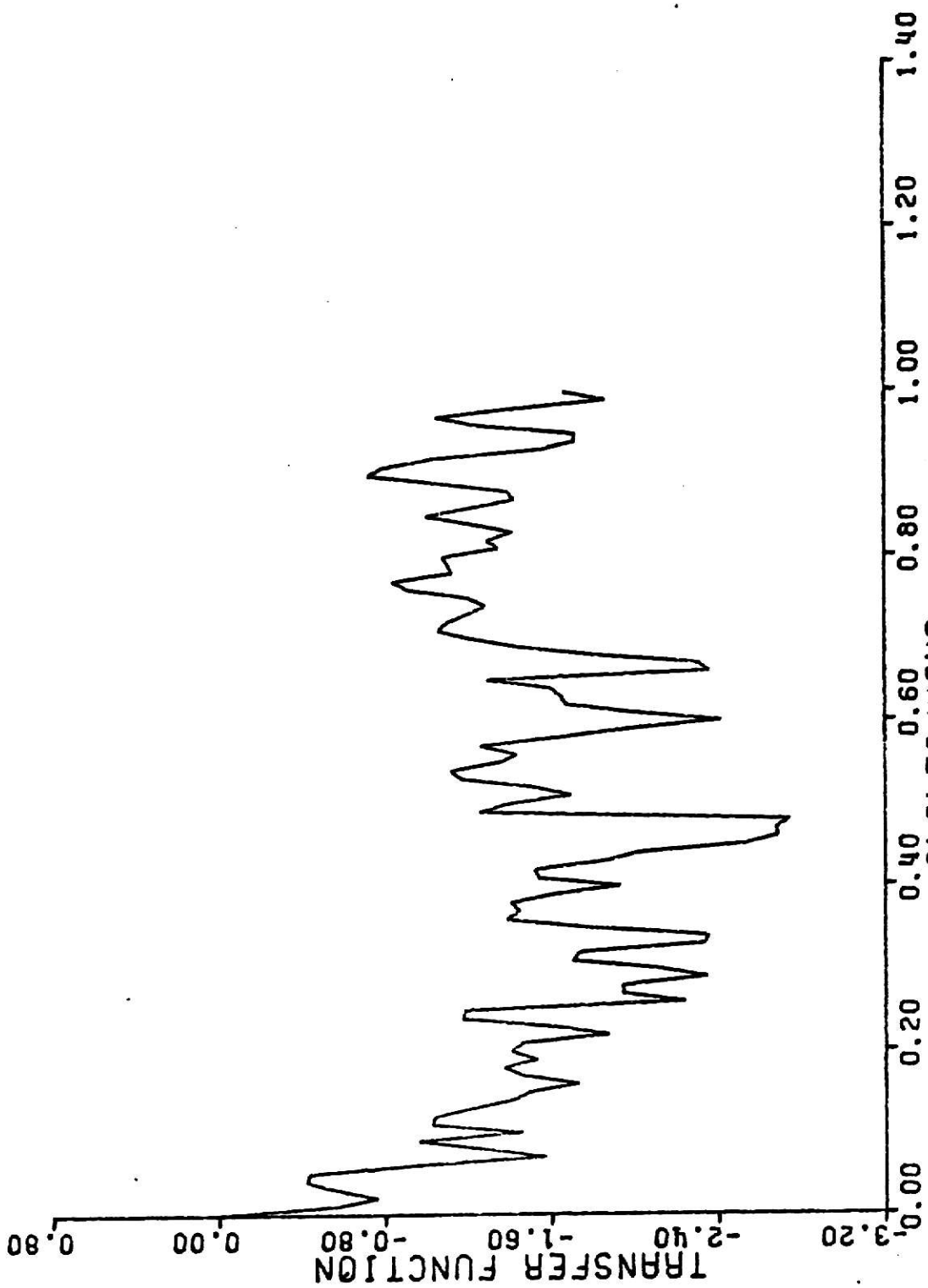

Fig. 5.56 Amplitude of transfer function, DC - temp., station 3.

coherency, phase and transfer function spectra. High values of coherency are seen at 0.04 cycles/hr., 0.09 cycles/hr, 0.24 cycles/hr and 0.37 cycles/hr. The first frequency refers to diurnal variation in temperature and dissolved oxygen. A high transfer function is observed at this frequency, implying that high diurnal variation in temperature will cause high variation in dissolved oxygen. High coherency at higher frequencies does not have any particular significance, since these variations are not dominant in the individual power spectrum. The phase spectrum seems to oscillate about 180^{o} and thus indicates a fixed angle lag relationship between the two series with temperature leading dissolved oxygen.

(b) Temperature and dissolved oxygen at station 3: Figures 5.52 to 5.56 show the crosscorrelation function, coherency, phase and transfer function spectra of the differenced data. High coherency is observed at low frequencies meaning that the long range fluctuations in temperature and dissolved oxygen are highly related. In particular, high values are seen in the band 0.03 cycles/hr to 0.05 cycles/hr corresponding to diurnal variation in each temperature. The transfer function is moderately high in this reg-on suggesting that high variability in DO is caused by high variability in temperature. The phase diagram is again seen to oscillate about 180^{o} suggesting a fixed angle lag. This may be expected as temperature and dissolved oxygen have a inverse relationship.

(c) Temperature station 4 and temperature station 3: Interstation cross-spectral analysis is used to investigate the variation in pollution at different points in a stream. As shown earlier in spectral analysis, the autocorrelation plots of temperature at station 4 and 3 damped very slowly

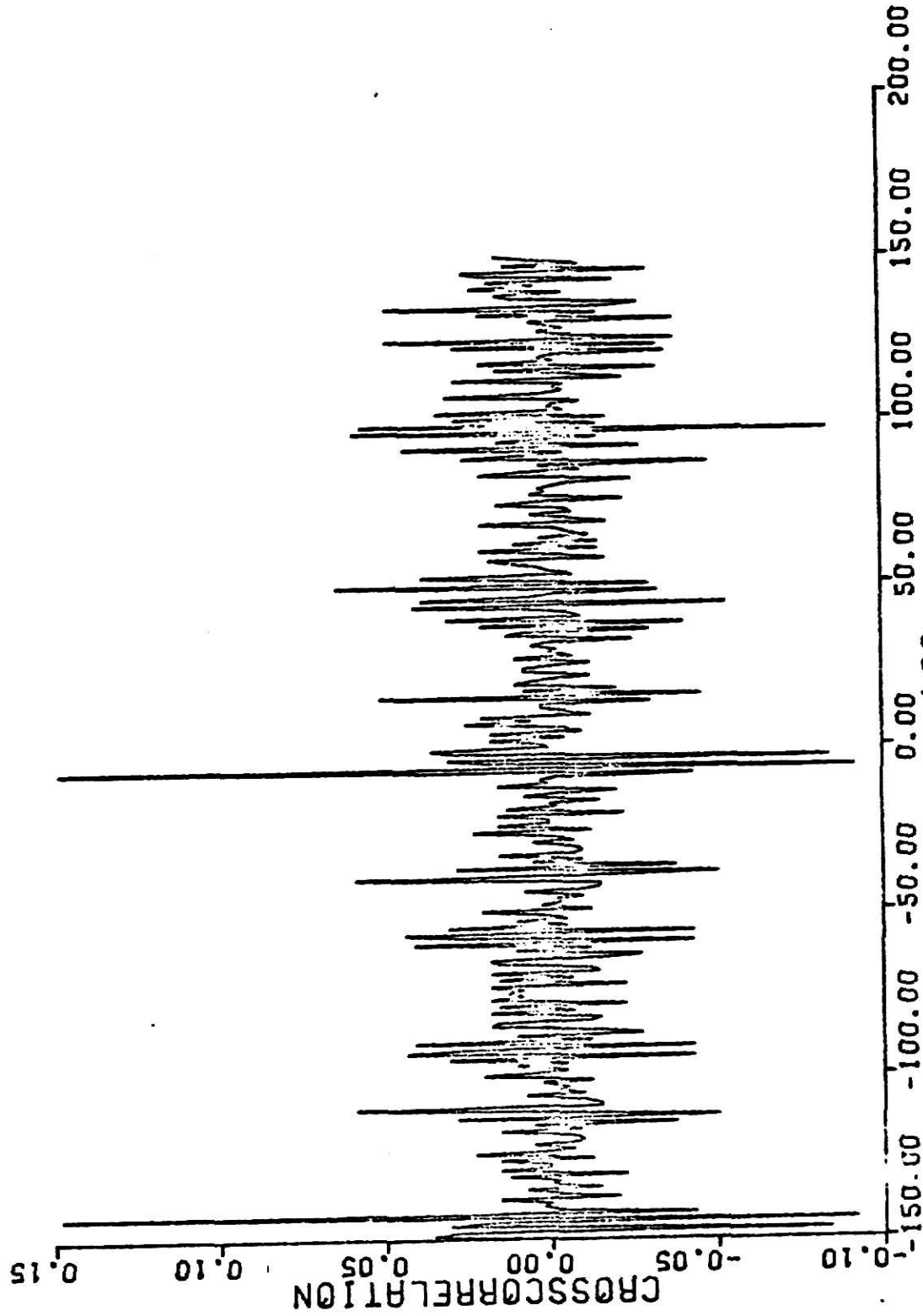

Fig. 5.57 Crosscorrelation of temp. station 4 - temp. station 3.

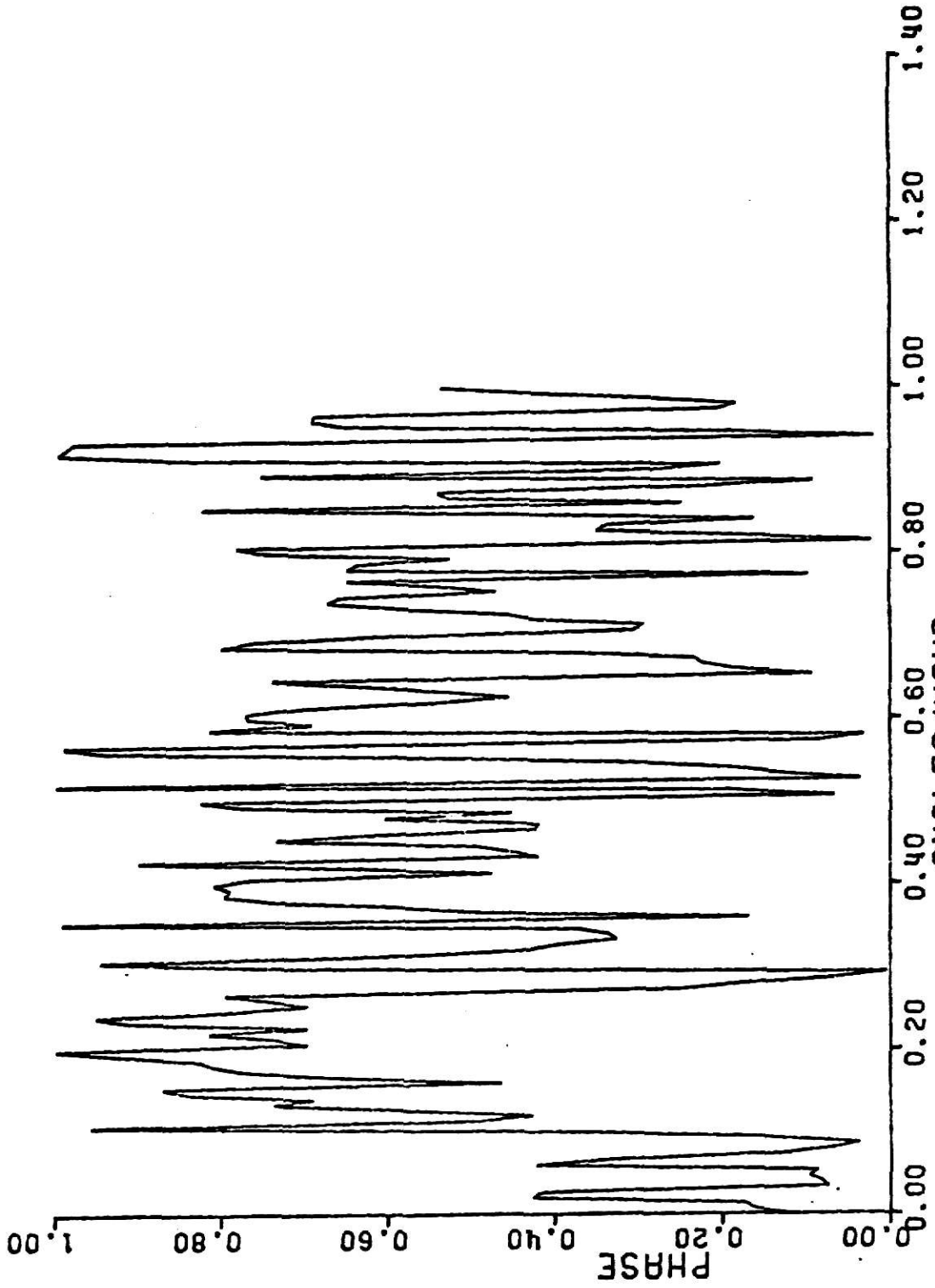

Fig. 5.58 Phase spectra, temp. station 4 - temp. station 3.

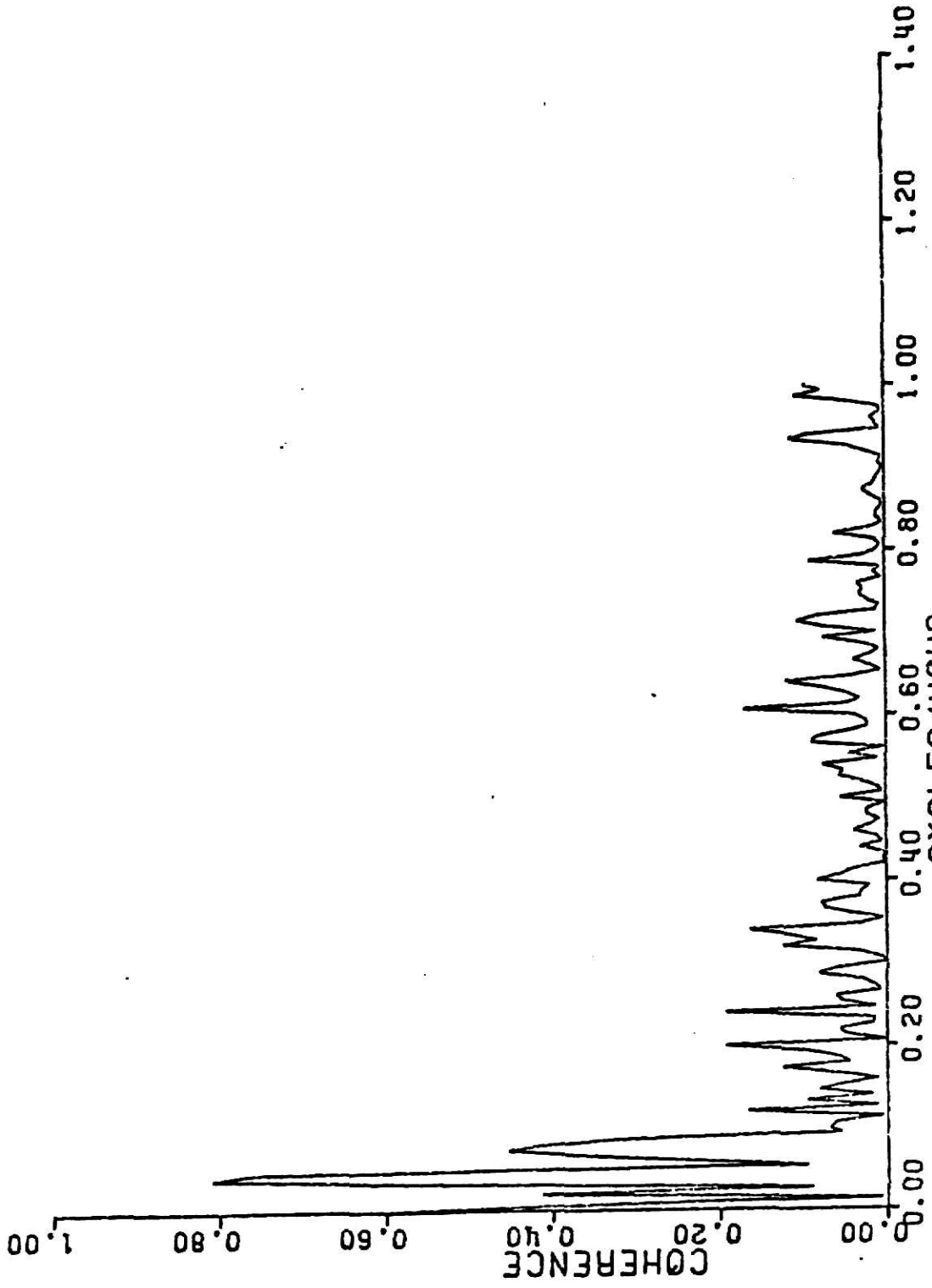

Fig. 5.59 Coherency spectra, temp. station 4 - temp. station 3.

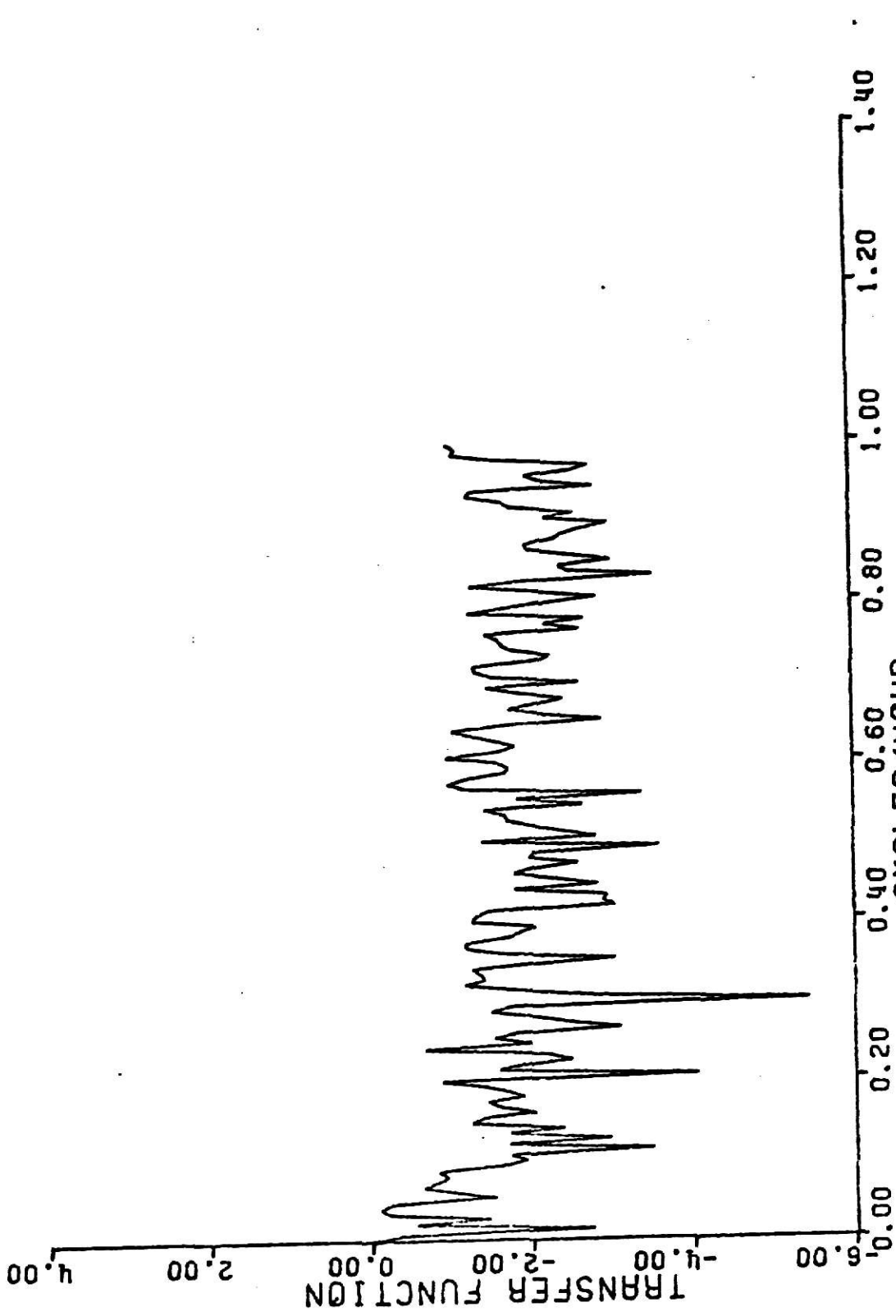

Fig. 5.60 Amplitude of transfer function, temp. station 4 - temp. station 3

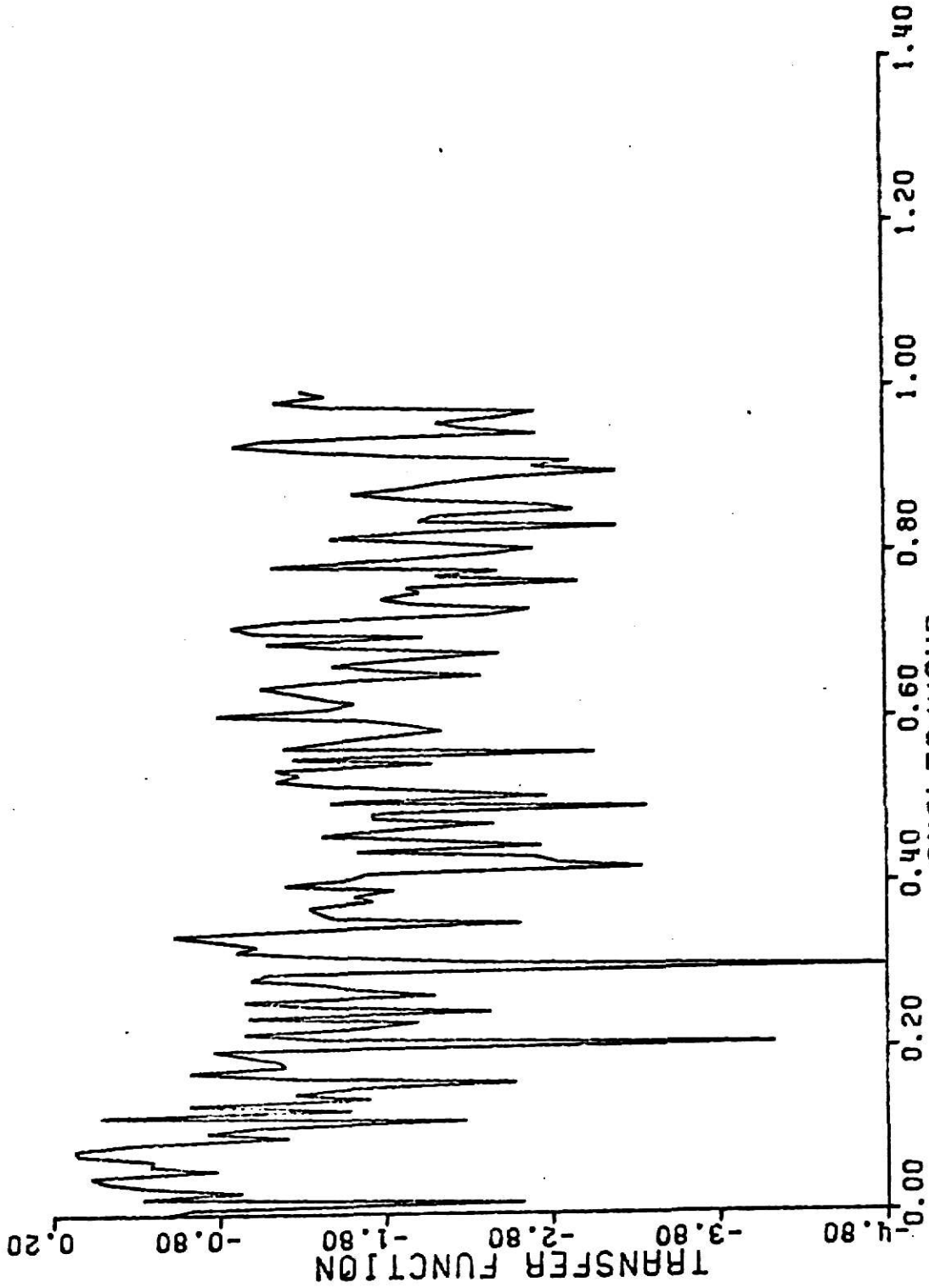

Fig. 5.61 Amplitude of transfer function, temp. station 3-temp. station 4.

indicating the desirability of filtered data for further analysis. Again, a simple difference filter was used. The crosscorrelation function, coherency, phase and transfer function are shown in Figures 5.57 thru 5.61. A high coherency is observed in the frequency band 0.040 cycles/hr - 0.046 cycles/hr, corresponding to 24 hrs period. The phase lag in this frequency band is around zero indicating that the two peaks occur simultanenously which may be expected for diurnal variation in temperature due to solar heating. Another high coherency is observed at zero frequency. This implies that the long range variations in temperature are also highly correlated. Coherency at all other frequencies is low. The transfer function is relatively low at all frequencies.

(d) DO station 1 and DO station 2:

Figures 5.62 thru 5.66 show the crosscorrelation, aligned coherency, phase and transfer function spectra of the differenced data. High coherence is observed at low frequencies implying thereby, the strong correlation of long periodic fluctuations. Low correlation is observed at 0.04 cycles/hr. (24 hrs period) and 0.08 cycles/hr. (12 hrs period). It was seen in the individual power spectral analysis that both the pollutants have significant diurnal variations. Low coherency at this frequency suggests that the causes of variation behind DO at station 1 and 2 are different. Since station 2 is near a waste discharge point, hence it seems that diurnal variation in BOD is a prime cause of the diurnal variation in DO at station 2 whereas diurnal variation in DO at station 1 is caused primarily by temperature and photosynthesis. A high transfer function is observed between station 1 and station 2 at low frequencies. It indicates

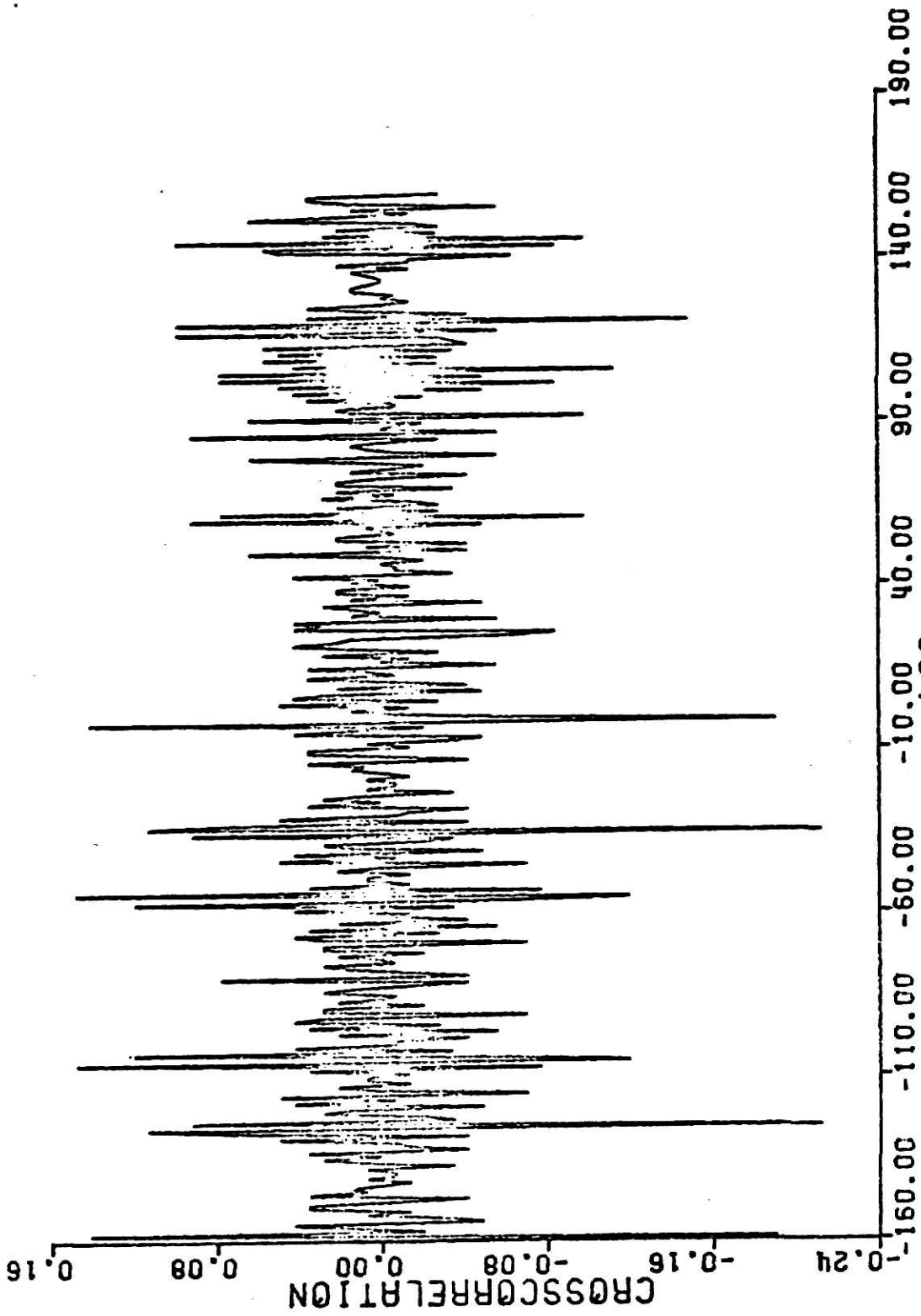

Fig. 5.62 Crosscorrelation of DO station 1 - DO station 2.

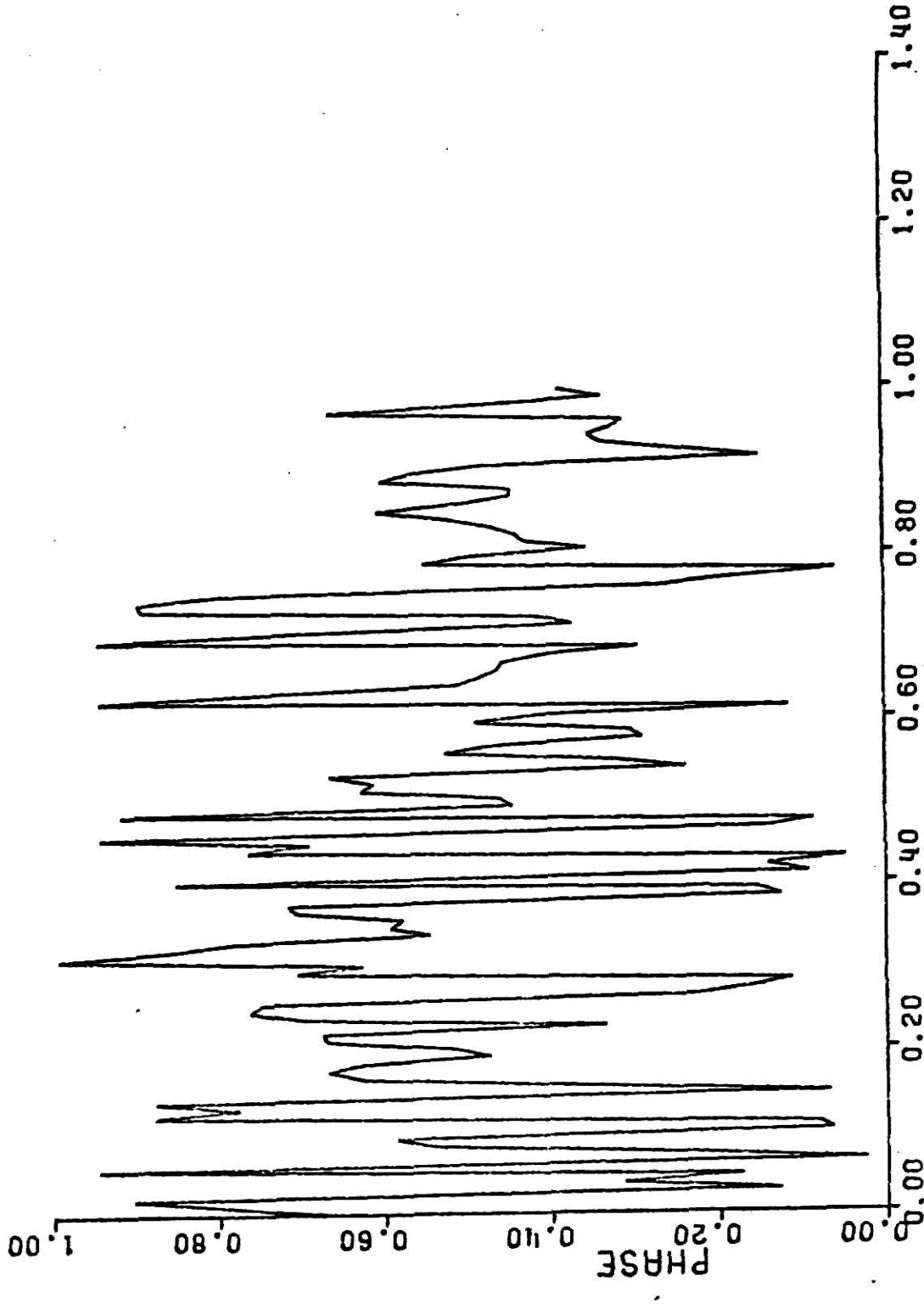

Fig. 5.63 Phase spectra, DO station 1 - DO station 2.

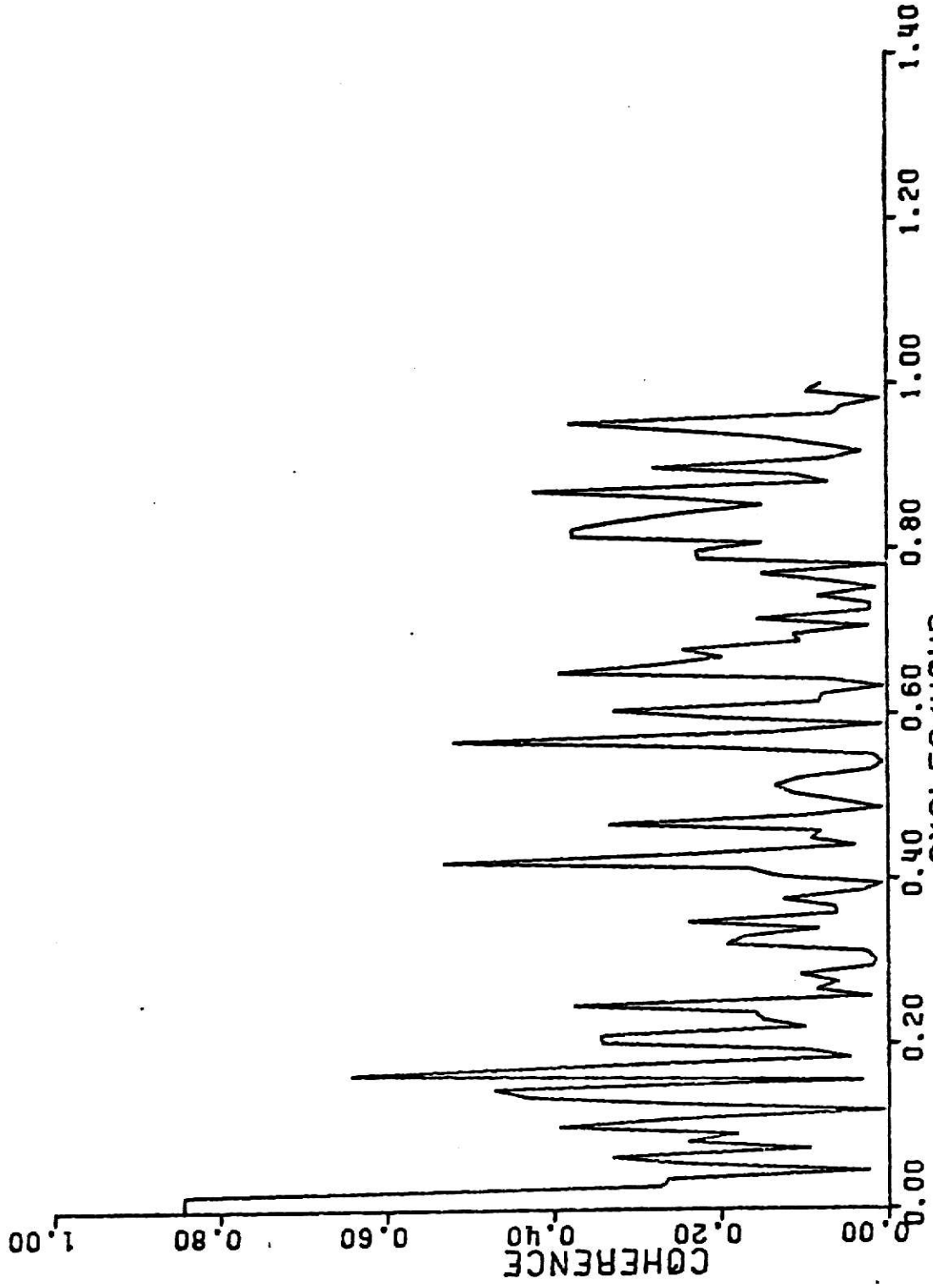

Fig. 5.64 Coherency spectra, DX station 1 - DC station 2.

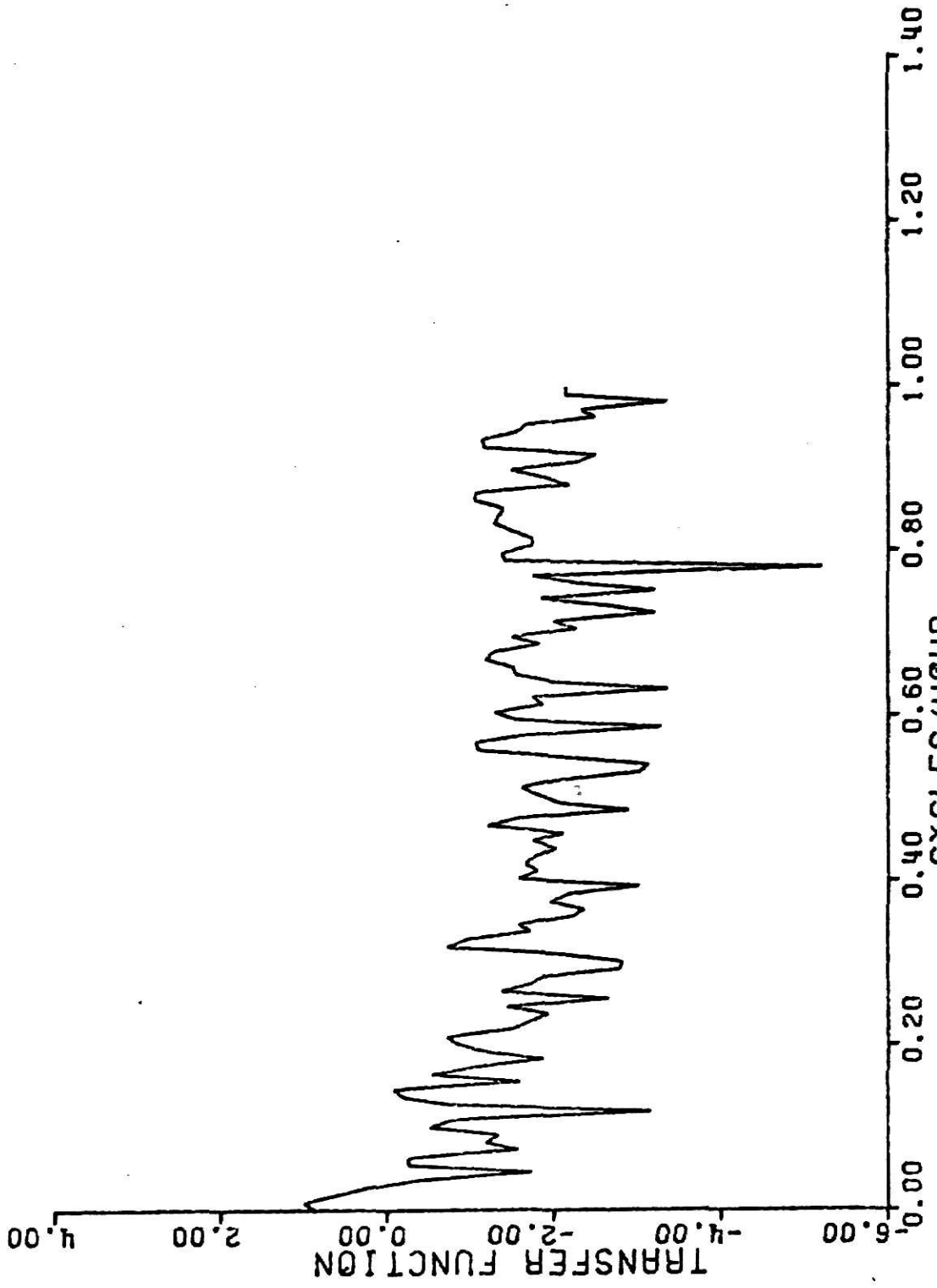

Fig. 5.65 Amplitude of transfer function, DO station 1 - DO station 2.

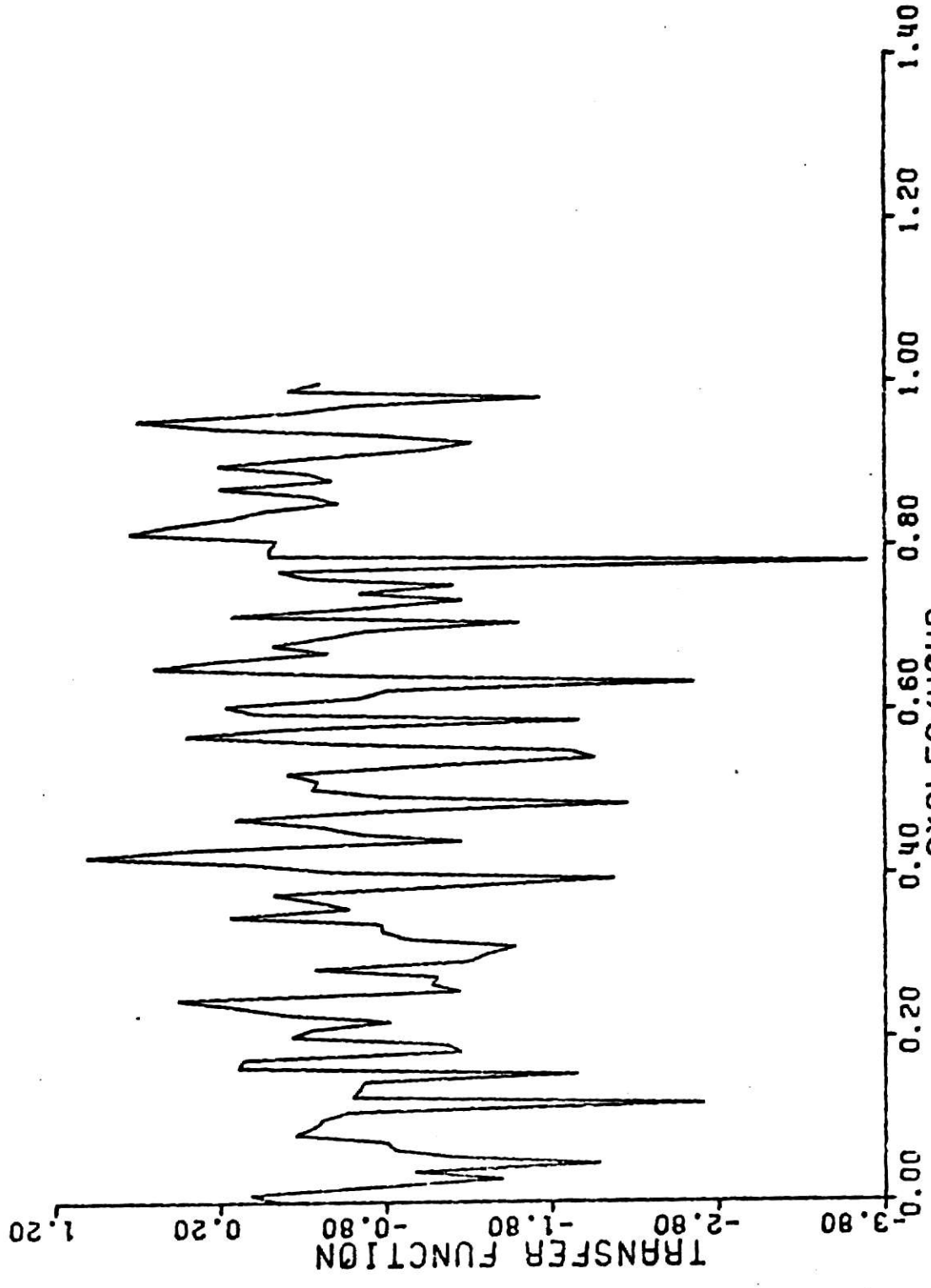

Fig. 5.66 Amplitude of transfer function, DC station 2 - DO station 1.

Table 5.13 Summary of Cross-Spectral Analysis of Temperature and Dissolved Oxygen at Stations 1,2,3 and 4

Series 1	Series 2	Frequency band corresponding to 24 hrs. period		Frequency band corresponding to 12 hrs. period		Any other frequency band		
		Coherency	Phase	Coherency	Phase	Period	Coherency	Phase
Temperature Station 2	DO Station 2	.147	180°	.124	180°		——	
Temperature Station 4	DO Station 4	.617	180°	High	180°	48 hrs.	.720	$.80^{\circ}$
Temperature Station 1	Temperature Station 2	.416	$79.2^{\circ} \pm 26^{\circ}$	Low	-	zero frequencies		
Temperature Station 1	Temperature Station 3	0.497	Approx. in phase	0.396	Approx in phase	zero frequency	.799	
Temperature Station 1	Temperature Station 4	.584	Series 4 lagging 1 by about 2 hrs.	.426	Approx. in phase	——		
Temperature Station 2	Temperature Station 3	.585	-	Low	-	zero frequency	.656	
Temperature Station 3	Temperature Station 4	.684	-	Low	-	——		
DO Station 1	DO Station 3	Low coherency throughout the frequency range						
DO Station 1	DO Station 4	Low coherency throughout the frequency range						
DO Station 2	DO Station 3	Low coherency throughout the frequency range						
DO Station 2	DO Station 4	Low coherency throughout the frequency range						
DO Station 3	DO Station 4	.751	In phase	.554	About 5 hrs. lag.	zero frequency	0.528	

that the variance in DO at station 2 will be high at low frequencies if the DO at station 1 were the only factor affecting it.

Similar cross spectral studies were conducted between other pairs of series and the corresponding results are summarized in Table 5.13.

5.3 Analysis of Potamac River Data at the Great Falls Station:

5.3.1 Introduction: Temperature, dissolved oxygen, biochemical oxygen demand and chloride records for this station were analyzed using spectral analysis and cross-spectral analysis. Predictive models were obtained for each pollutant using parametric modeling. About 2% of the observations were found missing in all the records. As suggested in [27], the mean for each record was subtracted from the whole series and the missing values were replaced by zero. Figures 5.67 to 5.70 show the records of the four pollutants. Inspection of the records does not reveal any obvious trends. The presence of an annual frequency is indicated in the plots.

5.3.2 Harmonic Analysis: As the pollutants show a tendency to have an annual cycle, it was decided to carry out harmonic analysis to investigate more about these cyclic fluctuations. Tables 5.14 thru 5.17 show the results of harmonic analysis. The annual component (7th harmonic) alone accounts for about 75 - 85% of the total variance for temperature and dissolved oxygen. In case of the chloride record, the annual component accounts for 30% of the total variance. Another 17% of the variance is caused by the 2nd harmonic corresponding to a 3 years cycle. Biochemical oxygen demand does not show a dominant effect for any particular frequency. It's variance is evenly distributed over the whole frequency range.

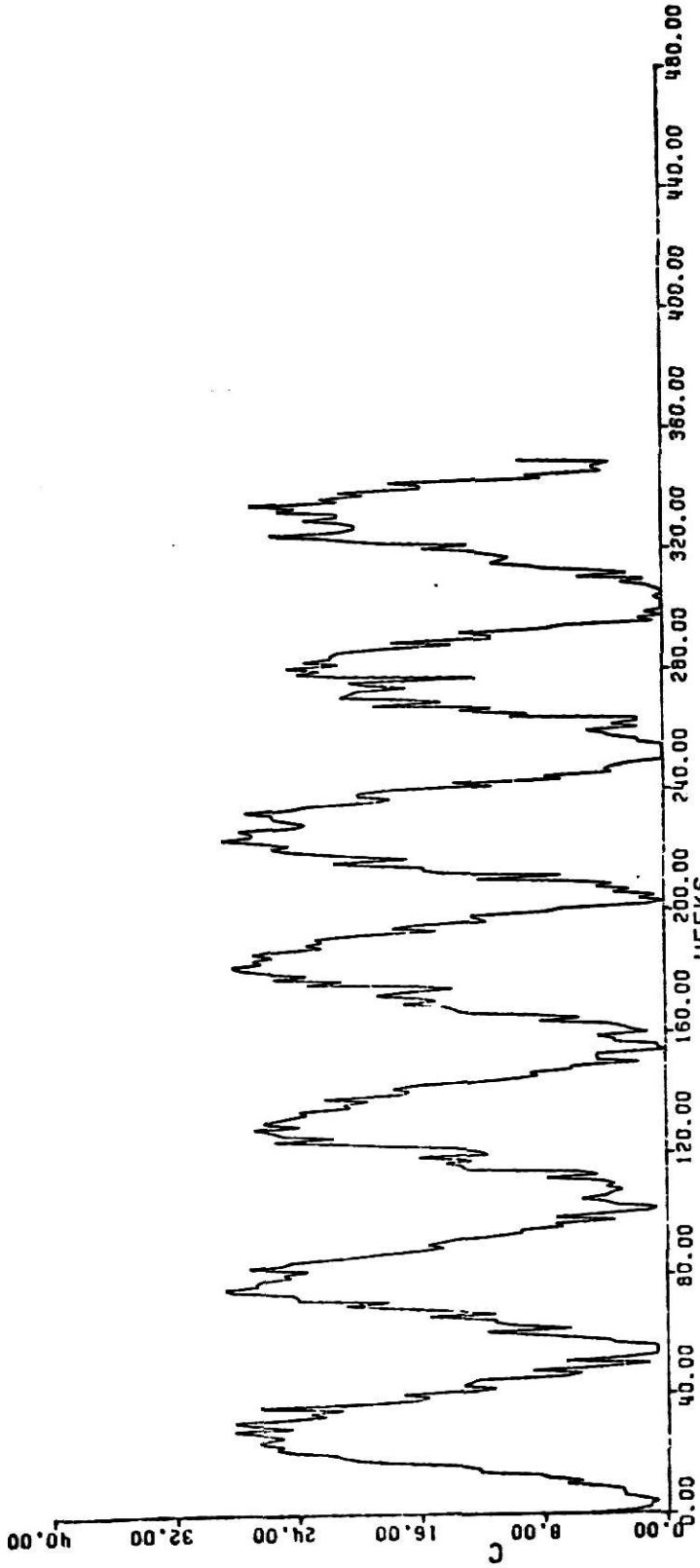

Fig. 5.67 Temperature record - Great Falls station

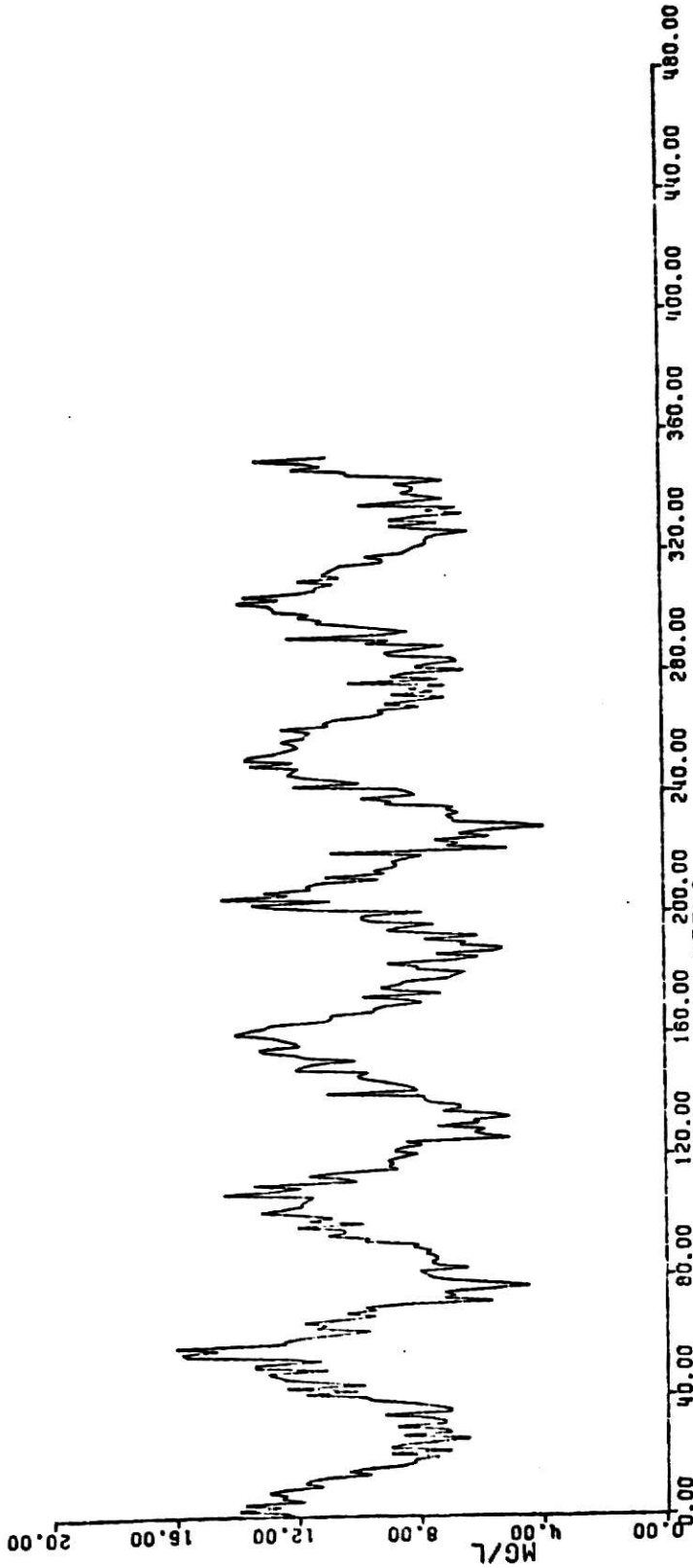

Fig. 5.68 Dissolved oxygen record - Great Falls station.

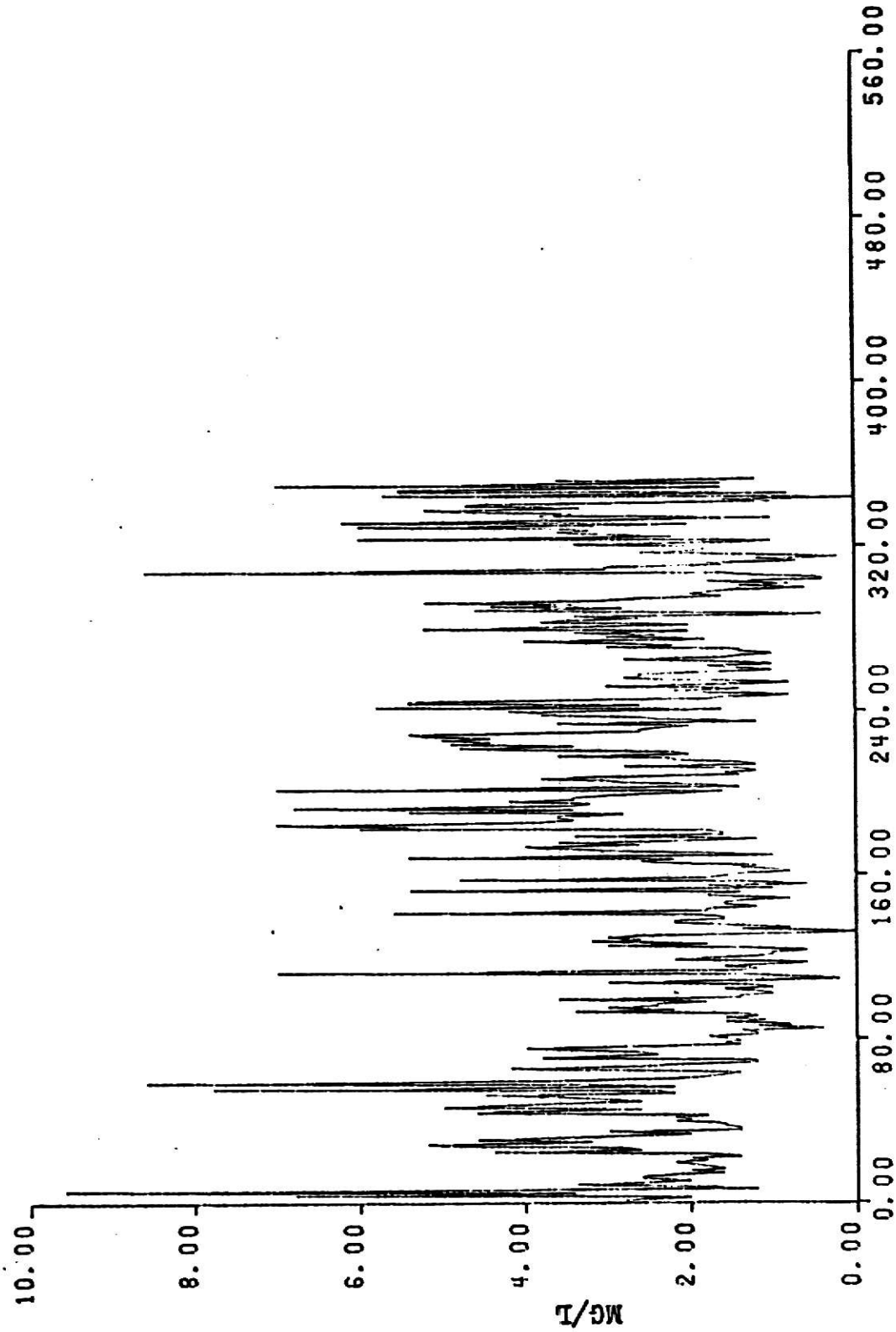

Fig. 5.69 Biochemical oxygen demand record - Great Falls station.

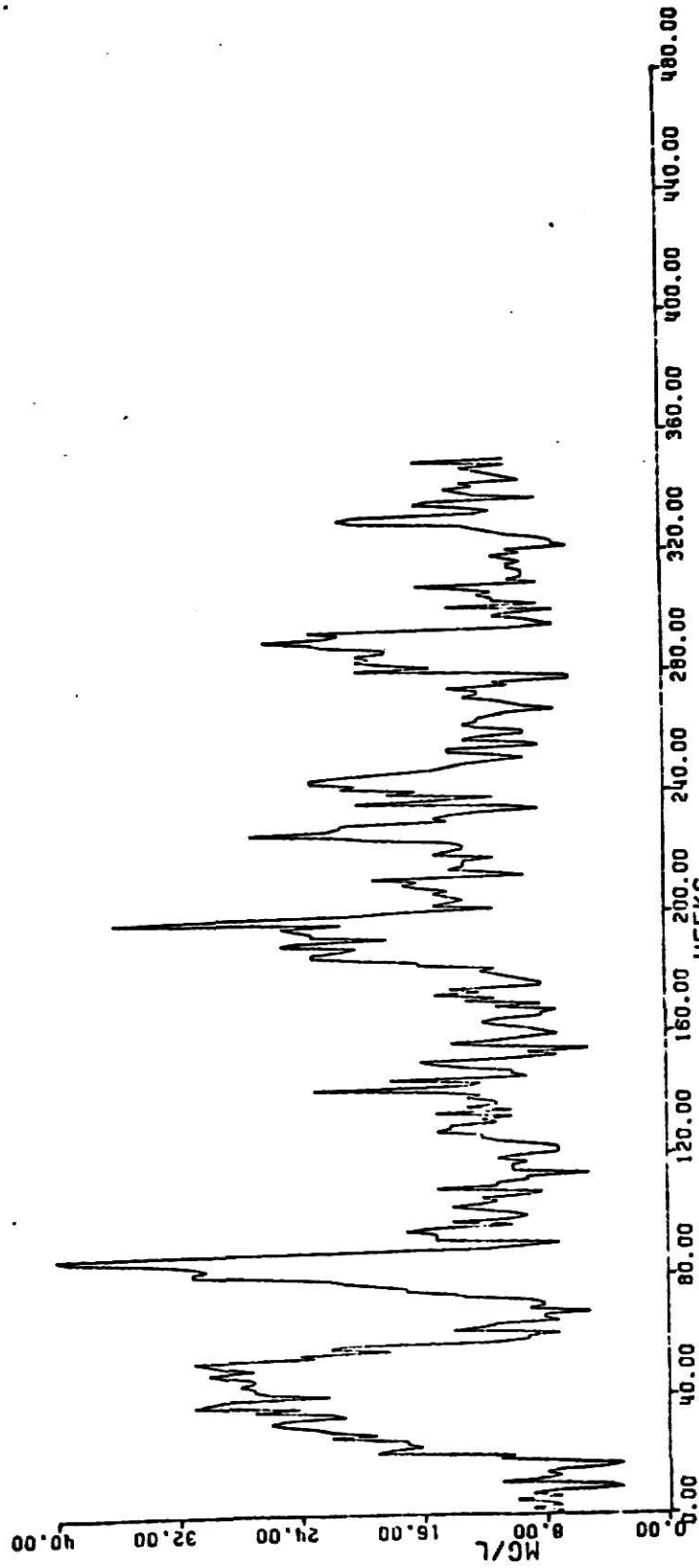

Fig. 5.70 Chloride record - Great Falls station.

Table 5.14 Harmonic Analysis - Temperature Great Falls Station

Mean = 13.49°C Total Variance = 77.58 $(^{o}C)^{2}$

Source	Amplitude ^{o}C	Phase (in degrees)	Percentage contribution to total variance
2nd harmonic	1.002	71.8	2.59
7th harmonic	5.81	39.22	87.18

Table 5.15 Harmonic Analysis - Dissolved Oxygen Great Falls Station

Mean = 9.53 mg/ℓ Total Variance = 5.36 $(mg/\ell)^2$

Source	Amplitude mg/ℓ	Phase (in degrees)	Percentage contribution to total variance
fundamental	0.192	16.1	1.38
3rd harmonic	0.207	-54.4	1.60
7th harmonic	1.421	35.9	75.28
14th harmonic	0.125	62.7	0.58
27th harmonic	0.168	-60.1	1.05

Table 5.16 Harmonic Analysis - Biochemical Oxygen Demand Great Falls Station.

Mean = 2.608 mg/ℓ Total Variance = 2.38 $(mg/\ell)^2$

Source	Amplitude (mg/ℓ)	Phase (in degrees)	Percentage contribution to total variance
fundamental	0.164	-55.5	2.28
2nd harmonic	0.302	49.7	7.71
5th harmonic	0.118	-84.9	1.17
6th harmonic	0.232	-16.7	4.54
7th harmonic	0.195	76.3	3.20
8th harmonic	0.161	-57.2	2.18
14th harmonic	0.124	34.5	1.30
15th harmonic	0.119	68.9	1.19
16th harmonic	0.173	-83.2	2.54
21st harmonic	0.172	10.9	2.49
23rd harmonic	0.135	9.1	1.55
25th harmonic	0.143	-1.9	1.72
27th harmonic	0.133	42.4	1.49

Table 5.17 Harmonic Analysis - Chloride Great Falls Station.

Mean = 14.46 mg/ℓ Total Variance = 43.89 $(mg/\ell)^2$

Source	Amplitude (mg/ℓ)	Phase (in degrees)	Percentage contribution to total variance
fundamental	0.518	80.7	1.22
2nd harmonic	1.908	-74.9	16.60
3rd harmonic	1.070	10.7	5.22
4th harmonic	0.709	-29.3	2.30
5th harmonic	0.592	23.6	1.60
6th harmonic	0.587	26.5	1.57
7th harmonic	2.525	-83.3	29.06
8th harmonic	0.582	-37.3	1.54
9th harmonic	0.584	42.2	1.55
10th harmonic	0.953	-52.8	4.15
12th harmonic	0.746	-82.1	2.54
13th harmonic	0.742	8.9	2.51
15th harmonic	0.497	-50.9	1.13
17th harmonic	0.886	28.5	3.58
24th harmonic	0.579	72.5	1.53
39th harmonic	0.534	22.7	1.30

After having obtained some idea about the behaviour of each pollutant by harmonic analysis, we proceed to spectral analysis.

5.3.3 Spectral Analysis: Figures 5.71 to 5.78 show the autocorrelation and power spectral estimates of the temperature, dissolved oxygen, biochemical oxygen demand and chloride respectively. In general, the autocorrelation function dies off quickly for all pollutants. A high value at 52 lags for DO, temperature and chloride indicates the presence of an annual cycle. The power spectrum for temperature shows a high peak at .011 cycles/week, corresponding to an annual cycle. Other less dominant peaks are observed at periods of 26 weeks, and 16 weeks which may be attributed to seasonal changes in temperature. An annual peak is also evident in power spectrum for dissolved oxygen. This may be correlated with the annual variation in temperature. Other dominant peaks are seen at periods of 30 weeks and 14 weeks. These may also be linked with the seasonal variation in temperature. The power spectrum for biochemical oxygen demand shows dominant peaks at 52 weeks, 26 weeks, 16 weeks, 7 weeks and about 18 days. The exact cause of variation of BOD over such a large frequency range is not known. As BOD is expected to follow cyclic fluctuations with periods shorter than one week, hence effects due to aliasing will be present. This may distort the spectrum, especially, at high frequencies. The chloride power spectrum shows dominant peaks at periods of 52 weeks, 12 weeks and 8 weeks. These variations in chloride content may be correlated with the variations in the flow rate in the river.

Harmonic regression was performed to remove the dominant harmonics as indicated by the harmonic analysis. The residuals, thus obtained

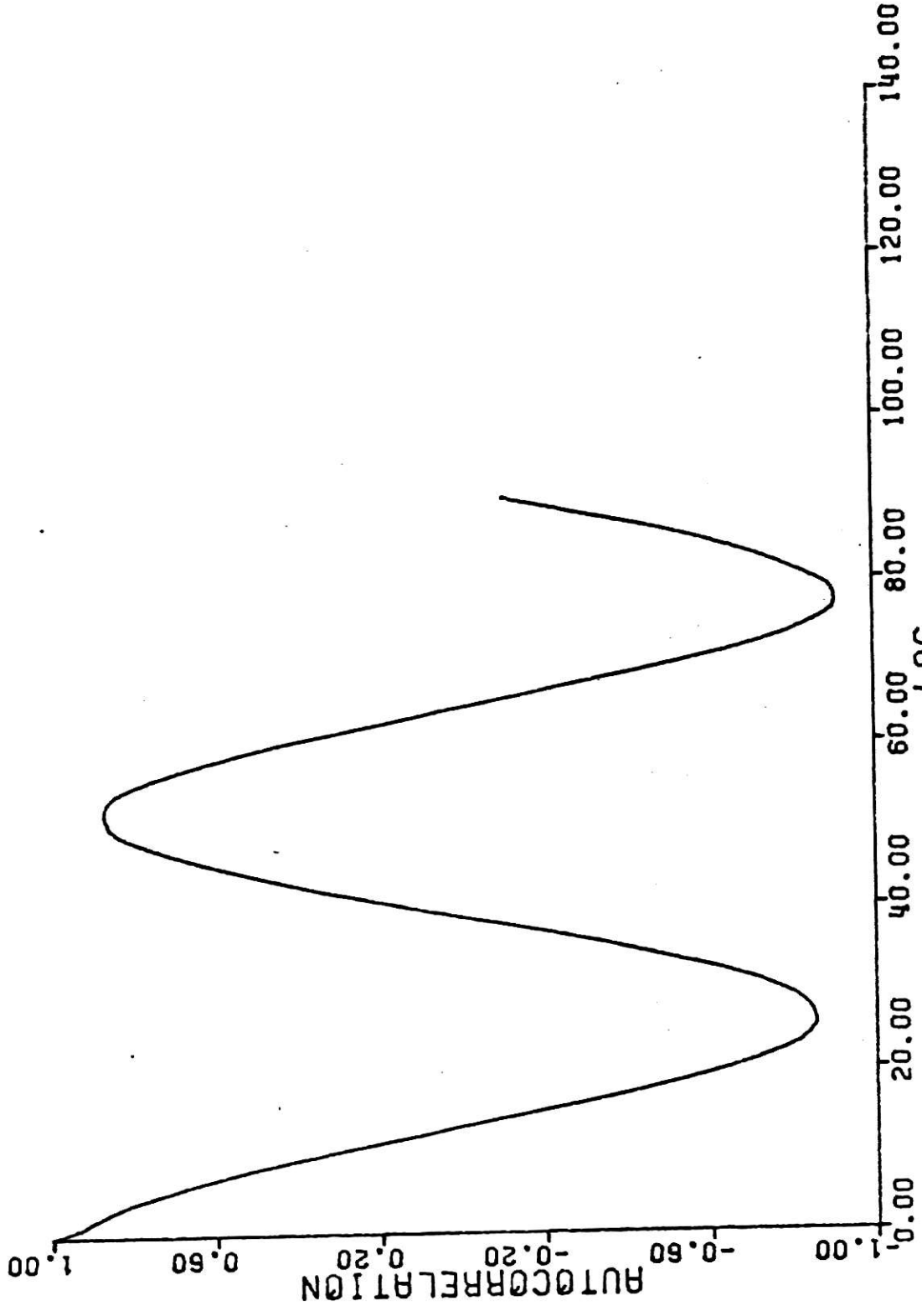

Fig. 5.71 Autocorrelation of temperature record.

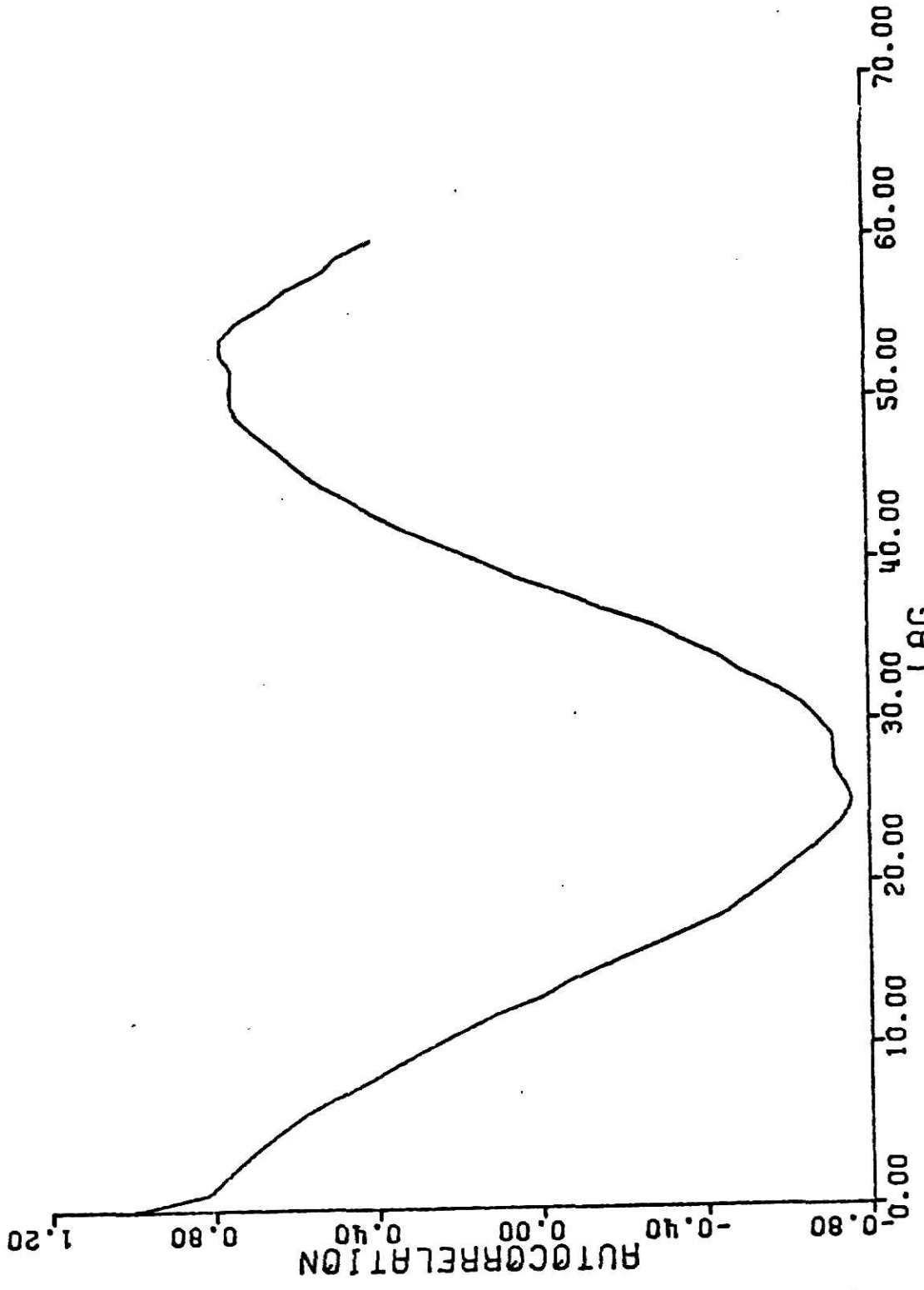

Fig. 5.72 Autocorrelation of DC record.

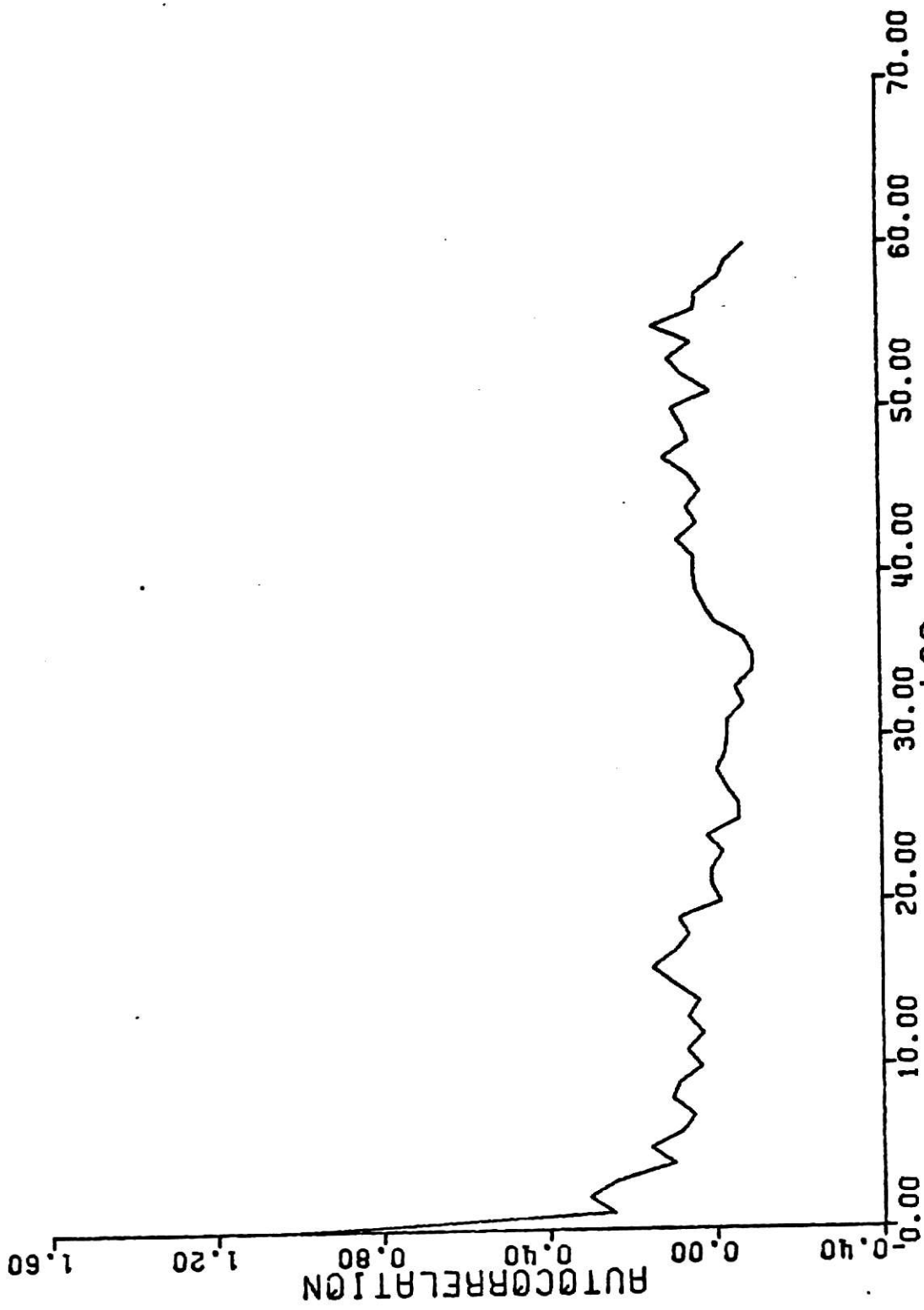

Fig. 5,73 Autocorrelation of BOD record.

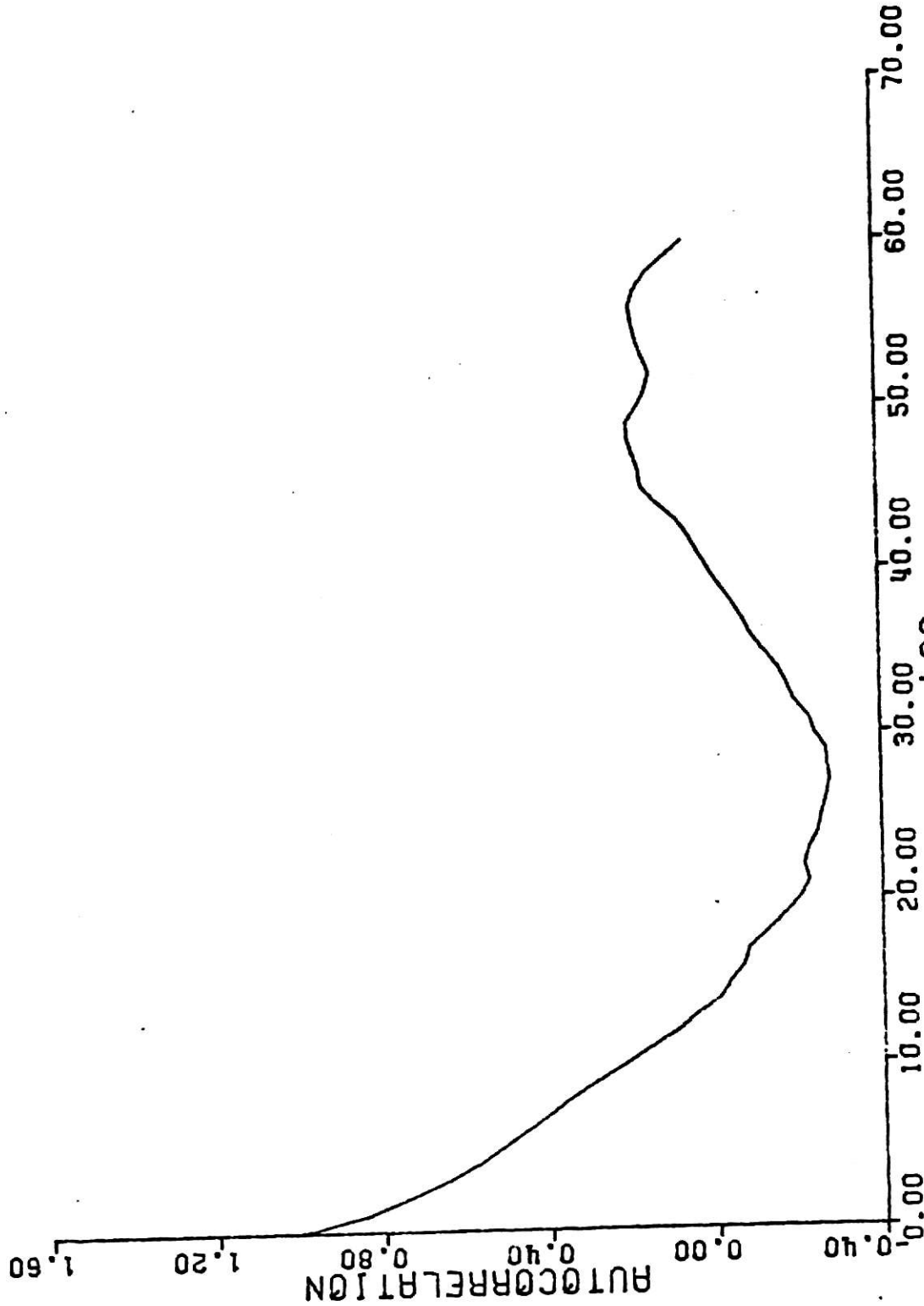

Fig. 5.74 Autocorrelation of Chloride record.

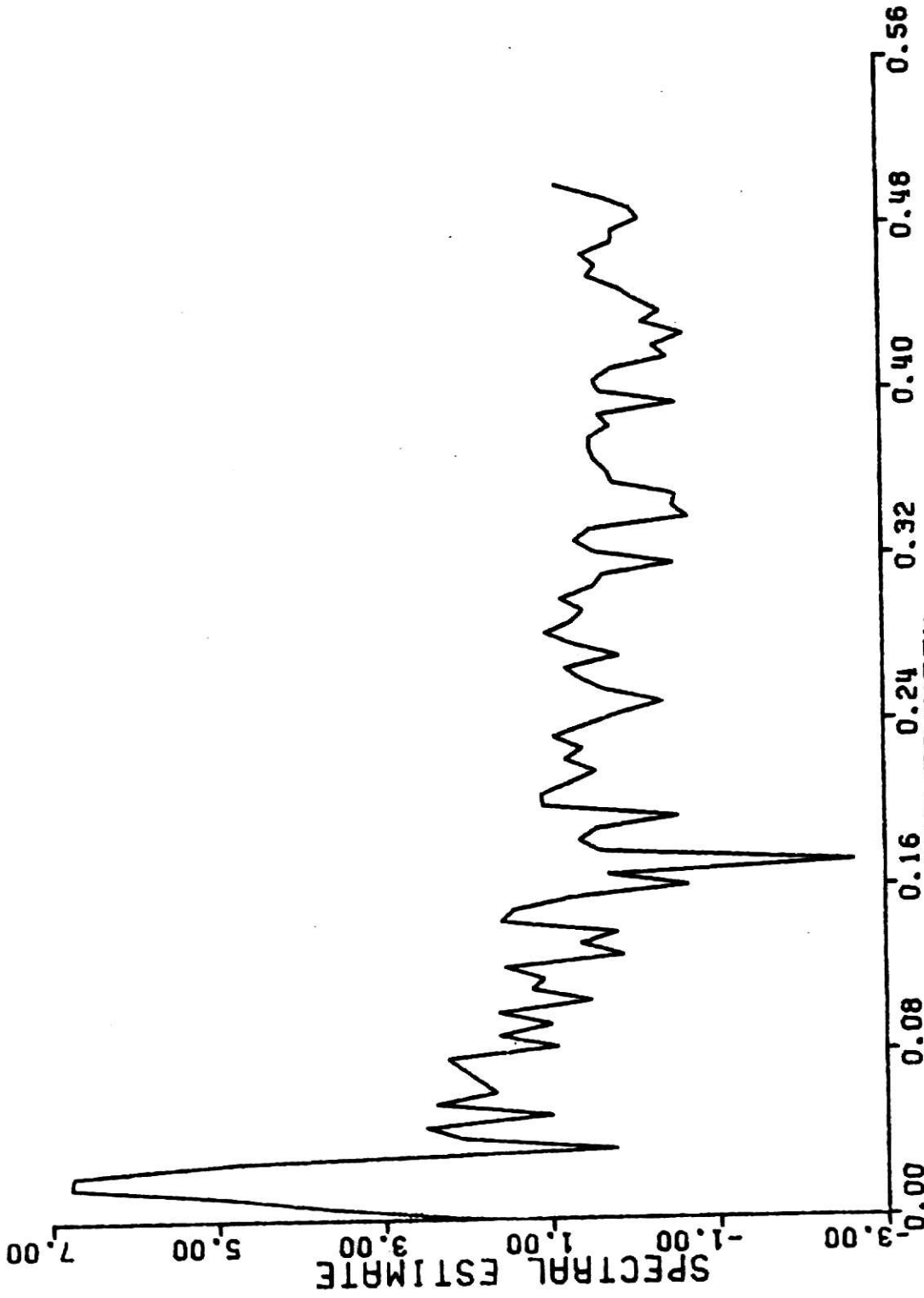

Fig. 5.75 Spectral estimate of temperature.

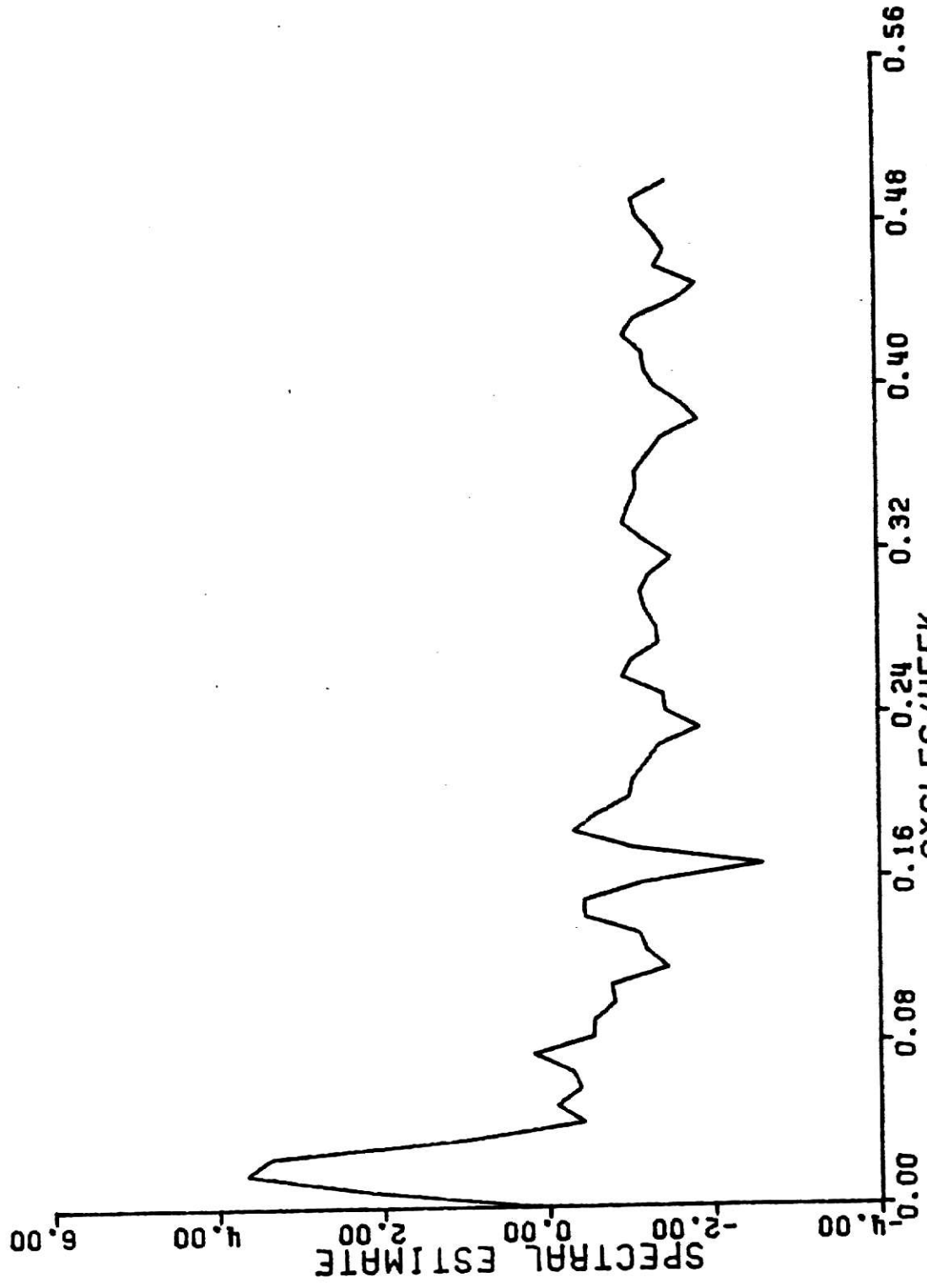

Fig. 5.76 Spectral estimate of ω.

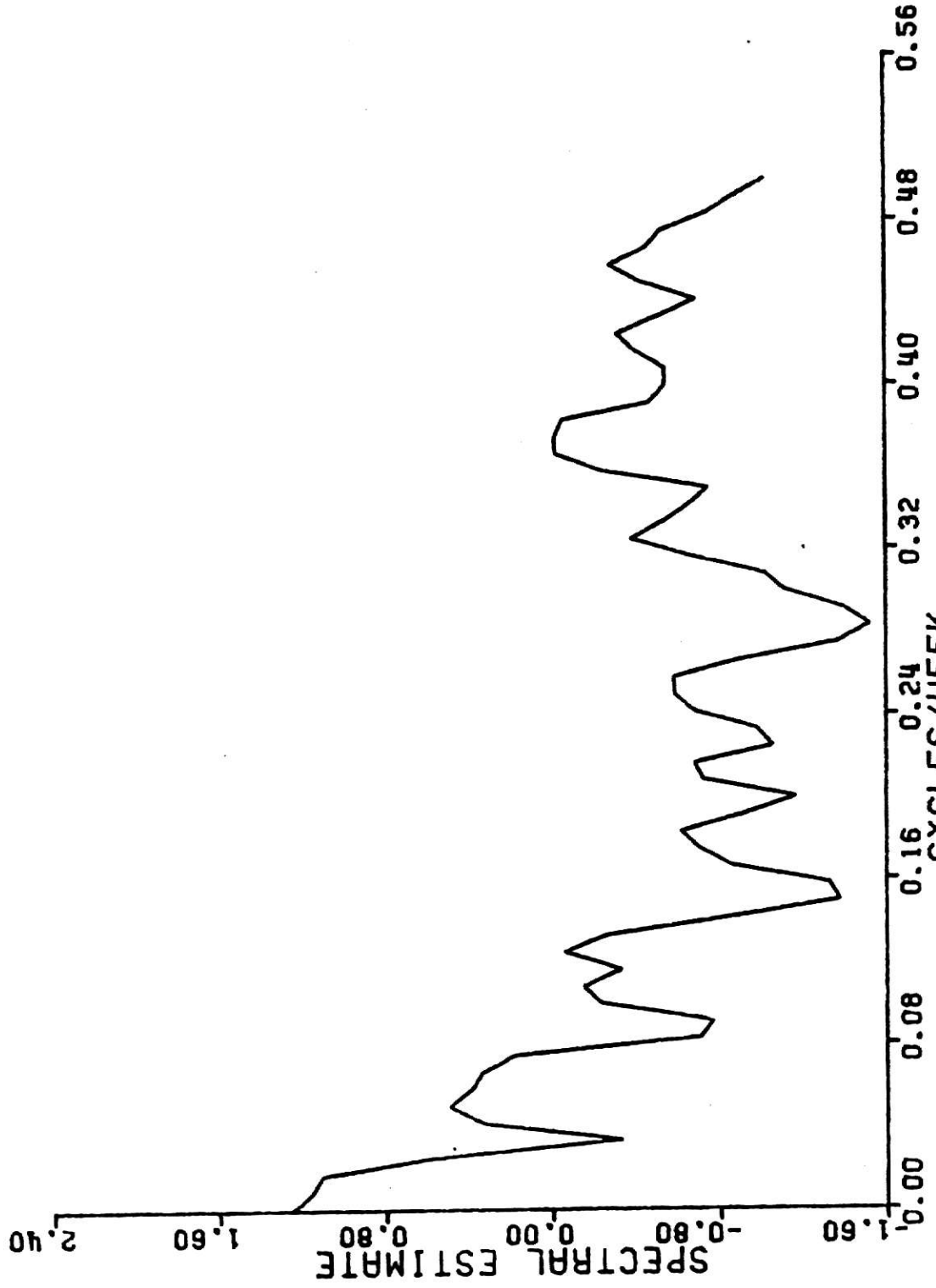

Fig. 5.77 Spectral estimate of BOD.

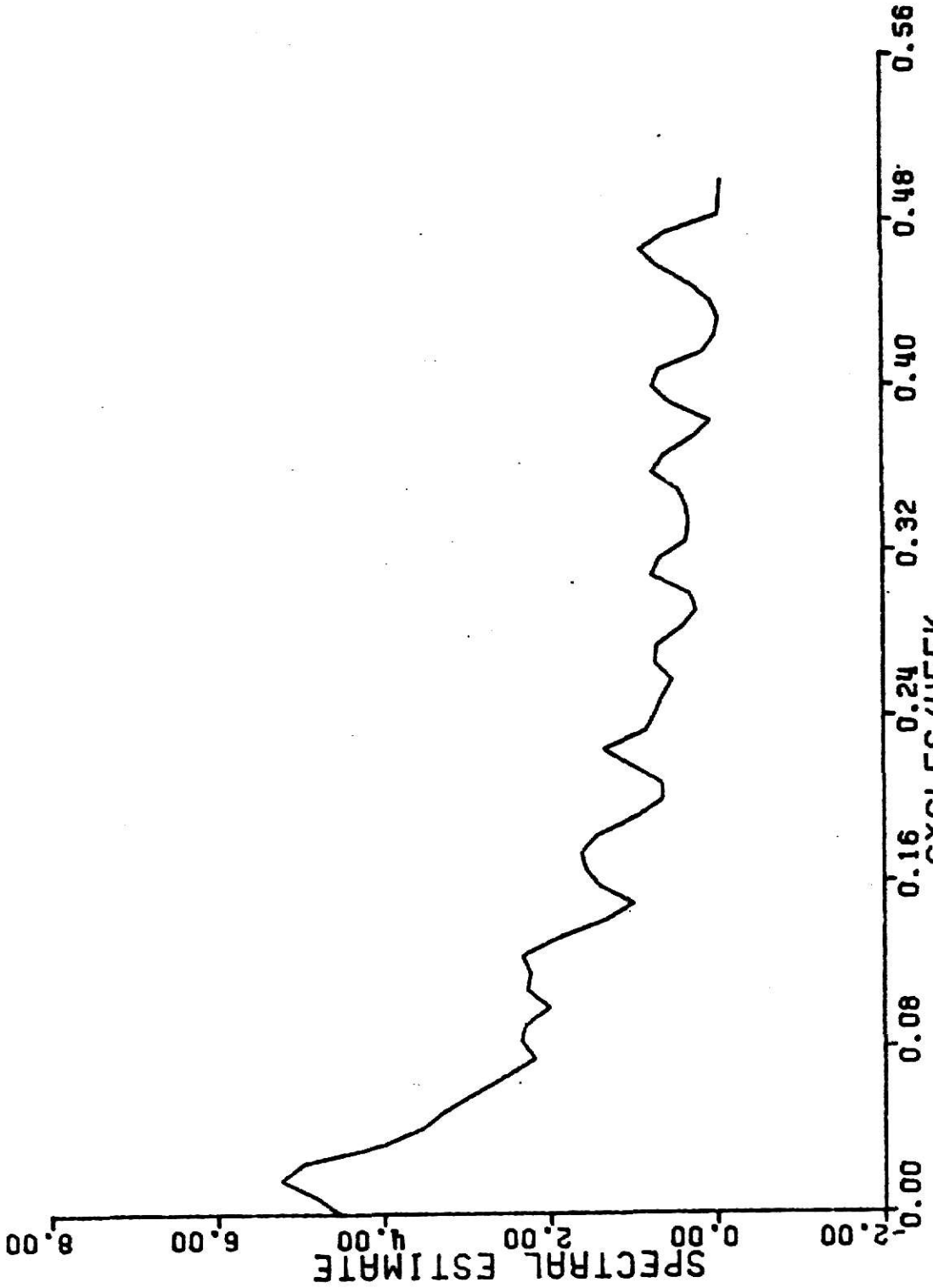

Fig. 5.78 Spectral estimate of chloride.

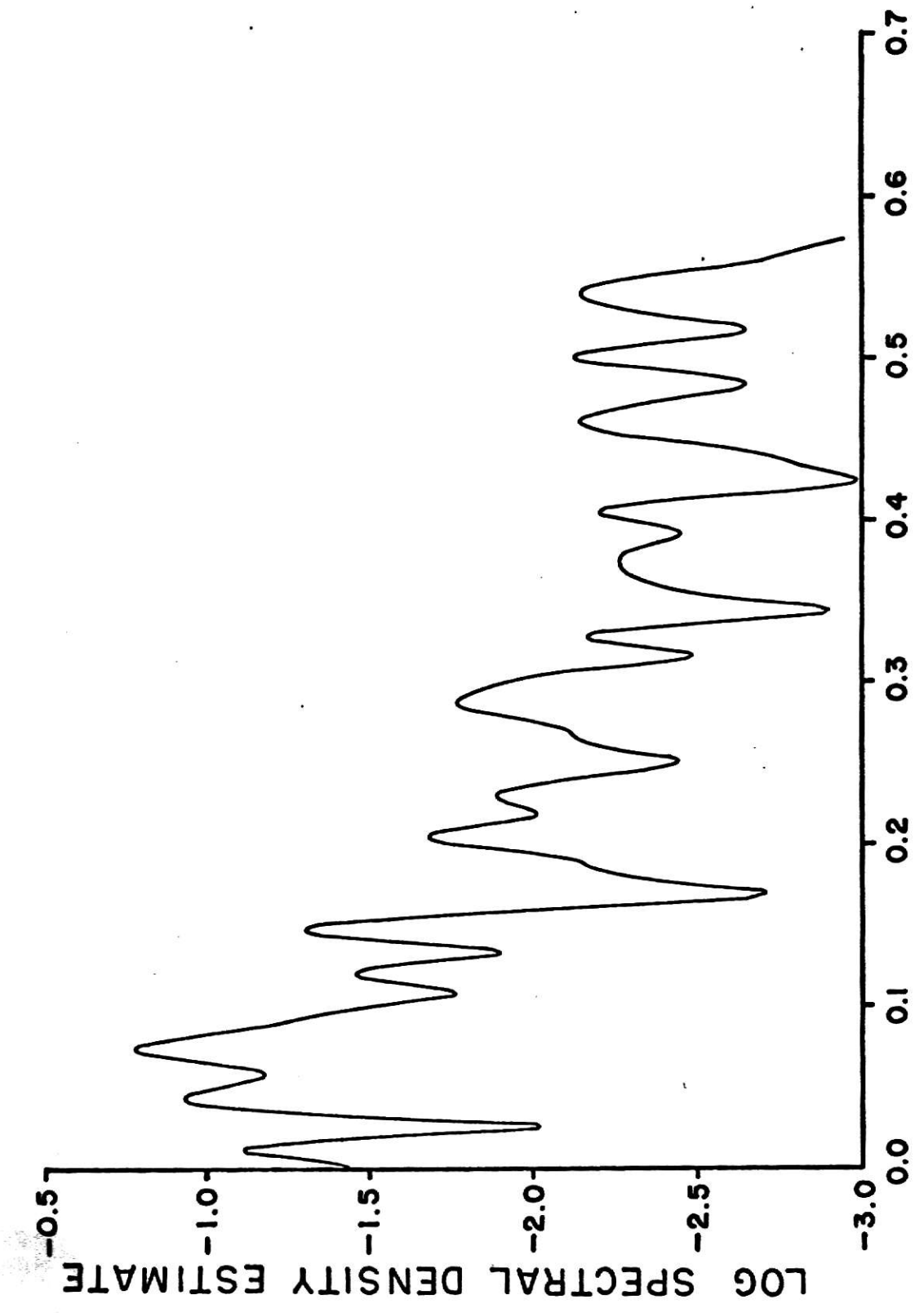

Fig. 5.79 Spectral estimate of temperature (residuals).

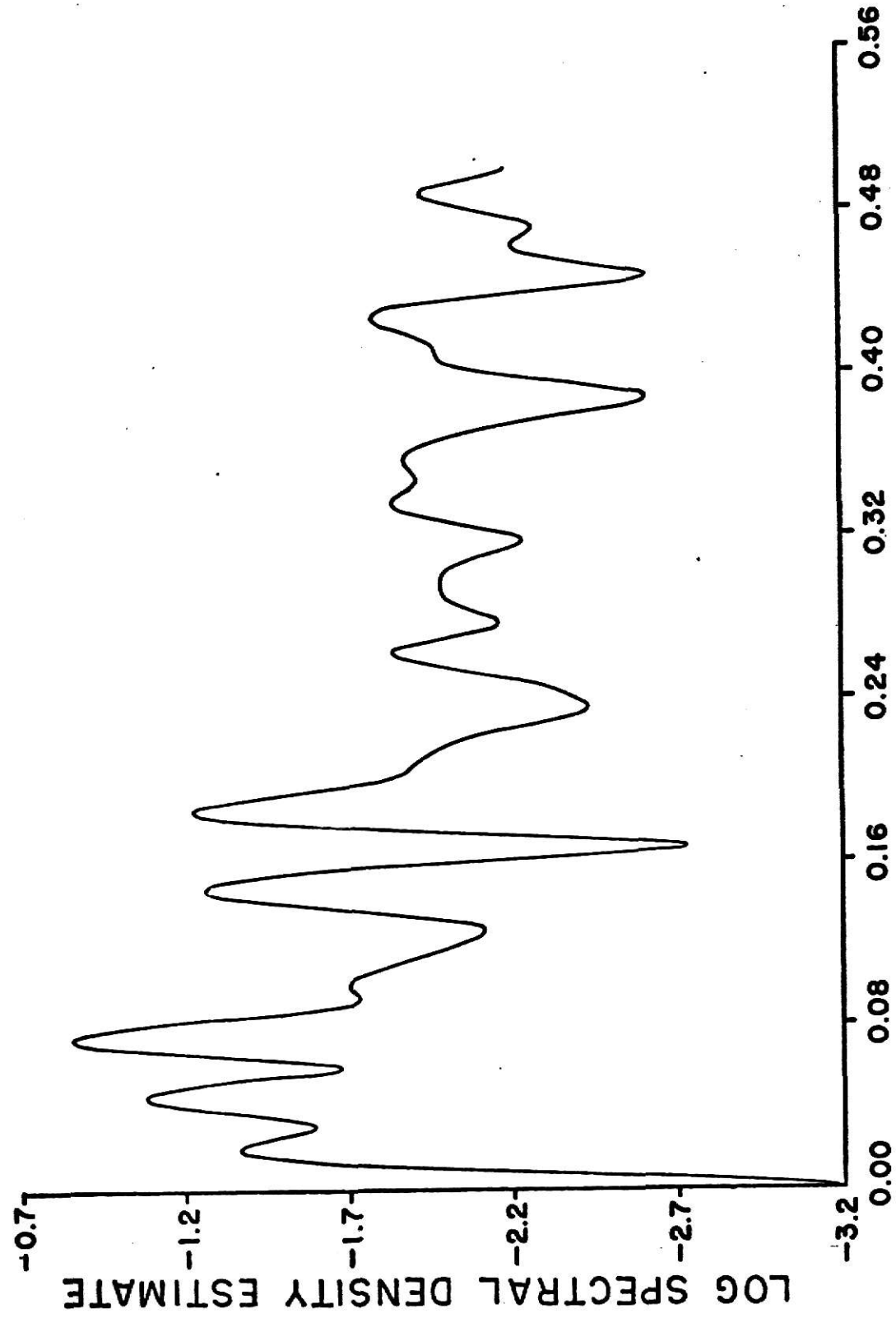

Fig. 5.80 Spectral estimate of DO (residuals).

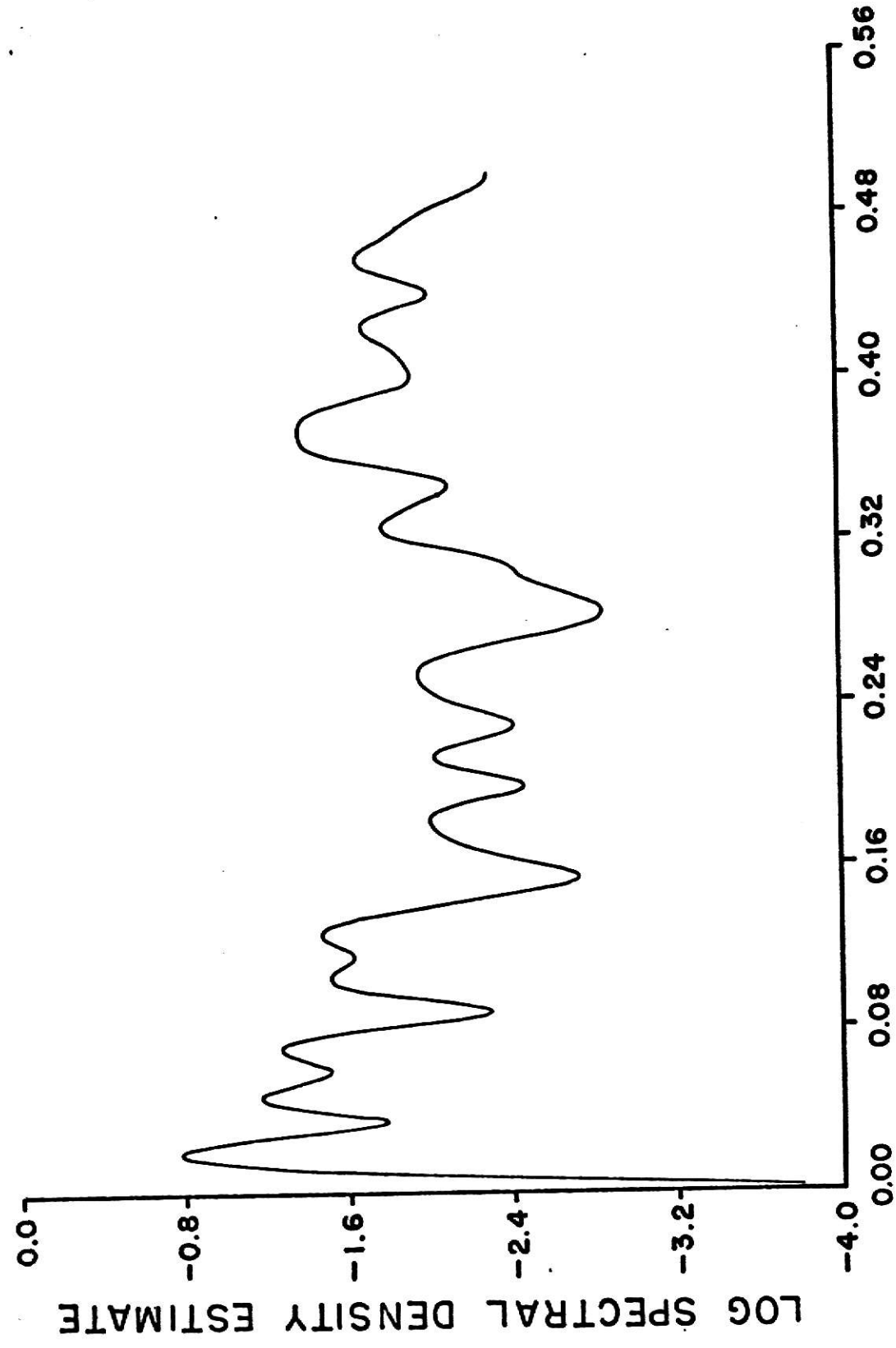

Fig. 5.81 Spectral estimate of BOD (residuals).

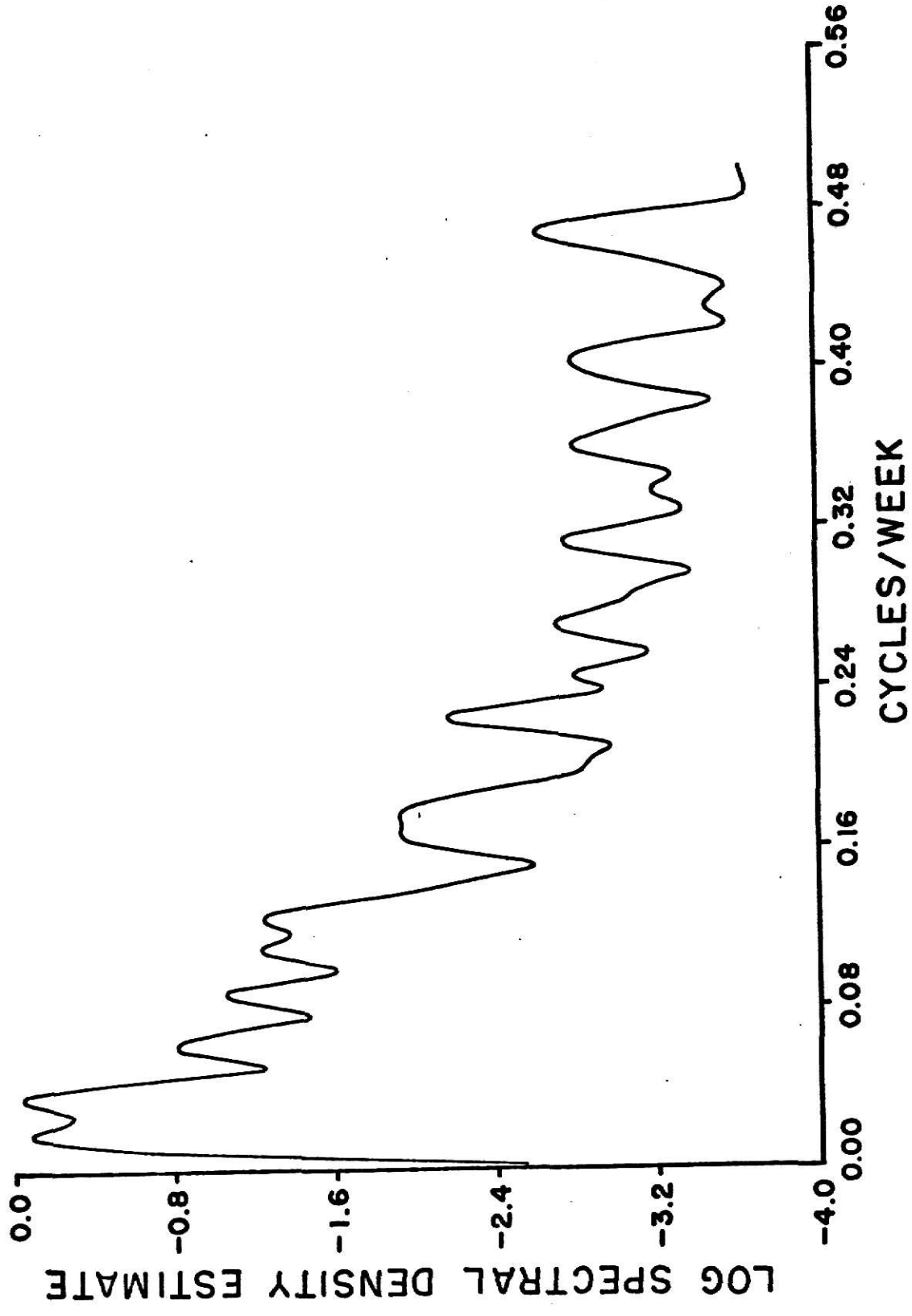

Fig. 5.82 Spectral estimate of chloride (residuals).

were again subjected to spectral analysis. Figures 5.79 thru 5.82 show the corresponding power spectrum estimates. A comparison of the variances before and after removal of the harmonics as evidenced by the magnitude of peaks in the spectral estimates shows that it has been reduced considerably throughout the whole frequency range for all pollutants.

The regression equation for each pollutant is given below. In each equation the mean has been subtracted from the data.

(a) Temperature:

Annual frequency alone gives a high multiple correlation coefficient. Thus, the model is

$$T_t = -0.0811 - 10.57 \cos \frac{2\pi t}{52} - 4.98 \sin \frac{2\pi t}{52} .$$

The mean of the temperature record should be added to it for obtaining the actual temperature.

with $R^2 = .936$

and $F(2,3\ 58) = 1275.0$

(b) Dissolved oxygen:

In this case also, a sufficiently high multiple correlation coefficient is obtained the annual frequency alone.

Thus, the model can be given as

$$DO_t = 0.0205 + 2.64 \cos \frac{2\pi t}{52} + 1.08 \sin \frac{2\pi t}{52}$$

with $R^2 = .868$

and $F(2, 358) = 548.69$

(c) Biochemical oxygen demand:

The model is

$$BOD_t = -0.3176 + 0.0018t - 0.1956 \cos \frac{2\pi t}{52}$$

$$-0.2664 \sin \frac{2\pi t}{52} - 0.2479 \cos \frac{2\pi t}{17}$$

$$-0.2306 \sin \frac{2\pi t}{17} + 0.2514 \cos \frac{2\pi t}{26}$$

$$+0.0740 \sin \frac{2\pi t}{26} - 0.0640 \cos \frac{2\pi t}{13}$$

$$+0.2024 \sin \frac{2\pi t}{13} + 0.3922 \cos \frac{2\pi t}{180}$$

$$+0.5636 \sin \frac{2\pi t}{180} + 0.1845 \cos \frac{2\pi t}{361}$$

$$-0.0727 \sin \frac{2\pi t}{361}$$

with $R^2 = .418$

$F(13, 347) = 5.68$

This model is not efficient for prediction purposes as evidenced by a low value of the multiple correlation coefficient.

(d) Chloride:

As indicated by harmonic analysis the regression equation consisted of 14 independent and 1 dependent variable. But a sufficiently high multiple correlation coefficient was obtained using only 6 terms. The model is

$$C_t = 0.0017 + 3.6680 \sin \frac{2\pi t}{180}$$

$$- 2.0997 \cos \frac{2\pi t}{120} - 5.0503 \sin \frac{2\pi t}{52}$$

$$- 1.2155 \cos \frac{2\pi t}{36} + 1.6463 \sin \frac{2\pi t}{36}$$

$$+ 1.6218 \sin \frac{2\pi t}{21}$$

with $R^2 = .754$

$F(6, 354) = 77.68$

5.3.4 Autoregressive Moving Average Models: Table 5.18 summarizes the results obtained for the parametric models for each pollutant.

5.3.5 Cross-Spectral Analysis: A cross-spectral study was made to study the behaviorival relationship of the pollutants. This involved the study of six pairs of series, viz.,

(i) Temperature and dissolved oxygen

(ii) Biochemical demand and dissolved oxygen

(iii) Temperature and chloride

(iv) Temperature and biochemical oxygen demand

(v) Dissolved oxygen and chloride

(vi) Biochemical oxygen demand and chloride.

Some of the important results of the cross-spectral analysis are presented below:

(a) Temperature and dissolved oxygen:

Figures 5.83 thru 5.86 show the cross correlation, coherency, phase and transfer function for temperature and DO series. The cross-correlation function oscillates around zero lag and shows a high

Table 5.18 Summary of Autoregressive Moving average Models for Great Falls Station.

	Series	Model	Parameters	Variance of raw data	Variance of residuals
(A)	Temperature	ARMA(0,2,2)	$\theta_1 = 1.257$ $\theta_2 = -0.37$	77.58	9.72
		ARMA(1,2,1)	$\phi_1 = -0.20$ $\theta_1 = 0.88$		10.07
		$T_t = -0.086 - 10.56 \cos \frac{2\pi t}{52} - 4.98 \sin \frac{2\pi t}{52} + x_t$			
		x_t: ARMA(2,0,0)	$\phi_1 = .442$ $\phi_2 = .147$		6.86
		ARMA(0,1,1)	$\theta_1 = 0.77$		7.20
		ARMA(1,1,1)	$\phi_1 = 0.318$ $\theta_1 = 0.92$		6.64
(B)	Dissolved oxygen	ARMA(0,1,1)	$\theta_1 = 0.445$	5.36	1.59
		ARMA(0,2,2)	$\theta_1 = 1.43$ $\theta_2 = -0.515$		1.68
(C)	Biochemical oxygen demand	ARMA(1,1,1)	$\phi_1 = 0.063$ $\theta_1 = 0.869$	2.38	1.21
(D)	Chloride	ARMA(0,1,1)	$\theta_1 = 0.20$	43.89	12.7
		ARMA(1,2,1)	$\phi_1 = -0.18$ $\theta_1 = 0.979$		13.2

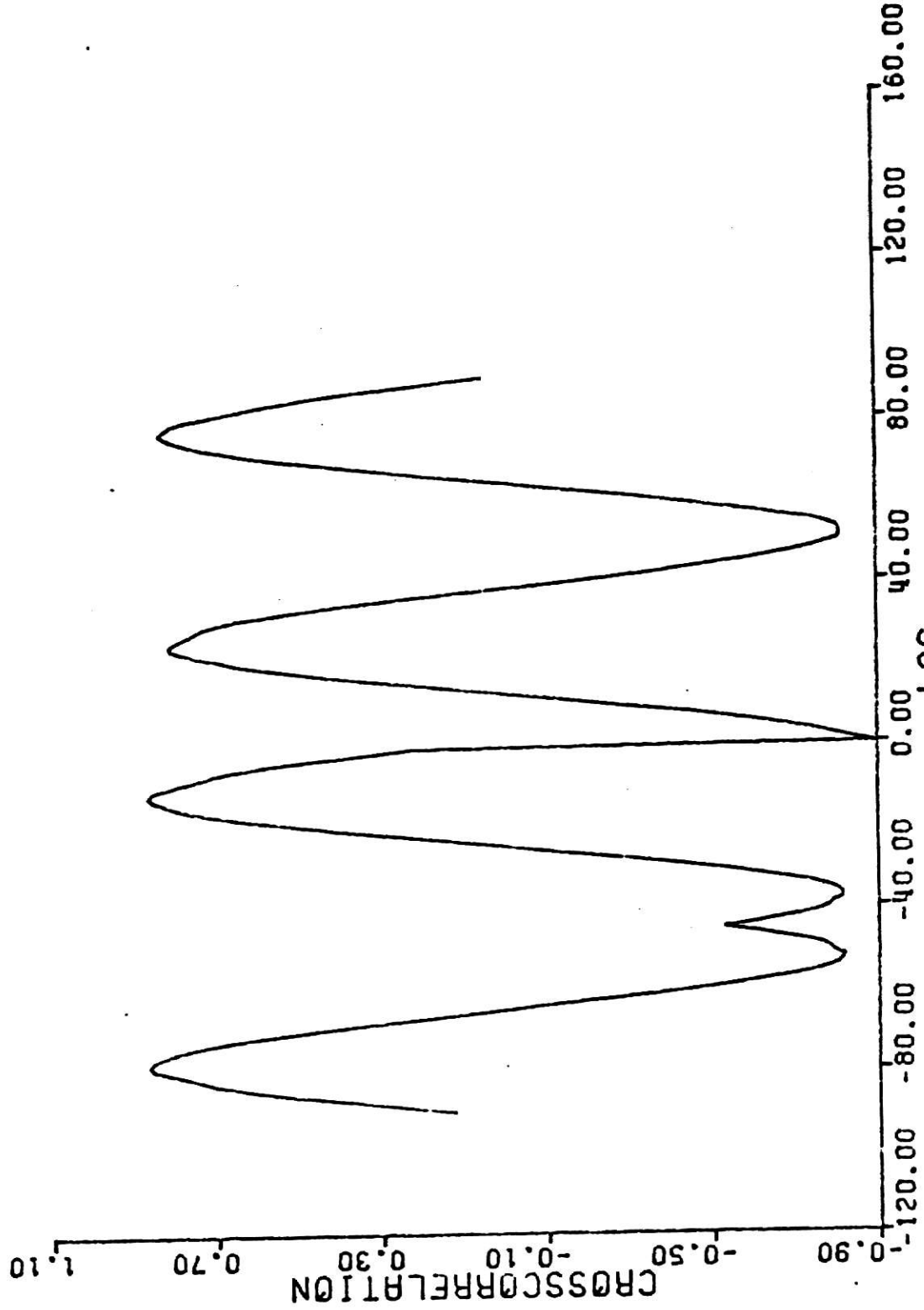

Fig. 5.63 Crosscorrelation of temp. - DO.

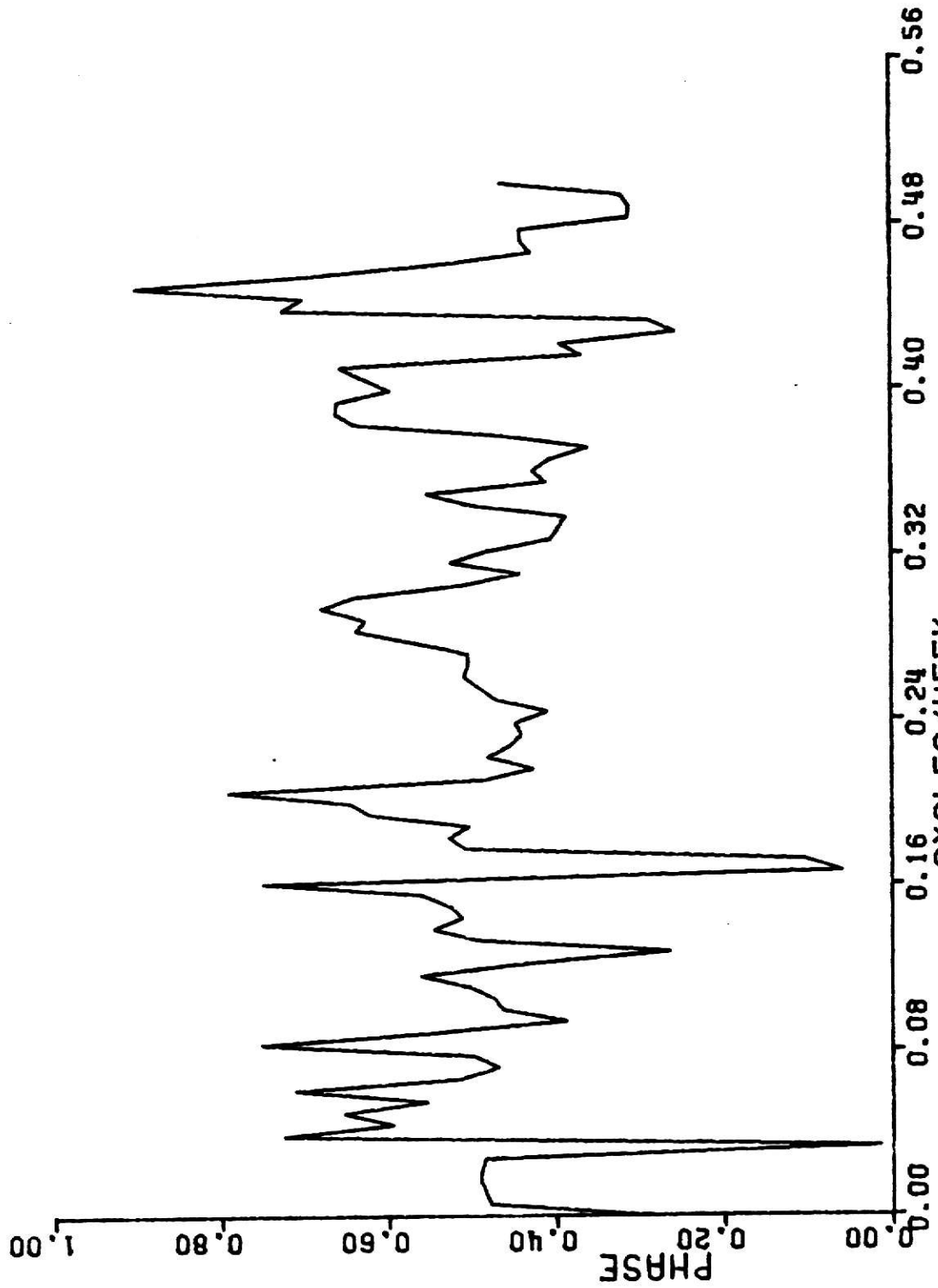

Fig. 5.84 Phase spectra of temp. - DO.

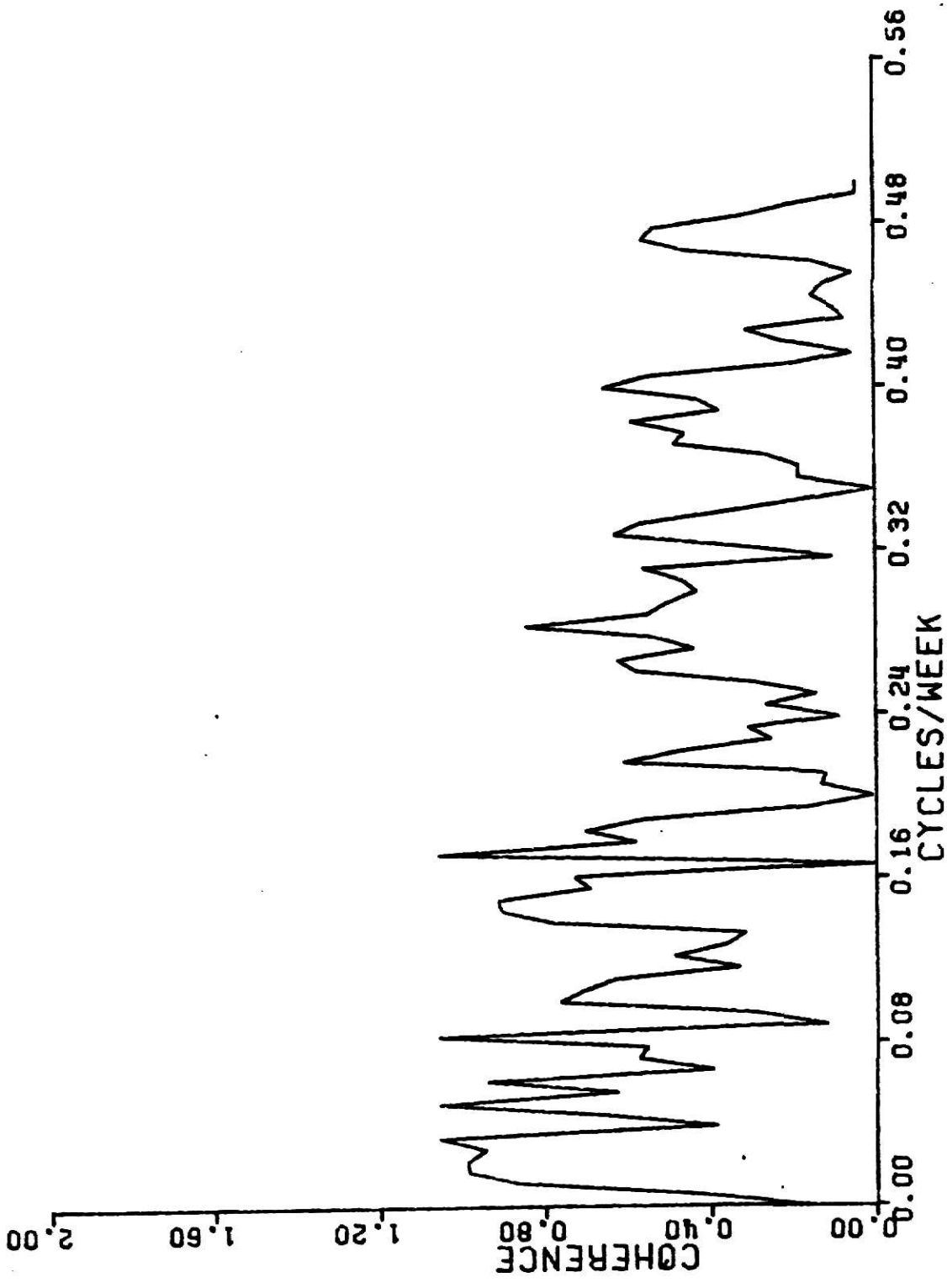

Fig. 5.85 Coherency spectra of temp. - IO.

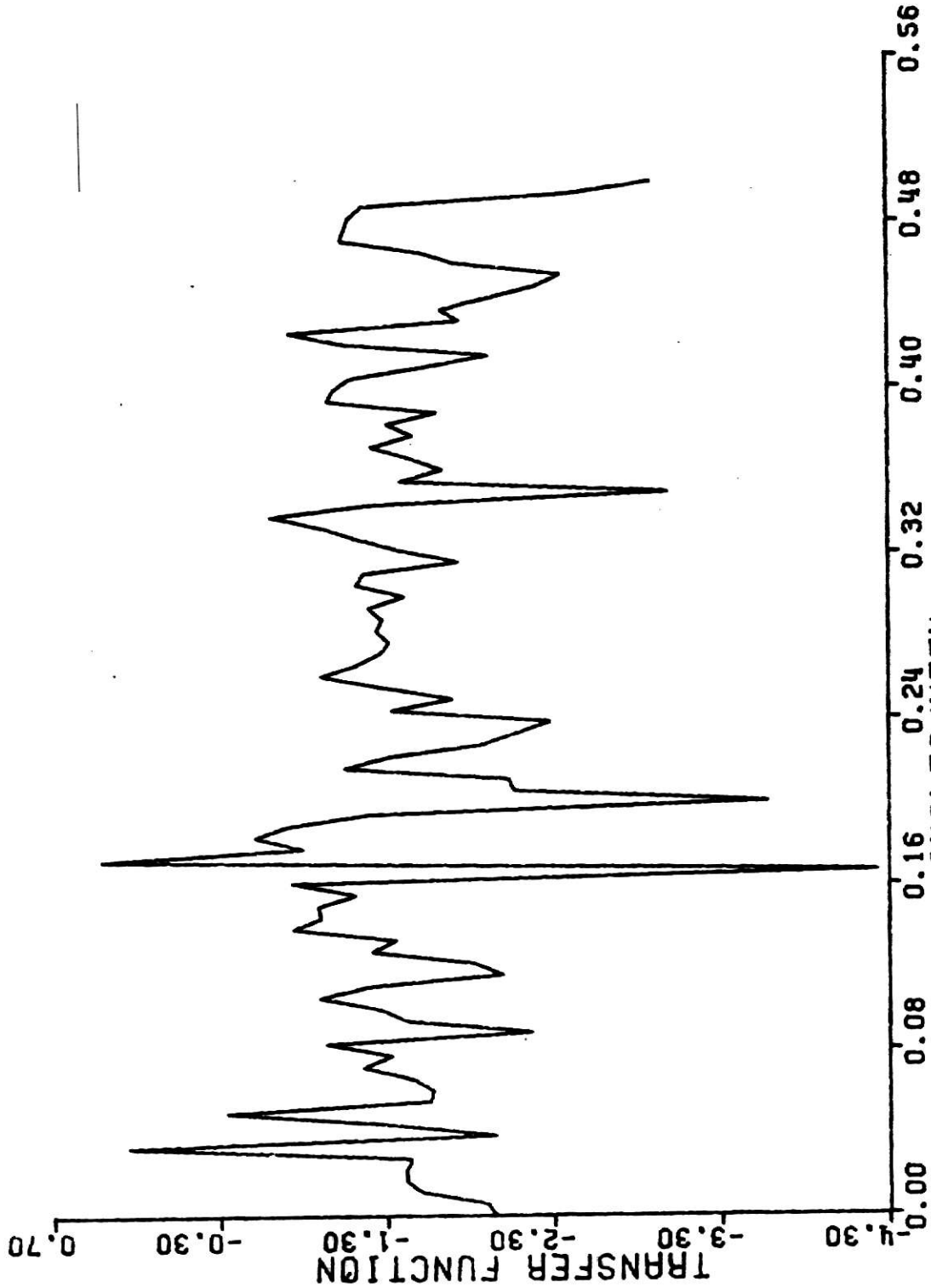

Fig. 5.86 Amplitude of transfer function of temp. - DO.

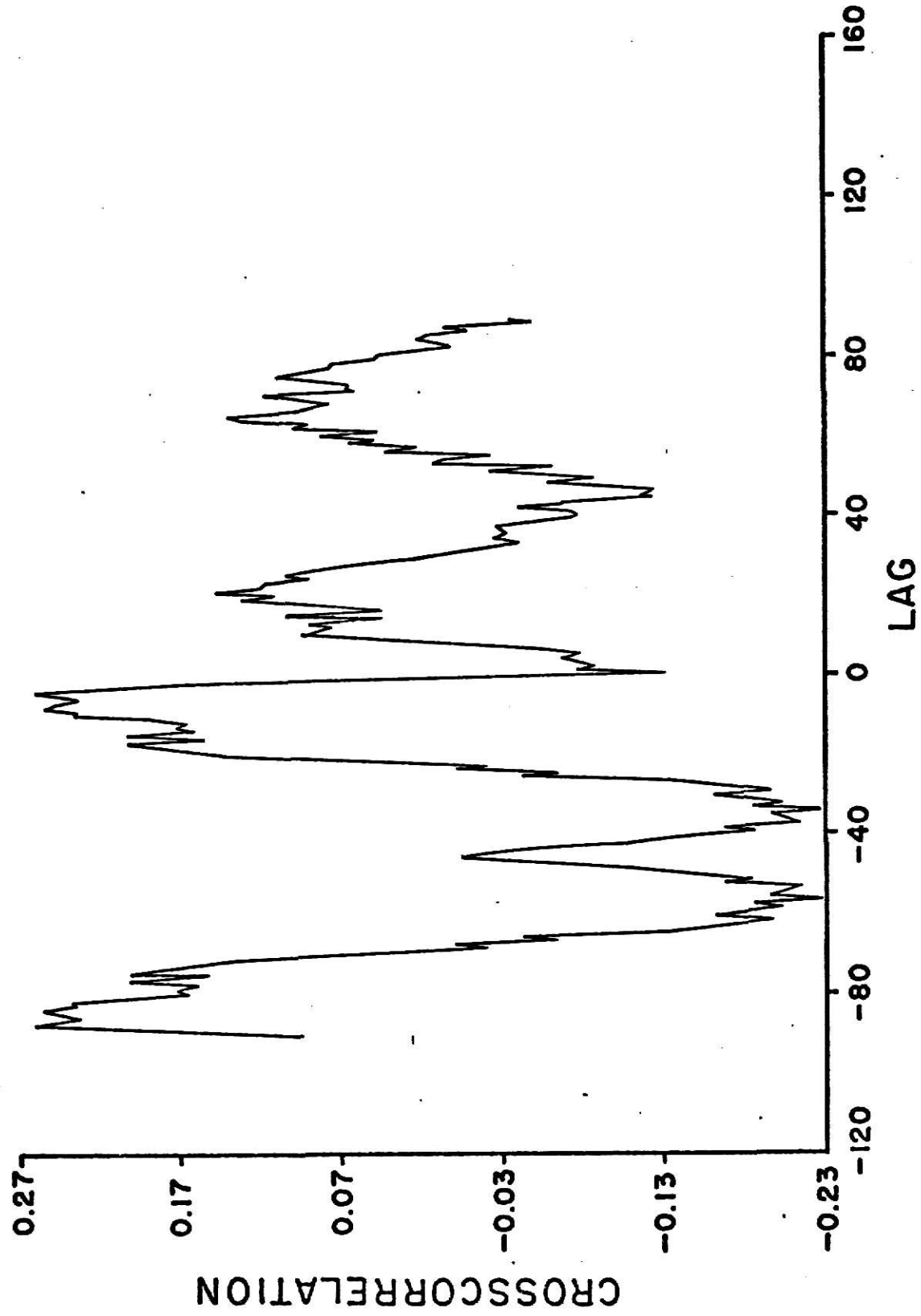

Fig. 5.87 Crosscorrelation of BOD - DO.

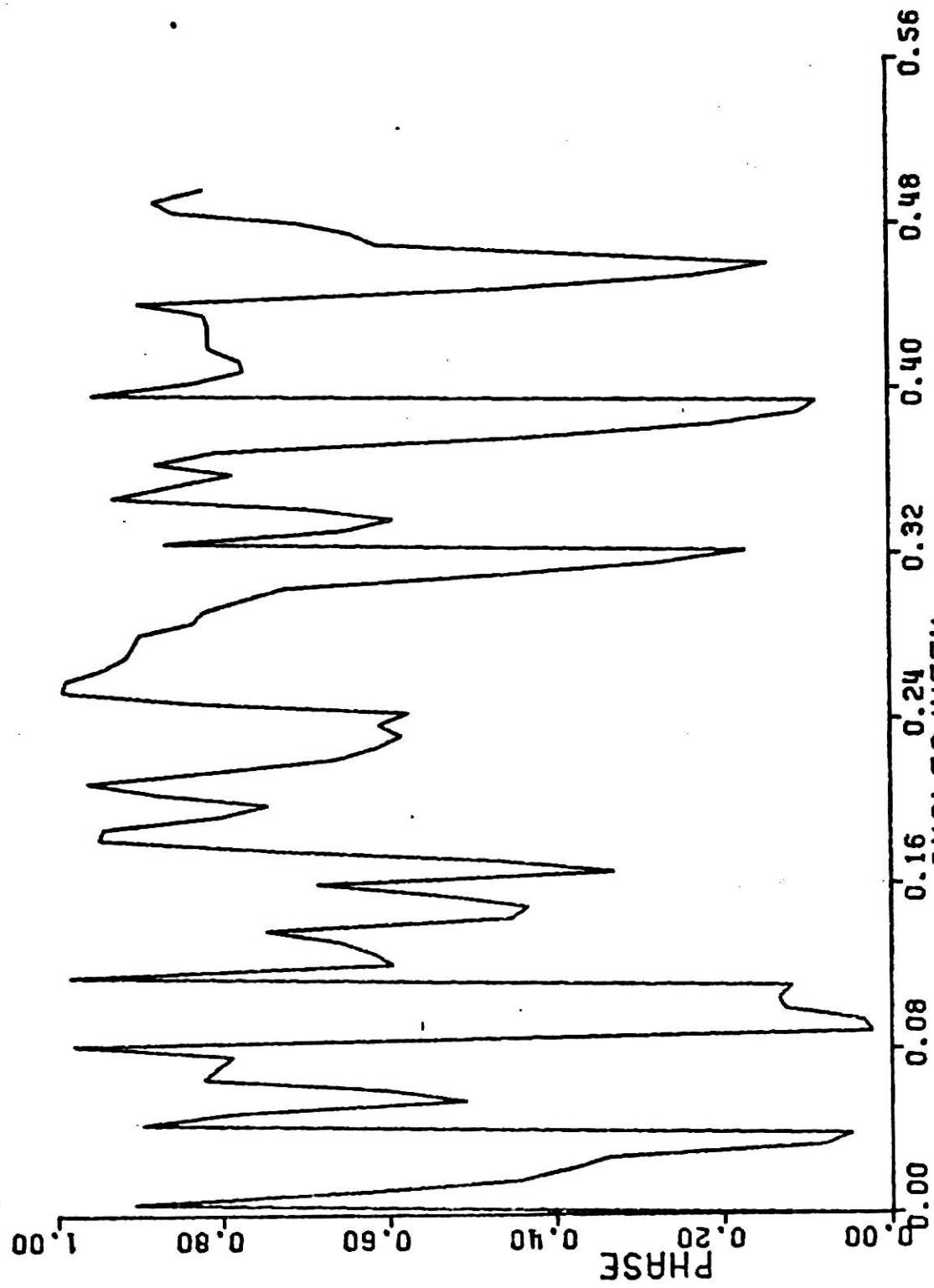

Fig. 5.86 Phase spectra of BOD - DO.

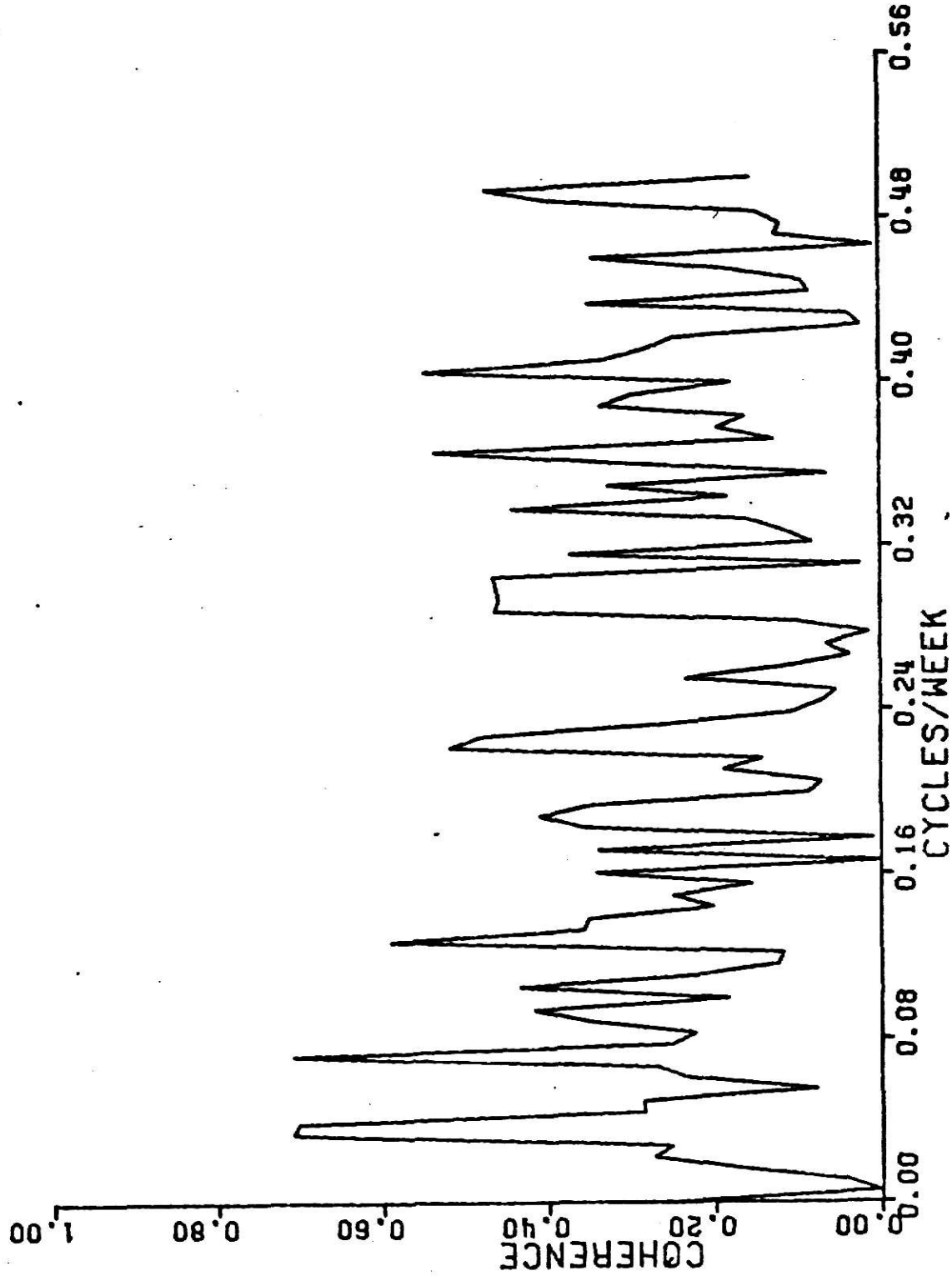

Fig. 5.89 Coherency Spectra of BOD-DO.

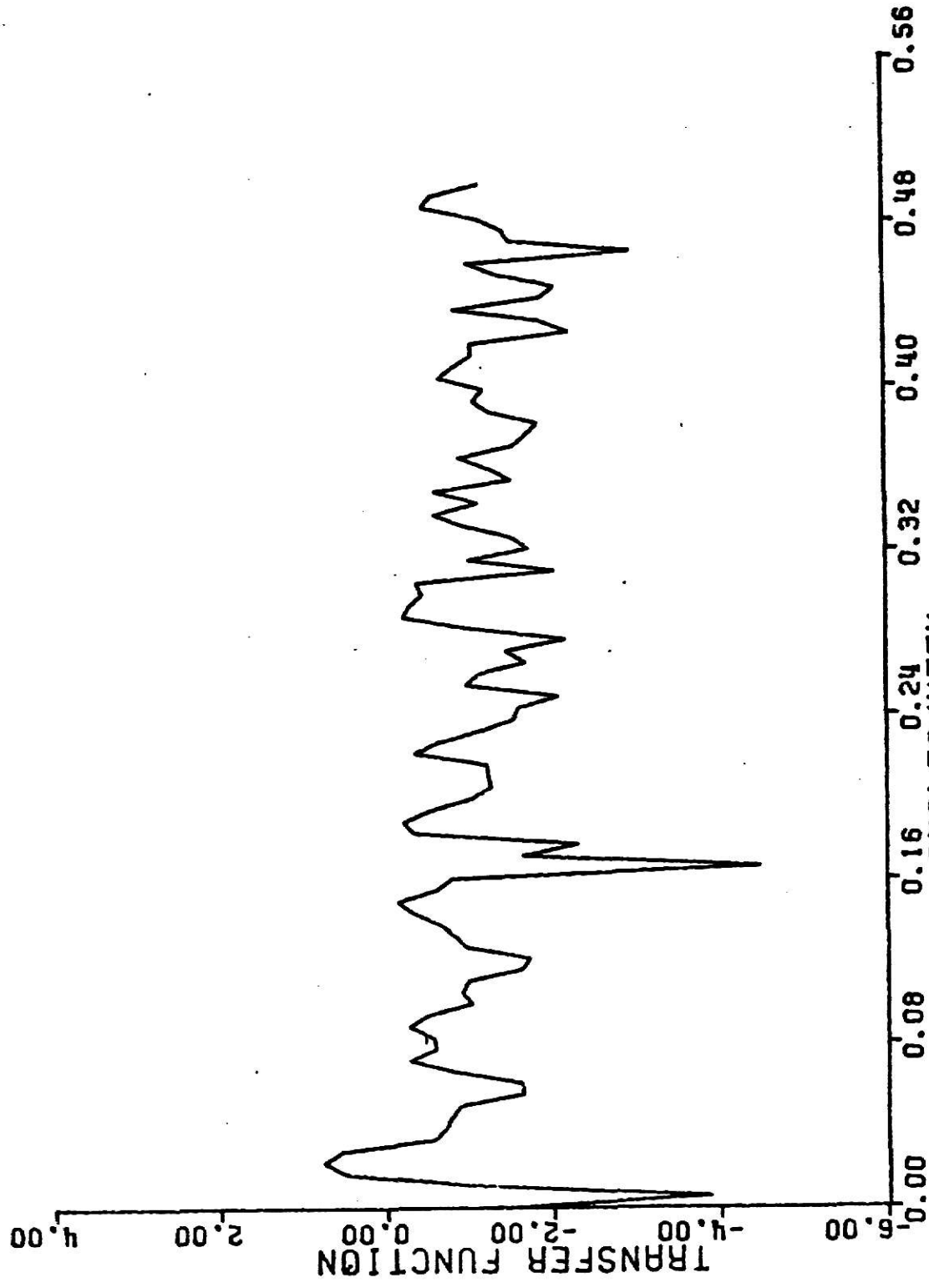

Fig. 5.90 Amplitude of transfer function of BOD - DO.

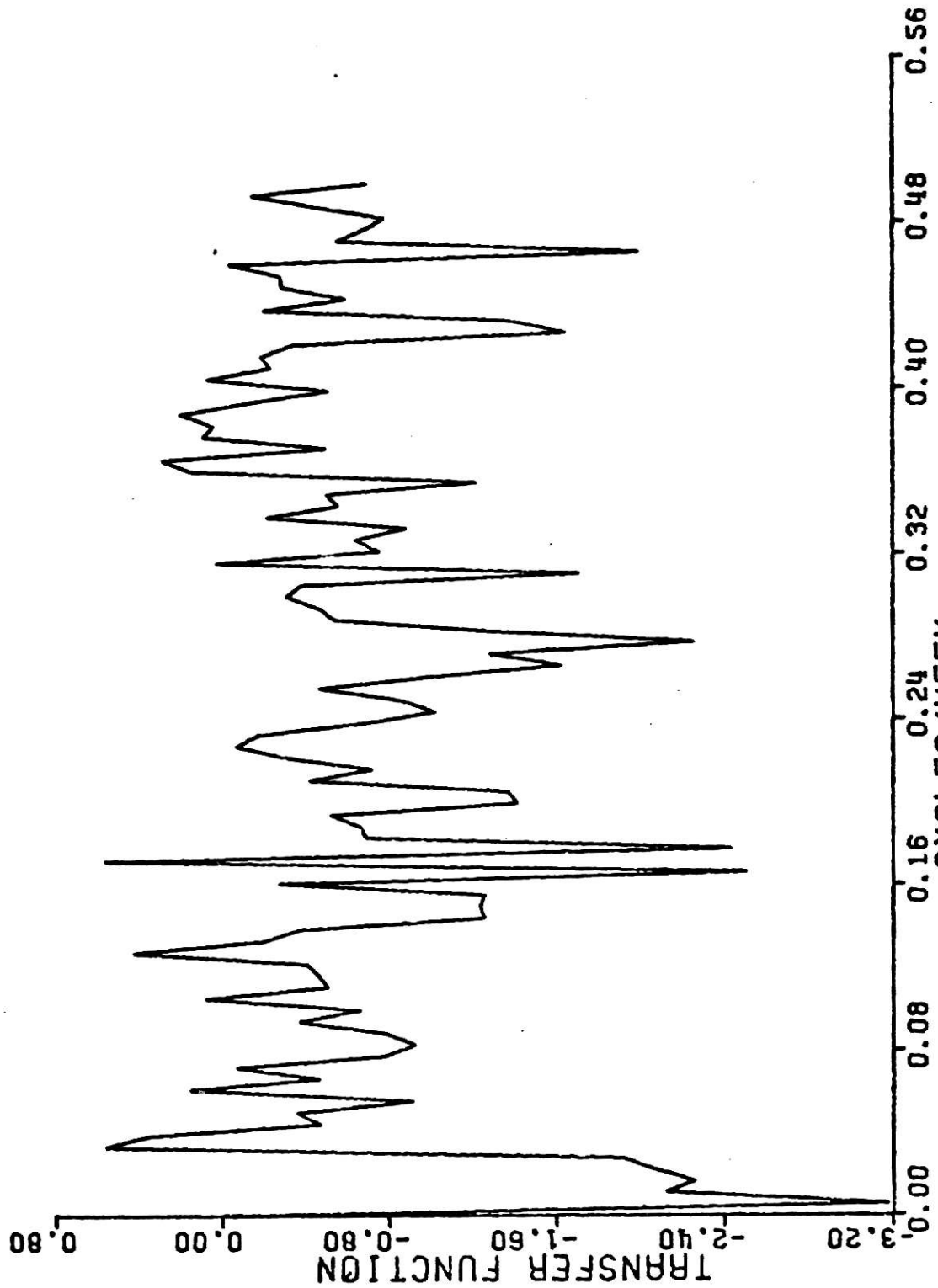

Fig. 5.91 Amplitude of transfer function of DO - BOD.

correlation at 52 lags. High coherency is observed in the frequency bands 0.011 cycles/week - 0.0333 cycles/week, 0.0444 cycles/week - 0.061 cycles/week, 0.1000 cycles/week - 0.111 cycles/week, 0.1389 cycles/week - 0.1611 cycles/week. These correspond to annual fluctuations, semi-annual and other seasonal variations in temperature and dissolved oxygen. The phase difference in these frequency bands is around 180° which is in accordance with the physical relationship between the two pollutants. In general the phase spectra oscillates about 180°. A high transfer function is observed from dissolved oxygen to temperautre.

(b) Biochemical oxygen demand and dissolved oxygen:

The crosscorrelation, coherency, phase and transfer function spectra of BOD and DO are shown in Figures 5.87 to 5.91. The crosscorrelation has a oscillating behaviour. High coherence is observed at 0.038 cycles/week and 0.0722 cycles/week corresponding to six monthly and seasonal changes in BOD and DO. The phase spectra shows a phase difference of about 1 week at these frequencies.

Table 5.19 summarizes the results for the cross-spectral study of this data.

Table 5.19. Summary of Cross-Spectral Analysis at Great Falls Station.

Series 1	Series 2	Conclusions
Temperature	BOD	Low coherency throughout the frequency range
Temperature	Chloride	Low coherency throughout the frequency range
BOD	Chloride	Low coherency throughout the frequency range
DO	Chloride	Low coherency throughout the frequency range

References

1. Streeter, H. W. and E. B. Phelps, "A Study of the Pollution and Nautral Purification of the Ohio River" Public Health Bulletin 146, Washington D.C. (1925)

2. Dobbins, W. E., "BOD and Oxygen Relationships in Streams" Jour. San. Eng. Div., Proc. Amer. Soc. Civil Engr., 90, SA3, 53 (1964).

3. Thoman, R. V., "Mathematical Model for Dissolved Oxygen" Jour. San. Eng. Div., Proc. Amer. Soc. Civil Engr., SA5, 30 (1963).

4. O'Connor, D. J., "The Temporal and Spatial Distribution of DO in Streams" Water Resources Research, 3, 1, 65 (1967).

5. Pence, G. D., J. M. Jeglic and R. V. Thomann, "Time Varying Dissolved Oxygen Model" Jour. San. Eng. Div., Proc. Amer. Soc. Civil Engr., 94, SA2, 381 (1968).

6. Thayer, R. P. and R. G. Krutchkoff, "A Stochastic Model for BOD and DO in Stream" Jour. San. Eng. Div., Proc. Amer. Soc. Civil Engr., SA3, 59 (1967).

7. Custer, S. W. and R. G. Krutchkoff, "A Stochastic Model for BOD and DO in Estuaries" Jour. San. Eng.Div., Proc. Amer. Soc. Civil Engr., 865 (1969).

8. Thomann, R. V., "Time Series Analysis of Water Quality Data" Jour. San. Eng. Div., Amer. Soc. Civil Engr.,SA1, 1 (1967).

9. Wastler, T. A. and C. M. Walter, "A Statistical Approach to Estuarine Behaviour" Jour. San. Eng. Div., Amer. Soc. Civil Engr., SA6, 1175 (1968).

10. Matalas, N. C., "Time Series Analysis" Water Resources Research, 3, 3, 817 (1967).

11. Julian, P. R., "Variance Spectrum Analysis" Water Resources Research, 3, 3, 831 (1967).

12. Thomann, R. V., "Variability of Waste Treatment Plant Performance" Jour. San. Eng. Div., Amer. Soc. Civil Engr., SA3, 819 (1970).

13. McMichael, F. C. et al, "Stochastic Modeling of Temperature and Flow in Rivers" Water Resources Research, 8, 1, 87 (1972).

14. Wallace, A. T. and D. M. Zollman, "Characterization of Time Varying Organic Loads" Jour. San. Eng. Div., Amer. Soc. Civil Engr., SA3, 257 (1971).

15. DeMayo, A. "The Computation and Interpretation of the Power Spectra of Water Quality Data" Canadian Water Quality Bulletin.

16. O'Connor, D. J., "Oxygen Balance of an Estuary" Jour. San. Engr. Div., Amer. Soc. Civil Engr., SA3, 35 (1960).

17. Li, W. H., "Unsteady Dissolved Oxygen Sag in a Stream" Jour. San. Eng. Div. Amer. Soc. Civil Engr., SA3, 75 (1962).

18. Thirty First Progress Report Committee on Sanitary Engg. Research." Effect of Water Temperature on Stream Aeration." SA6, 59 (1961)

19. O'Connor, D. J. and W. E. Dobbins," The Mechanism of Reaeration in National Streams." Jour. San. Eng. Div., Amer. Soc. Civil Engr., 123, 655 (1958).

20. Churchill, M. A., H. L. Elmore and R. A. Bukinghum, "The Prediction of Stream Aeration Rates." Jour. San. Eng. Div., Amer. Soc. Civil Engr., SA4, 1 (1962).

21. Thomann, R. V. and M. J. Sobel, "Estuarine Water Quality Management and Forecasting." Jour. San. Eng. Div., Amer. Soc. Civil Engr., SA5, 9 (1964).

22. Gunnerson, C. A., "Optimizing Sampling Intervals in an Estuary." Jour. San. Eng. Div., Amer. Soc. Civil Engr., SA2, 103 (1966).

23. Ignago R. and C. F. Nordin, "Some Applications of Cross-Spectral Analysis in Hydrology; Rainfall and Runoff." Water Resources Research, 5, 3, 608 (1969).

24. Yu, L. S. and W. Brutsaert, "Stochastic Aspects of Lake Ontario Evaporation." Water Resources Research, 5, 6, 1256 (1969).

25. Ward, J. C., "Annual Variation of Stream Water Temperature." Jour. San. Eng. Div., Amer. Soc. Civil Engr., 1 (1963).

26. Thomann R. V., "Recent Results for Mathematical Model of Water Pollution Control in Delaware." Water Resources Research, 3, 1, 349 (1965).

27. Jones, R. H., "Spectrum Estimation with Missing Observations." Annals of the Institute of Statistical Maths., 23, 3, 387 (1971).

28. Thomann, R. V., "Systems Analysis and Water Quality Management." Environmental Research and Applications, Inc., New York, N.Y. (1972).

29. Thomann, R. V., D. J. O'Connor and D. M. Ditoro, "The Management of Time Variable Stream and Estuerine Systems." Chemical Engineering Progress Symposium (1968).

30. Anderson, P. W. and J. S. Zogorski, "Analysis of Long Term Trends in Water Quality Parameters." Proceedings of the Fourth American Water Resources Conference at Urbana, Ill. (1968).

31. Tirabassi, M. A., "A Statistically Based Mathematical Water Quality Model for a Non-Estuarine River System." Water Resources Bulletin, 6, 7, 1221 (1971).

32. Kothandaraman, Veeraswauny, "Analysis of Water Temperature Variations in Large Rivers." Jour. San. Eng. Div., Amer. Soc. Civil Engr., 97, 19 (1971).

33. Jenkins, G. M. and D. G. Watts, "Spectral Analysis and its Applications." Holden Day, San Francisco (1968).

34. Granger, C. W. J. and M. Hatanka, "Spectral Analysis of Economic Time Series." Princeton University Press, Princeton (1964).

35. Fishman, G. S., "Spectral Methods in Econometrics." The Rand Corporation, California (1968).

36. Blackman, R. B. and J. W. Tuckey, "The Measurement of Power Spectra" Dover Publications Inc. N. Y. (1958).

37. Wastler, T. A., "Spectral Analysis." U.S. Department of Interior, Washington (1969).

38. Draper, N. R. and H. Smith, "Applied Regression Analysis." John Wiley and Sons Inc., N. Y. (1966).

39. Box, G. E. P. and G. M. Jenkins, "Time Series Analysis, Forecasting and Control." Holden Day, San Francisco (1970).

40. Stralkowski, C. M., "Low Order Autoregressive - Moving Average Stochastic Models and their use for the characterization of Abrasive Cutting tools" Ph.D. Thesis, University of Wisconsin, (1968).

41. "Proceedings in the Matter of Pollution of the Interstate Waters of the Potomac River." U.S. Department of Interior (1969).

42. Feigner, K. D., "Potomac River Data". Environmental Protection Agency, Seatle, Washington. Private communication.

43. Environmental Protection Agency. "Stochastic Modeling for Water Quality Management". Project No. 16090 DUH.

TIME SERIES ANALYSIS OF WATER QUALITY DATA

by

NAVIN KUMAR BHARGAVA

B.Sc.(Engg.), University of Delhi, Delhi, India, 1969

AN ABSTRACT OF A MASTER'S REPORT

submitted in partial fulfillment of the

requirements for the degree

MASTER OF SCIENCE

Department of Industrial Engineering

KANSAS STATE UNIVERSITY

Manhattan, Kansas

1974

ABSTRACT

As water flows down a river basin, it undergoes many physical, chemical and biological changes thereby varying its quality characteristics. In this report mathematical models have been formulated using Time Series Analysis to study the behaviour of various water quality indicators such as temperature, dissolved oxygen, BOD etc. for some stations on Ontario River and Potomac River.

Water pollution data have been known to exhibit cyclic variations due to several factors like seasonal changes in surrounding weather, photo synthesis, semidurinal tides etc. Spectral analysis was applied to determine these cyclic variations. Using the information thus obtained, regression analysis was performed to obtain the prediction model for each pollutant.

The interaction of one pollutant with another pollutant at the same station and the relationship of the same pollutant at different stations was studied using cross-spectral analysis. Coherency spectrum was drawn for each case.

Another approach to model the time series data is through the use of Autoregressive Moving Average methods. In this study, autocorrelation function of each series was used to identify the models. The values of the parameters of the suggested models were estimated and diagnostic checks were made to ascertain the appropriateness of each model.