Project 361
R + D Project Record
361/4/NW

Analysis, Storage and Archiving of Water Quality Data

Cremer and Warner Ltd.

R&D Project Record 361/4/NW

NRA

National Rivers Authority

Dissemination and Uptake Note

This note provides details of how the attached output from the R&D programme is to be disseminated to the end-user and details of how the customer wishes the output to be taken up.

1. Project

Commission: WATER QUALITY

Topic Area: A15/A10 Instrumentation

Project Number: AK 361

Project Leader: T.C. GASKELL

Region: NW

2. Output

Title: Analysis, Storage and Archiving of Water Quality Data.

Reference: Project Record 361/4/NW

Output Type[1]:

Distribution Instructions:

INSTRUMENTATION GROUP LEADER
TOPIC LEADER
COMMISSIONER

Dissemination Status — Internal: RELEASED TO REGIONS — External: PUBLIC DOMAIN

3. Uptake

Responsibility:

Details (should include any seminars, training sessions or Working Group considerations as necessary)

The document outlines options for automatic data validation and incident detection. It highlights the inherent variability of water quality data and recommends areas for further study.

4. Core Function/Customer Authorization

Project Leader:

Topic Leader:

Commissioner:

Group Chairman[2]:

Notes:
[1] Reference to paper R&D(92)3 Definition of R&D Outputs.
[2] Signature of Chairman of Working Group if appropriate.

Date Output Sent: / /

Analysis, Storage and Archiving of Water Quality Data

Cremer & Warner Ltd
17 Angel Gate
City Road
LONDON
EC1V 2PT

Project Record 361/4/NW

National Rivers Authority
Rivers House
Waterside Drive
Almondsbury
BRISTOL
BS12 4UD

Tel: 0454 624400
Fax: 0454 624409

Dissemination status
Internal: Released to regions
External: Released to public domain

Research Contractor
This document was produced under R&D Contract 361 by:

Cremer and Warner Ltd
17 Angel Gate
City Road
London
EC1V 2PT

Tel: 071 278 8338
Fax: 071 833 9090

NRA Project Leader
The NRA's Project Leader for R&D Contract 361
T C Gaskell - North West Region

Additional Copies
Further copies of this document may be obtained from Regional R&D Coordinators or the R&D Section NRA Head Office.

Project Record 361/4/NW

EXECUTIVE SUMMARY

In Phase I, a questionnaire was designed and despatched to six NRA regions. This was followed up by interviews with appropriate officers in the regions to gather supplementary information and seek clarification of some of the answers number of key requirements for the handling, alarming, analysis, storage and archiving of water quality data. These requirements fell into three broad areas:

- data monitoring
- incident detection
- water quality planning

In review of practices in other industries and application areas revealed that, often, a more sophisticated approach to data management is adopted compared with current techniques used in the NRA. These techniques made fuller use of the data by correlation or comparison with other data and the use of rate of charge alarms and two or more levels of alarm threshold. Alternatively, in the process industry, a voting system is often used, in which three probes monitor the same parameter. This allows cross checking between probes for the ready detection of a fault probe and a reduction in the loss of data. Techniques used in other application areas are rule and knowledge based methods, pattern recognition techniques and neural systems. Analysis of the data exploited statistical packages, mathematical tool boxes and Geographical Information Systems.

Based on this review, a number of recommendations are presented. A more sophisticated approach to incident detection and data validation is recommended on the lines described above. The main objective is to reduce the incidence of false alarms. The recommendations for assisting the water quality planning function concentrate on the use of the water quality data in planning and real time models.

Consideration has been given to the method of transmitting data from the monitoring station. The costs and benefits of detecting the alarm at a central facility or at the monitoring station need further examination. Storage technology does not impose any limits on the way water quality data should be stored. A suitable quality assurance system should be devised.

limits on the way water quality data should be stored. A suitable quality assurance system should be devised.

Phase II was designed to test out a number of techniques on automatically monitored water quality data. Initially, exploratory data analysis was used to gain an insight into the nature and quality of data. Hence, a large number of time history graphs were generated which allowed visual cross-comparison of the different parameters to understand their behaviour under different operating and water quality conditions.

Spike detection algorithms were developed to remove calibration and cleaning cycle data.

Kalman filtering techniques were used to remove calibration spikes, and also diurnal variation in data.

It was clear that on many occasions, false data could be removed by comparison against a threshold value or from changes in the mean and standard deviation. However, on other occasions false data was less easily discernible from genuine pollution incidents.

The correlation of parameters was investigated using regression analysis and two-parameter plots during periods when data remained at background levels i.e when there were no unusual incidents evident.

This analysis revealed strong correlations between dissolved oxygen, pH, conductivity and temperature. From the analysis conducted with the two parameter plots it appears that pollution incidents may be identified and distinguished from natural phenomena such as storm events which do not carry man made pollution into the river system.

Two-parameter plots which compare incident data with historical stable or background data are a simple form of pattern recognition for the detection of outliers. The analysis provided limited success in recognising more subtle events. An initial exploration of principal component analysis was unsuccessful.

The report concludes that single parameter events can be detected with high reliability and low incidence of false alarms, using suitable software. Outline specifications for such software are provided.

KEY WORDS

Archiving, automatic, monitoring, pollution, rivers, water quality.

LIST OF TABLES

LIST OF FIGURES

After Page

APPENDICES

1. INTRODUCTION

Cremer & Warner (C&W), part of Simon Environmental, were commissioned by National Rivers Authority (NRA) to carry out a research project on Data Communication, Processing and Storage of automatically recorded water quality monitoring data. In carrying out the research, C&W were supported by BA^{e}SEMA as sub-contractors. The work was completed by Simon Hydrotechnica Ltd, another component part of Simon Environmental.

The work was reviewed by Dr Gordon Woo, an independent mathematical consultant. After this review, the section on Kalman filtering was included. This section benefitted greatly from discussion with Professor Peter Young, from the Centre for Research on Environmental Systems, University of Lancaster.

The overall objective of the project is to establish and specify widely applicable methods for receiving, processing and archiving data from automatic water quality monitors. More specifically the objectives are as follows:

1. Review the current techniques used in the NRA and other industries.

2. Evaluate the scope of statistical variability in automated water quality monitor data and characterise the sources of this variability.

3. Find a range of summary statistics to optimise output of routine information at the same time as storing data in as compact a form as practicable.

4. Identify a method of recording water quality events that is compatible with objective number 2.

5. Develop statistically valid methods for accommodating missing values in summaries.

6. Analyse data to find best technique for raising alarms during operation of monitors. Emphasis should be placed on dynamic and multivariate alarms.

7. Provide recommendations for best present practice/techniques and changes to current NRA practices and guidelines for future R&D research.

The project has been divided into three phases:

Phase I Review of Current Practice for Data Storage, Summarising and Alarm States. Research under Phase I has been divided into two distinct tasks:

1. A review of current practices within the different NRA regions.
2. A review of practices in other industries.

Phase II Evaluation of Identified Techniques for Data Storage, Summarising and Alarm States within the context of the NRA.

Phase III Reporting including recommendations for best present practice and techniques, changes to current NRA methodologies and an outline of future R&D requirements.

Section 2 examines current practice within six of the ten NRA regions and identifies the key requirements for data management whilst Section 3 of the Report reviews practices in other industries. Section 4 puts forward preliminary recommendations based on the reviews conducted and described in Sections 2 and 3. Section 5 assesses the recommended techniques for data management by application to automatically monitored water quality data sets collected from sites in the Severn Trent region.

Conclusions and recommendations are presented in Section 6.

Appendix A presents the questionnaire used. Graphical plots of data used are given in Appendices B and C. Appendix D presents a review of the report made by Dr. Woo.

2. NRA REQUIREMENTS

2.1 Introduction

Practices for data handling, storage, analysis and alarm setting vary from region to region within the NRA. This is mainly because the systems were inherited from the regional water authorities when the NRA was created. New developments have taken place since them, but always on a regional basis, responding to local circumstances and requirements. A sample of six regions were selected for a review of their practices; Severn Trent, Thames, Anglian, Wessex, Southern and North West. The review was restricted to six regions due to time and budget constraints. A questionnaire was designed to obtain information on current methods of acquisition, processing, storing, archiving and use of data in the regions. Views were sought on the current system and likely changes in the system to meet future requirements. A copy of the questionnaire used in the survey is given in Appendix A. The questionnaires were sent out to each of the six selected regions and followed up by a visit to gather supplementary information and clarify the answers given in the questionnaire. In particular, an important part of the visit was to understand how the data was ultimately used and whether the systems fulfilled the user requirements. The following sub-sections summarise the findings from the questionnaire and interview survey, in terms of data management requirements.

2.2 Kev Uses

The use of the automatic quality data is not consistent between the regions. However, five key areas of use may be identified:

1. Incident detection in rivers
2. Water quality data monitoring
3. Water quality planning
4. Special projects
5. Provision of information

2.2.1 Incident Detection

Monitors are used, amongst other information sources, to alert the NRA to pollution incidents within rivers and estuaries. This is currently achieved by simple threshold level alarms on each of the parameters monitored. When the measured levels exceed or fall below the threshold level the operator takes action accordingly. Once preliminary checks have been undertaken to ensure the alarm is valid, further monitoring is initiated either by taking water samples, manually, for laboratory analysis or placing mobile monitors upstream and downstream of the suspected point of pollutant discharge. In addition, an alarm condition may also trigger the automatic taking of samples for subsequent laboratory analysis. Samples for laboratory analysis are taken as evidence for prosecuting the offender. The use of automatically monitored water quality data for a prosecution has not yet been tested in the courts.

The NRA may also inform the water abstractors of pollution incidents to prevent the contaminated water reaching potable or process water supplies. There is no legal obligation to inform abstractors as they monitor their own intakes and have responsibility for the quality of the water supplied.

The NRA are also able to take remedial action to alleviate the impacts of a pollution incident such as increasing dilution or aerating the water to increase the dissolved oxygen.

2.2.2 Data Monitoring

Water quality data is monitored, on a computer screen, by the operator for instrument faults, general validation of the data and for the changes in water quality. A short time period of the order of a few days or less is generally examined. This occasion is often used to manually clean-up the data and identify and remove unwanted features, such as calibration cycles and data from faulty instruments. When a faulty instrument is suspected, engineers are deployed to examine and repair or replace the instrument. There is no standard procedure for cleaning-up and treating the water quality data.

2.2.3 Water Quality Planning

The long term water quality data is used to monitor changes in the quality of rivers and estuaries over time. This allows rivers to be classified and environmental quality objectives (EQO) set for different parts of the river system. The automatic data is also used as input to computer models which can simulate pollution loads under different environmental conditions and for different levels of discharge into the river system. This assists the water quality planner to set consent conditions which progress towards meeting the EQO. Flow and rainfall data are also used in modelling and for water quality planning to understand and control pollution loads under the differing flow conditions.

2.2.4 Special Projects

Mobile water quality monitors are often used to investigate particular discharges or sections of a river which are suffering from unacceptable quality due to an unknown source of pollution. Data is monitored in periods ranging from one week to a few months depending on the circumstances. The data are analysed in detail in order to understand the nature of the problem, frequency and time of occurrence of incidents, type and point of discharge etc. The objective is to identify and stop the discharge of pollutants, prosecuting where necessary. This is a long term analysis of data similar to water quality planning, but clearly with different final objectives.

2.2.5 Provision of Information

Under the Water Resources Act, 1991, the NRA makes water quality data available to researchers, environmental organisations or other interested parties through the Water Quality registers. The automatic quality data potentially provides a valuable source of such data. The form of the data provided will vary with the recipient and the purpose for which it is required.

2.2.6 Summary

Examination of the uses to which the data are put shows that these fall into two kinds: incident detection and monitoring, and various long term requirements.

Incident detection and monitoring require timely and reliable data - accuracy is a secondary concern, since (at least at the moment) automatic measurements will be supplemented by additional samples. It is important that all incidents are detected, and that the number of false alarms is minimised. The data set used will be incomplete, since data from future time are not available.

For longer term uses such as reporting statistics, and planning, accuracy is much more important. Delays in processing of a few hours, or even days, are less important. The data set is complete for each period analysed.

These different priorities need to be incorporated in the data handling strategy.

2.3 Requirements

This sub-section examines the NRA data requirements under each of the key uses of that data. Again, this information has been derived from the interviews with the NRA regions. Once the requirements have been defined, a specification or recommendation for data handling, analysis and storage can be presented.

2.3.1 Incident Detection

The system for detecting incidents and raising the alarm needs to be precise, accurate and reliable to reduce the incidence of false alarms. This requirement does not necessarily imply that the water quality monitors and the communication of the monitored data to user needs to be precise, accurate and reliable or that the quality of the raw data is high. The dissolved oxygen and ammonia data drift out of calibration and require regular re-calibration either automatically or manually. Drift between calibrations can be as high as ±10-15%. Hence, the alarm detection mechanism needs to be capable of assimilating this type of data and keeping the incidence of false alarms to a minimum. There is a clear need for an alarm detection method which is more sophisticated that the simple threshold currently used, in order to detect alarms under different environmental conditions. Once an alarm condition has been detected the system is required to communicate that information to the user immediately so that appropriate action can be taken. Also, generally during alarm conditions, data will need to be monitored more frequently than for other uses of the data in order to monitor how an incident is progressing.

The user will wish to have the facility to examine and edit the data in a similar manner to that described under "Data Monitoring" (Section 2.2.2). In addition, particular users will need to be able to alter, manually, the parameters for alarm detection and the frequency of monitoring.

2.3.2 Data Validation and Interpretation

Validation and Interpretation

The operator will wish to have the facility to manually screen and edit the most recent water quality data covering relatively short time periods e.g. three consecutive days of data. The operator will also wish to edit erroneous data and remove unwanted features in the data e.g. calibration cycles. Visual screening of the data will require the ability to display the data in graphical and tabular form. The operator will also need to be able to identify when monitors or the communication system are no longer functioning. These are also important requirements when the data are being used under alarm conditions.

Access

The operator will require ready access to the water quality data for screening and analysis. Multi-user access to the validated data is generally required in order that different areas within the NRA region can examine the data directly and take action locally. Access to historic data may be necessary for the operator to verify an alarm condition.

2.3.3 Water Quality Planning

Water quality planners need to examine data over long time periods in order to observe trends in water quality. This may require, for example, comparing a full year's data with that of a previous year so that seasonal variations may be examined. Flow and rainfall have an influence on water quality and the user will therefore require ready access to hydrometric data to support the analysis. Again the user will need to display the information in graphical form for visual comparison of different parameter, time period and outstation data sets. The water quality planner will also need to interface the data with the analysis and predictive tools at his/her disposal e.g. trend analysis, time-series analysis, simulation models.

Clearly the user will require ready access to the longer term data. Also, it is important that the data the user requires have unique identifying parameters so that consistent data sets are retrieved for comparison. Hence, a system which allows traceability and identification of files in storage and archive is required. This would be a necessary part of any quality assurance system for the database.

2.3.4 Special Projects

The user requirements for handling and analysis data from special projects is dependent upon the nature of the investigation. In general, the requirements are the same as for data monitoring and water quality planning given above and there are no unique requirements for special projects data.

2.4 Summarv of Interview Findings

The interviews with the six NRA regions visited, revealed a number of interesting differences in the way which the data are monitored, transmitted to a central computer, analysed, stored and archived. Nevertheless, there was also a lot of commonality between some regions. This report does not give a detailed account of our interviews with the various regions as this would add little benefit to the reader. However, Table 2.1 summarises the key facts from our interviews and serves to highlight the differences and commonality between regions. The headings follow the layout of the questionnaire sent to the regions. Most regions monitored six water quality parameters using fixed monitoring stations with mobile monitors for special investigations or incident investigations. The quality of the data from the dissolved oxygen and ammonia monitors is poor and require frequent calibration. Generally, data validation is carried out manually with the user displaying the parameter values graphically against time. Through inspection the data is edited by the user to remove faulty or unfeasible data, calibration and cleaning cycles. The capability to produce multi-parameter, multi-location and multi-period plots is not widely available.

Where it is used, alarm detection relies upon a single threshold value. The user or operator is alerted to the alarm condition through the visual display and takes action according to a set of written rules. At this stage the operator does not validate the alarm.

Simple statistical techniques are used to analyse the water quality data, usually means, minimum, maximum and standard deviation. Commonly, data tables and X-Y plots are used to examine and analyse the data. Hydrometric data are also used in the analysis of water quality data.

There are a variety of methods used for storing and archiving data. Data are usually stored on hard disc on either IBM PC compatibles or a centralised computer. Archiving is usually to magnetic tape or diskette or both.

In general, most regions believed that when planned expansions had been implemented, there would be sufficient monitoring stations in their region. The implementation programme may be limited to financial resources. The parameters monitored allow for detection of organic pollutants which was considered sufficient for predominantly agricultural regions with little heavy industry. However, if the technology were available at reasonable cost the monitoring of metals and red list substances would be desirable for specific investigations. A number of new data analysis, storage and archiving systems are currently being installed or are planned. These are all a great improvement on the previous systems. For example, the ability to produce multi-parameter plots to identify, visually, parameter correlations; the use of a two-level threshold alarm; a graphical editor. Nevertheless, as might be expected, they contain a number of very useful features but have inadequacies in other areas. These inadequacies were often recognised.

2.5 Summarv of Reguirements

The information gathered through interviewing the selected six NRA regions revealed a number of key uses and user-requirements for the handling, alarming, analysis, storage and archiving of water quality data. For ease of reference, the key user-requirements are summarised below:

1. data monitoring

   - ready access
   - short term data
   - screening and editing of data
   - visual and graphical analysis tools
   - multi-user access

2. incident detection

   - good quality data which has been validated
   - an alarm detection system which minimises false alarms
   - the means to raise the alarm immediately

3. water quality planning

   - long term data
   - access to hydrometric data
   - trend and predictive analysis tools
   - simple interface with computer models

4. other

   A system needs to be devised which will ensure the traceability and identification of files stored and archived.

Table 2.1 - Summary of Interview Findings

	Severn Trent	Thames	Anglian	North West	Wessex	Southern
Number of Monitors*	15F, 5M	25F, 3M	12F, 2M + 10	6F	12M	6F
DATA VALIDATION Calibration method	Manual except NH_3, pH	Manual except NH_3, NO_3	Manual except NH_3	Automatic trigger of samples, Manual except NH_3, pH, DO	Manual	Manual except NH_3
Treatment of Calibration data	Manual no treatment	Separate files, correct on tideway O_2, manual		Not stored or recorded, intend to correct manually	Not stored or compensated	Left in, not stored or compensated
Cleaning-up data	Flag out of range	Elementary check	Manual	Manual	Manual	Manual
Correlation	None	None	Manual	Multi-graphic flow on simple stations	Manual	No/Manual
ALARMS	Threshold and Historic	Threshold, ARGOS on raw data	Low and high threshold, rate of increase (at centre)	None	Threshold	2 levels of importance and out-of-range
Other Analysis	None	None	None	Use other data for storm conditions	None	Manual Analysis
ANALYSIS Stats	Mean, max, min %iles manual	Mean, %iles S.D.	PC/Supercalc data tables	Mean, max, min, S.D. commercial packages	Data tables	Lotus 123, Data tables
Graphics	X-Y	Frequency half tide	Histogram X-Y	Time series, clean-up on PC, transfer to mainframe	X-Y Plots	Histogram X-Y
Other Data	Flow/Cumbersome to download		Flow		Flow/rainfall	Flow/Manual examination
STORAGE	File Meteorburst-PC	Clipper database on HP hard disc/tape	Quasar model PDP11 Winchester	Data archive database on mainframe	Lotus 123 Worksheets	386 PC file storage, floppy

* F - Fixed Monitors
M - Mobile Monitors

Table 2.1 - Summary of Interview Findings (Continued)

	Severn Trent	Thames	Anglian	North West	Wessex	Southern
Number of Monitors*	15F, 5M	25F, 3M	12F, 2M + 10	6F	12M	6F
ARCHIVE	Minworth data - HP disc/tape; meteorburst data -PC floppy disc	To tape	To tape	Tape database	VAX (raw data)	VAX/tape, disc
USES Planning	WO planning, classification system, extreme events, trends in river pollution	WQ Planning	WQ Planning	WQ Planning, Consent setting		WO Planning
Incident Detection	Pollution control	Real-time Monitoring	Response to major pollution incidents		Indicators	
Special Investigations	Special investigations		Modelling, Special investigations	Monte Carlo simulation, Special investigations	Prosecution, Investigation particular discharge, Damage limitation	Prosecution, Identify pollution source using computer simulation (SIMCAT)
Other	Regional communications system	Quality Reports Tideway Monitoring		Annual Reports		National report, Incident statistics, Estuary modelling

3. REVIEW OF PRACTICES IN OTHER INDUSTRIES

The review has drawn upon the experience of consultants within C&W and BAeSEMA with knowledge of practices for data handling, storage, analysis and alarm setting in different industries. A number of specific industries and applications were examined as follows:

1. National radiation monitoring

2. The nuclear industry

3. Air emission monitoring

4. The process industry

In many instances, standard systems are employed and therefore the relevant information on the capabilities of these systems has been obtained from the manufacturer's literature. Section 3.5 reviews the literature for different techniques which have been used in a variety of application areas.

3.1 <u>Nuclear - RIMNET System</u>

3.1.1 Type and Quantity of Data

By June 1992, the RIMNET phase 2 system will provide continuous gamma radiation dose rate monitoring at 92 fixed sites throughout the UK. At present, measurements from these fixed sites are known as ''basic'' data. Monitors will be polled routinely once per hour by the RIMNET central computer, although they can be interrogated more frequently should the need arise. The gamma radiation dose rate supplied comes in the form of a single reading from each of the fixed sites. Information on the operating condition of the monitor is also supplied. This information, together with time, date and location is transferred to a central database facility (CDF) held on the central computer.

The CDF will also hold a second set of data, known as ''supplementary'' data for radiological measurements. The majority of ''supplementary'' data would consist of spectrometric measurements, where the concentrations of a number of individual radionuclides are recorded.

It will be possible for approved bodies to enter supplementary data direct to the CDF using public data network links.

3.1.2 Validation and Calibration

Data from the automatic gamma radiation dose rate and additional monitors will trigger alerts on abnormal radiation. In order to control the action of users and the differentiation of tests, alerts and multiple alerts, the CDF will operate in several system states depending on the progress of an alert or incident. Readings from the automatic radiation monitors will each be checked at the time of receipt against alert thresholds to validate the nature of the rating. This checking will be carried out using an alert activation algorithm. To support the alert activation algorithm, provision will be made for the CDF to hold upper and lower error thresholds (i.e. the feasible range of a parameter's values) and normal radiation levels. Other factors such as the supporting evidence factor, sustained increase factor and the number of values for a running average will also be taken into consideration before an alarm is activated. There will also be access to change alert threshold parameters.

3.1.3 Alarm Detection

The central database facility routinely analyses the ''basic'' data coming in from the fixed sites, and raises alarms in the offices of the DOE if any abnormal increases are detected. Each high basic data reading needs to be checked in 3 ways:

1. Historically, i.e. was the previous reading also high.

2. Geographically, i.e. are any neighbouring monitors also producing high readings

3. Sustained, i.e. is there an upward trend in the readings from the monitor.

If any of these situations are found to be true, then an alert condition must be signalled and the alarm sounded. The reason for the alert must be investigated and the system changed to "incident" or the monitor reading marked as faulty and the system "alert" status changed to "normal".

3.1.4 Storage, Transmission and Analysis of Data

The "basic" data monitors are linked to the RIMNET central computer systems by a telephone line communications system based on a combination of BT's Packet Switch Stream (PSS) network and dedicated leased data lines. "Supplementary" data entry will be via interactive terminals or computer installations linked to the CDF through the PSS network or Electronic Mail. The CDF will have a back-up system. Both the original and the back-up systems will consist of dual DEC VAX 6000 series computers. There will be automatic switchover to the back-up in the event of main systems failure.

RIMNET phase 2 uses the INGRESS relational database management system for data storage and DEC decision software for analysis. System functions are based largely on VAX VMS operating system software and other DEC VAX software packages. Terminals are DEC 3100 VAX stations, equipped with a range of powerful computer graphics and mapping facilities for interrogation and analysis of data.

Geographical background data will include national grids, coastlines, major rivers and lakes, major towns and cities, road networks, county boundaries, height contours and locations of installations such as analytical laboratories. Foreground data will include basic data values and supplementary data values, normal radiation levels, monitor sites and sampling locations, a facility to form a contour map from interpolation of data points, scaling facilities, graphical data storage and sequenced displays.

RIMNET will provide a dedicated internal electronic mail system for communication between the various government bodies that would be involved in the response to an overseas nuclear accident. The Telecom Gold electronic mail system will be used for communication with non-government bodies. The system provides links with CEEFAX and PRESTEL to enable information concerning an incident and the Government's response to be communicated directly to the public in their homes and workplaces.

RIMNET phase 2 will only be used at a level approaching its maximum capability in the event of a serious nuclear accident, where large amounts of monitoring data have to be collected and stored over a relatively long period of time. The extent of use at other times will be much more limited.

3.1.5 Users

Users of the system will include:

Incident Report Room, Technical Co-ordination and Information Centre Management, Government Departments and agencies, Accredited local authorities, Nuclear Industry, Water Suppliers, HMIP and Health Authorities.

3.2 Nuclear Power Stations

3.2.1 Type and Quantity of Data

Gamma radiation dose rate monitoring is provided at the perimeter fence of most nuclear installations in the UK. The monitoring systems (i.e. monitors, data transfer and data analysis systems) are provided by commercial companies as a "package". The gamma radiation monitors are continuous monitors with an adjustable sampling period. Generally an integrated sample of one hour is employed and this data, together with wind speed, wind direction, time date and location data, is transmitted, by dedicated line, back to a central computer every hour.

3.2.2 Checking Validation of Data

Laboratory Implex Ltd provide a typical radiation monitoring system. The operator display utilises high resolution colour graphics and provides animation for such items as meteorological indicators for wind speed and direction and for gamma dose rate.

Erroneous data may arise due to a fault probe monitor, a break in the communication line or software errors. Changing symbols both in colour and shape show alarm status with the corresponding measurement alongside: green denotes normal conditions, grey a fault or monitor ''out of service''. A threshold level is used to detect faulty probes. The threshold is set at a level which is significantly less than normal background radiation or less than a low level radiation source which can be placed next to the probe.

Calibration factors for all monitoring positions are kept in one file. During dose rate calibration, up to 20 points may be stored for each detector. These can be used to allow system checks of the calibration factors to ensure the instruments are well calibrated.

3.2.3 Alarm Detection

Two threshold alarm levels are used with amber and red displays providing lower and upper level alarm alerts. The lower level amber alarm produces an audible and visual alarm signal, which has to be accepted by the operator, before the system is able to advance to the next alarm level. This ensures that higher level alarms are not triggered by false data. The action converts the flashing signal into a fixed signal. When an amber light is raised the monitoring frequency is increased automatically to a pre-set level by polling the outstations from the central computer. The alarm threshold levels may be reset by the operator at the central computer. This is to allow for higher levels of background radiation caused by particular meteorological conditions.

The display identifies key areas of the site peripheral buildings, muster points, risk areas and the location, status and reading of the detectors.

At the highest security level, the user is able to select alarm thresholds for each detector at any monitoring point. All events are logged to disk and a hardcopy event log is automatically generated.

The monitored data is often stored in a database which interfaces with a computer model for simulating the dispersion of radioactivity in the atmosphere. Meteorological parameters are also measured on site.

3.2.4 Analysis

Live values or integrated results for a user-determined period may be displayed in data tables. Graphical displays include:

1. Live updates of logged data.

2. Recall data by minute/hour/day/month/year.

3. Results from all positions at a fixed point in time.

4. Results from up to four positions with time.

5. Auto/manual scaling.

6. Results in integrated dose and dose rate units.

3.2.5 Storage

The monitored data is stored in a database within the central computer. The database provides for rapid access and setting of data for presentation as tables or in graphical form (see Section 3.2.4).

3.3 Air Emission Monitoring

3.3.1 Type of Data

There is a variety of equipment available on the market for monitoring "air" emissions on a continuous basis. The types of monitors frequently encountered are those for process emissions, accidental releases, site perimeter monitoring, mine gases, health and safety in the workplace and landfill gas. Each application has its own specific requirements with respect to computer hardware and software, sensors, alarm systems, the way information is displayed, the data handling capabilities and types of user. Additionally the number of sampling points, the length of term of monitoring and the ambient air environment will vary considerably. A number of examples of Manufacturer's equipment and their uses, gathering and storing large quantities of data are discussed in this section.

3.3.2 Monitoring Equipment

There are several manufacturer's systems for monitoring gas concentrations in ambient air. Software packages can establish a complete, co-ordinated monitoring system with a network of gas monitors, including alarm triggering. However, systems are often limited to a small number of monitors and are often for single gas measurement only.

Drager manufactures the Polytron Evaluation Unit, a modular gas detection system. The system consists of a field transducer with selective sensors for direct measurement and a digital display for simple "on site calibration". There are separate display and alarm facilities for up to 12 different gases. The modular concept allows for ease of interchangeability and expansion of the system.

The Polytron modular gas warning system is a rack assembly with a continuous display for gas concentration at each channel, and an LED display of channel status. The statuses are:

Green:	Correct Operating Voltage
Yellow:	Short or open circuit break in power supply or missing sensor. Channel flashing with 1Hz; calibration switch is activated.
Red:	1st and 2nd alarm levels. The 1st level is a pulsing at the raising of a new alarm. After acknowledgement the light changes to constant (2nd alarm level) until the gas concentration falls below alarm level.

There are two fail-safe alarm relays and one fail-safe alarm fault for each channel. All electronic sub-assemblies undergo temperature stress testing over several days. The alarms can also be set for increasing or decreasing levels.

For monitoring the emissions of landfill gas there are a variety of continuous monitoring systems available. Permanent landfill gas monitors are located to detect the escape of landfill gas from landfill sites where the risks presented by landfill gas migration are particularly high.

3.4 Process Industries

3.4.1 Voting System

On safety critical systems it is important to avoid false alarms and ensure all incidents are detected. Typical examples are fire alarms, smoke detectors, toxic gas alarms, water sprays on a venturi gas stream discharge, CO/CO_2 ratio with titanium dioxide processing. In all of these cases a voting system is used in which there is built-in redundancy through the use of multiple probes. A three probe system is often used in which cross checks can be made between the output from the three probes. A faulty probe is therefore readily detected with confidence and excursions from normal will be detected by all three probes and confirmed by cross-checking. The increased cost of built in redundancy is justified for safety critical systems.

3.4.2 Alarm Detection

Discussions were held with the On-Line Analysis and Measurement section of ICI regarding practices in the process industry.

The process industry currently makes use of thresholds and rate of change of a parameter's value to detect excursions from normal operating practice to raise the alarm. Inferential methods are also used to confirm an alarm condition by cross checking a parameter's value against another. The automatic validation of an alarm condition is followed by a manual checking procedure.

The analysis of effluent quality is not conducted in real time. Data is usually downloaded from the monitoring system onto a portable personal computer for analysis at a future time.

3.4.3 Research and Development

A research project, funded by the European Commission, is examining a more integrated approach to plant control through a linking together of information from each part of the production cycle. The system makes use of intelligent sensors and digital communication between sensors to provide monitoring, feedback and control. An expert system, based on the rules operators apply, is used to alert operators to an incident and provide advice on its control. Figure 3.1 shows a diagrammatic representation of the system.

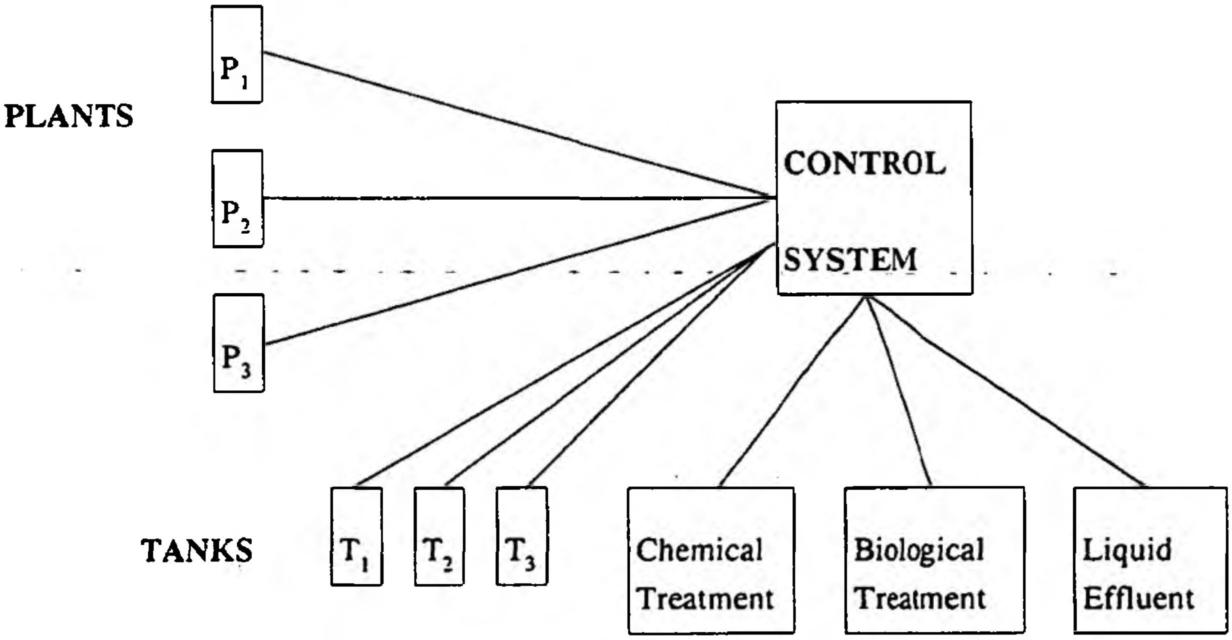


Figure 3.1 Diagrammatic Representation of a Plant Control System.

The central system provides real-time monitoring of the condition of the various processing plant, storage tanks and treatment plant to maintain an efficient, balanced operation. The control ensures that each unit is working within its design capacity and that, through feedback, waste is minimised.

The sensors are tested to established their fault modes which are then reflected back for fault detection through on-line validation of the measurements and their precision. For example, a typical fault is when the sensor readings fall to zero. In these circumstances, there is a marked change in the mean parameter value and the standard deviation becomes zero. Both of these data attributes are tested to detect a faulty sensor. (See also Yung and Clarke, 1989). Yung and Clarke, 1989 present a number of techniques for local sensor validation which are capable of dealing with unexpected data values in time series data e.g. rate of change, spikes and changes in the variance of the data. Kalman filters would also seem to be an appropriate technique for validating time series data.

Principal component analysis (PCA) is the main technique used to validate measurements and to calculate confidence limits. PCA transforms a set of variables into a new set of variables on the basis of the variance and correlation within the original variables. The new variables are defined by the eigenvalues and eigenvectors of the covariance matrix of the original variables. The eigenvector corresponding to the largest eigenvalue defines the coefficients of the first principal component:

$$y_1 = a_{11}\ x_1 + 1_{12}\ x_2 \ \ldots.\ a_{1p}\ x_p$$

The first principal component represents the linear combination of the original variables that accounts for the largest amount of the variation in the data.

The second principal component accounts for the next largest amount of variability, and is uncorrelated with the first (after the eigenvectors are orthogonal). There are a total of *p* principal components for *p* original variables. However, it is usually found that it is possible to account for most of the variance in the original data (75%) with a smaller number of principal components than the original number of variables.

Through an understanding of the parameters which account for the greatest variability in a system better control of process can be achieved.

3.5 Intelligent Automation Techniques

Current approaches to water quality monitoring clearly involve a significant amount of manual data processing and analysis. The most obvious way to provide a degree of useful automation would therefore be to partly or wholly automate a well chosen subset of the functions performed by water quality experts by encoding the procedures and skills that he uses, for supporting his analysis activities.

An automatic water quality monitoring (AQM) system should be able to provide some or all of the following main functions:

- alarming operators in a timely fashion when water quality is behaving in an unexpected or unacceptable way;
- advising operators of characteristics and possible interpretations of the abnormal situation;
- advising operators on possible explanations for causes of the detected abnormal situation;
- advising operators on appropriate corrective action which should be taken;
- where relevant, automatically collecting samples of water for more extensive, offline analysis and to provide evidence.

A system which provides or supports most or all of the functions listed above would be an example of an Intelligent Process Monitoring and Control System. Intelligent Process Monitoring and Control is a well researched area and a number of software tools and systems have been developed.

Thus by studying the issues related to the development of this more general class of systems, it is likely that useful lessons can be learnt in respect of:

- possible choices for the overall system architecture;

- possible choices for the decision support technologies to use to provide the higher level AQM functions:

- appropriate forms of human computer interaction and allocation of function between man and machine.

A brief review of the literature in the general area of intelligent process monitoring and control has highlighted a number of publications which are relevant to the above topics including:

- papers describing top level designs and toolkits for constructing general purpose process monitoring and alarm processing systems which are able to deal with large quantities of data in real time:

- papers dealing with knowledge representation and reasoning techniques for process control and alarm processing (e.g. the roles and uses of statements of fact about the domain and the processes involved; the inductive processes used by the system; the integration of deep and shallow knowledge, i.e. including some ''understanding'' of the problem domain in the analysis of the data and methods of coupling numerical and symbolic, i.e. logic processing).

- papers on human factors issues (e.g. assessing the impact of intelligent automation on humans, problems in man machine communication, use of natural language interfaces and comparison of different approaches to data display).

As time and space do not permit a detailed analysis of these references the following sections contain a synthesis of the material.

3.5.1 Techniques of Potential Interest for AQM

The water quality expert's approach to problem solving may be divided up into the following groups of top level functions:

- Data Validation: validating the integrity of the data that is to be used for subsequent analysis (e.g. the data could be from a faulty sensor or manually injected data may have been entered incorrectly);

- Data transformation: if necessary, manipulating the acquired and validated data into forms which can readily reveal the absence or presence of environmental states of interest;

- Feature Extraction and Abnormality Detection: extracting standard features of interest from transformed data and using these to detect states of interest (especially alarm stages);

- Diagnosis and Prognosis: determining the causes of detected environmental states of interest or the effects and consequences of current stages;

- Advice and Action Generation.

These functions are generally performed in the order listed above as data are transformed from signals to features to symptoms to causes, effects and explanations and finally to advice and action.

The following paragraphs provide a brief review of some techniques of relevance to the automation of each of the five functional groups listed above.

3.5.2 Data Validation

Before any analysis can be performed, it is clearly necessary to have in place reliable (automated) procedures for validating automatically acquired sensor data. It would also be desirable, but perhaps less essential, to have automatic procedures for verifying the correctness of any manually entered data.

Because of this pivotal role that sensor data validation plays in many systems automation developments, it is a well studied problem.

Hardware redundancy of sensing (and processing) is often exploited, particularly in mission critical and high value context where robustness to sensing failures is essential.

Software based methods for sensor validation have also been developed and deployed in a wide variety of domains ranging from aircraft flight control to the maintenance of power generation equipment. Software based sensor validation schemes can, by and large, be classified into one or the other of the following types:

- algorithmic
- knowledge based

Knowledge based sensor data validation schemes exploit information such as:

- temporal characteristics of individual signals; e.g. the maximum change between two successive readings which is physically feasible
- empirical and model based relationships between different sensors;
- sensor failure characteristics.

As might be expected, this heavy reliance on application specific information means that knowledge based sensor data validation schemes are typically developed in a bespoke fashion for each application.

3.5.3 **Data Transformation**

Mechanising the visualisation procedures used by experts represents a simple way of automating their task but one which could:

- reduce the expert's monitoring workload;
- accelerate the process of discovering new knowledge (i.e. empirical relationships in the problem data) if provided as part of a more extensive suite of data analysis and visualisation tools (see below);

Many software tools are commercially available for manipulating spatial and temporal data in standard ways, visualising the results and extracting basic summarising features. Examples include descriptive statistical analysis tools such as STATGRAPHICS, UNISTAT, SAS and IMSL; Geographical Information Systems (GIS) such as TERRAIN-GIS and mathematical toolboxes such as MATLAB and NAG.

More generally, the graphical presentation of information is a standard topic in the design of information processing systems, and principles and approaches have been derived which are well founded in terms of their accommodation of the human factors associated with the visual assimilation of information.

It is also worth noting the potential value of inductive and similar structured learning methods in providing possible means for automating or assisting the expert with the process of discovering new empirical relationships in large bodies of data.

3.5.4 Feature Extraction and Abnormality Detection

Mechanising the perceptual skills exhibited by humans is a hurdle common to man intelligent task automation problems. Vision and Speech are the most general examples of perception, but expert interpretation of more specialised one, two or three dimensional data is a commonly studied topic.

Partly because of the diversity of data forms which exists, many pattern classification methods have been developed for modelling how humans, in different problem domains, extract information from (i.e. ''understand'') signals.

Particular domains in which the perceptual abilities of an expert have been modelled include the interpretation of:

- oil well logs;
- seismic data;

- acoustic signals;

- manufacturing process parameters (pressures, temperatures, flows etc);

- machine condition monitoring indices (e.g. vibration);

- patient monitoring (e.g. in an Intensive Care Unit)

- medical imaging;

- surveillance imagery (e.g. CCTV, aerial photography, remote sensing).

There has recently been a resurgence of interest in fundamental research and practical applications of neural computing methods, particularly to perceptual/pattern classification problems.

The key features of neural systems are that:

- they are trainable - by giving the neural systems sufficient examples of related input patterns (which could be a time series of water quality data) and outputs (e.g. alarm state) the system can adapt its behaviour to give the desired response.

- they offer some robustness to missing input data - outputs can be produced even when some parts of the input data are missing.

- while training a neural network can require significant computing resources. However, applying the resulting model to data is a simple process which can be done with very limited computing resources.

- their structure is biologically based and thus could be claimed to more closely model human perception.

Particular examples of neural control systems include:

- real time control;

- speech recognition;
- image classification;
- target tracking;
- sonar classification;
- credit risk assessment;
- commodity and currency dealing.

Particular examples of neural control systems include:

- real time control
- speech recognition;
- image classification;
- target tracking;
- sonar classification;
- credit risk assessment;
- commodity and currency dealing.

An analysis of the methods used in training neural systems indicates that most of the algorithms commonly used are direct analogues of traditional pattern classification methods. However, neural systems can be developed and modified more readily than these other approaches because of the ease with which they can be updated with new training examples. Neural training algorithms are also better supported in terms of software tools and, now, even dedicated hardware systems that are sufficiently powerful to allow the parallel processing algorithms to be executed in real time.

A common problem in modelling perception is getting the expert to explain what it is, in the data, that enables him to classify it. Thus, for example, a water quality expert might not be able to adequately explain why he draws a particular conclusion from a given plot of data. Given enough training data, neural systems are well suited to addressing such ''Recognition and Understanding'' problems.

It is worth noting that, following this stage, it is quite usual to use data compression techniques to reduce the volume of storage required, and that in general data compression can be achieved using both numerical and symbolic oriented methods.

Numerical data compression algorithms are generally concerned with identifying redundancy in the raw signal and coding methods have been developed and applied to both one dimensional, time varying data (e.g. speech) and two dimensional data (e.g. video). Popular data compression algorithms include LPC, RLC and DCT. More recently, methods derived from work on fractals have been also been exploited for two dimensional compression.

Using numerical methods, the original low-level signal can be completely reconstructed. By contrast symbolic techniques aim to summarise the essential features of data required for making decisions by abstracting out the relevant signal characteristics and throwing the rest away. With this approach reconstruction of the low level signal is not, in general, possible.

A variety of symbolically oriented signal modelling and representation techniques have been developed, including linguistically based approaches (e.g. fuzzy pattern classification) wherever descriptions of the signal are contained in a natural language grammar. Some examples might be:

- ''the total rainfall in February was well above average for the time of year''

- ''the average flow rate was slightly above normal during March''

- "at about 5 pm there was a sudden increase in pH measured at station X, it reached a peak at about 5.30 pm and then steadily declined, returning to its normal value after about six hours"

This approach is particularly attractive if the data is to be used by human experts or expert systems which attempt to capture the imprecise (fuzzy) manner in which human knowledge is often expressed.

3.5.5 Diagnosis and Prognosis and Advice and Action Generation

Automation of Diagnosis/Prognosis and Advice/Action Generation (i.e. moving from symptoms to causes, effects and explanations and then to advice or actions) is a huge area in terms of:

- research into relevant decision support techniques;

- practical applications of available methods.

Despite this a brief literature survey has revealed little work in other domains on intelligent automation problems comparable to water quality monitoring. The two most noteworthy projects are:

ENVIRONS - a rule based system which can be used to assess the impact of environmental pollution on the health and safety of workers, and which can assist with the processes of detecting the pollutant and its source of origin, and the steps necessary to effect pollution management;

TEN-PRO - an ESPRIT project which has recently begun and whose aim is to develop a comprehensive system to effect Total ENvironmental PROtection. The main market for the system would be operators of large production processes (e.g. chemical works) which discharge by-products and effluents into the marine environment. The envisaged system has a hierarchical organisation extending from sensor data collection and validation, through to system control and environmental protection management.

3.6 Conclusions

A review of practice in other industries and application areas has revealed that, often, a more sophisticated approach to incident detection, data validation and analysis is adopted compared with current techniques used within the NRA. Techniques commonly used for incident detection and validation are:

- two or more levels of alarm threshold
- rate of change alarms
- comparison against adjacent outstations
- comparison against correlated parameters
- built-in redundancy through the use of multiple probes at one site.

More sophisticated techniques are available and have been used in other application areas to detect trends or patterns in data for both data validation and incident detection:

- rule and knowledge based methods which exploit temporal characteristics of individual signals, empirical and model based relationships between different sensors and sensor failure characteristics
- pattern recognition techniques
- neural systems using neural training algorithms.

The other areas examined are generally concerned with simple systems (e.g. radiation monitoring), or complex systems where the normal behaviour is well understood (e.g. the process industry). In both these areas, the human expert can readily distinguish normal and abnormal behaviour, and implement corrective action.

Under these circumstances, automation of the task can be envisaged as possible.

The behaviour of river water quality data is potentially complex, depending on internal and external events. There is no clear definition of abnormal behaviour (what constitutes a pollution event?).

Therefore since the task cannot yet be performed by a human expert, automation of the task is still distant. However, techniques in use elsewhere can ease the problem of data processing, particularly in data validation.

The editing, access, analysis and interpretation of data may be assisted through the use of commonly used displays of data and analysis tools:

- x-y graphs
- histograms
- statistical analysis tools e.g. STATGRAPHICS, UNISTAT, SAS, IMSL
- mathematical tool boxes e.g. MATLAB, NAG
- area maps showing outstation location and parameter values
- Geographical Information Systems (GIS), e.g. TERRAIN - GIS
- graphical editors, combined with GIS or maps

In general data is stored on a central database which provides access to current and historical data in a variety of forms and the form of the database depends on the characteristics of the system: volume of data, speed of response required, single or multiple user access etc. The choices fall between data storage mechanisms developed specifically for the application and proprietary database management systems. Except where the data is of low volume and simple structure or where there are very specialised requirements, e.g. specialist hardware, bespoke data storage systems will probably have unacceptably high development costs. Of the various types of database management system, the choice depends on the required performance and data volumes and the flexibility required in accessing the data. Where the volume and performance requirements are not too stringent and the data may be extracted from the database according to a variety of criteria which cannot be completely pre-defined, both of which criteria would be satisfied in an AQM system, a relational database, such as the one used for RIMNET would be the appropriate choice.

4. DATA ANALYSIS TECHNIQUES - RECOMMENDATIONS

4.1 Introduction

In presenting our findings we have concentrated on recommending techniques which could be applied by the NRA regions for the handling, analysis, alarming, storage and archiving of water quality data. We have not made recommendations for particular hardware and software to be used as each of the regions had already installed different computer systems.

Our recommendations for storage and archiving the data are a consequence of the user requirements and quality assurance considerations.

In general, there is not a requirement for particularly sophisticated analysis of the water quality data, at this stage, possibly because the users require the basic analysis tools to be developed as a priority. The techniques we have recommended are standard, readily available, well tried and tested. They increase the flexibility and efficiency of the system rather than the complexity.

4.2 Incident Detection

4.2.1 Exceptions and Incidents

An *exception* can be defined as a measurement or series of measurements that is in some way unusual. An *incident* is something that requires action such as follow up sampling, which will be flagged by the raising of an *alert*. At some intermediate stage, alarms may be used. An *alarm* is a means of drawing the operator's attention to an exception that requires action. Not all exceptions will be incidents.

As a part of the incident response process, it is necessary to detect exceptions, assess them, and raise an appropriate alarm.

4.2.2 Types of Exception

Four kinds of exception can be distinguished:

1. Instrument/communications failure - no information about the parameter value is being received;

2. Parameter has exceeded a critical value (test on single parameters only);

3. A combination of parameters has occurred which is unusual, although the individual parameters themselves are not unusual;

4. Unrepresentative value has been produced as a result of normal instrument operation (e.g. a calibration cycle).

Of these exceptions, type 1 always requires action (repair instrument), while type 4 always needs no action. The action required for type 2 and type 3 will depend on severity and on policy. For instance a dissolved oxygen reduction may indicate a pollution event which may warrant investigation with respect to prosecution of an offender, or dissolved oxygen may have fallen to a level where fish are stressed and emergency response is necessary (whether or not this is the result of pollution).

Type 1 exceptions (instrument failure) may be detected in a number of ways. The instrument itself or the communications channel may detect a hardware fault and send an error code to the monitoring system. This is clearly independent of the monitoring system. However, type 1 errors may be more subtle than total failure, and therefore depend on the monitoring system for detection. Examples of such errors may be unfeasible results (e.g. temperature outside the range 0-30°C), very sudden changes, or periods of completely constant readings (indicating a blocked sampling tube).

Type 2 exceptions (single parameter values) may be easy to detect (e.g. turbidity responses), or they may be masked by other phenomena. For example, dissolved oxygen has a strong diurnal cycle in summer, and the average dissolved oxygen value is seasonally dependent. In this case, thresholds for the detection of single parameter events need to incorporate expected behaviour, they must be adaptive.

Type 3 exceptions need a multi-parameter approach. This may be by correlation, or voting, or some more sophisticated method.

Type 4 exceptions can be handled in a number of ways. Ideally, calibration data should be identified by the sending instrument, so that they can be stored separately from 'real data'. This is now available with more modern instruments, but there is still a backlog of historic data to be processed. Thus a method of filtering is necessary. Filters vary from manual (accurate, but labour intensive), through averaging, to frequency dependent filters and custom de-spiking algorithms. The last two are easily realisable with present computers. De-spiking algorithms could be built into the instrumentation itself.

4.2.3 Alarms

After detection of an exception, the operator needs notification in an appropriate way.

Three levels of alarm are recommended:

Flashing Green	-	instrument or communications failure (constant green light indicates fully operational)
Amber	-	initial detection alarm
Red	-	full alarm conditions

An amber alarm is indicated on first detection. It is frequently impossible to determine if a single exception represents some kind of erroneous data point, or the beginning of a trend. Thus the alarm should require confirmation by the operator, probably on receipt of the next data point. On confirmation, the alarm status is increased to red and action is taken - the exception has become an incident.

This imposes a delay in the implementation of action of at least one measurement interval, nominally one hour. If this is unacceptable, when alarm status amber exists the measurement station should be instructed to increase the measurement frequency.

C&W recommend that the status of an alarm is validated by reference to a number of different measures:

1. The rate of change of parameter value.

2. Maximum and minimum feasible values.

3. Correlation against other parameter values.

4. Compensation for expected natural variations in a parameter's value.

C&W recommend that operators are alerted to the alarm status through audible and visual signals at the computer terminal. Different signals should be used for each alarm status as recommended above. We believe that the most effective alarm display is a map of the NRA region showing the river system and location of the water quality monitoring stations. The alarm status and readings would be displayed at the relevant station on the map. A graphics editor would be used to select the outstation of interest and display the data leading up to, and triggering the alarm. The graphical display of data is described in Section 4.3.

4.3 Data Validation

It is acknowledged that the fixed site ammonia monitors readily drifts out of calibration and therefore need to be calibrated automatically every 12 hours. The dissolved oxygen and pH monitors are calibrated weekly or fortnightly and the other monitors every 3 months or more. Under incident conditions, where readings are likely to be taken every 15 minutes it is important to obtain the most accurate data. Analysis of water quality data taken in the field will reveal whether calibration drift follows a consistent pattern. If a clear pattern is evident then it will be possible to compensate for this drift. It is likely that the calibration drift will be unique to a particular instrument and therefore a correction algorithm will need to be derived for each instrument.

We, therefore, recommend that calibration data is stored in a separate computer file so that the magnitude of drift can be automatically monitored. The predictive algorithm for compensating for drift may then be updated in the light of the new calibration data. Specific techniques were used to investigate real data in Phase II of this work, reported in Section 5.

Missing data may arise due to a fault in the monitor, data logger, outstation power supply or the communications system. We recommend that these data are automatically flagged in the computer system and in the stored data so that they will be detected and dealt with appropriately during analysis and display. The detection of faulty data will use the techniques described in the previous section on incident detection.

In order that further data validation can be undertaken by the user we recommend that the facility to display, graphically, parameter value against time is provided with the following features:

1. Selection of time period for display of at least 10 days data;

2. Selection of up to 6 parameters on one plot (including hydrometric data);

3. Two parameter scatter plots, with recent data flagged (different colour);

4. Selection of up to 6 outstations on one plot;

5. Selection to plot calibration data for up to six outstations;

6. Mark data as instrument error or calibration data using a graphical editor.

4.4 Water Quality Planning

The manually sampled data are currently used as the main data source for water quality planning. Long run statistics of the mean, maximum, minimum and 95th percentiles are produced. However, the quantity of manually sampled data is small and the statistics, therefore, do not have a high level of confidence. There is a desire to make greater use of the, higher volume, automatic water quality data for planning purposes. Water quality standards are set as the 95th percentile values for the various water quality parameters, i.e. 5% or less of values taken within a year must not exceed this level. The use of automatic quality data will allow a more accurate determination of the mean and 95th percentile for comparison with the standard. However, the planner will wish to examine data which comprises the upper or lower 5th percentile to analyse the severity, period, timing and location of events. It is only through such analysis that plans for improving water quality may be drawn up. We, therefore, recommend that the facility to display data in the form of a cumulative distribution function showing the 5% tail of the distribution is provided. The data will need to be sorted in order of increasing parameter value to produce this plot. We recommend that the user has the facility to gain ready access to this data set for examination and further analysis of individual data values, if required. These facilities are more readily achieved by use of a database-graphics computer package (see following section). Furthermore, it is important that the datasets have unique identifying parameters so that consistent datasets are retrieved for comparison. Hence, a system which allows traceability and identification of files in storage and archive is required. This is a necessary part of any quality assurance system.

Computer models are used to assist the water quality planner. Two main types of model are used or likely to be used in future:

- water quality planning model

- real-time control models.

The water quality planning model is a Monte Carlo simulation which uses long term (1-5 year) data, input in the form of probability distributions. It is used, mainly for examining water quality under varying flow and temperature conditions and pollution loads to study the effect of a new discharge or abstraction on overall water quality compared with the standard. Consent limits are set based on the results of this analysis.

For ready interface with the models it is more efficient to store the water quality information in the form of a probability distribution function. A simple two parameter distribution function, such as a beta function, may be fitted to the cumulative data distribution function (CDF) and converted into a probability distribution function (PDF). If a more complex fit is required a four parameter distribution, such as the log logistic distribution may be used. The advantage of both the distributions cited above is that they have finite maxima and minima thus avoiding the possibility of sampling very extreme values.

Cumulative distribution function plots of data against a fitted distribution will reveal whether a good fit has been achieved and show the degree of data uncertainty.

A number of issues arise with this type of modelling mainly relating to the number of data points sampled and the level of confidence in the predicted result. Techniques, such as importance sampling, are available for defining the extremes of the output distributions and converging on a consequence prediction more quickly.

Real time models are used to predict the downstream effects of continuously changing flow and water quality conditions. These may be used to predict the migration of a pollution incident downstream so that action may be taken to reduce the impact of the incident e.g. opening sluice gates to increase dilution. Alternatively, such models may be used to back-predict the location and time of the pollution incident so that more detailed analysis may be mobilised. The successful use of real time control is dependent upon a sufficient number of monitoring stations being available to gather the data for the model. Clearly an efficient model interface is required so that the predictions may be updated immediately more data becomes available.

4.5 Data Transmission

In order to meet the requirement for incident detection some processing of the data must occur to determine if an alarm condition is to be raised. If an alarm condition is detected then this must be reported at some central point, for example a regional control room. In order for the alarm to serve its purpose, it must be detected and reported soon after the incident occurs.

The current monitoring stations which provide the capability for incident detection (i.e. are capable of communicating with a central site) are of the following two types:

1. The data is transmitted immediately after being measured, for example, by radio.

2. Data is stored at the monitoring station and transmitted as a batch at intervals, typically daily over the PSTN. The monitoring station also has the capability to contact the central system at any time to raise alarms.

With type 1, the detection of alarm conditions can be done at the central site. Relatively sophisticated alarm detection algorithms could be used and both historical information on the behaviour of parameters at that monitoring site and current information for other sites, where relevant, could be used in alarm detection. With type 2, however, the alarm must be detected at the monitoring station so that communication with the central site can be initiated. As the processing capability of the controllers at the monitoring stations is limited, the sophistication of the alarm detection algorithms could be limited and only current information for that monitoring station would be available to the alarm detection algorithm.

The disadvantages, for incident detection, of batch transmission of data to the central site could be overcome by providing greater data processing capability at the monitoring stations and facilities whereby data used in the detection of alarm conditions could be downloaded automatically to the monitoring station processor. The relative costs and benefits of these two approaches need to be examined further.

A combination of the techniques may be used, with the monitoring station normally communicating in batch mode, but switching to real time transmission when possible exception is detected. Further central processing of the data would be used to investigate the exception. After the exception is cleared, the monitoring station could be instructed to return to batch mode.

4.6 Storage and Archiving

After initial viewing and validation, the data may subsequently be used for a number of purposes. Recent data might be used for investigation of the current state of a river, older data might be required for examination of long term trends. Data may also be required for input to models or to examine particular incidents. Because of these different requirements it is recommended that the full datasets should be stored in validated but unsummarised form. The raw data should be preserved also, with a status flag to identify such conditions as calibration cycle, missing value, interpolated point, etc. For maximum flexibility in use, it should be stored as a time series, i.e. time+value pairs, rather than as tables or other elaborate formats. Such formats as daily summaries (in which we received some of the data) impede use for other purposes, and should be generated on output rather than on input.

Given that the regions interviewed had no plans to increase the number of monitoring stations significantly, the volume of data likely to be generated is not great in comparison with the capacity of current storage devices. Storage of the data from a single parameter monitor, measuring at 15 minute intervals, will not generate more than 2000 bytes of data per day (this allows for recording time to the nearest second, a raw data value, a validated data value, and a status flag to describe the validation process). Thus a station measuring 8 determinands will collect 5-6 Mbytes of data per year. Fairly simple file optimisation techniques could reduce this by 50%. Present removable disk technology stores up to 600 Mbytes per disk, online disks exceed 1 Gbyte. Storage technology does not, therefore, impose any limits on the way in which water quality data should be stored and all data for a region could be kept on-line, thus removing the need to archive the data. Archive copies would need to be made to ensure data security and, if required, to transfer the data to a national database.

Data management software will be required to select, extract and process data from this store. It is understood that the NRA plans the implementation of a High Resolution Data Store (HRDS) which will amongst other items, store the water quality data. The requirements on the data management software should be considered for the whole of the HRDS rather than for the water quality data alone.

5. APPLICATION OF TECHNIQUES

5.1 Introduction

The second phase of the research project was designed to apply the ideas and techniques discussed in Sections 3 and 4, to automatically monitored water quality data. It is only through the examination of genuine data sets that the nature and variability in data can be appreciated and hence the true performance of the techniques tested.

The data sets chosen for investigation were taken from the James Bridge site on the River Tame, and the Thurmaston site on the River Soar, both in the Severn Trent region. These water quality datasets were supplemented by flow data from nearby hydrometric stations. This allowed comparisons to be made between different types of river. The Tame at James Bridge drains urban and industrialised areas of the West Midlands and is significantly polluted. It runs in an artificial channel at James Bridge. The Soar at Thurmaston drains a catchment with mixed land use, a large agricultural area but including the city of Leicester.

Data were obtained for the year from 1 July 1991 to 1 July 1992, so that variations over summer and winter could be investigated. Some smaller datasets for various other periods were also obtained, containing measurements of particular interest. Representative data are presented as time series plots in Appendix C. Further data analysis centred around the Thurmaston site so that comparisons could be made with a known pollution incident which was detected at Thurmaston, Sileby and Kegworth on the River Soar (see Section 5.5).

Generally, preliminary work relied upon Exploratory Data Analysis in order to understand patterns, correlations and features of the data. This involved an iterative approach of presentation in graphical form, visual analysis followed by proposals for further manipulation, analysis and presentation of the data.

5.2 Distinguishing Features of Data Sets

Appendix C contains the following data from the River Tame and the River Soar:

- temperature
- conductivity
- pH
- dissolved oxygen
- ammonia
- suspended solids/turbidity
- flow

Data are presented for August 1991, January 1992, May 1992 for the River Soar, and January 1992, April-May 1992 for the River Tame. These are typical of the complete record.

Examination of these plots shows a number of features:

A **diurnal cycle** in dissolved oxygen, pH and temperature during the summer months (April-October). The daily temperature cycle is a result of solar heating during the day, and radiative cooling at night. It is most pronounced during cloud free weather. The canalised nature of the river at James Bridge makes the diurnal temperature variation more pronounced. At this time of year with strong sunshine and warm temperatures there will be a high level of photosynthetic activity which produces the diurnal cycle in dissolved oxygen. Photosynthesis also has the effect of removing carbon dioxide which in turn reduces the acidity of the water thus increasing pH. A correlation between dissolved oxygen, temperature and pH is therefore expected at this time of year. Slight diurnal cycles are also sometimes visible in conductivity and ammonia - these are also attributed to photosynthesis.

Large positive **calibration spikes** in the ammonia data, roughly twice a day, caused by the automatic calibration of the sensor. The calibration spikes in the ammonia data from Thurmaston are much larger than the real measurements. Less frequent calibration spikes are also evident from other sensors at the Thurmaston site. These spikes may be positive or negative. Other **random spikes** are present in some of the data.

Rapid **increases in flow**, especially of the River Tame. Rapid run-off from the surrounding industrial area will have contributed to the rapid increase in flow. These rapid increases in flow are matched by an increase in suspended solids, decrease in conductivity and decrease in pH. The reduction in conductivity is probably caused by the diluting effect of the rainwater on the total dissolved solids. Also the large volume of rainwater has had the effect of modifying the background pH level of the river, although the changes in pH are not of great significance. The River Soar also responds to rainfall with rapid rise and a slow fall, but in a more subdued fashion than the River Tame. This reflects its larger catchment and small proportion of impermeable area.

It should also be noted that there are occasions of increased suspended solids and changes in conductivity which are not associated with a rapid increase in flow e.g. 19th/20th May. These may be caused by a pollution incident or change in the discharge from a sewage treatment works. However there is no clear change in ammonia or dissolved oxygen during the same period.

The flow records from both sites, and the conductivity data from James Bridge, show the presence of statistical noise. This is a random fluctuation in the measurement, and in the case of the flow measurement is in part a consequence of the turbulent flow over the weir. In the Thurmaston flow record, noise may be as much as 50% of the signal at low flows.

Further examination of the ammonia data (see Figure 5.1) shows unrealistically high values in the period 10th - 12th May and this may indicate a sensor failure. A period of constant readings between 12th - 13th May certainly indicates a measurement problem. Whilst the plot provides a clear visual indication of this sensor failure it is more difficult to automatically screen out these events. This will be discussed further in Section 5.5. A more subtle fault can be seen in the flow record for James Bridge in April 1992 (Figure C26), where there is a clear minimum recorded value well above instrument zero, suggesting some kind of transducer failure. The Thurmaston pH data (Figures C13-C15) are different in character between 1991 and 1992, showing that the sensor was changed and a different, higher resolution sensor was installed around November 1991. This will affect the calculation of statistical data.

Examination of the full datasets shows that missing values occur. These may extend for large periods of time. Correction of calibration cycles, spikes and instrument faults are also a source of missing values.

5.3 Processing of Data

5.3.1 Requirements

From the above examination of the data from two sites, a number of processing requirements can be identified. These are: spike removal, missing value replacement, filtering of periodic variation, change detection, and identification of correlated events.

5.3.2 Spike Removal

Most of the spikes result from calibration cycles. The easiest solution to removing these is not to record the spikes in the first place. This will depend on the capabilities of the monitoring stations and whether they can be programmed to write the calibration data to separate file for transmission to the central computer at a future time. Retaining the calibration data in a separate file will allow analysis and possible correction for calibration drift. Examination of the capabilities of the monitoring stations is outside of the scope of this study but should be seriously considered by the NRA.

This will not remove the spikes in the existing data archive, and therefore some automated method is necessary.

Calibration spikes occur regularly but not at exact times and therefore cannot be eliminated on timing alone. The value of the spikes is variable and in general is within the range of possible data, so the spikes cannot be eliminated on value alone. The calibration spikes are characterised by a sudden increase in parameter value from one reading to the next and a very similar sudden decrease in value in the succeeding reading. The values preceding and succeeding the calibration value are usually at a similar level. Random spikes have the same character. Occasionally spikes will have a shoulder, either the preceding or succeeding measurement is also affected to some degree.

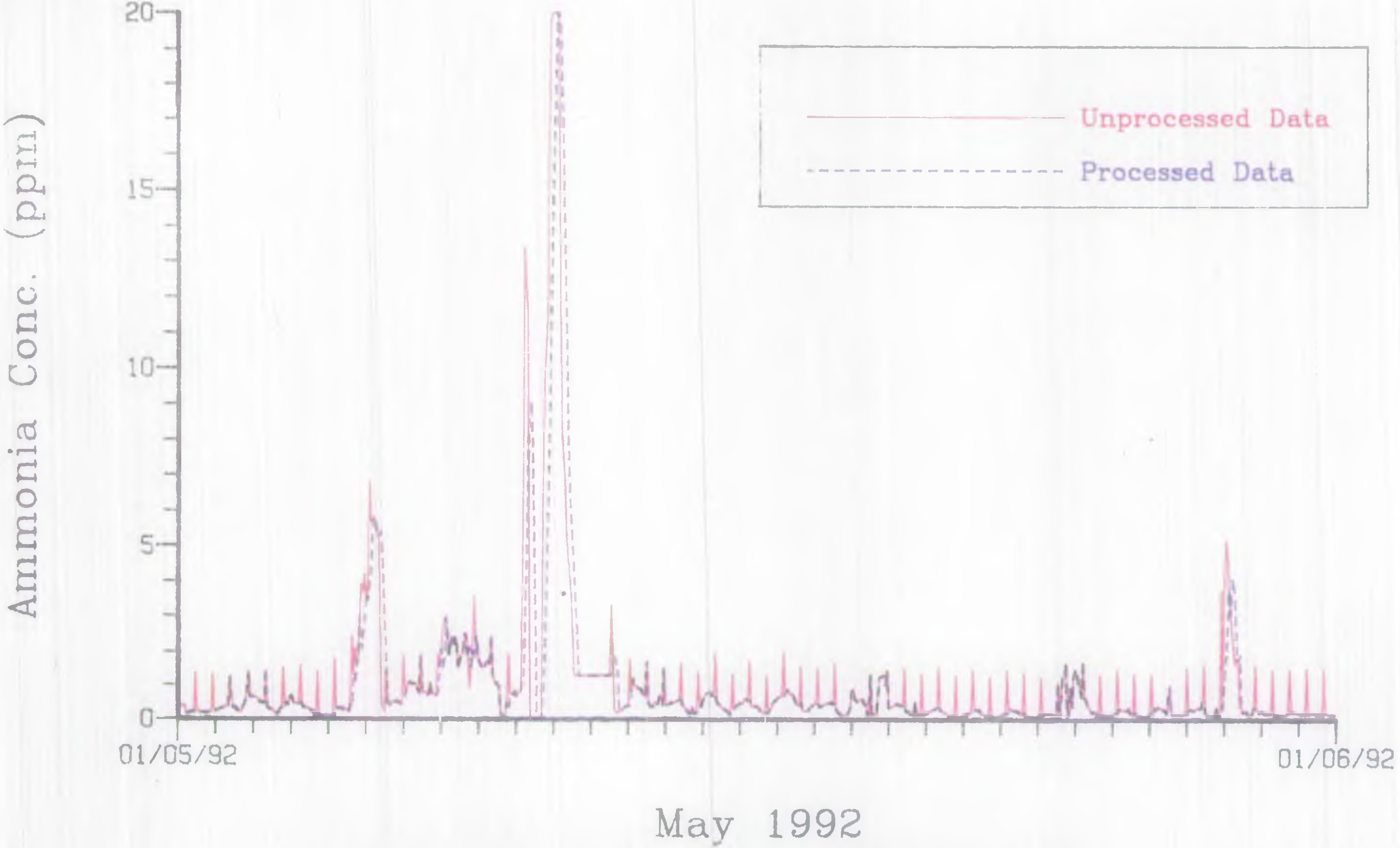


Figure 5.1 - De-spiking using 5 hour Moving Average, James Bridge Ammonia

The first approach to removing the spikes used a moving average. A five hour moving average was used to detect these sharp rises and fall in parameter value as follows:

```
set $i, j = 1$

$y_j = \frac{1}{5} \sum_{i=1}^{5} x_{i+j-1}$

set $i = 5$

$z_j = \frac{1}{6} (5y_j + x_{i+j+1})$

IF $x_{i+j} - y_j > B$

AND $x_{i+j} - z_j > B$

THEN $x_{i+j} = \frac{1}{2}(y_j + z_j)$

repeat for $j = j + 1$
```

where: x_i *is the raw value of the parameter*

y_j, z_j *are moving averages*

B is value selected for the parameter and based on 'typical' average values of the parameter after calibration data has been removed.

This algorithm detects abnormal results by comparison with the average over the previous 5 hours, and interpolates a 6 hour average through the suspected spike. Figure 5.1 shows a plot of ammonia data for May 1992 at James Bridge after applying the algorithm, given above, compared with the raw data. A value of 1.0 was used for B. Clearly the value of B will depend on the parameter and the time of year. The algorithm is successful in removing most spikes, but there is a penalty in a slight shift of the data.

An alternative algorithm used the fact that the changes either side of the spike are opposite and nearly equal.

At point *i*

$$G1 = x_i - x_{i-1}$$

$$G2 = x_{i+1} - x_i$$

$$G3 = (x_{i+1} - x_{i-1})/2$$

$$B = \text{tolerance } x \; |G3|$$

Data is spike *if*

$$|G1| > B$$

and

$$|G2| > B$$

and

$$|G1 + G2| < 2 \text{ x } B$$

Interpolate by

$$x_i = x_{i+1} - G3$$

Tolerance is adjusted for best performance.

The tolerance needs some adjustment for optimum response. The detected spikes may be replaced either by a linear interpolated value, or by a missing value flag. The results for the ammonia data for James Bridge and Thurmaston are shown in Figures 5.2-3. The algorithm removes nearly all spikes from the James Bridge data, even those on sharply rising parts of the data. It is less successful with the Thurmaston data. Examination of the Thurmaston data shows that the spikes escaping detection have shoulders and are thus effectively two values wide - the algorithm is very specifically set up for spikes only one value wide. A different algorithm would be needed to improve success rate on this data.

5.3.3 **Filters**

Digital filters were investigated for removing cyclic components of the data (e.g. diurnal changes), and high frequency parts of the data (e.g. noise and spikes). Several approaches are possible: time series of finite length can be transformed into the frequency domain using Fourier transforms, filtered, and transformed back; filtering can be done directly using a correctly designed weighted moving average; or recursive estimation techniques can be used to produce an adaptive filter, this can be likened to a weighted moving average with variable terms. The last technique is Kalman filtering. Kalman filters are a class of filters which use some model of the data behaviour to estimate its value. These models include an element of stochastic behaviour and are thus very good at describing real measurement systems.

Kalman filtering techniques can be applied to finite sets of data to perform smoothing, which estimates missing values. Because the filter embeds the statistical variability of the data and a model of its structure, the estimated missing values preserve the variability and cyclicity of the data and the error in the estimate can be calculated.

The model used in this investigation comprised two or three terms. In general, an observation $y(k)$ is expressed as

$$y(k) = t(k) + p(k) + e(k)$$

where t is a trend

p is a periodic component describing cyclic fluctuations about t

e is a zero mean, serially uncorrelated white noise component.

The periodic component was not used in some models.

Each component of the observation equation is described in terms of discrete time state equations for each component. These state equations are updated with each data point and therefore allow the equations to describe non-linear and non-stationary (changing mean) processes.

The trend model used was an integrated random walk (IRW), characterised by a variance σ^2_t. The noise component is characterised by variance σ^2_n. The behaviour of the system is then characterised by the noise variance ratio (NVR) σ^2_t/σ^2_n.

Possible periodic models include general transfer functions (similar to the ARMA models of Box and Jenkins) and dynamic harmonic regressions (DHR). The first are useful for quasi-periodic series, while the second are suited to strongly periodic data. DHR models the periodic component as a series of sine and cosine terms of specified frequencies. In the case of some of the data sets examined here, the periodicity is known to be related to daylight hours and therefore the DHR technique is more appropriate. Periodograms were prepared to identify the frequencies to use; in addition to the expected 24 hour diurnal period, harmonics of this were identified as a consequence of the variation being not exactly sinusoidal.

Algorithms using these techniques are incorporated in the package *micro*CAPTAIN. This package was used to examine three datasets, ammonia data from James Bridge and Thurmaston, and dissolved oxygen data from Thurmaston.

The ammonia data from Thurmaston contain frequent, large calibration spikes. As shown previously, spike removal algorithms can detect and eliminate most, but not all, of these spikes. These data were smoothed using an IRW model with NVR=0.05, Figure 5.4. This has not succeeded in removing the spikes, but there is a considerable reduction in amplitude, from around 1 mg/l to around 0.2 mg/l. A real event, of around 0.8 mg/l, is now clearly distinguishable.

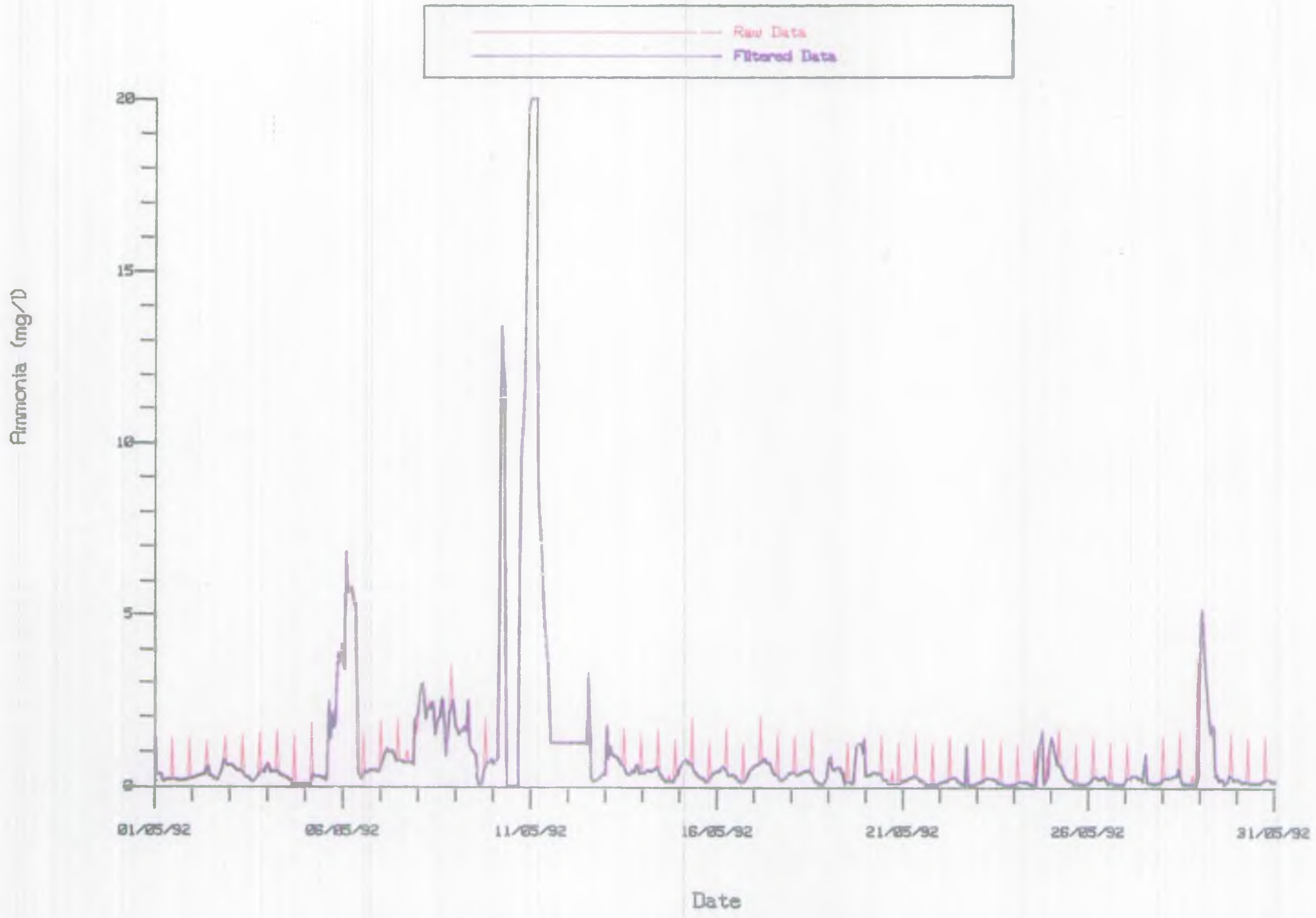


Figure 5.2 - De-spiking using 3 point Algorithm, James Bridge Ammonia

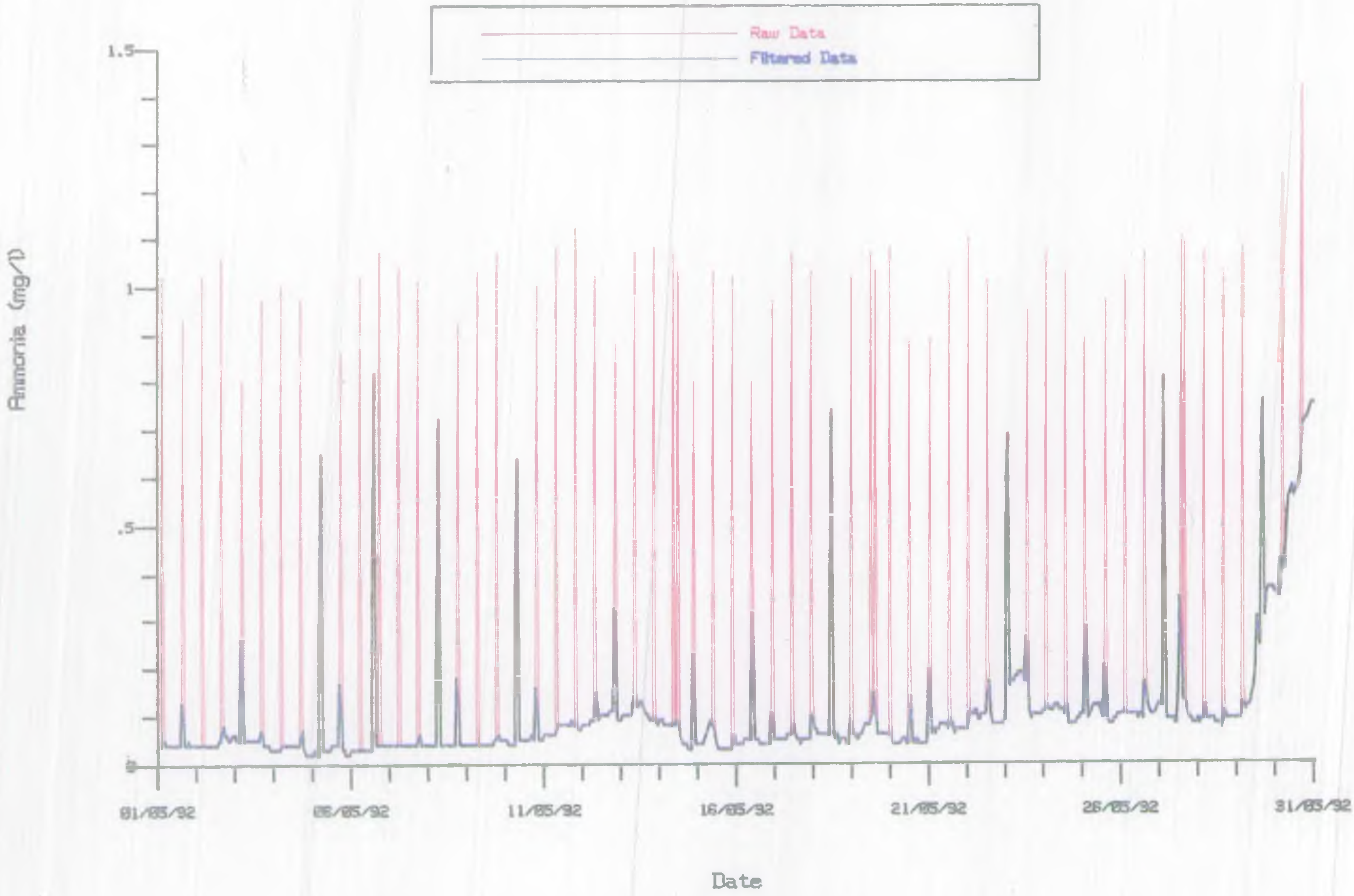


Figure 5.3 - De-spiking using 3 point Algorithm, Thurmaston Ammonia

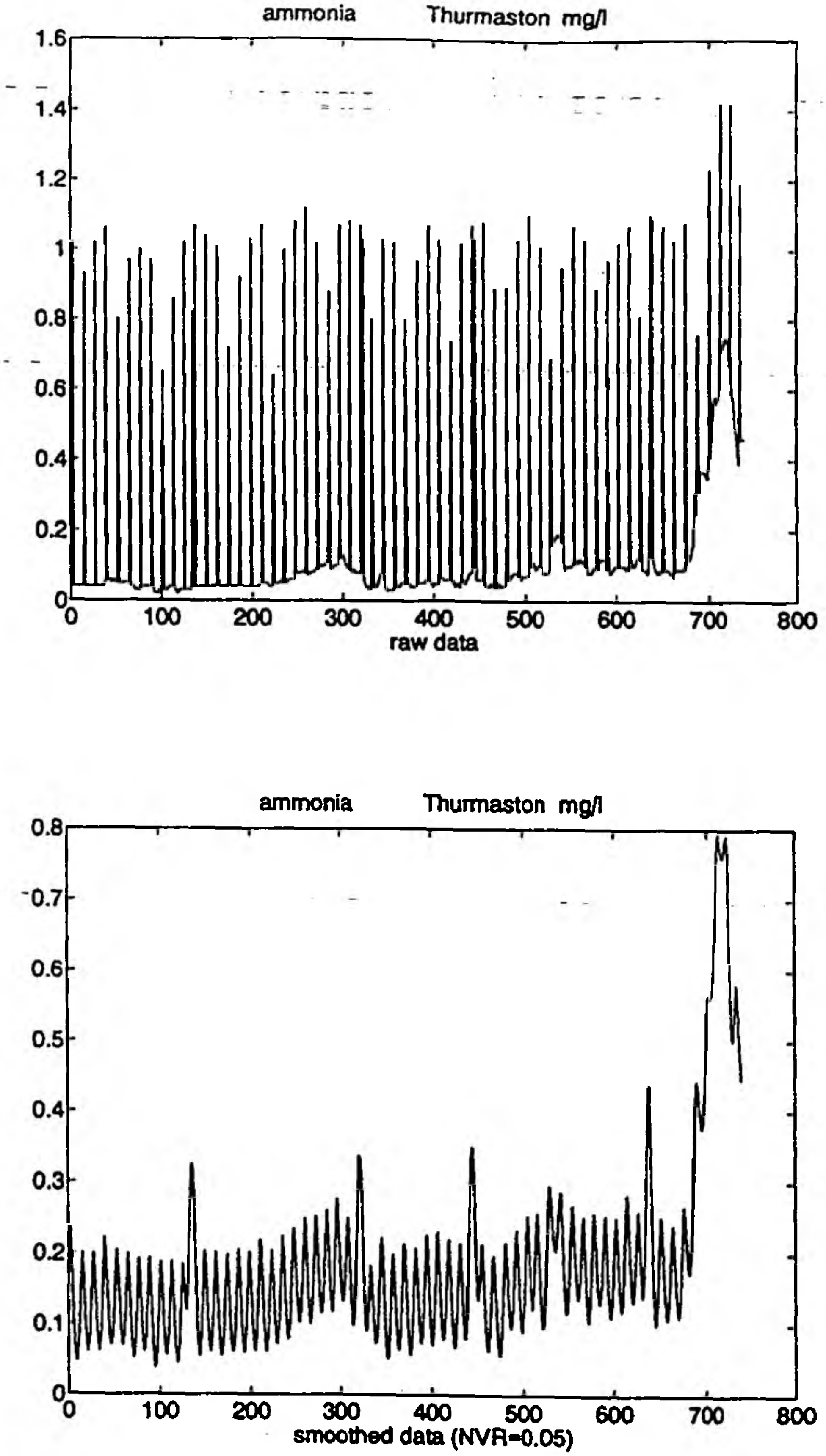


Figure 5.4 - IRW Smoothing of Thurmaston Ammonia Data

To test an alternative approach, the data from James Bridge were de-spiked by hand, leaving a large number of missing data points. These data were then interpolated using IRW. The results are shown in Figure 5.5. Application of smoothing to the raw data is very successful in this case. To illustrate the success, Figure 5.6 shows the model residuals, i.e. what was removed by smoothing. The calibration spikes are clearly seen, but some of the large, possibly real, event is also removed.

The dissolved oxygen data from Thurmaston are clearly periodic, but also contain some noise and three negative spikes, Figure 5.7. These data were processed into a three component model comprising trend, periodicity, and residuals. These three components are separated in Figure 5.7. The underlying trend is now clear. It may be considered that this trend itself shows some periodicity now that the diurnal fluctuation has been removed. A periodogram showed a component with a period of around three days to be present. This may represent the variation of insolation with weather patterns. This is a particular case of a frequent dilemma with filtering; what to leave in, and what to leave out? The answer to this will depend on the use to which the data are to be put. There is a good case for preserving the original data, as well as filtered data.

Filtering of periodic data can be achieved to some extent by using a sufficiently long moving average filter. However, this acts very indiscriminately, removing all high frequencies and severely attenuating short term events. Long averages also have the disadvantage of either producing long time delays, or requiring future data. Some of these affects can be seen in Figure 5.8. The Kalman technique overcomes these disadvantages.

5.3.4 Change Detection

The simplest method of change detection is a simple threshold test - a change is indicated if any measurement is above or below a predetermined value. This is very simple to implement and responds as soon as such a value is detected. It has the disadvantage of being very vulnerable to spikes, and also needs careful adjustment of the threshold to take account of seasonal change. Adjustment of seasonal change can be done with respect to trends isolated by Kalman filtering.

A more sophisticated approach is a cumulative sum test, such as the Hinkley test. This effectively counts ups and downs and changes value only when ups or downs are occurring consistently. The sensitivity can be varied in relation to mean value. The insensitivity of this method to spikes can be demonstrated with respect to the James Bridge ammonia data, Figure 5.9. It was unsuccessful with the Thurmaston ammonia data, where spike amplitude exceeds data amplitude, but works well when spike amplitude is first decreased using IRW smoothing, Figure 5.10. A further application is shown with respect to the dissolved oxygen data from Thurmaston, where the Hinkley test is applied to the trend from DHR smoothing, Figure 5.11. The effect of the trend separation is clearly seen here, since the Hinkley parameter does not always change in the same sense as the immediate change in data.

One penalty of this technique is that there is some delay. This is a inevitable consequence of an algorithm that waits for some more data to be sure that a change is occurring.

5.3.5 Correlation Analysis

Examination of the available datasets has shown that there are a number of correlations within them. Changes in these correlations may indicate exceptions, even though the parameter values themselves are not unusual.

Two kinds of correlation were seen:

- diurnal cycles (only seen in the summer half of the year). Temperature, dissolved oxygen, and pH are all positively correlated;

- high flow events. Turbidity is positively correlated, conductivity negatively correlated. Temperature, pH, and dissolved oxygen show weak negative correlations in summer for the River Soar.

Example correlations are given in Tables 5.1, 5.2. The diurnal cycles in temperature, dissolved oxygen and pH arise from a common origin but there is no direct relation between them and therefore the correlation may change. Changes in correlation can be examined in a number of ways. The simplest is by examination of scatter plots;

other techniques are principal component analysis and the calculation of regression probabilities. Scatter plots are quite informative, particularly if the points are plotted in time order, but they are not amenable to automatic event detection.

Principal component analysis was investigated as a more automated technique. For two variables, the first principal component captures the greatest variability in the data, and the second the variation orthogonal to this. The two principal components are presented in Figures 5.12, 5.13. Unfortunately, there is no clear difference in their nature, and nothing that may be called an exception. The period May-September was examined visually, without detecting anything unusual.

Table 5.1 Correlation matrix

3rd - 9th May 1992, Raw Data from James Bridge

		DEPENDENT					
		Am	**pH**	**SS**	**Cond.**	**DO**	**Temp**
INDEPENDENT	**Am**	1.00	0.03	0.01	0.04	0.14	0.05
	pH		1.00	0.01	0.01	0.20	0.10
	SS			1.00	0.01	0.01	0.05
	Cond.				1.00	0.02	0.09
	DO					1.00	0.56
	Temp						1.00

Am -	Ammonia
SS -	Suspended Solids
Cond -	Conductivity
DO -	Dissolved Oxygen
Temp. -	Temperature

Table 5.2 Correlation matrix

13th - 18th May 1992, Raw Data from James Bridge

		DEPENDENT					
		Am	**pH**	**SS**	**Cond.**	**DO**	**Temp**
INDEPENDENT	**Am**	1.00	0.29	0.05	0.01	0.37	0.57
	pH		1.00	0.00	0.04	0.57	0.23
	SS			1.00	0.04	0.01	0.01
	Cond.				1.00	0.07	0.20
	DO					1.00	0.53
	Temp						1.00

Am -	Ammonia
SS -	Suspended Solids
Cond -	Conductivity
DO -	Dissolved Oxygen
Temp. -	Temperature

The connected scatter plot, Figure 5.14, reveals more of the reason for this. It is clear that there is a very strong correlation of dissolved oxygen with temperature for short periods of time (a few days) and the slope of the correlation is reasonably constant. However, every few days the intercept changes considerably. Effectively, although daily variation of temperature and dissolved oxygen are correlated, variations over a few days are not. This is not surprising since water temperature depends on air temperature as well as insolation, whereas oxygen production does not depend strongly on air temperature but it is reduced by cloudy weather.

Significant changes in flow are produced by rainfall events. There is also more river flow in winter than in summer, this being particularly pronounced for the Tame. High flow events are associated by a reduction in conductivity as a result of dilution. The duration of a low conductivity event is closely linked to the duration of the high flow event. High turbidity events are usually linked with high flow events, but in this case the duration of the event is less than that of the flow event. This is probably caused by surface runoff, which lasts for only for the duration of the rainfall event.

A reduction of pH, temperature and dissolved oxygen occur during high rainfall events in summer. This may be an indirect effect, caused by the cloudiness associated with the rainfall event. The River Tame shows an increase of dissolved oxygen during winter high flow events, over a rather low background value. This probably represents the increased turbulent entrainment of air during high flow events.

Thus it becomes necessary to identify those changes that are correlated as a result of natural processes, and those that are correlated or not correlated as a result of unnatural processes.

5.4 Natural Phenomena

As mentioned in Section 5.2, it is important to distinguish between changes in water quality caused by natural phenomena, such as thunderstorms, which raise the level of suspended solids, and pollution incidents which may cause damage to aquatic life or contaminate drinking water supplies. It is acknowledged that storms can cause pollution incidents through sewage overflowing into storm water outlets, bunds or slurries overflowing or general run-off from polluted surfaces.

Hence, changes in suspended solids, conductivity and pH which are associated with rapid changes in flow should not necessarily be used for raising alarms. However ammonia and dissolved oxygen data for this period should not be removed from the analysis for the reasons given above.

5.5 Incident Detection

5.5.1 Analysis of a Known Pollution Incident

A known pollution incident, involving release of sewage into the River Soar, occurred at the end of June 1992 and was detected by measurements from Thurmaston, Sileby and Kegworth water quality monitoring stations. Time histories of the pollution incident for seven parameters are given in Figures 5.15-5.21 (for convenience the remaining figures in this section follow page 70). Data for this pollution incident were provided by Severn Trent NRA together with matching flow data from the hydrometric station at Pilling, the closest site to Thurmaston.

Examination of the flow data (Figure 5.21) shows a large and rapid increase in flow in the evening of June 30th followed by a smaller broader peak in flow in the second half of July 1st. There is then a third sharp increase in flow which lasts somewhat longer from the afternoon of July 3rd until 24 hours later. The increase in flow is matched by:

- an increase in turbidity: the exceptional rise in turbidity on 30th June at Thurmaston may be due to a storm following a long dry period. All high flow periods increase the turbidity at Sileby and to a lesser extent at Kegworth.

- a decrease in temperature caused by rainwater run-off, evident at all three sites.

- a decrease in conductivity caused by the diluting effect of the rainwater, evident at all three sites.

- a decrease in pH which is most pronounced during the first high flow period.
- a decrease in dissolved oxygen which shows a very close correlation with flow at Sileby.
- an increase in ammonia which is closely correlated with flow at Sileby and, about 36 hours later, at Kegworth.

The increase in rainfall and flow will cause an increase in turbidity from run-off and greater turbulence in the river. The diluting effect of the rainwater will also decrease conductivity and temperature. As discussed below, a fall in temperature will have an impact upon dissolved oxygen and then pH. However, some of the impact may be due to pollutants entering the river. The increase in ammonia is particularly evident at Sileby and Kegworth which are both downstream of the Wonlip sewage works. An increase in storm water will lower the efficiency of the sewage works and therefore increase the level of ammonia in discharges. It would appear that discharges from storm tanks has increased the level of ammonia detected at Sileby. This would explain why the second high flow event, whilst peaking at a flow 70% of the first event, produces a higher ammonia concentration.

This event was evaluated by comparison with the normal behaviour of the river during stable weather conditions.

A three month period, June-August 1989, was selected for further analysis. Calibration cycles were removed using the method described in Section 5.3.2. Time history plots of each parameter (Figures 5.22 - 5.28) were compared to identify 'stable' periods containing good quality data but excluding natural or man made pollution incidents as described above. The following stable periods were selected:

June 8th, 14:00	-	June 16th, 06:00
June 17th, 01:00	-	July 6th, 24:00
July 11th, 01:00	-	July 17th, 24:00
July 20th, 01:00	-	July 21st, 24:00
July 30th, 01:00	-	August 18th, 24:00

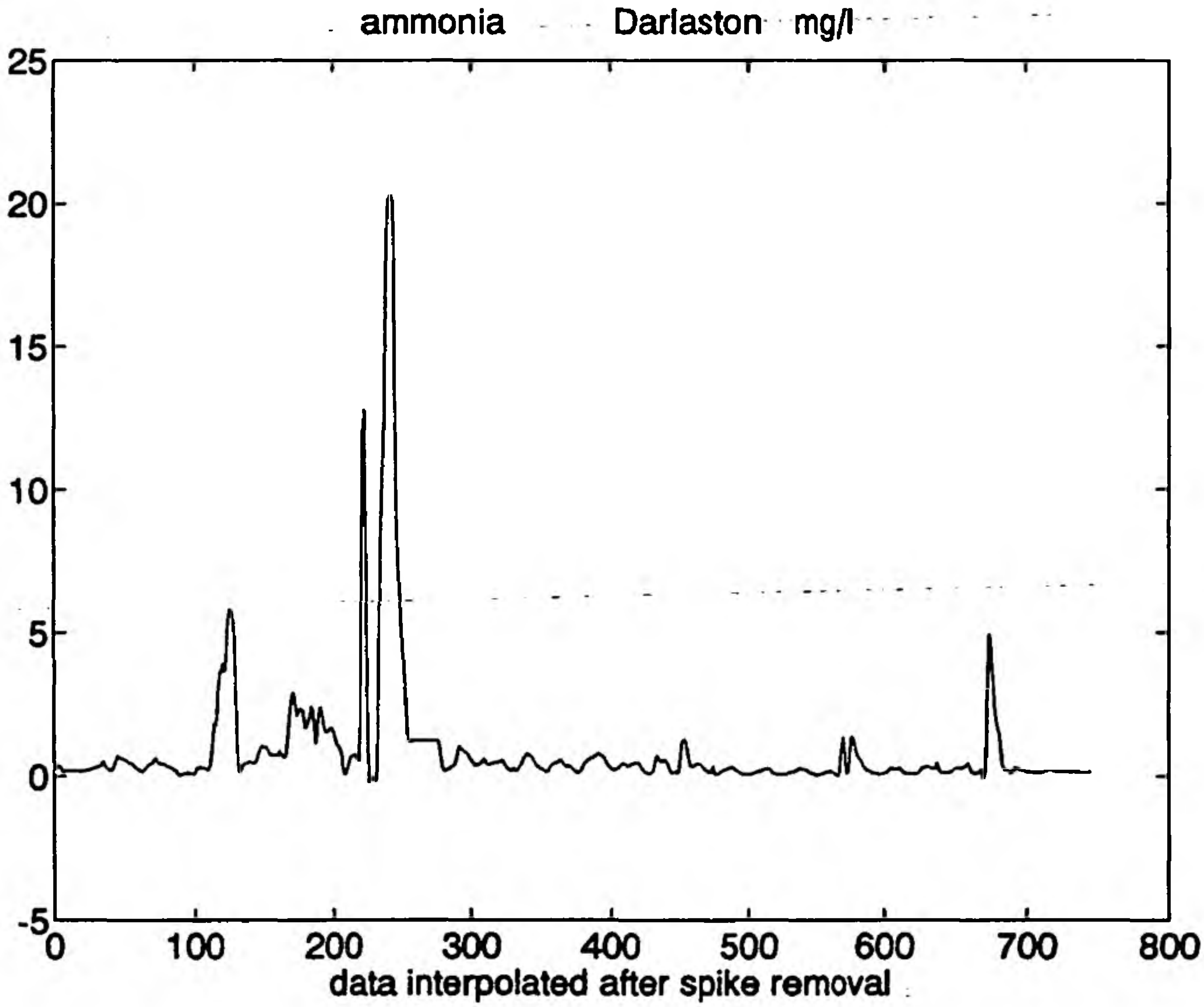


Figure 5.5 - IRW Interpolation James bridge Ammonia Data

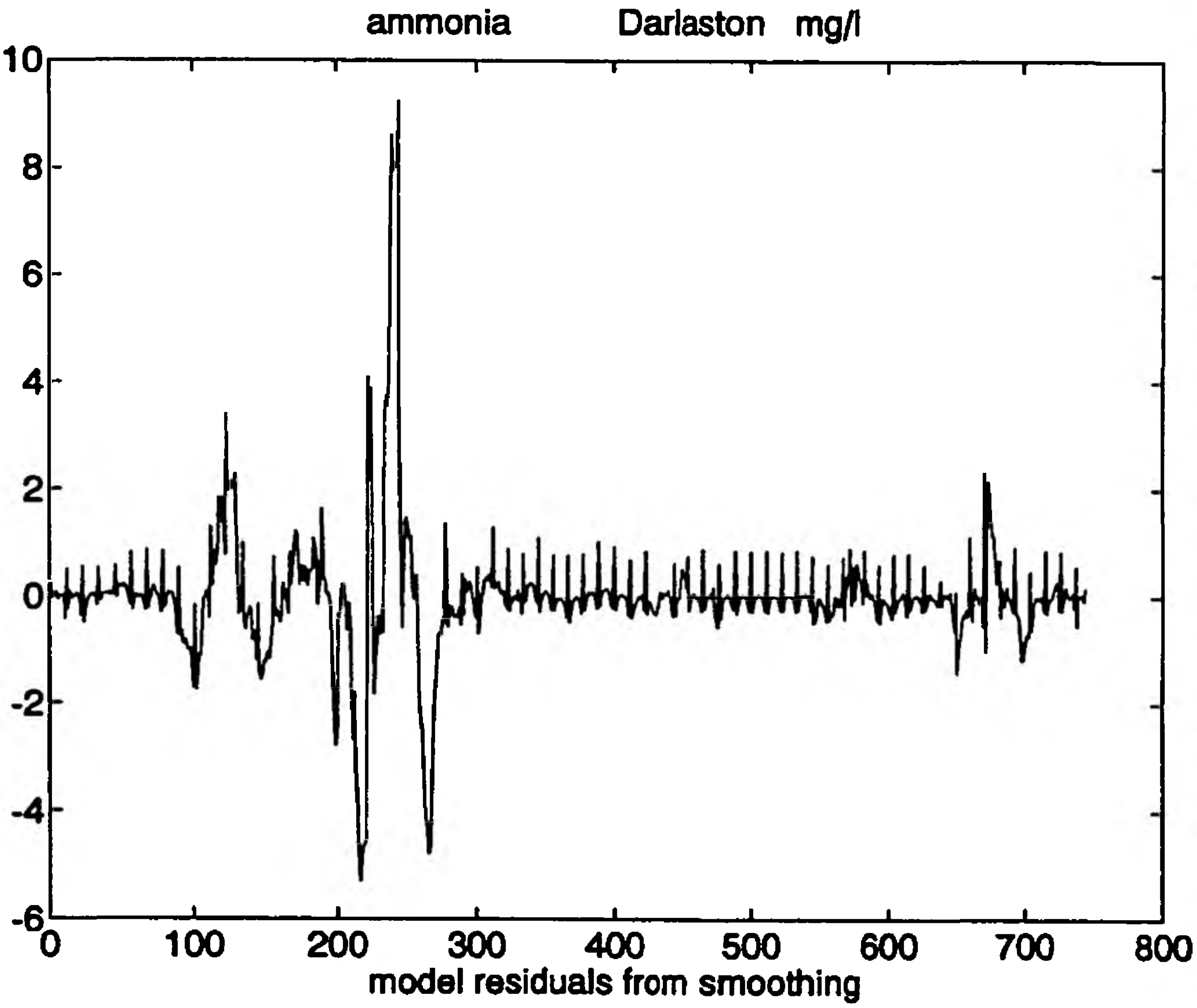


Figure 5.6 - IRW Smoothing Residuals from James Bridge Ammonia Data

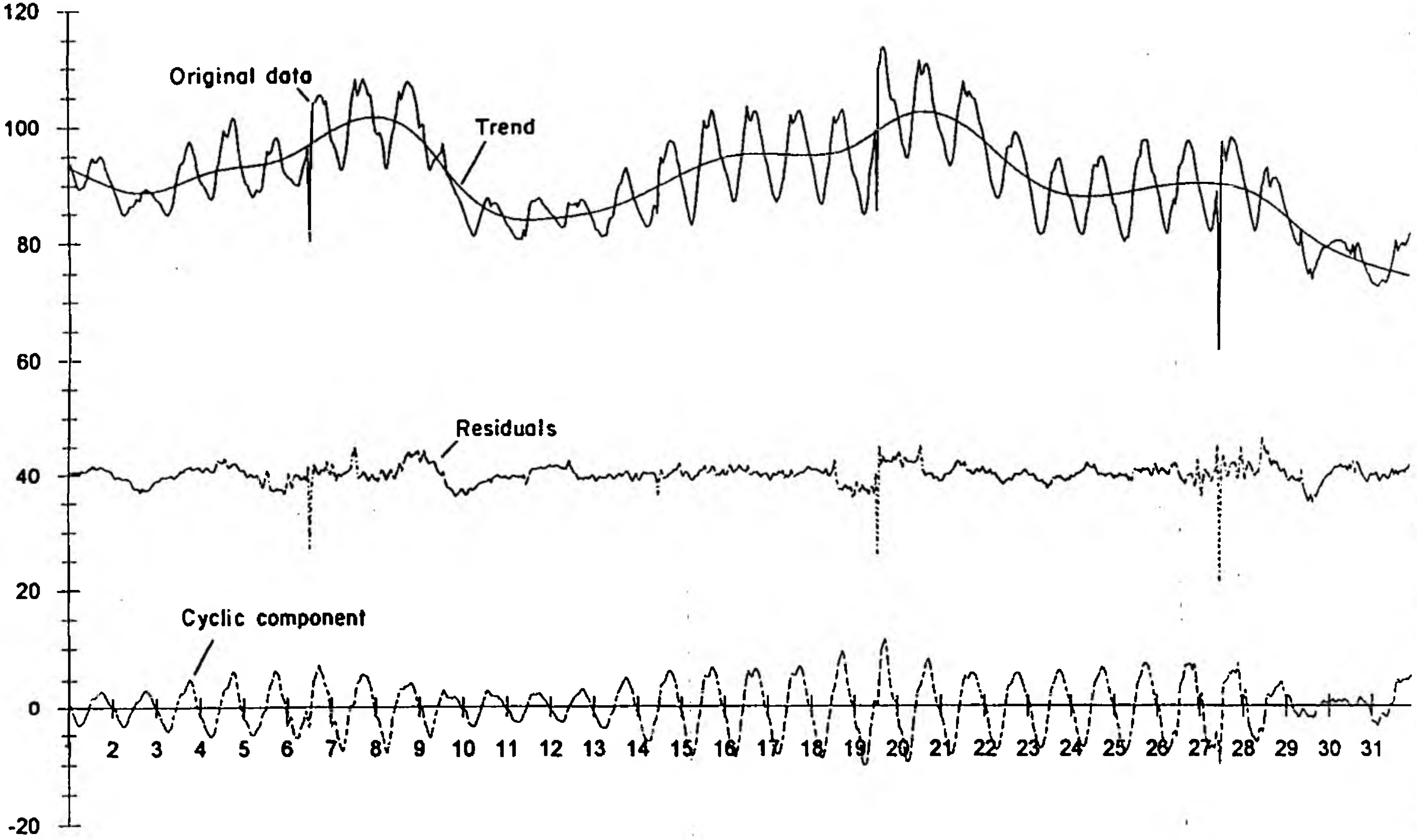


Figure 5.7 - DHR Analysis of Thurmaston Dissolved Oxygen Data, May 1992

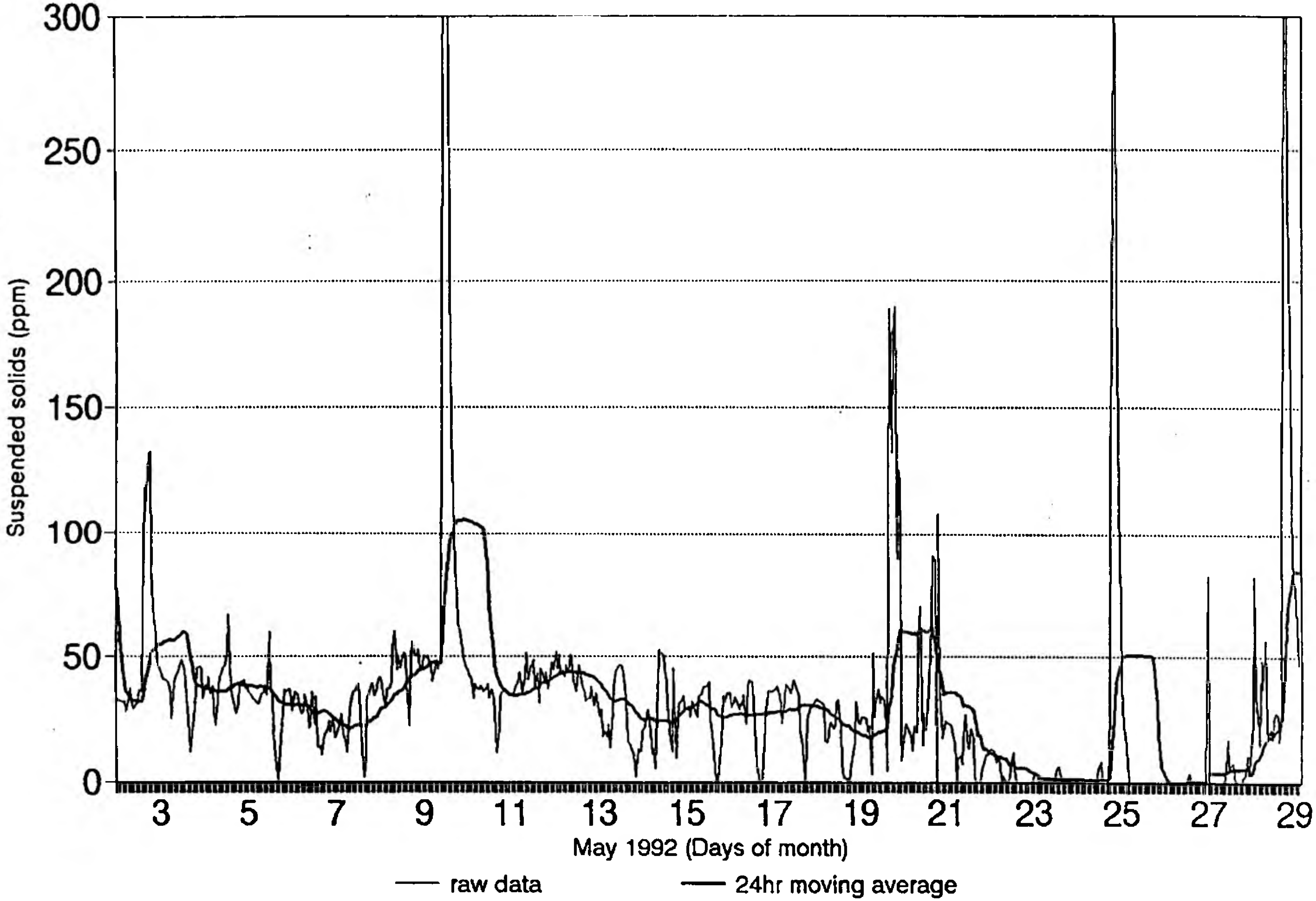


Figure 5. 8 **Time history of suspended solids, James Bridge, May 1992, s howing 24 hour moving average**

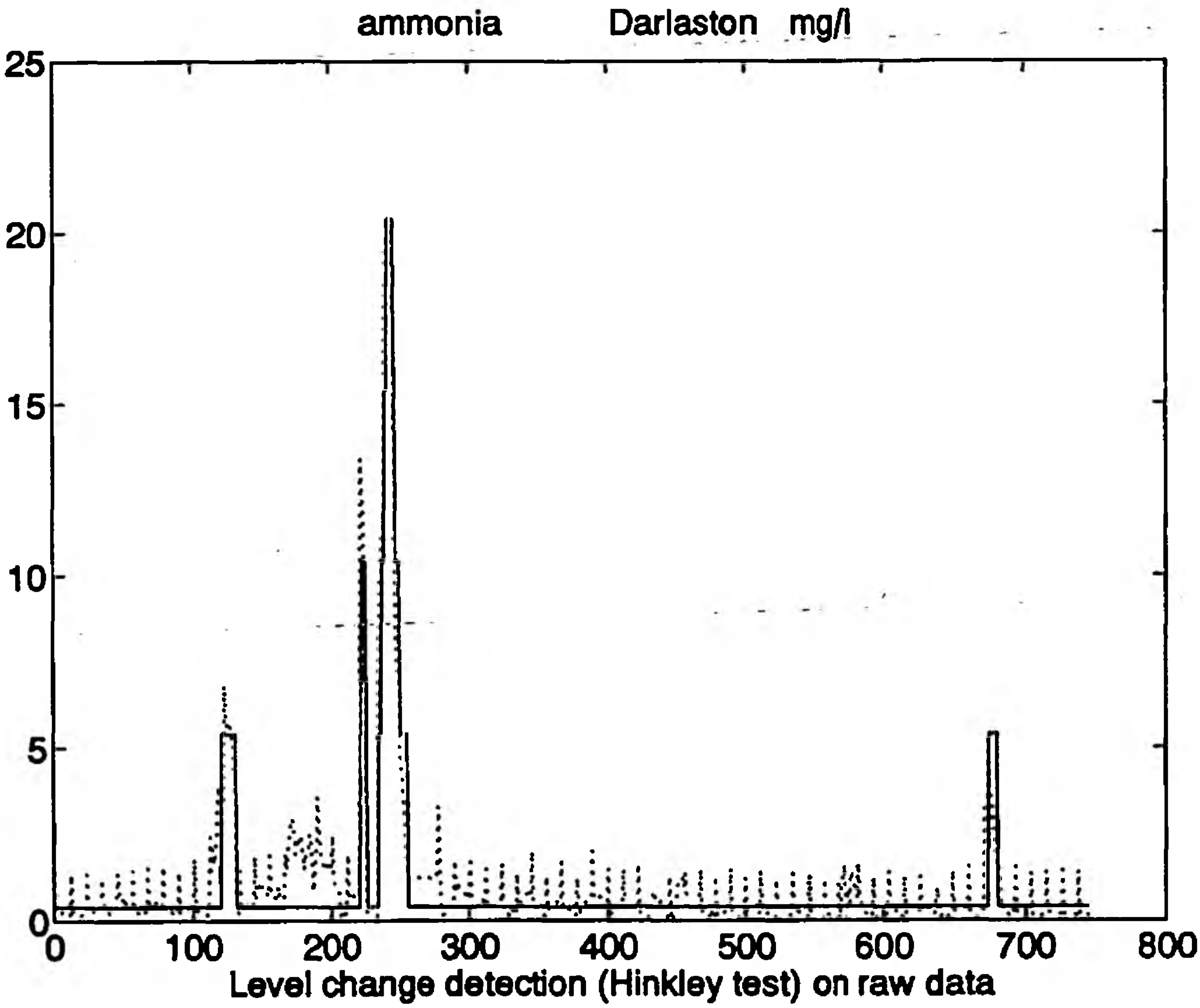


Figure 5.9 - Hinkley Test on James Bridge Raw Ammonia Data

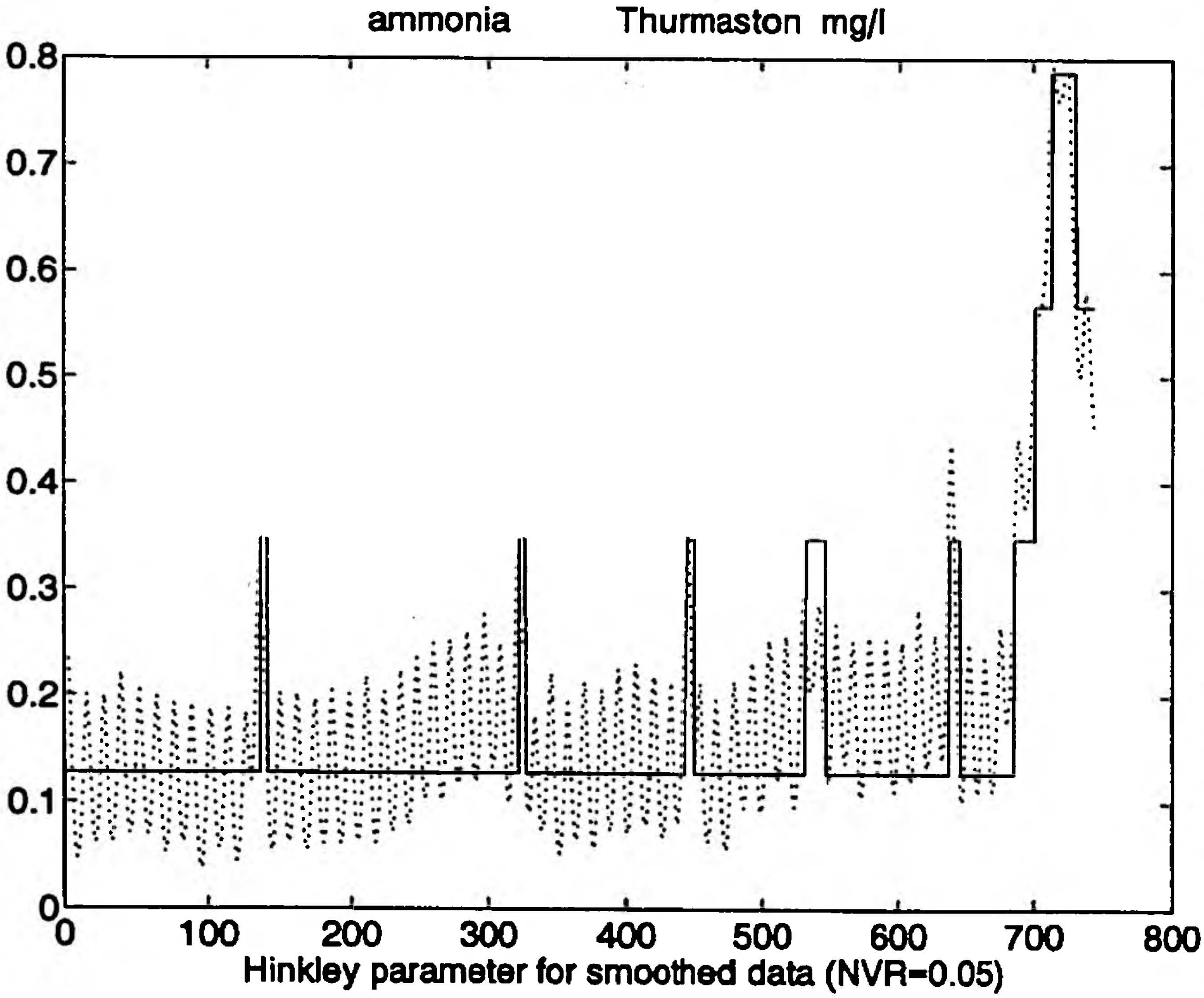


Figure 5.10 - Hinkley Test on IRW Smoothed Ammonia Data for Thurmaston

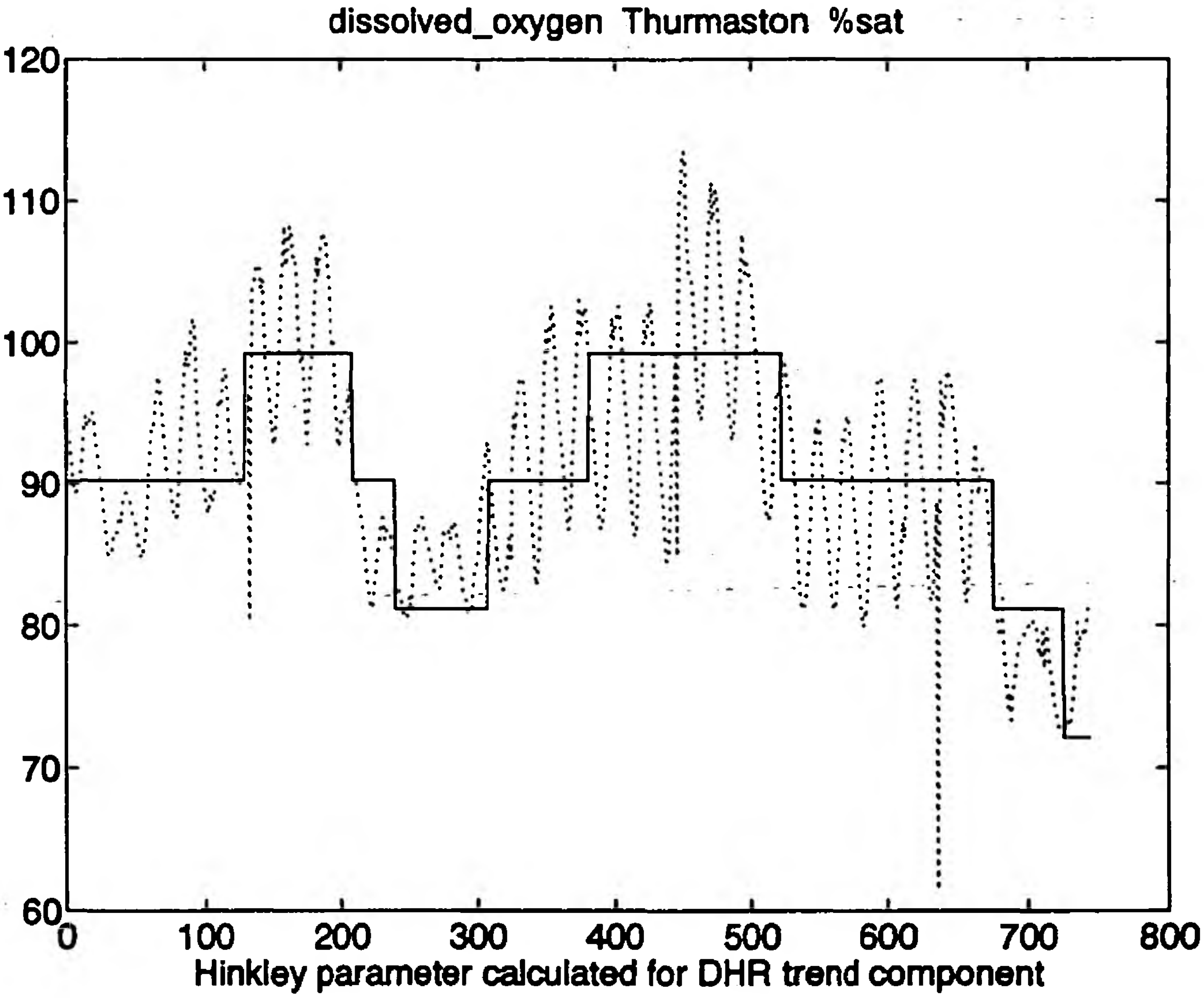


Figure 5.11 - Hinkley Test on Trend from Thurmaston Dissolved Oxygen Data

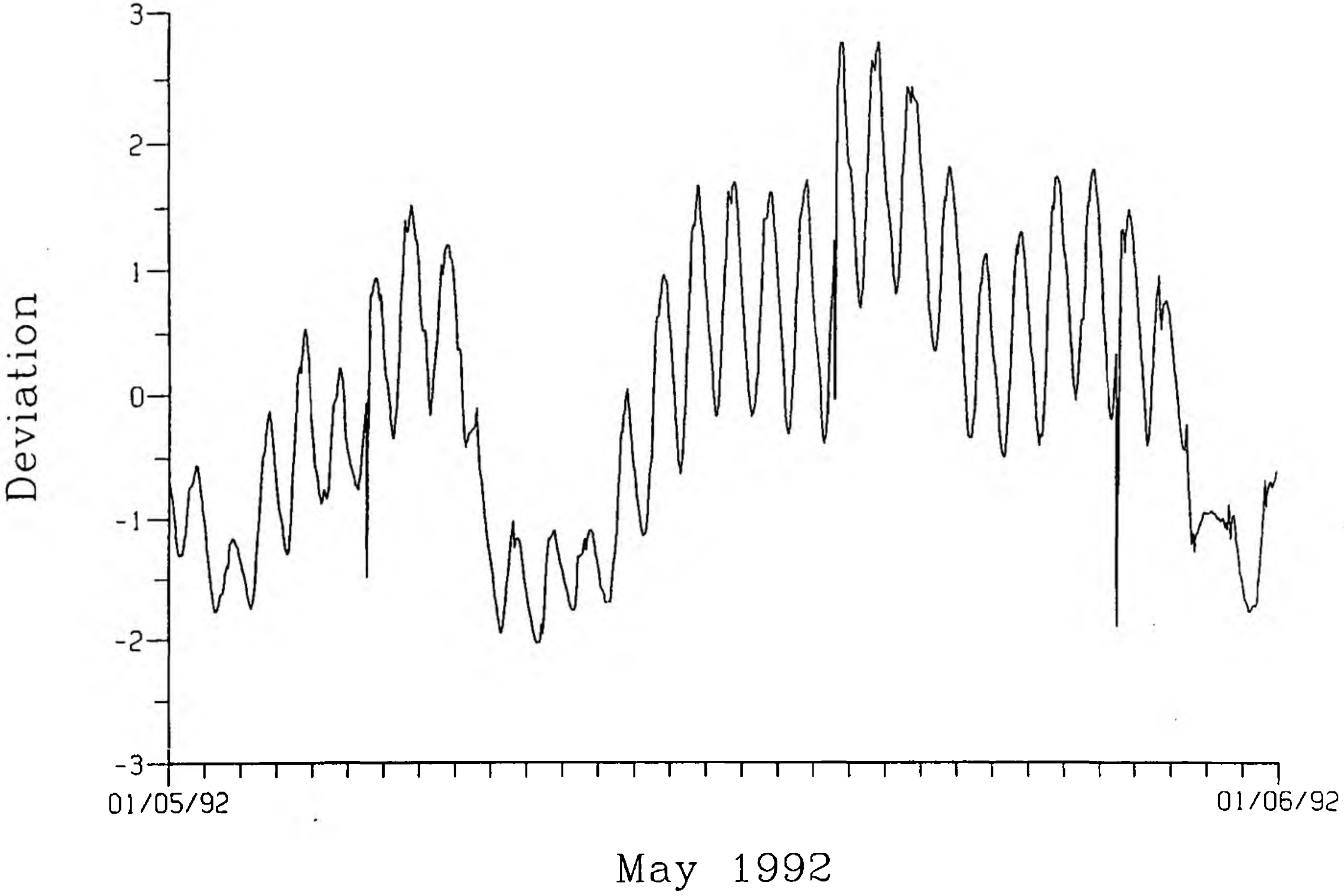


Figure 5.12 - First Principal Component for Dissolved Oxygen and Temperature, Thurmaston

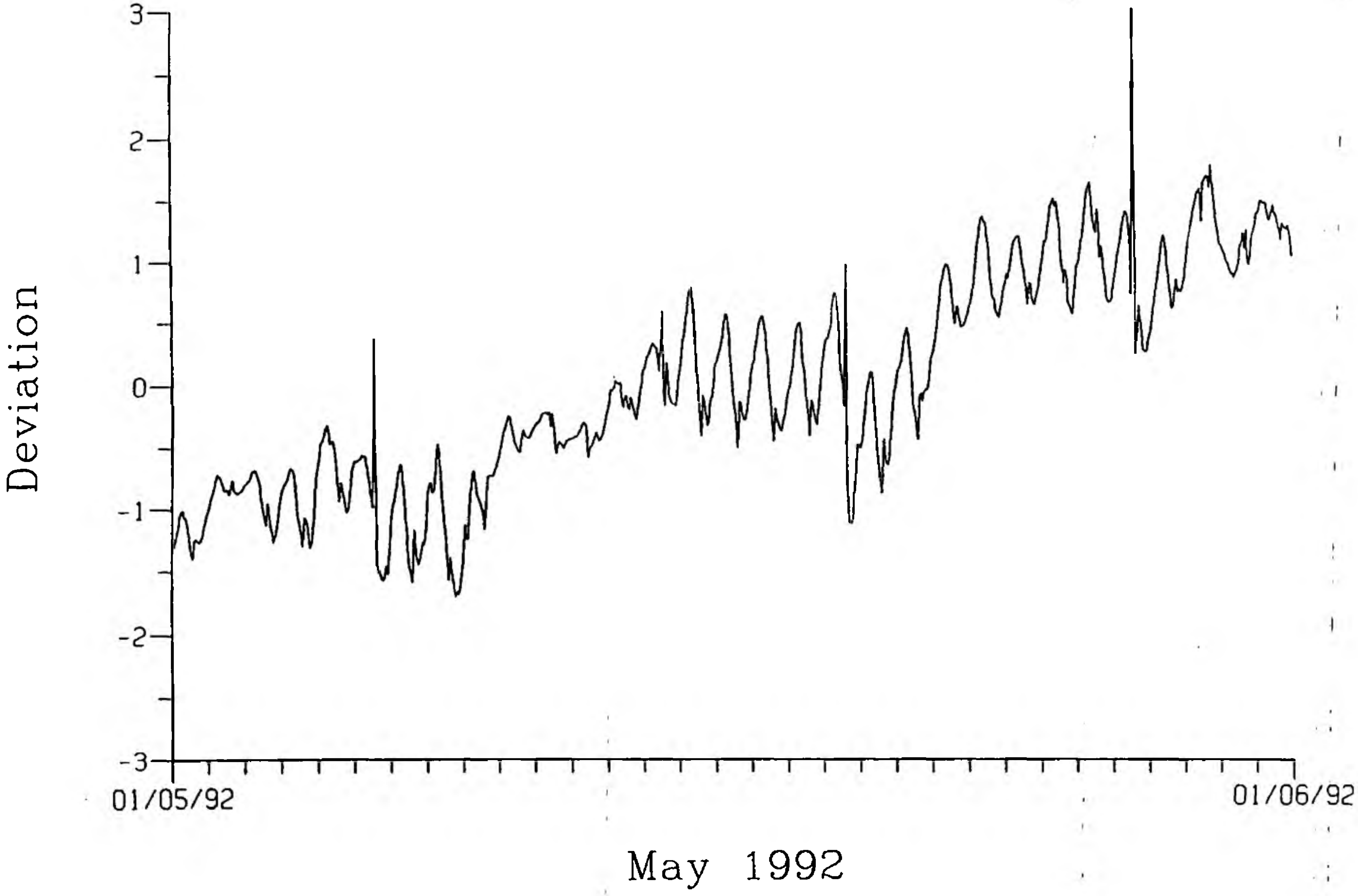


Figure 5.13 - Second Principal Component for Dissolved Oxygen and Temperature, Thurmaston

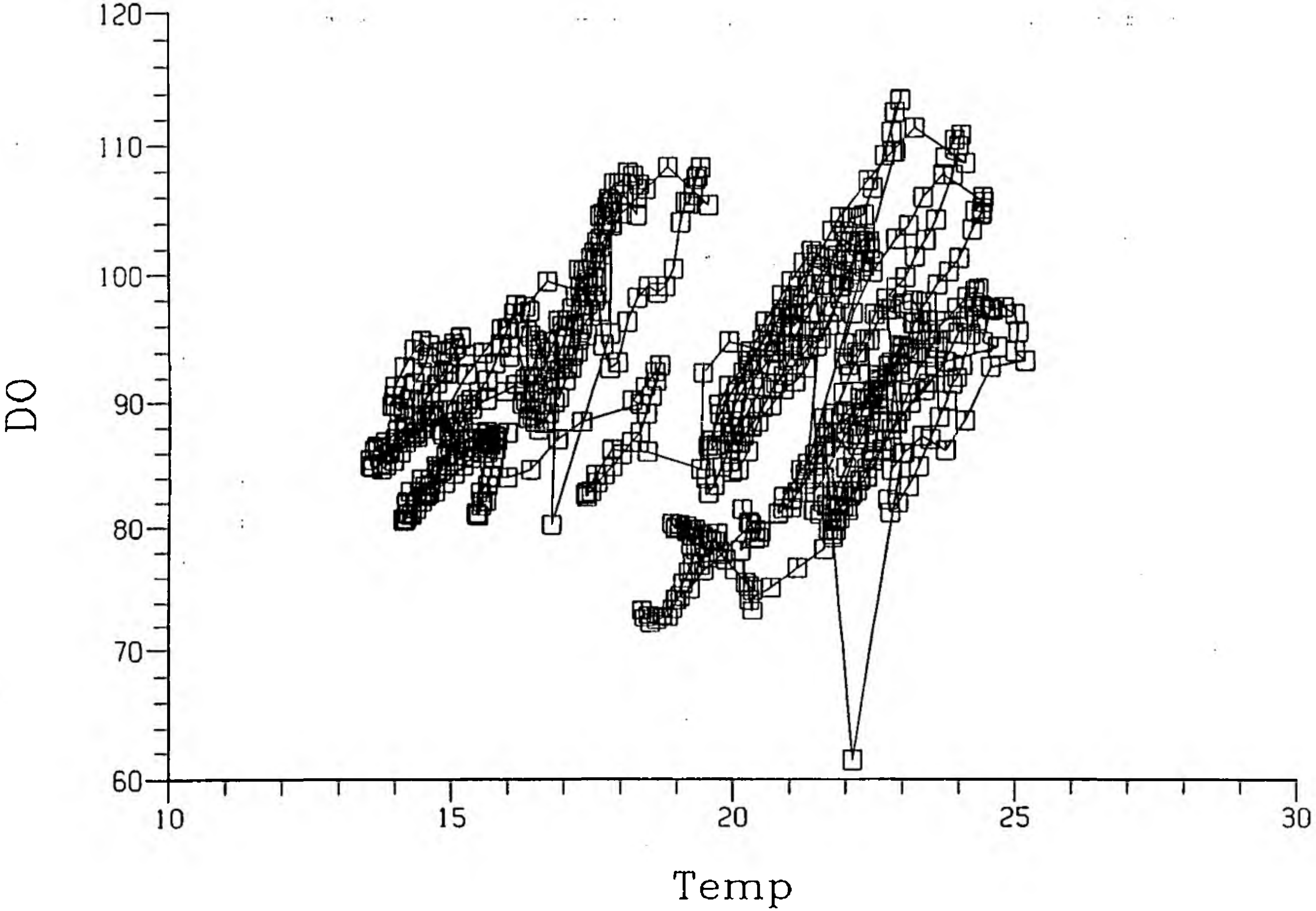


Figure 5.14 - Corrected Scatter Plot, Dissolved Oxygen and Temperature, Thurmaston

August 20th, 01:00 - August 25th, 24:00
August 30th, 01:00 - September 1st, 24:00

Correlations between pairs of parameters were explored further by plotting each parameter against all other parameters for the table data. These plots define an envelope of points, which represents a 'normal' or typical pollution state of the River Soar at the Thurmaston site. Temperature is relatively independent of a pollution incident except where thermal pollution occurs. Hence, temperature as one of the parameters in a two parameter plot facilitates comparisons with incident data. Examination of the plots showed a clear relationship between the following parameters and temperature (Figures 5.29 - 5.31):

- dissolved oxygen
- conductivity
- pH

These correlations are also evident in comparisons of pH vs dissolved oxygen, dissolved oxygen vs conductivity and conductivity vs pH as would be expected. Hence, during the stable periods, there is no underlying relationship with the other parameters and turbidity, flow and ammonia. In addition, a regression coefficient was calculated for each parameter pair and the results are presented in Table 5.3. The table confirms the strong correlations revealed by the plots.

The pollution incident data for Thurmaston has also been plotted on Figures 5.29 - 5.34. The parameters correlated with temperature are of particular interest and the following features are evident for the pollution incident:

- **dissolved oxygen vs temperature (Figure 5.29):** the pollution incident data lies well outside the envelope of stable data with dissolved oxygen lower than would be expected for a given temperature. With the passage of time the dissolved oxygen levels return to normal within the stable envelope. However it should be noted that most of the incident data is within the normal range of temperature and dissolved oxygen values and it is only a two-parameter plot which reveals clearly the excursion from the norm.

- **conductivity and temperature (Figure 5.30):** a similar situation to that described above with lower values of conductivity than would be expected for a given temperature. The incident data points also reflect the time history of the development of the incident progressing from the high conductivity/high temperature values down to the low conductivity/temperature values.

Table 5.3 Correlation Matrix: Stable Periods, Thurmaston

June - August 1989

		DEPENDENT					
		Am	pH	Turb.	Cond.	DO	Temp
INDEPENDENT	Am	1.00	0.07	0.01	0.04	0.07	0.20
	pH		1.00	0.02	0.54	0.73	0.65
	Turb.			1.00	0.01	0.00	0.01
	Cond.				1.00	0.21	0.51
	DO					1.00	0.56
	Temp						1.00

Am -	Ammonia
Turb. -	Turbidity
Cond -	Conductivity
DO -	Dissolved Oxygen
Temp. -	Temperature

- **pH and temperature (Figure 5.31):** a similar picture to those described above with lower pH than would be expected for a given temperature.

Comparison of incident data with data from the stable periods appears to show that significant change to water quality parameters can occur due to storms which increase the flow in a river whilst not necessarily creating a pollution incident. It is therefore important to examine flow data in the first instance.

A pollution incident may be detected from an increase in ammonia concentration above the threshold value, or rate of change of concentration, and/or from the two-parameter plots of dissolved oxygen vs temperature and confirmed against the pH vs temperature plot.

The two-parameter plot method shows promising results when used for a major pollution incident. Further tests for potential pollution events were performed during the period June - August 1989 at Thurmaston i.e. using some of the 'non-stable' data. (see next sub-section).

5.5.2 Analysis of other Incidents

During the period June - August 1989, at Thurmaston, four different events are evident from the sharp increases in turbidity (Figure 5.27).

- 7th - 8th June
- 7th - 11th July
- 22nd - 29th July
- 26th - 28th August

These events were analysed following the same procedure described in Section 5.5.1. Two-parameter plots of the stable and incident data together were produced for the parameters listed in Table 5.4. The figures are given in Appendix B.

Table 5.4 Two Parameter Plots for Four Events in the Period

June - August 1989

Date	Figure Numbers
7th - 8th June	Figures B.1 - B.7
7th - 11th July	Figures B.8 - B.14
22nd - 29th July	Figures B.15 - B.21
26th - 28th August	Figures B.22 - B.28

Table 5.5 below summarizes the results of the analysis using the following notation:

T - temperature	N - normal values
C - conductivity	H_1 - higher than normal values
DO - dissolved oxygen	H_2 - much higher than normal
NH_3 - ammonia	L_1 - lower than normal
t - turbidity	L_2 - much lower than normal
V - velocity	

Parameter values during an event are evaluated relative to values preceding and succeeding the event in addition to average values during the event. For the correlations the evaluation relates dissolved oxygen, pH and conductivity to temperature. e.g. L_2 against the DO vs T column means that dissolved oxygen is much lower than would be expected for a given temperature.

The interpretation of each event is discussed below. Note that the flow data plotted in Figure 5.29 are daily averages and may therefore mask short severe thunderstorms.

Table 5.5 - Summary of Analysis of Four Events in the Period

June - August 1989 at Thurmaston

Event No	Period	Parameter Condition							Correlations		
		T	C	pH	DO	NH_3	t	V	DOvsT	pHvsT	Kvs T
1	7-8 June	L_1	L_2	N	N	N	H_2	?	H_1	H_1	N
2	7-11 July	L_2	L_2	L_1	L_1	L_2	H_2	H_2	N	N	L_1/N
3	22-29th July	N	L_2	L_2	L_2	H_1	H_2	N	L_2	L_2	L_2
4	26-29 Aug	L_1	L_1	L_1	L_1	N	H_1	H_1	N	N	N

Event 1 (7 - 8 June): Unfortunately, during this period no flow data is available. However, the sudden rise of turbidity combined with a drop in temperature and conductivity indicates heavy rainfall leading to high flow. pH and dissolved oxygen remain at near normal levels but with a reduction in the amplitude of the diurnal variation in dissolved oxygen reflecting a similar reduction in diurnal variation of temperature. pH and dissolved oxygen do not fall to match the fall in temperature. It is possible that the initial run-off and river turbulence increases dissolved oxygen.

Ammonia remains at normal levels although it does show greater variability possibly because of the reduced efficiency of sewage works or similar discharges upstream.

It is concluded that because there is no marked increase in ammonia or decrease in dissolved oxygen/pH, taking temperature variations into account, this is a natural event which has not lead to a reduction in water quality.

Event 2 (7 - 11 July): In this event, high flow has increased the turbidity, lowered the temperature and lowered the conductivity of the river. Associated with the fall in temperature, dissolved oxygen and pH have also fallen.

The two-parameter plots show that the reduction in dissolved oxygen and pH is purely a consequence of meteorological conditions, as evidenced by the fall in temperature, since they still remain within the stable envelope of the plots. Much of the conductivity change can be explained by the diluting effect of the rainwater run-off, as shown in the conductivity temperature plot.

Ammonia levels rise sharply during this event which may be due to storm water overflows from treatment plant upstream of Thurmaston, brought about by a particularly severe storm.

It is concluded that a storm has created the increase in turbidity and has lead to pollution entering the river from storm water overflows. The pollution has increased ammonia but has not affected dissolved oxygen.

Event 3 (22 - 29 July): In this event there is again a sharp increase in turbidity and a decrease in conductivity associated with a small increase in flow on 22 July. Temperature falls towards the end of the period (around 25 July), but both pH and dissolved oxygen decrease significantly and much earlier (22 July). This becomes clearer when the two-parameter plots are examined i.e. dissolved oxygen and pH are lower then would be expected for a given temperature. The changes are quite marked for the small increase in flow. It could mark a period of hot, sultry weather with heavy overcast. However, here is a moderate increase in ammonia on 22 July which decreases only slowly.

It is concluded that there may have been a release of pollutants into the River Soar on 22 July. Turbidity is increased, dissolved oxygen reduced, and ammonia increased. All but the ammonia increase have possible natural explanations but the pattern of behaviour is generally unusual.

Event 4 (26 - 28 August): This is similar to Event 1 but with less severe conditions. A moderate increase in flow has increased the turbidity and lowered the conductivity and temperature. The corresponding falls in dissolved oxygen and pH can also be seen and confirmed in the two-parameter plots to be in the normal range. Ammonia levels remain normal.

It is concluded that turbidity has increased due to an increase in flow. Rainfall appears to have continued for 2 or 3 days as indicated by the fall in temperature. No reduction in water quality has resulted from the event.

5.6 Summarv Statistics

The NRA report on Guidance and Methods for Data Quality Control (National Rivers Authority, 1992) issued by the Steering Group on Data Handling has been reviewed. The report contains sections on:

- the Presentation of summary Statistics
- Methods for handling outliers

The methods for calculating and presenting summary statistics are satisfactory and we refer the reader to the NRA report.

Certain rules are presented for excluding and accommodating outliers. Various tests are used to identify outliers and methods are described for accommodating the outliers when calculating the mean, standard deviations and percentiles. Multiple outliers can be tested for with a recommended upper ceiling of n/10 outliers, where n is the data population. The text deals mainly with the identification of absolute outliers whereas the water quality data is principally time series data and, therefore, the techniques presented are not as powerful as they could be.

Once outliers have been identified, certain rules are presented for accommodating outliers. For example, when calculating a monthly mean, false data should be excluded from the calculation of summary statistics. This will not be affected by the daily time series of the data. However when calculating a daily mean, a more accurate interpolation is achieved if the time series is taken into account.

The work presented here has produced automatic methods for removal of outliers caused by spikes and equipment malfunctions, replacing them by missing value flags. A powerful estimation technique is available for interpolating these missing values, using Kalman filtering. This also produces an estimation error. This estimation error can be used to determine when interpolation should be discontinued.

5.7 Conclusions

Using expert knowledge of the behaviour of individual parameters, it is possible to assist the task of the water quality monitor operator. Since normal operation of the various sensors is well know, relatively simple rules or algorithms can be used to eliminate problems such as spikes, and to flag abnormal sensor operation. A simple Kalman filter using an IRW model is also adequate to remove calibration spikes for alarm purposes, and has the advantage of making very few assumptions about normal system behaviour.

A more sophisticated Kalman filter, incorporating a cyclic component, incorporates the knowledge that diurnal fluctuations are normal (at least for the two rivers examined) and allows the underlying trend to be examined with much greater sensitivity.

Examination of two parameter plots showed that the system can behave in a complex way, such that it is far from clear when an event constitutes an exception. This makes it very difficult to automate detection of such events. However, the ability to compare present data with data from the past does give the human operator more support when deciding whether an exception requires an alert.

The analysis in Section 5.5 appears to have been successful in identifying a known pollution incident and distinguishing less severe incidents from natural events. The most distinctive parameters are turbidity and ammonia. In summary the following steps should be taken:

1 - identify an event using a simple threshold value on turbidity;

2 - compare turbidity against flow, again using a simple threshold value;

3 - compare ammonia against a simple threshold;

4 - detect for dissolved oxygen, pH and conductivity values which fall below (temperature on X- axis) the envelope of stable values in the two-parameter plots for the few days preceding.

It would be possible to automate this procedure for alarm detection and generate appropriate graphics for examination by the operator as a back-up.

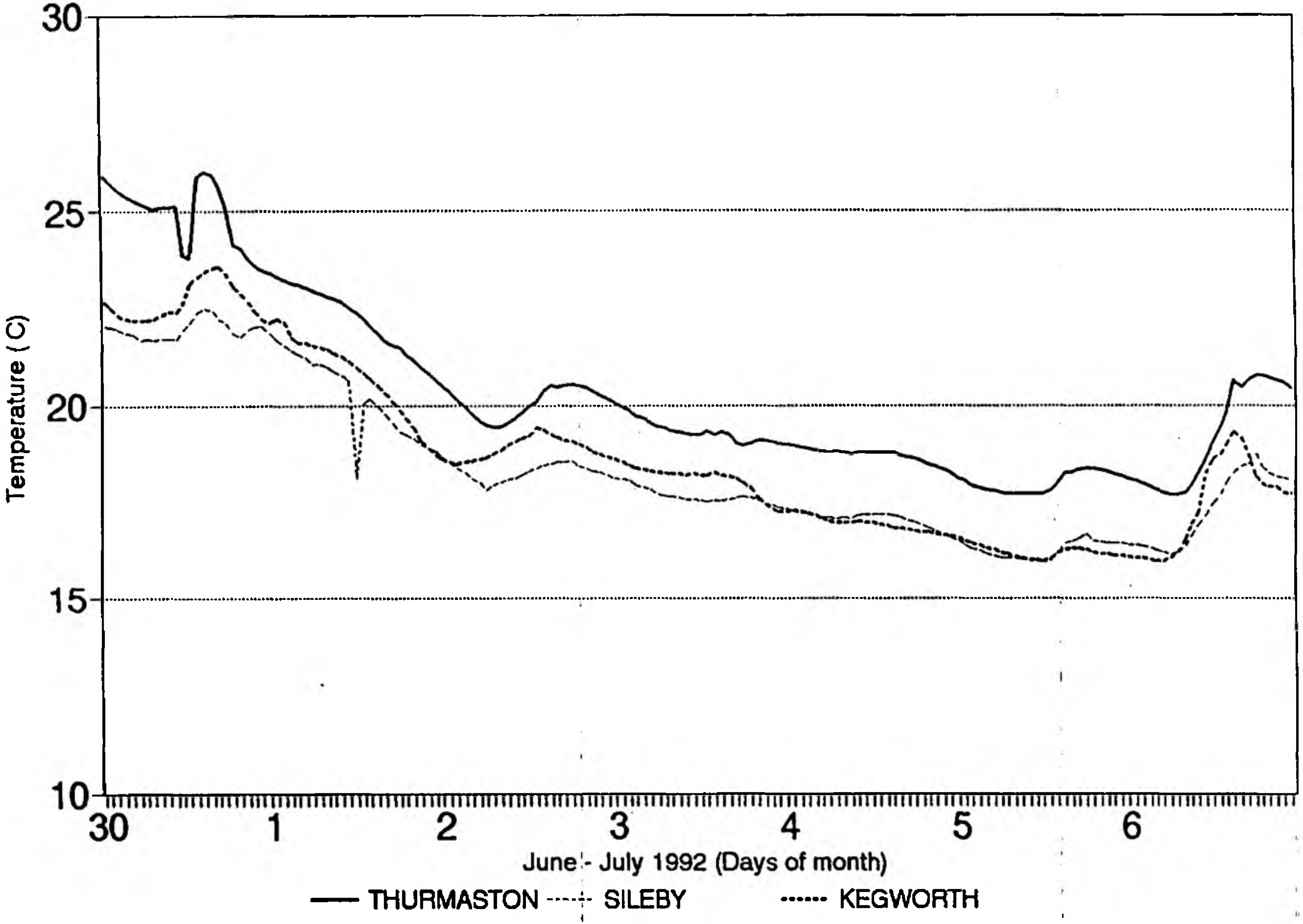


Figure 5. 15 Time history of temperature from sites on the River Soar, during a pollution incident (June 30th - July 6th 1992)

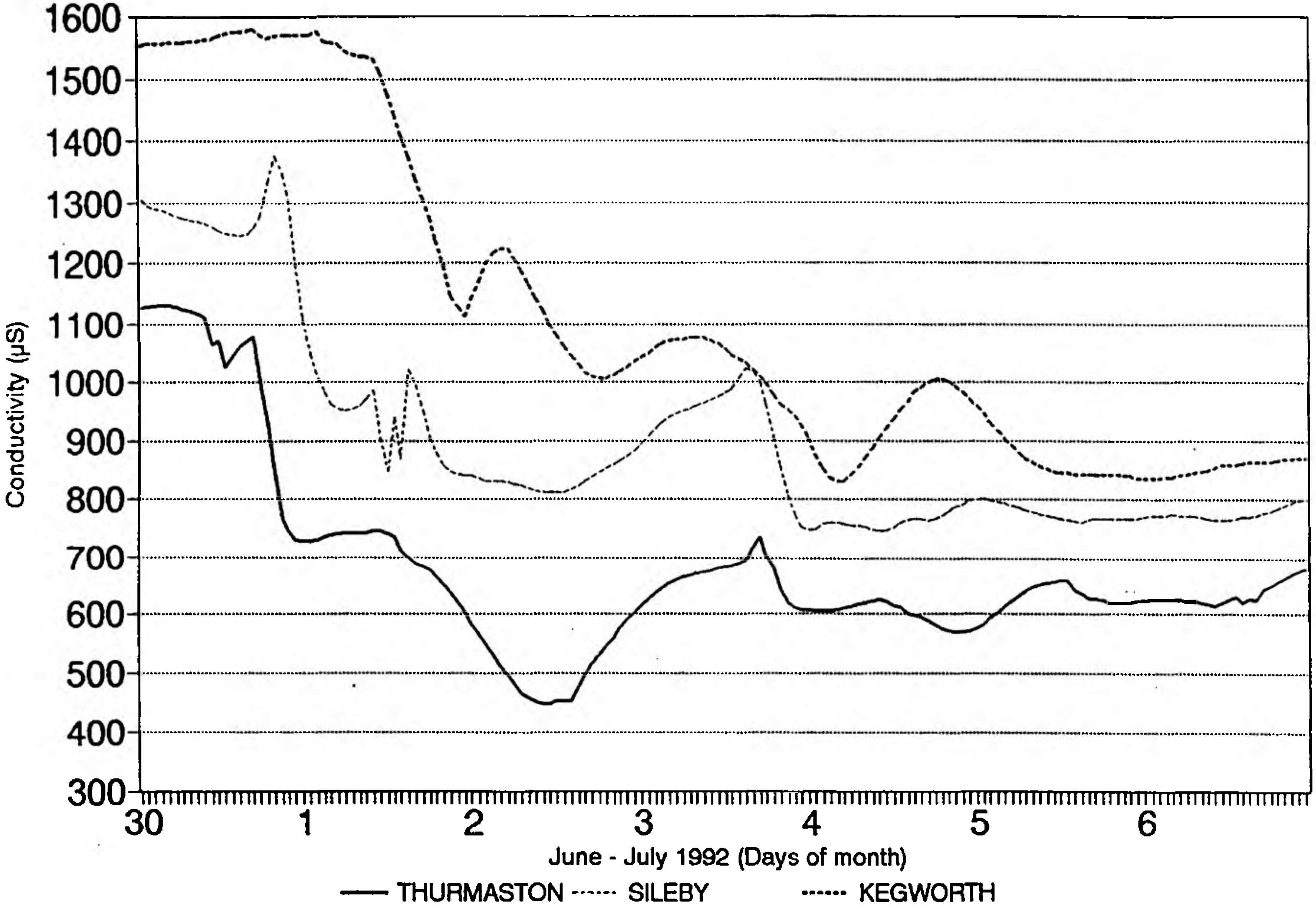


Figure 5.16 Time history of conductivity from sites on the River Soar, during a pollution incident (June 30th - July 6th 1992)

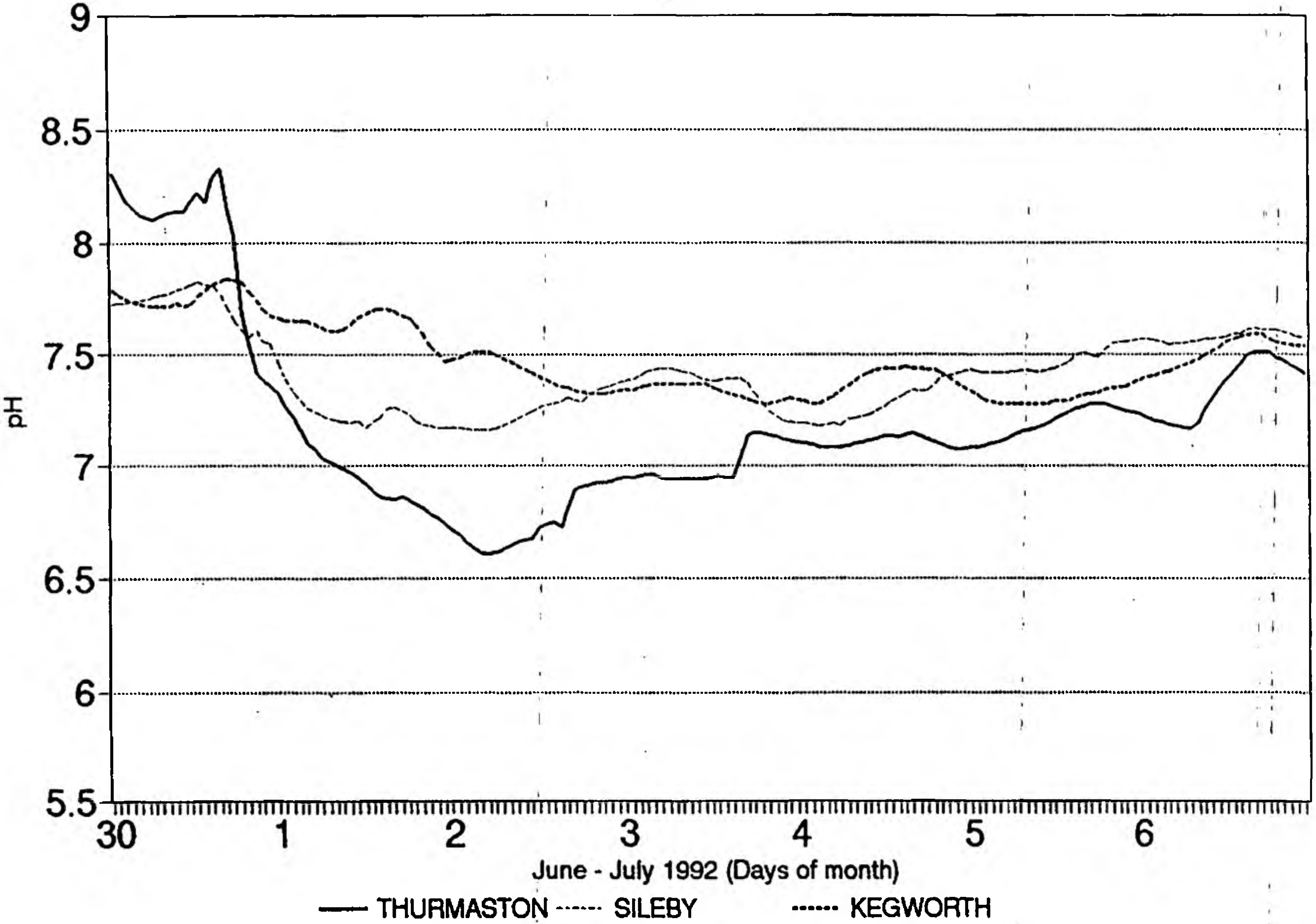


Figure 5.17 Time history of pH from sites on the River Soar, during a pollution incident (June 30th - July 6th 1992)

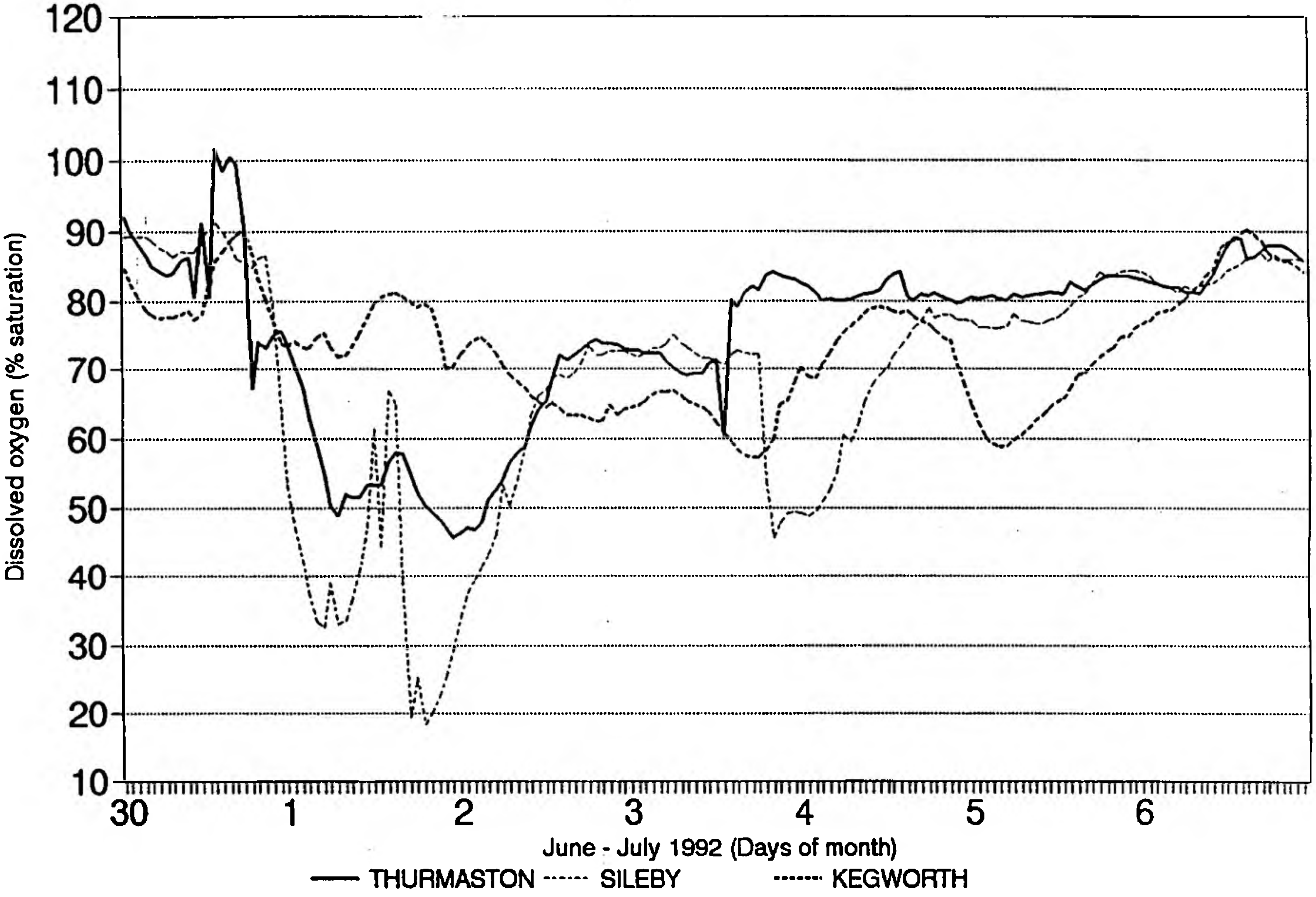


Figure 5. 18 Time history of dissolved oxygen from sites on the River Soar, during a pollution incident (June 30th - July 6th 1992)

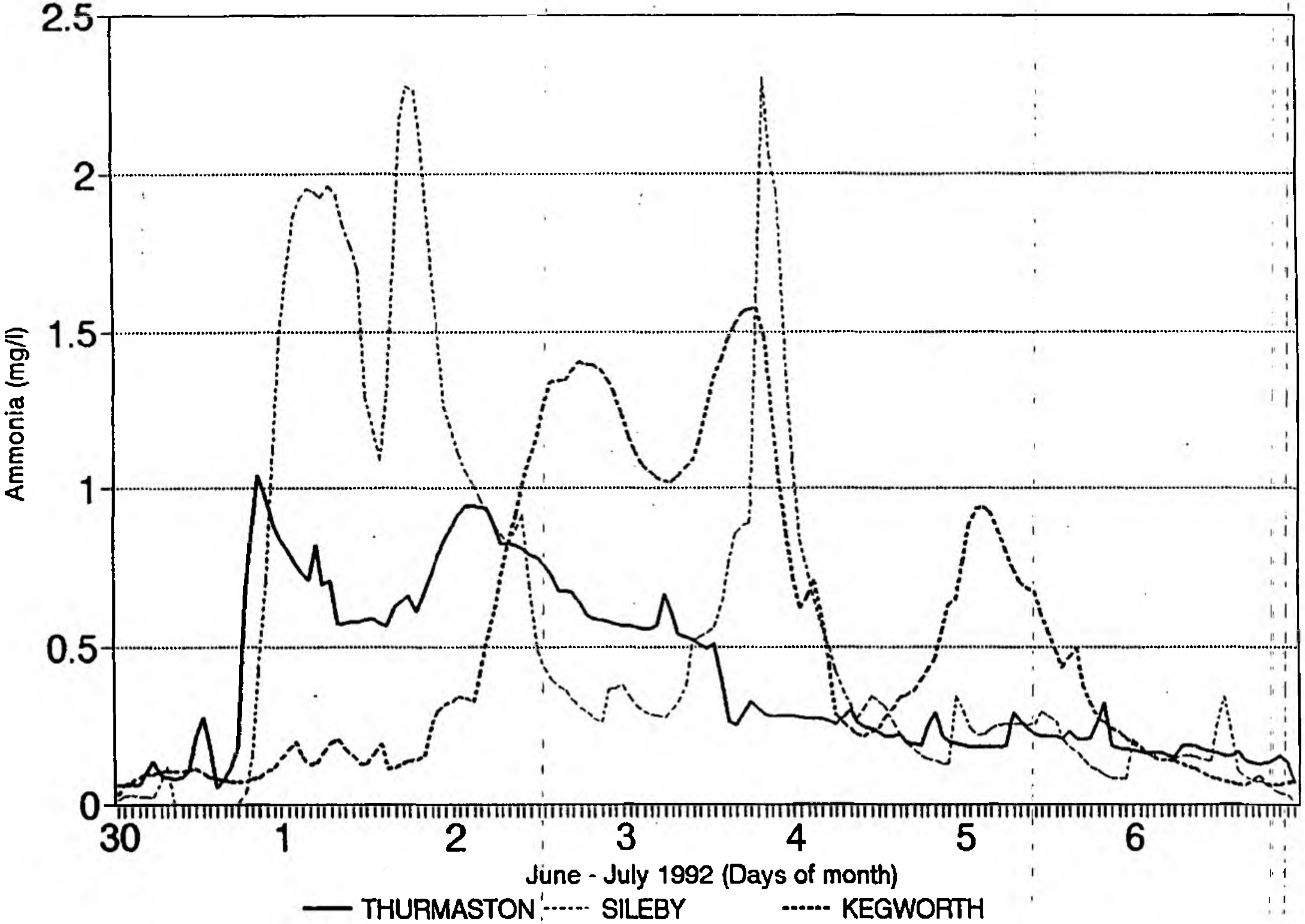


Figure 5.19 Time history of ammonia from sites on the River Soar, during a pollution incident (June 30th - July 6th 1992)

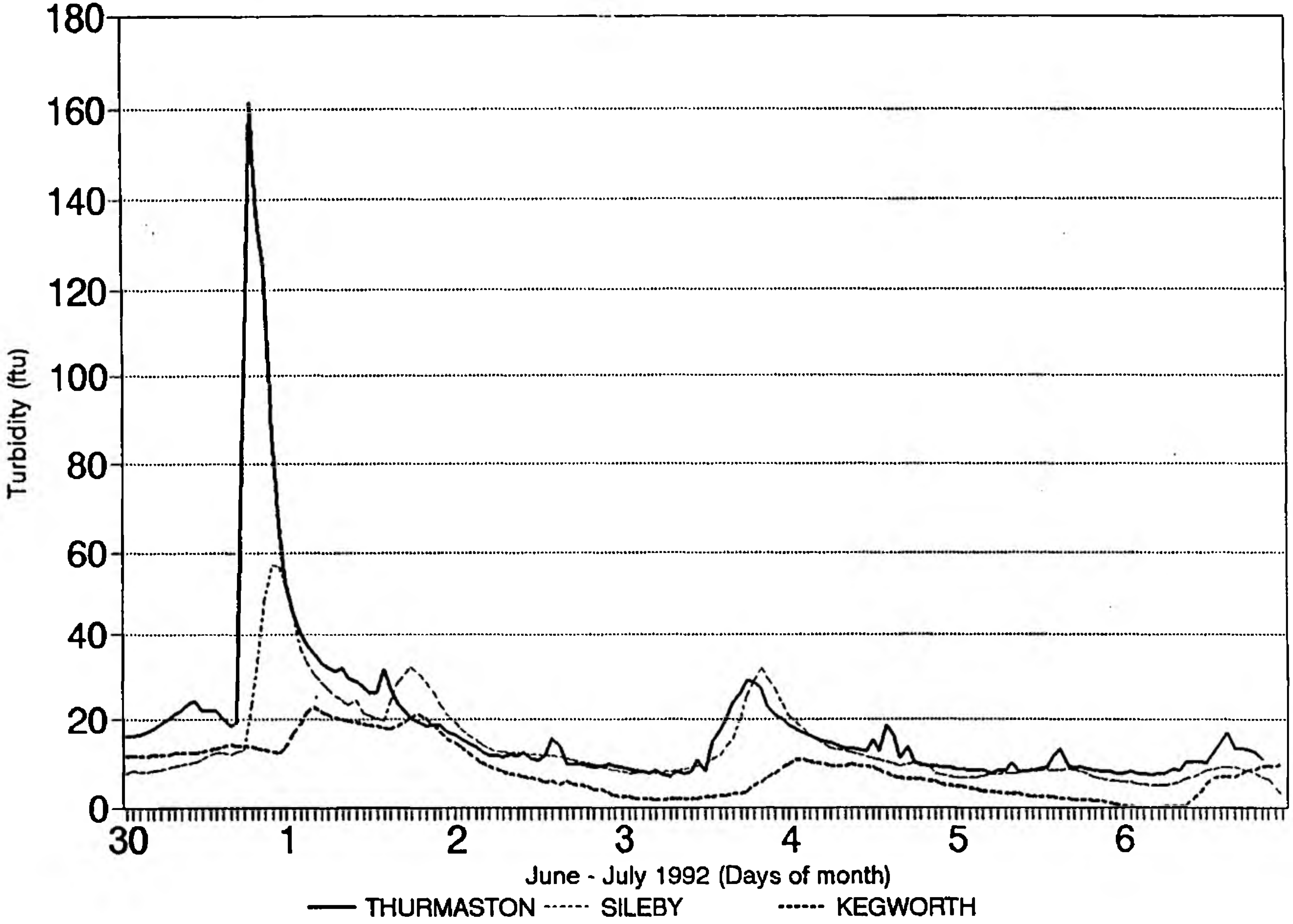


Figure 5.20 Time history of turbidity from sites on the River Soar, during a pollution incident (June 30th - July 6th 1992)

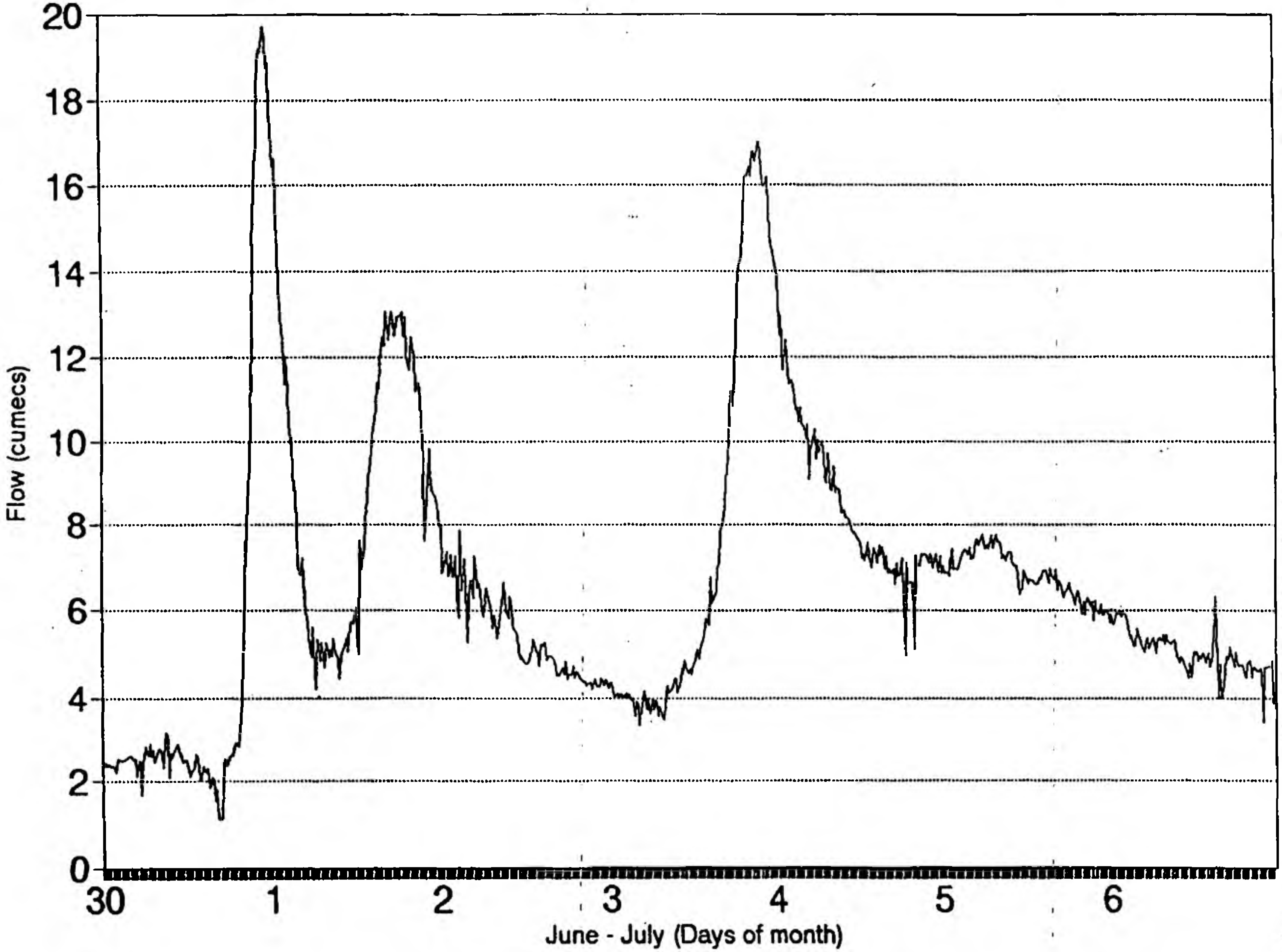


Figure 5.21 Time history of flow from Pillings Hydrometric Station, on the River Soar, during a pollution incident (June 30th - July 6th 1992)

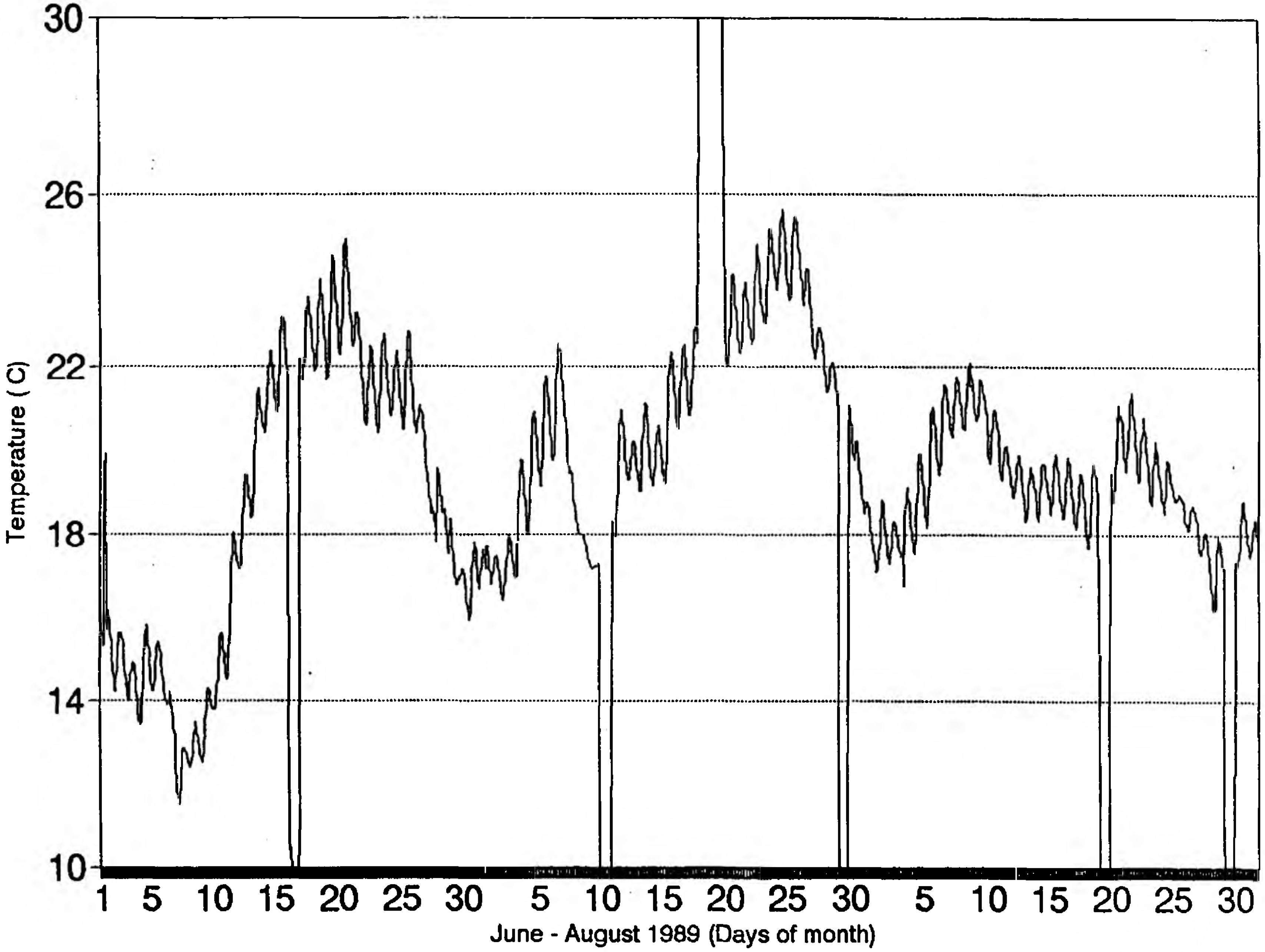


Figure 5.22 Time history of temperature, Thurmaston, June - August 1989

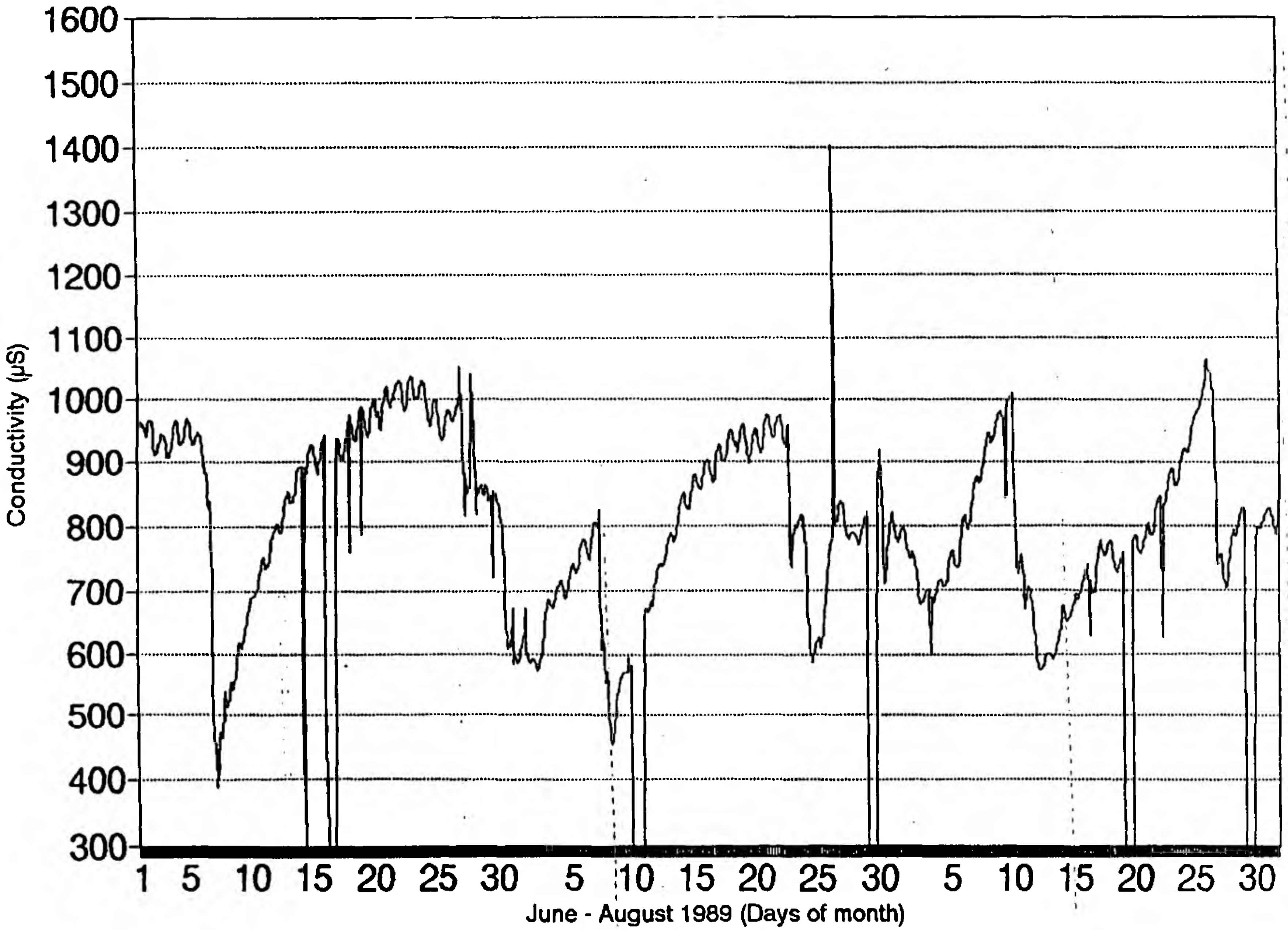


Figure 5. 23 Time history of conductivity, Thurmaston, June - August 1989

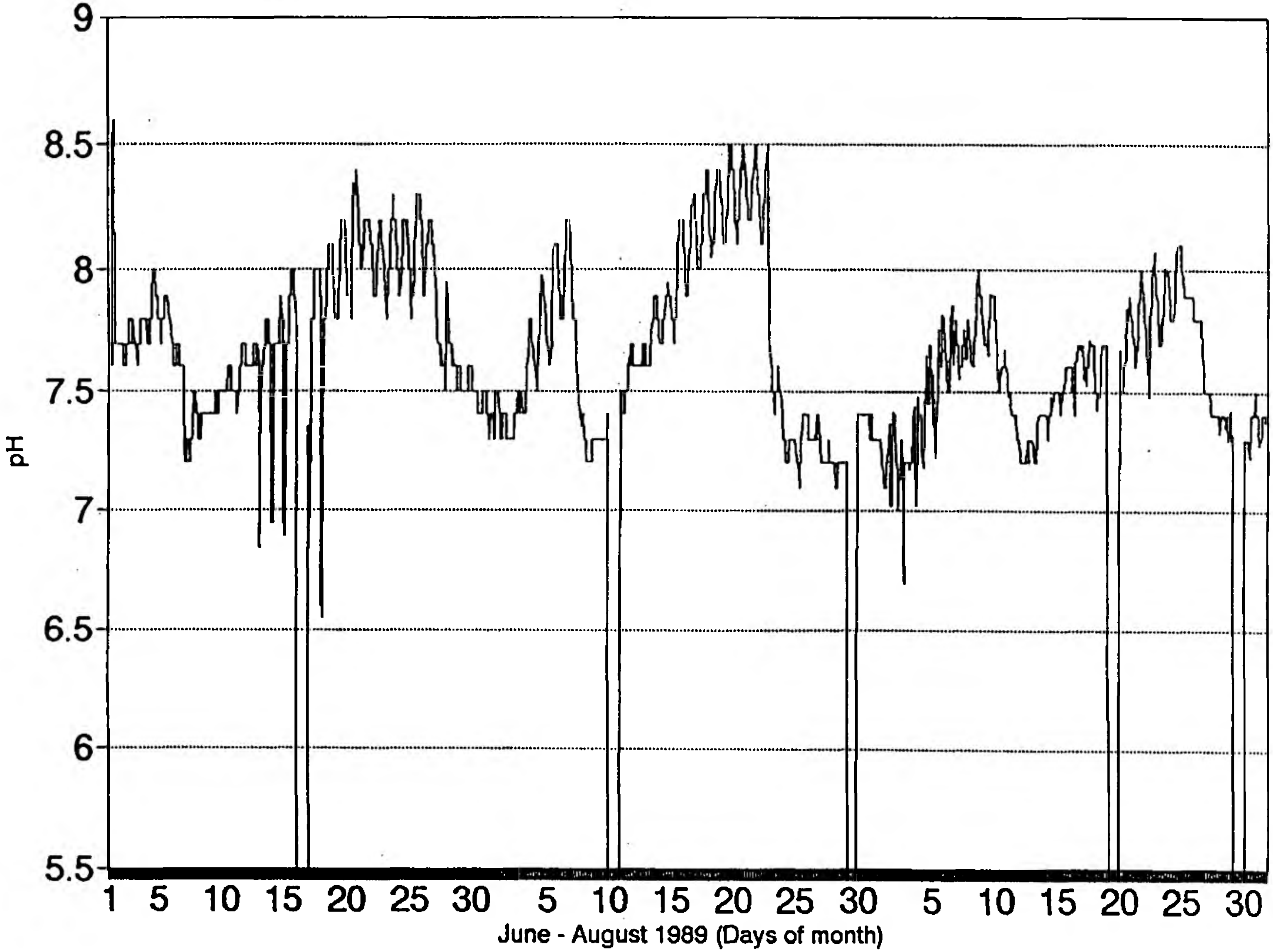


Figure 5.24 Time history of pH, Thurmaston, June - August 1989

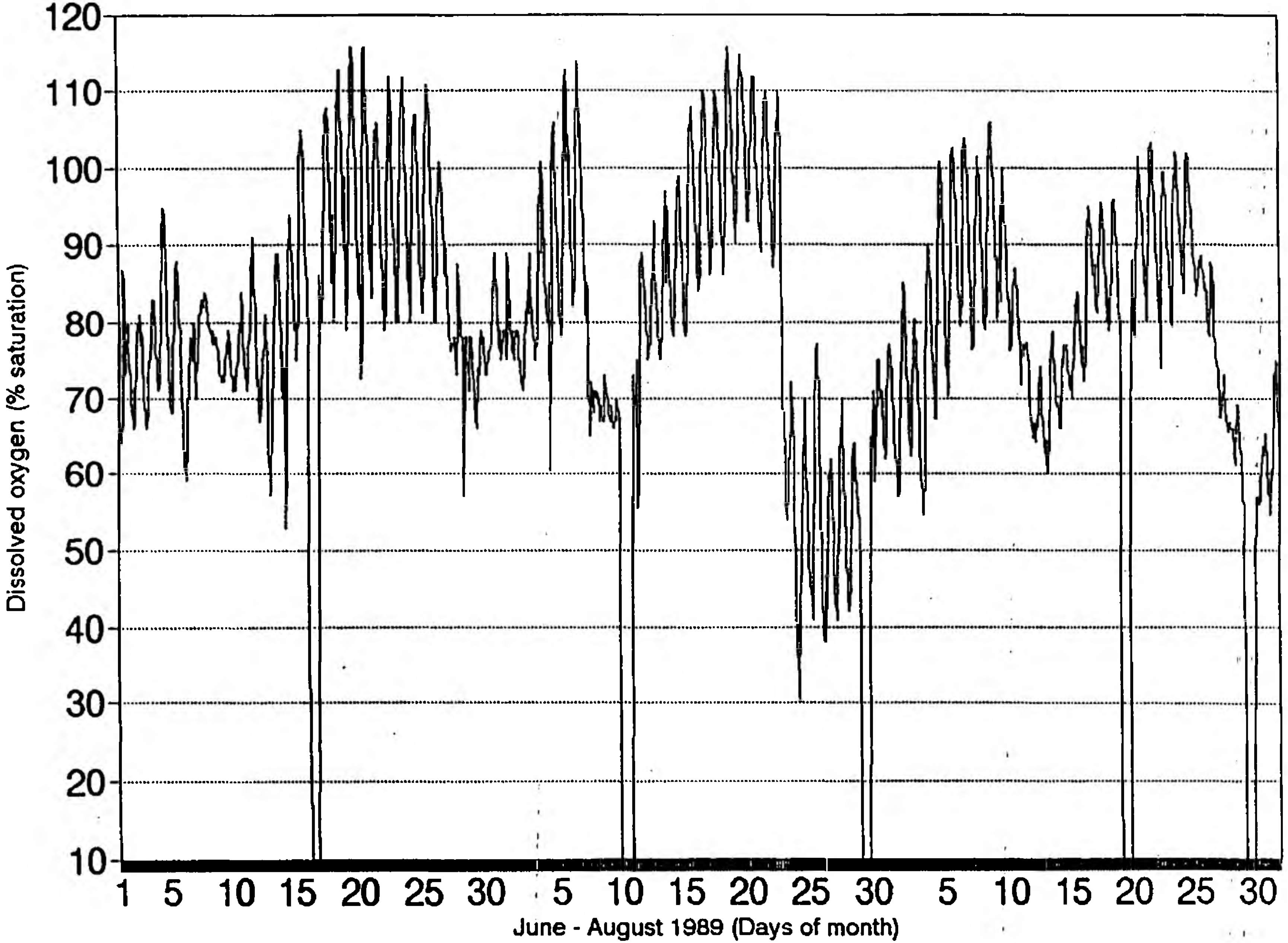


Figure 5.25 Time history of dissolved oxygen, Thurmaston, June - August 1989

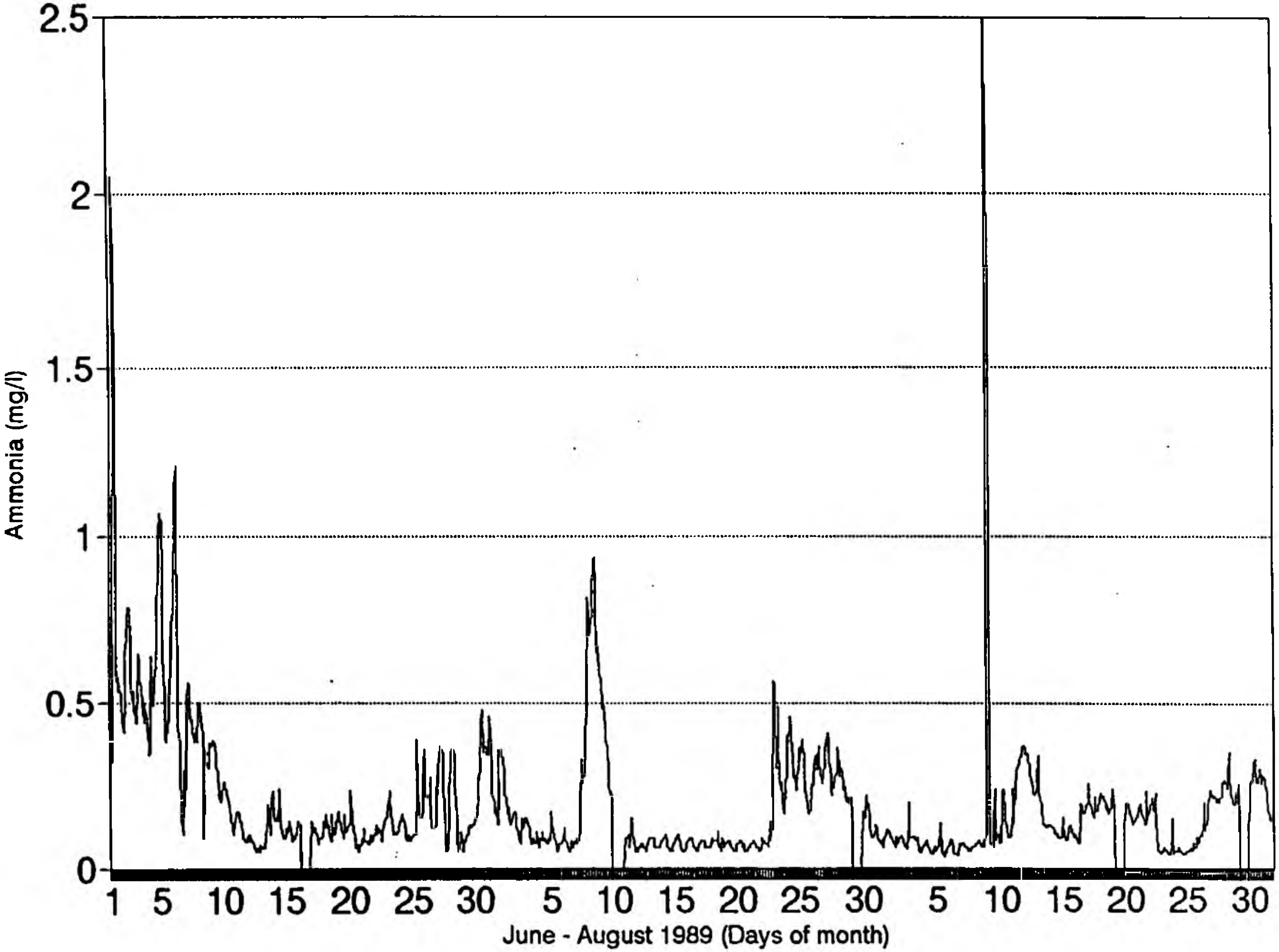


Figure 5.26 Time history of ammonia, Thurmaston, June - August 1989

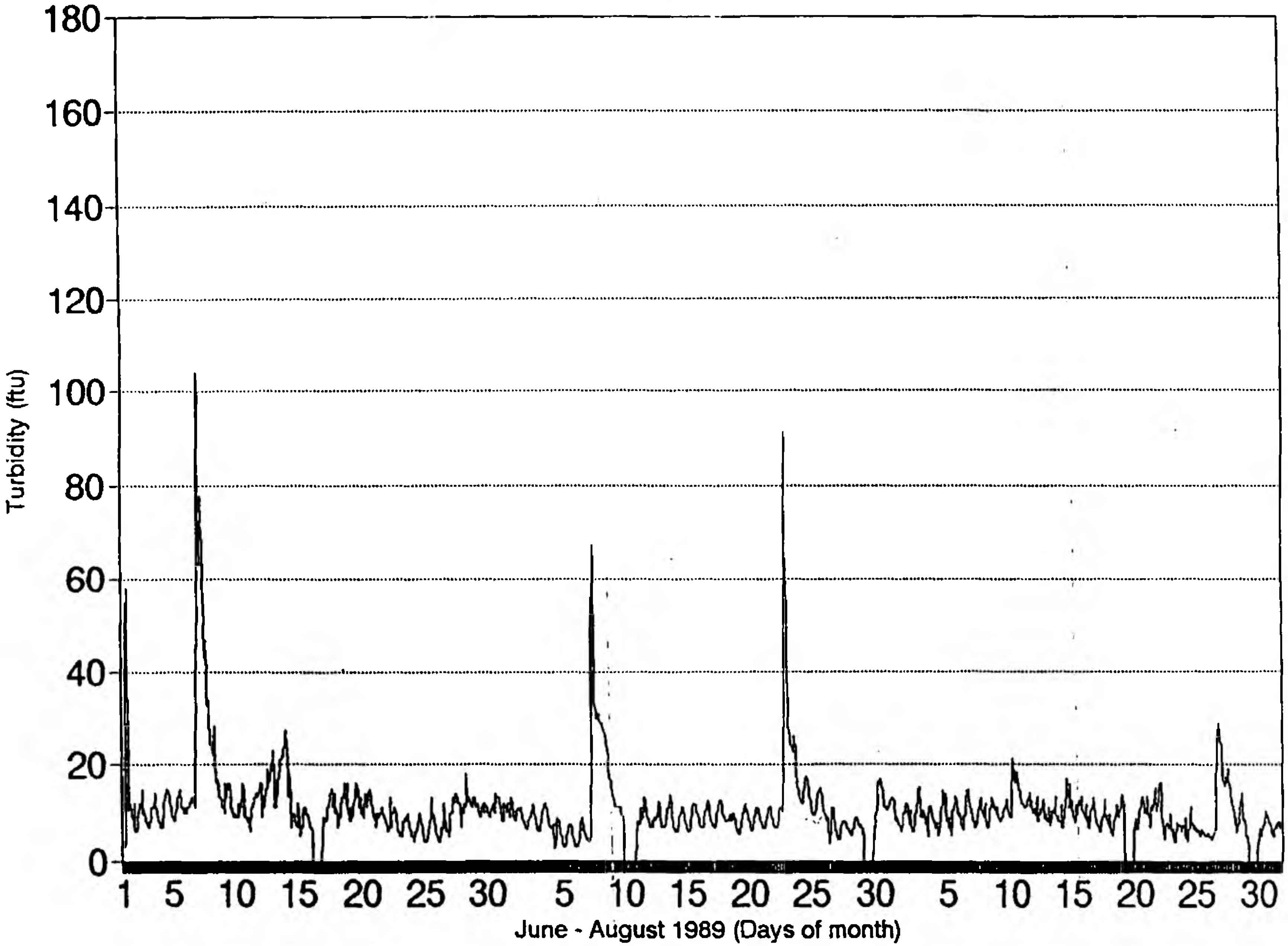


Figure 5.27 Time history of turbidity, Thurmaston, June - August 1989

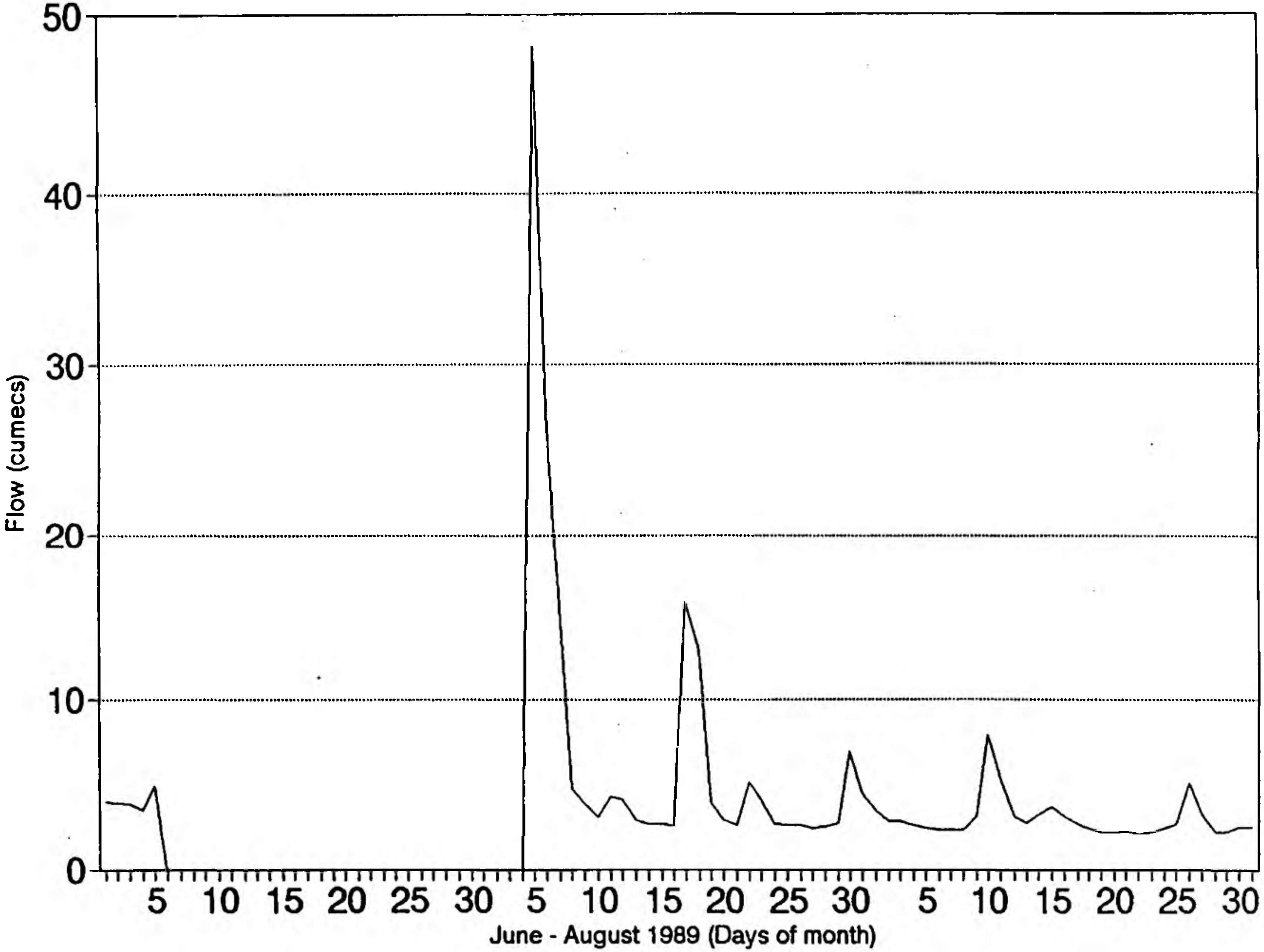


Figure 5.28 Time history of flow, Thurmsaton, June - August 1989

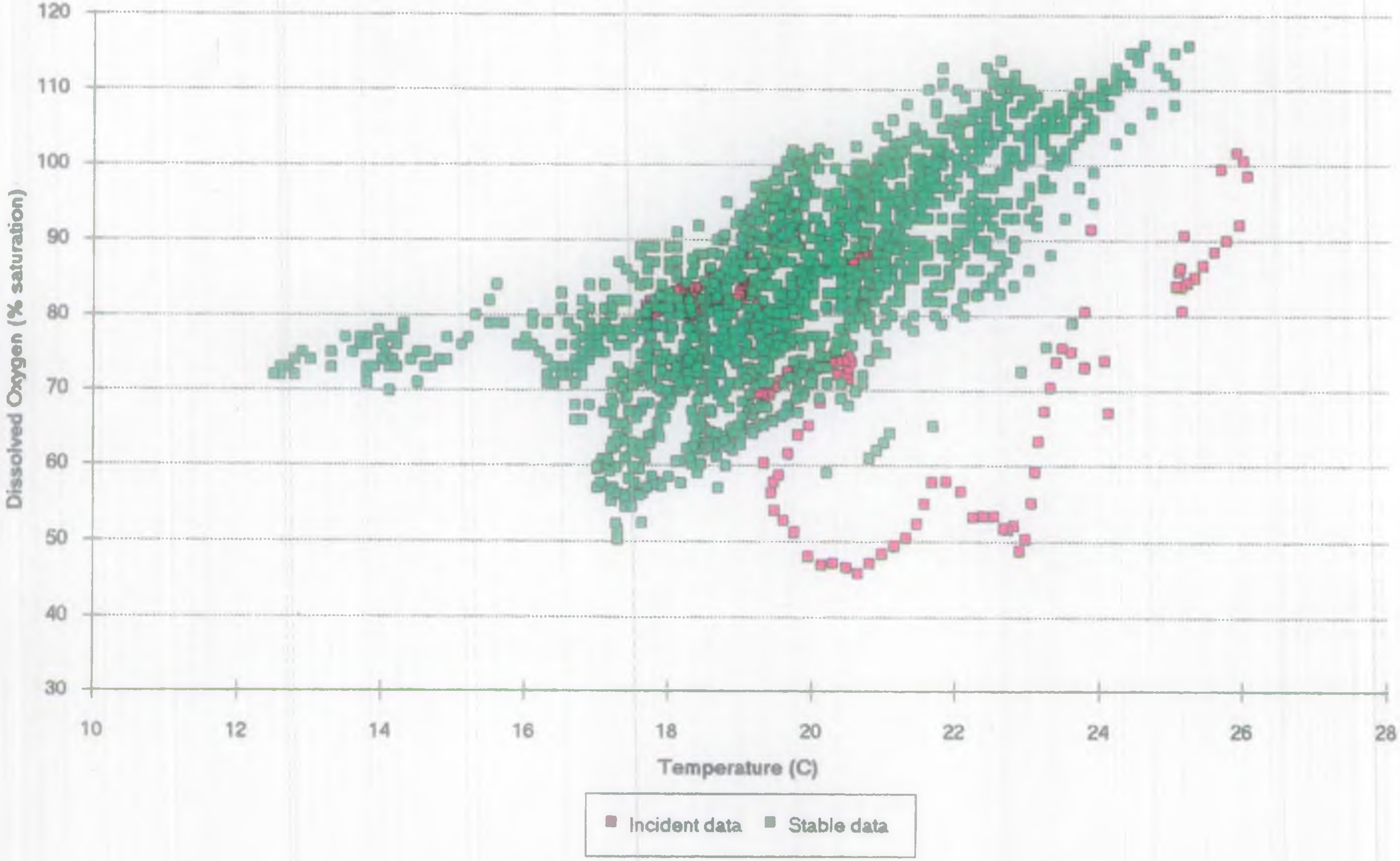


Figure 5.29 Graph of Dissolved Oxygen vs Temperature for Stable Data and Pollution Incident Data

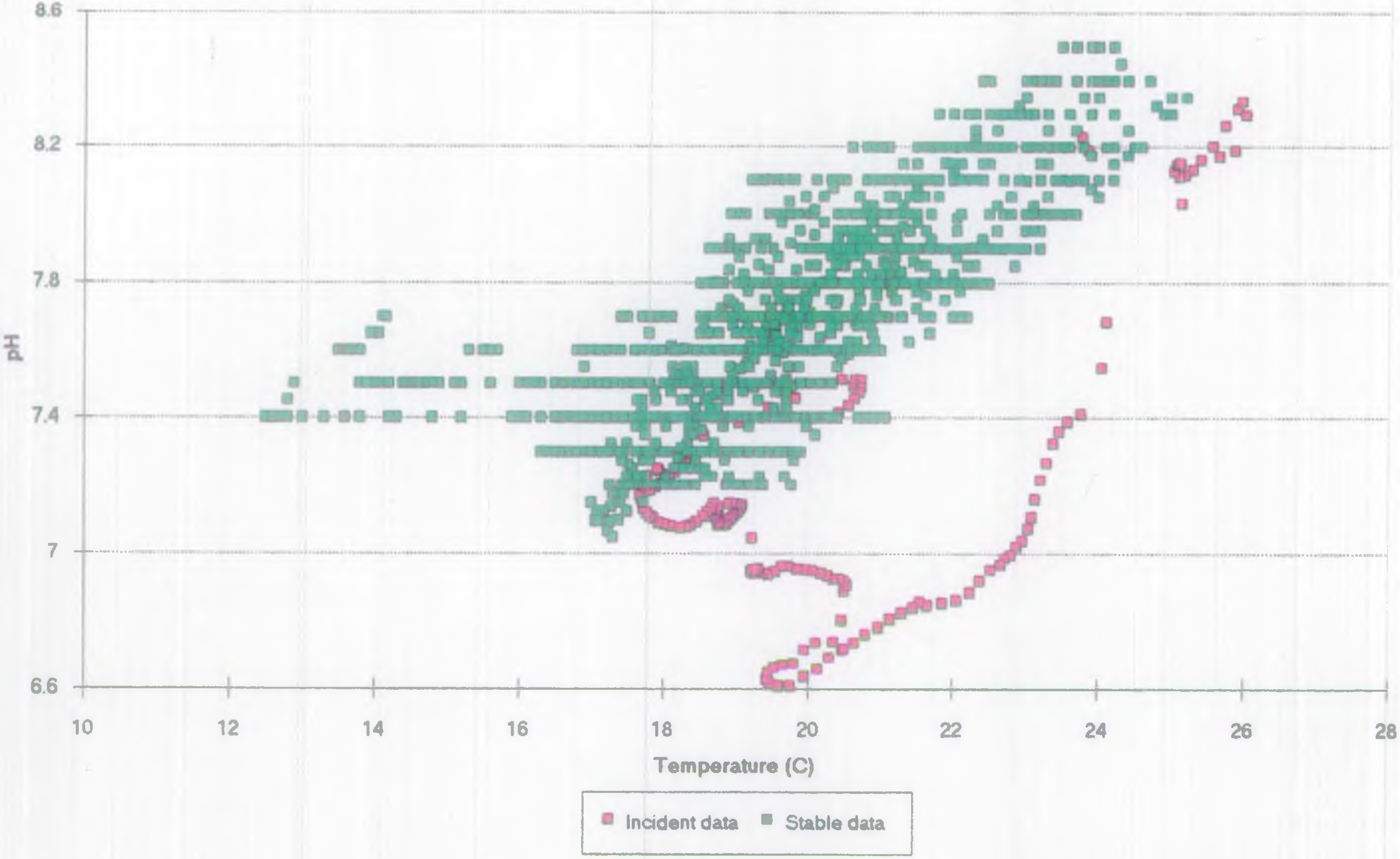


Figure 5.30 Graph of pH vs Temperature for Stable Data and Pollution Incident Data

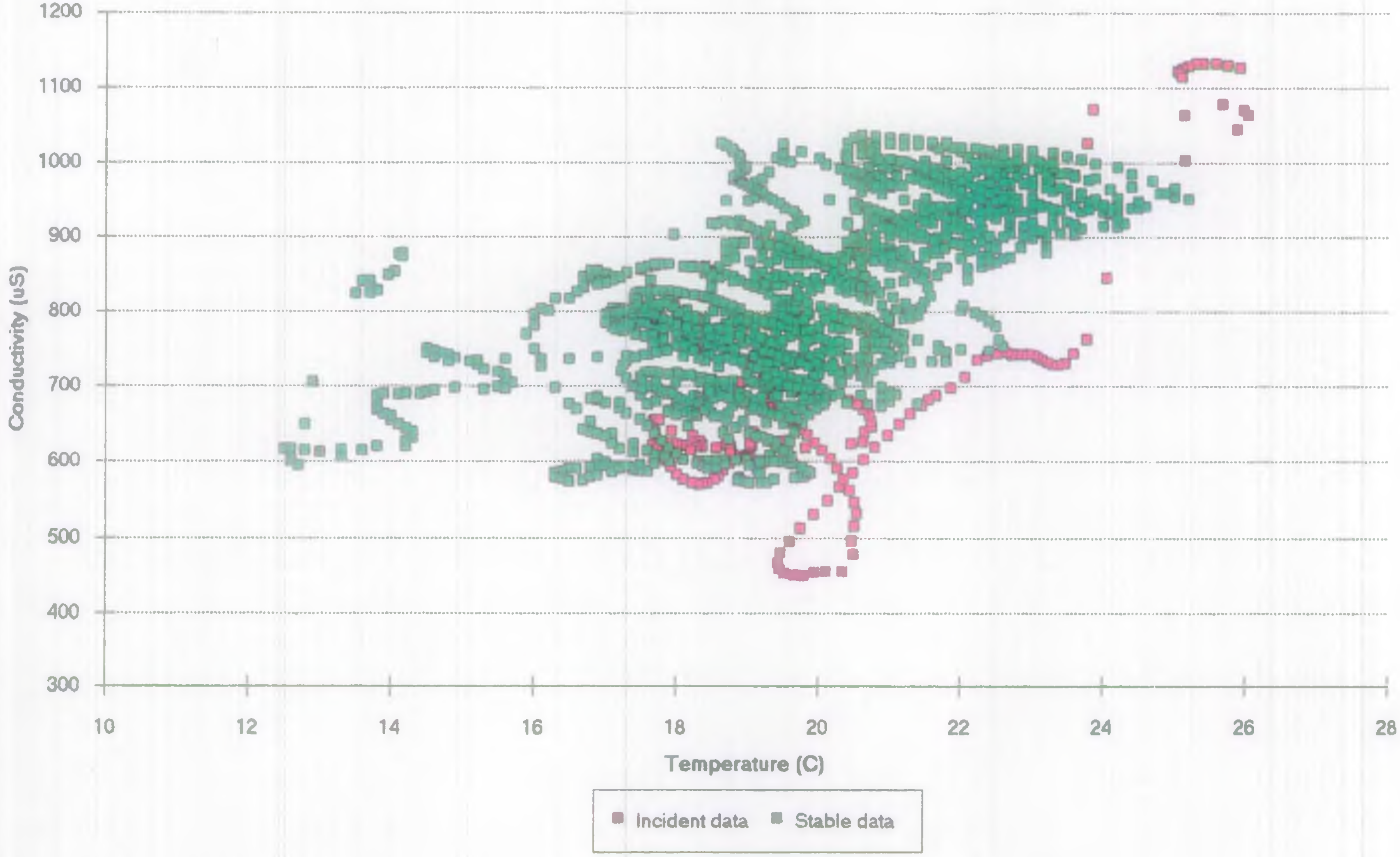


Figure 5.31 Graph of Conductivity vs Temperature for Stable Data and Pollution Incident Data

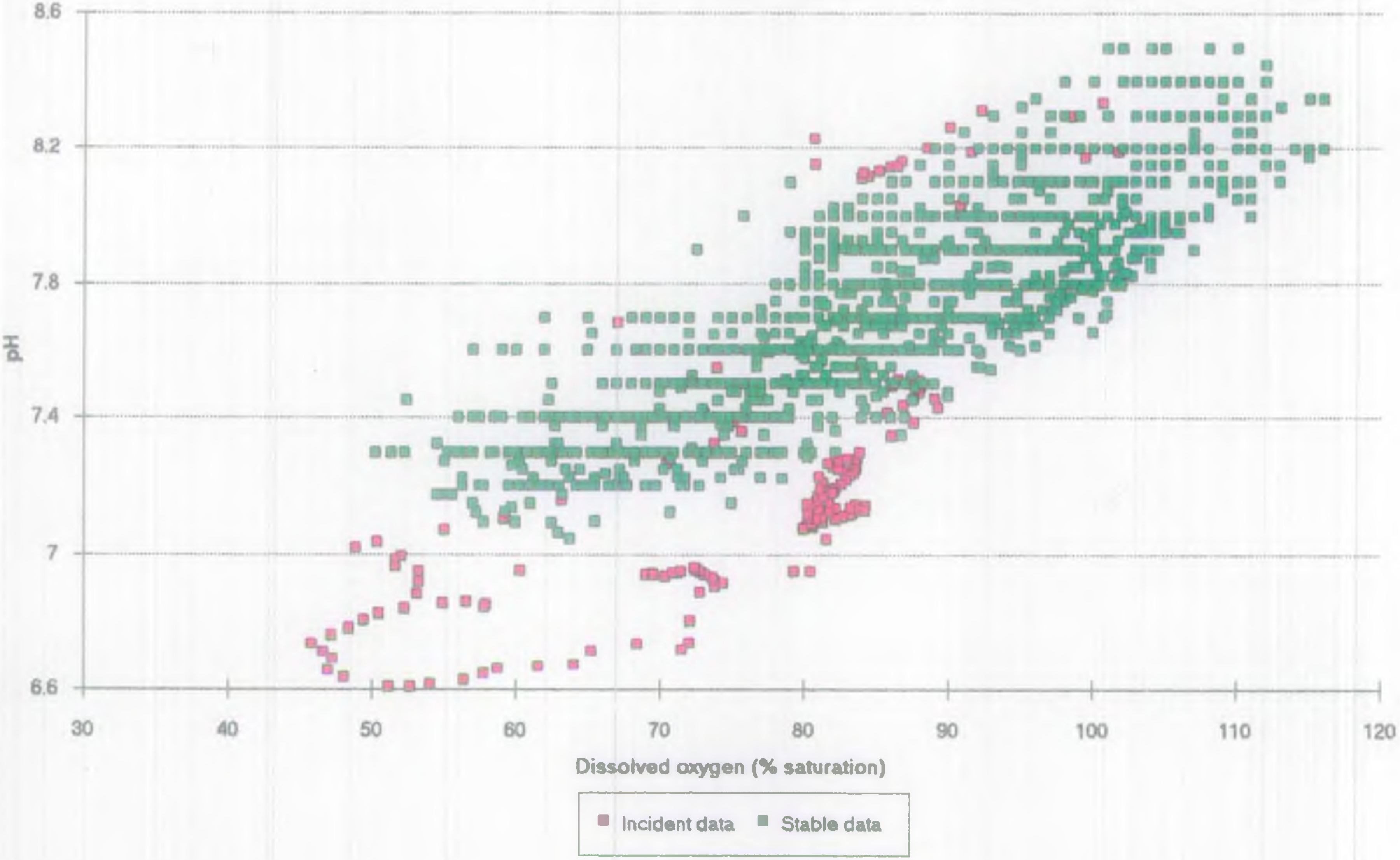


Figure 5.32 Graph of pH vs Disolved Oxygen for Stable Data and Pollution Incident Data

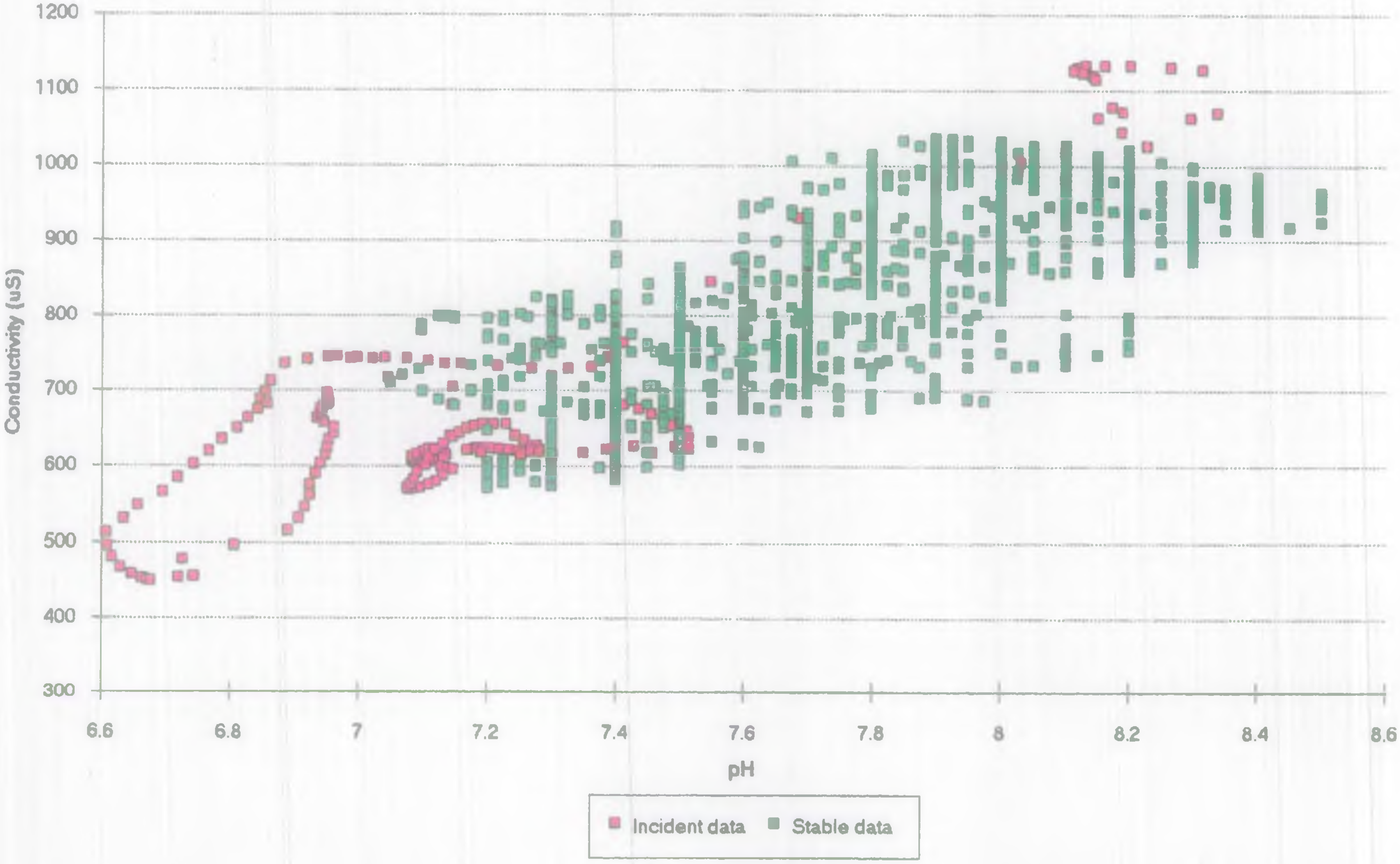


Figure 5.33 Graph of Conductivity vs pH for Stable Data and Pollution Incident Data

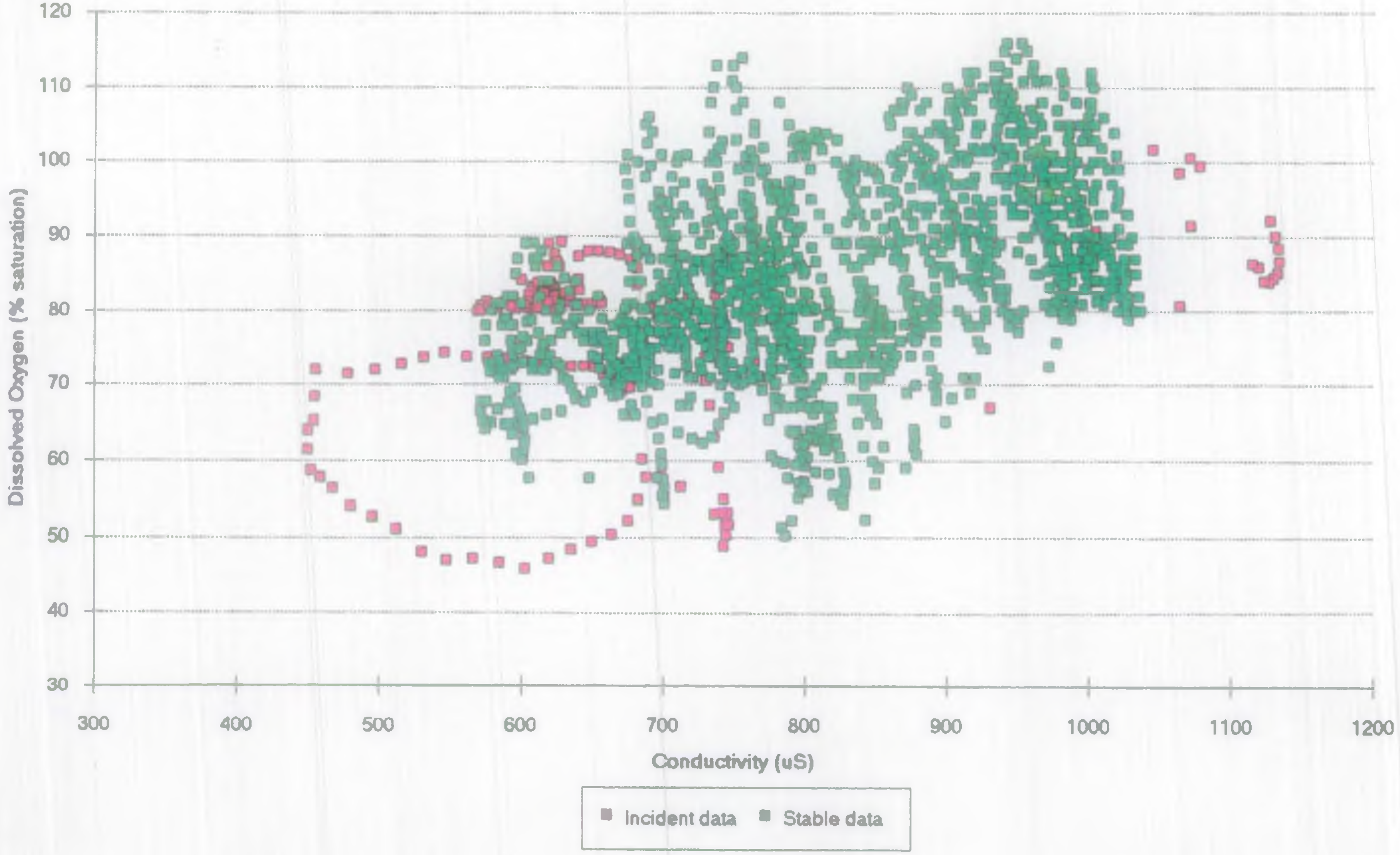


Figure 5.34 Graph of Dissolved Oxygen vs Conductivity for Stable Data and Pollution Incident Data

6. CONCLUSIONS AND RECOMMENDATIONS

6.1 Conclusions

The review of practices for data analysis, storage and archiving of water quality data in the NRA regions has served to identify the NRA requirements for the use of this automatically monitored data. A review of practices in other industries showed that a number of data analysis techniques already used by the NRA were also used in these industries. However, the review also revealed a number of more sophisticated techniques which could be applicable to water quality data. Preliminary recommendations were made, at the end of Phase I, for testing a number of these techniques on the data during Phase II.

Phase II allowed for a greater appreciation of the nature and quality of the data through the use of exploratory data analysis. Analysis concentrated on methods for data validation and techniques for detecting pollution incidents. Algorithms were developed for removal of calibration spikes. Filtering using Kalman filters was used as an alternative method of reducing spikes, and was also used to remove periodic behaviour. Further refinement of this technique would help the identification of outliers.

Two-parameter plots were produced and appeared to show some success in identifying pollution incidents where there was a significant change in dissolved oxygen and pH. In particular, they showed the underlying correlation between pH, dissolved oxygen and temperature, and that this correlation was short term (a few days) and limited to summer.

An initial exploration of principal component analysis did not improve understanding of the data.

Other characteristic patterns of behaviour in response to high river flow events were recognised.

Ammonia and turbidity were recognised as particularly clear indicators of pollution incidents, although turbidity is correlated with flow.

In general, the identification of multi-parameter exceptions proved difficult, even using human expertise. This implies that this identification will be difficult to estimate. Further work in this area, using multi-parameter techniques or non-parametric statistics may prove to be more powerful. The identification of pollution incidents using two-parameter plots is one technique which relies upon comparisons with past experience (i.e. historical data) to define different exception modes. Testing more sophisticated pattern recognition techniques would be a natural progression from this work.

It is clear that major events, such as high ammonia or low dissolved oxygen, of such an intensity that they represent problems in their own right, are capable of reliable detection with a low false alarm rate. More subtle incidents, which may be evidence of pollution but which are not a problem in their own right, are difficult to detect reliably without a high risk of false alarms.

Comparing rivers with other systems subject to automatic monitoring, rivers are seen to have a complex and poorly understood behaviour. For a high degree of automation to succeed, understanding of the complex behaviour will need to be improved. The understanding is likely to be site specific. A clear definition of 'pollution incident' is necessary for the successful automatic detection of these incidents.

6.2 Recommendations

6.2.1 Data validation

1 introduce methods, using local outstation intelligence, for recording calibration cycle data and cleaning cycle data to a separate file;

2 consider the cost and practicality of introducing a voting system using three probes for each parameter monitored. This may be cost-effective, apart from ammonia monitors, and would increase the reliability of the data thus reducing the incidence of false alarms.

3 develop appropriate data cleaning algorithms for each probe at each site. These are likely to include simple threshold detection of instrument failure, spike removal, and filtering by an IRW filter.

4 there are occasions where faulty data is readily detected by use of a simple threshold. However, examination of the data from James Bridge showed that on other occasions faulty data displays a pattern not too dissimilar from incident data. This type of data cannot readily be detected without additional information, such as operator knowledge, or by multiple probes.

6.2.2 Incident detection

1 adopt a two stage approach to exception detection where the first stage uses a coarse filter to identify outliers immediately. The second stage uses more complex techniques using further data to confirm the alarm. It may be prudent to instruct the monitor to increase the measurement frequency in such circumstances. Efficiency of exception detection without false alarms can be increased by filtering data to remove spikes and diurnal fluctuations, and using a cumulative sum approach rather than a simple threshold.

2 use simple thresholds and rate of change to raise the first level alarm.

3 examine the use of control charts to compare current data sets with recent historical data to check consistency over time.

4 further investigate the use of multi-parameter techniques for the identification of pollution events. These may allow measures of water quality to be compared with predicted values.

5 use Kalman filtering techniques for missing value estimation, and also to characterise the internal nature of data summarised by PDF's.

6.2.3 Implementation

Implementation requires changes in three areas: monitoring site hardware, control hardware and software, and NRA policy decisions.

Site hardware needs to be upgraded to record calibration separately, and relay multiple probes back to control (or perform voting on site). Two way communication is desirable so that monitoring frequency can be increased in response to exceptions. Given the importance of flow measurements in data interpretation, monitors should ideally be sited immediately upstream of flow gauging stations.

Suitable display hardware and associated software is needed for the monitor operator. The display hardware needs to be capable of high resolution colour graphics. The underlying software needs to incorporate:

- Multi-parameter time series plots for previous periods up to 1 year;
- A graphical editor allowing spikes and bad data to be marked (but not removed);
- Algorithms for spike removal and filtering before display;
- Threshold and cumulative sums exception detection;
- Two parameter cross-plots for specifiable data windows, with recent data points identifiable;
- A database of previous information, recording raw data, processed data, and processing flags;
- This database must be accessible by quality planners and other data users, as well as by the monitor operator.

The NRA needs to determine if detection of minor pollution incidents is a priority purpose of the monitoring system. Such use of the system will be difficult, since it is not clearly possible even on a sample basis, with present knowledge. However, prompt detection of major events that would require remedial action, is possible. The NRA needs to decide promptness of response. The present Severn Trent system cannot effectively raise an alarm in less than two hours.

6.2.4 Costs

Costs are highly dependent on national and local policy.

Upgrading monitoring stations is an ongoing process and it is assumed that the NRA is knowledgeable about such costs.

The display capabilities and processing requirements could be served by Unix workstations costing significantly less than £10 000 each if file system support is from a network server elsewhere. This network would be part of a region's corporate network and would be the mechanism of data distribution to users.

The basic software technology already exists and therefore development costs for this will not be incurred. There is a cost integrating this technology together into the robust application that would be needed for this purpose. The cost would depend on the specification, and on the degree of standardisation between NRA regions (currently small). Such software development costs may be of the order of £200 000, assuming that a single implementation will serve all regions.

REFERENCES

Ancellin J., Legaund P., "An Expert System for Nuclear Alarm Processing", *Proceedings of the Sixth International Conference on Expert Systems and their Applications*, Avignon 1986.

Barnett, M.W., Patry G.G., "An Expert System Network for Diagnosis and Control of an anaerobic Digestion Process", AIENG 92, *Applications of AI in Engineering*, Wessex Institute of Technology (1992).

Button D. et al, "AI Applications in Process Design, Operation and Safety", *The Knowledge Engineering Review*, Vol **5**, pp 69-95 (1990).

Comerford J.B. et al, "AI Approach to the Integration of Engineering Knowledge: Water Resource Case Studies", AIENG 92, *Applications of AI in Engineering*, Wessex Institute of Technology (1992).

Crommelynck V. et al, "QUALITEAU: an Operational Expert System to Optimize the Quality of Drinking Water Produced by an Industrial Processing Procedure", *Proceedings of the Eleventh International Conference on Expert Systems and their Applications*, Avignon 1991.

Fuller F.C., Tsokos C.P., "Time Series Analysis of Water Pollution Data", *Biometrics*, Vol **27**, pp 1017-1034 (1971)

Gallanti M., et al, "Integrating Shallow and Deep Knowledge in the Design of an Online Process Monitoring System", *International Journal of Man-Machine Studies*, Vol **27**, pp 641-664 (1987).

Glen J. A., Weir B., "Knowledge Based Sensor Data Validation", Proceedings 1990 IME Conference.

Hajek B.K. et al, "A Generic Task Approach to a Real Time Nuclear Power Plant Fault Diagnosis and Advisory System", *Proceedings of the International Workshop on AI for Industrial Applications*, Hitachi City, Japan (1988).

Hollnagel E., Mancini G. (eds), ''Intelligent Decision Support In Process Environments'', Springer Verlag, Berlin (1986).

ICI Chemicals, SEMA Group et al, TENPRO Project - Technical and Management Proposals (1991).

Kerby J.P. et al, ''A Prototype Expert System for Troubleshooting Well Water Contamination'', AIENG 92, *Applications of AI in Engineering*, Wessex Institute of Technology (1992).

Kowalik J.S., Kitzmiller C.T., ''Coupling Symbolic and Numerical Computing in Expert Systems II'', North Holland, Amsterdam (1988).

McKerchar A.I., Delleur J.W., ''Application of Seasonal Parametric Linear Stochastic Models to Monthly Flow Data'', *Water Resources Research*, Vol **10**, pp 246-255 (1974).

McMichael, F.C., Hunter J.S. ''Stochastic Modeling of Temperature and Flow in Rivers''. *Water Resources Research*, Vol **8**, pp 87-98 (1972).

Mohamed W.A., Simonovic S.P., ''Generous-Man: an Algorithm for Allocation of Limited Water Resources'', AIENG 92, *Applications of AI in Engineering*, Wessex Institute of Technology (1992).

Nann S.R., et al, ''A Decision Support System for Real Time Monitoring and Control of Dynamical Processes'', *International Journal of Intelligent Systems*, Vol **6**, pp 739-758 (1991).

Nelson, A.C. et al, ''Validation of Air Monitoring Data'', Report EPA-600/4-80-030. U.S. Environment Protection Agency, *National Technical Information Service*, Springfield, VA (1980).

Rao V.J., et al, ''Environs - an Expert System to Assess the Impact of Environmental Pollution on the Health of Industrial Workers and the General Public'', *Proceedings of the Annual Conference of AI Systems in Government*, Washington, DC (1989).

Sakakima S. Kojiri T., "Real Time Reservoir Operation with Neural Nets", AIENG 92, *Applications of AI in Engineering*, Wessex Institute of Technology (1992).

Slow T., Mathew R. "The Role of Intelligent Systems in Integrated Design Environments for Water Quality and Hydraulic Modelling", AIENG 92, *Applications of AI in Engineering*, Wessex Institute of Technology (1992).

Walley W.J., "Application of Bayesian Inference to Water Quality Surveillance", AIENG 92, *Applications of AI in Engineering*, Wessex Institute of Technology (1992).

Young P.C and Minchin P.E.H. "Envirometric Time Series Analysis: Modelling Natural Systems from Experimental Time Series Data", *Int. J. Biol. Macromol.*, Vol 13, pp 190-201 (1991).

Young P.C., Ng C.N., Lane K. and Parker D., "Recursive Forecasting, Smoothing and Seasonal Adjustment of Non-stationary Environmental Data", *Journal of Forcasting*, Vol **10**, pp 57-89 (1991).

Yung, S.K. and Clarke, D.W (1989) Local sensor validation, *Measurement and Control*, Vol **22**, pp 132-139.

APPENDIX A
SURVEY QUESTIONNAIRE

Our Ref: 69 C3679

Date: JANUARY 1992

WATER QUALITY
MONITORING
QUESTIONNAIRE

Prepared for	:	National Rivers Authority
Prepared by	:	Cremer & Warner

DEFINITION OF TERMS USED IN THE QUESTIONNAIRE

Instrument accuracy

This refers to an instrument's ability to measure the true value of a parameter. A well-calibrated instrument will be accurate.

Instrument precision

This refers to the inherent uncertainty in measured values and relates to instrument error e.g. ± 1%.

Name of data processing system

The name by which the data processing system is commonly referred i.e. locally used name.

Out-of-range values

Generally, only certain parameter values are feasible. Measured values can be validated against a feasible range of values, and those values outside of the range flagged up.

Data storage

Local storage of data which is in a readily accessible form e.g. on the hard disk of a computer.

Archived data

Data which is stored off-line and would need to be reloaded on to the computer to be accessed.

1.0 BACKGROUND INFORMATION

NAME: DATE:

POSITION: TELEPHONE:

NRA REGION: FAX:

Description of Responsibilities

2.0 DESCRIPTION OF CURRENT DATA STORAGE AND ANALYSIS SYSTEM

2.1 Data Acquisition

2.1.1 How many water quality monitoring sites are in the region?

2.1.2 What parameters are monitored automatically?

	Parameter	No. of Sites
(i)		
(ii)		
(iii)		
(iv)		
(v)		
(vi)		
(etc)		

2.1.3 What is the interval between measurements?

2.1.4 How frequently is data transferred?

2.1.5 How are the instruments calibrated for each parameter?

	<u>Parameter</u>	<u>Calibration</u> Method	<u>Frequency of</u> <u>Calibration</u>	<u>Accuracy</u> <u>and Precision</u>
(i)				
(ii)				
(iii)				
(iv)				
(v)				
(etc)				

2.1.6 Are there other automatic water monitoring sites in the region? YES/NO

(If NO go to 2.1.7)

(a) How many automatic hydrometric monitoring sites are there in the region?

(b) At how many sites is water quality monitoring undertaken?

2.1.7 Is manual sampling undertaken at automatic water quality monitoring sites?

(a) At how many sites?

(b) At what frequency are samples taken?

2.2 Data Processing

2.2.1 Name of data processing system.

2.2.2 Describe quality checking and validation of raw data.

(a) Treatment of missing data

(b) Detection of calibration cycles

(c) Detection and treatment of calibration drift

(d) Detection of cleaning cycles

(e) Detecting out-of-range (infeasible) values

(f) Interpolation of missing values

(g) Correlation with other parameters

(h) Other (please describe)

2.2.3 Describe alarm detection and validation methods

(a) Threshold level

(b) Comparison with previous reading(s)

(c) Correlation with other parameters

(d) Other (please describe)

2.2.4 What techniques are used to interpret the data?
(Attach example output)

(a) Summary statistics

(b) Data tables

(c) Statistical techniques

(d) Other (please describe)

2.2.5 What graphical output is generated?
(Attach example output).

(a) Histrograms

(b) Frequency plots

(c) X-Y plots

(d) Other (please describe)

2.2.6 Is any additional data required from other sources? (e.g. meteorological data)

(If NO go to 2.2.7)

(a) What is data source?

(b) How frequently is it updated?

(c) How is the data transferred from source?

(d) Where is data stored?

(e) How is the data accessed?

2.2.7 Are the other IT system which access the water quality data?

(a) Water quality archives

(b) Simulation models

(c) Regional warning system

(d) Other (please describe)

2.2.8 Describe any firm plans for new or upgraded processing systems.

2.3 Data Storage

2.3.1 What data is stored?

(a)

(b)

(etc)

2.3.2 How is the data stored?
(Files, database etc)

2.3.3 Where is the data stored?

(a) Type of computer

(b) Storage medium

2.3.4 What volume of data is stored in a readily accessible form?

2.3.5 Is data transferred elsewhere?

2.3.5.1 If so, where and for what purpose?

2.3.6 Is data accessible from other locations?

2.3.7 How is data archived?

2.3.8 How often is data archived?

2.3.9 How is data retrieved from archive?

2.4 <u>Data Uses</u>

2.4.1 Users:

	<u>Name</u>	<u>Responsibility</u>	<u>Type of User (Operational, Scientific, Other)</u>	<u>Uses</u> (see 2.4.2)
(a)				
(b)				
(c)				
(d)				

Organisation chart
(if appropriate)

2.4.2 Describe how the data is used (purpose) and frequency of use (See also 2.2.2 - 2.2.4)

	Use	Dataset	Frequency
(i)			
(ii)			
(iii)			
(etc)			

2.4.3 How is the data accessed?

3.0 VIEWS ON CURRENT SYSTEM

3.1 Are there enough monitoring stations?

3.1.1 If not, why are more needed and how many?

3.2 Are there enough parameters monitored?

3.2.1 If not, why are more needed and which parameters?

(a)

(b)

(etc)

3.3 Is the monitoring frequency adequate?

3.3.1 If not, what rate is required and why?

3.4 **Is the data quality adequate**

3.4.1 If not, what improvements could be made?

3.5 **Is the system for raising alarms adequate?**

3.5.1 If not, what are the problems?

3.6 **Is the statistical and summary data adequate?**

3.6.1 If not, what additional processing and graphics is required?

3.7 **Is storage and archiving adequate?**

3.7.1 If not, what are the problems?

3.8 **Is data readily accessible (means/response time)?**

3.8.1 If not, what are the problems

3.9 **Are the links to other IT systems adequate?**

3.9.1 If not, what are the problems?

3.10 **Are there any additional inadequacies in the performance of the current system?**

4.0 FUTURE REQUIREMENTS

4.1 What additional uses are envisaged for automatic water quality data?

4.2 What additional data requirements will this generate?

(a) Parameters

(b) Frequency

(c) Calculation

4.3 What additional data processing requirements will this generate?

(a) Statistical techniques

(b) Graphics

(c) Other (please describe)

4.4 What storage requirements will this generate?

APPENDIX B

TWO PARAMETER PLOTS OF WATER QUALITY DATA

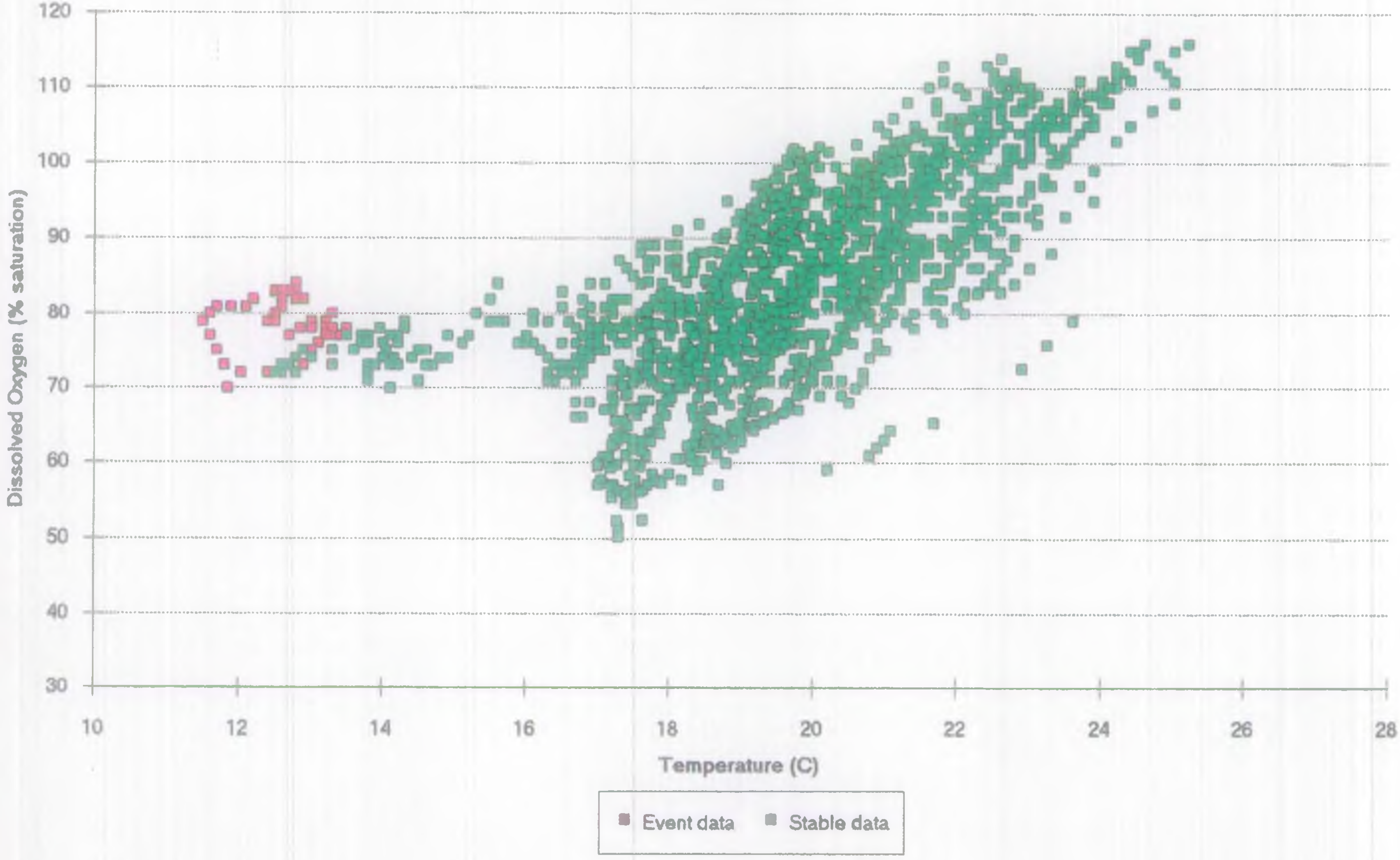


Figure B1 Graph of Dissolved Oxygen vs Temperature for Stable Data and an Event on the 7-8th June, 1989

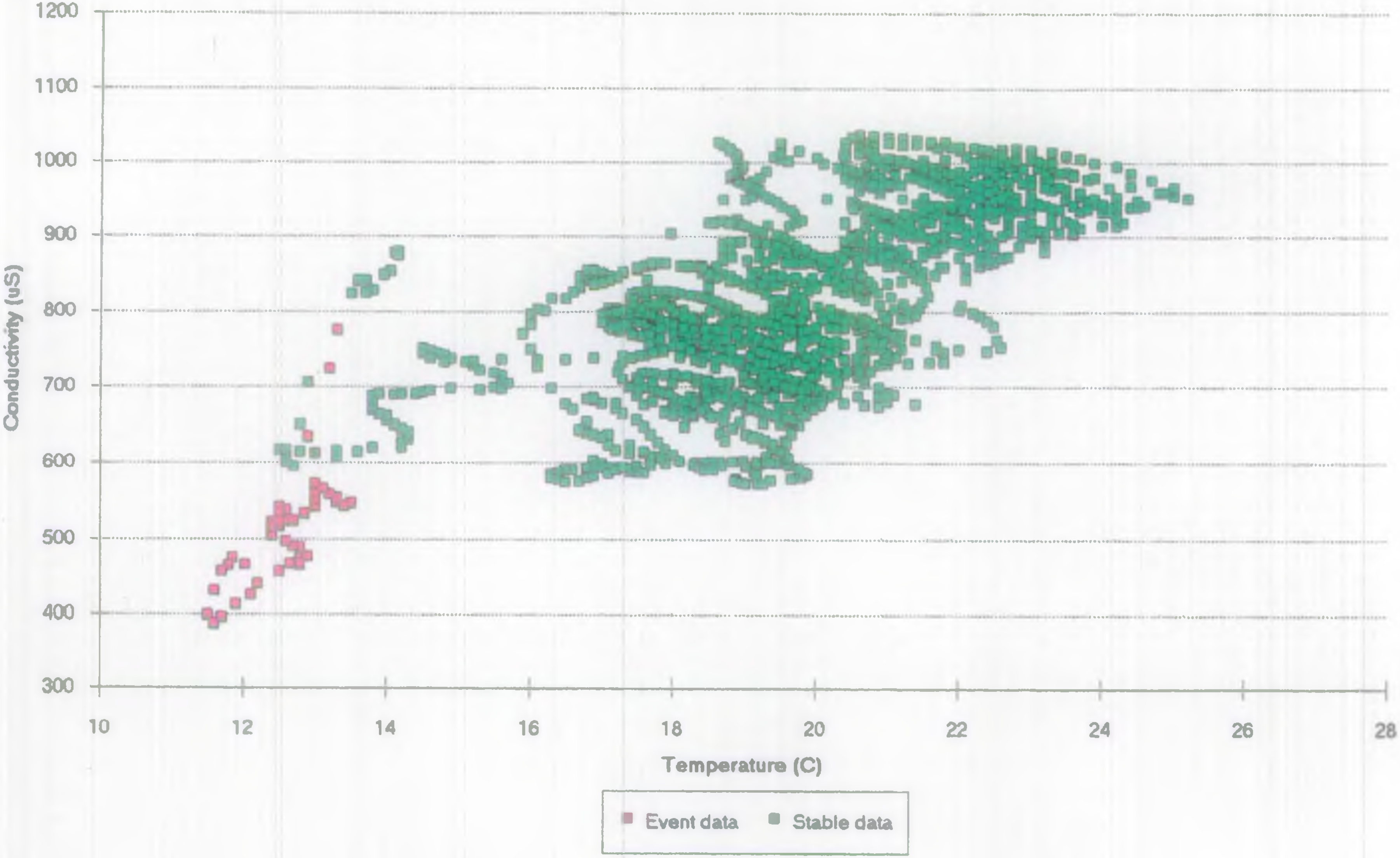


Figure B2 Graph of Conductivity vs Temperature for Stable Data and an Event on the 7-8th June, 1989

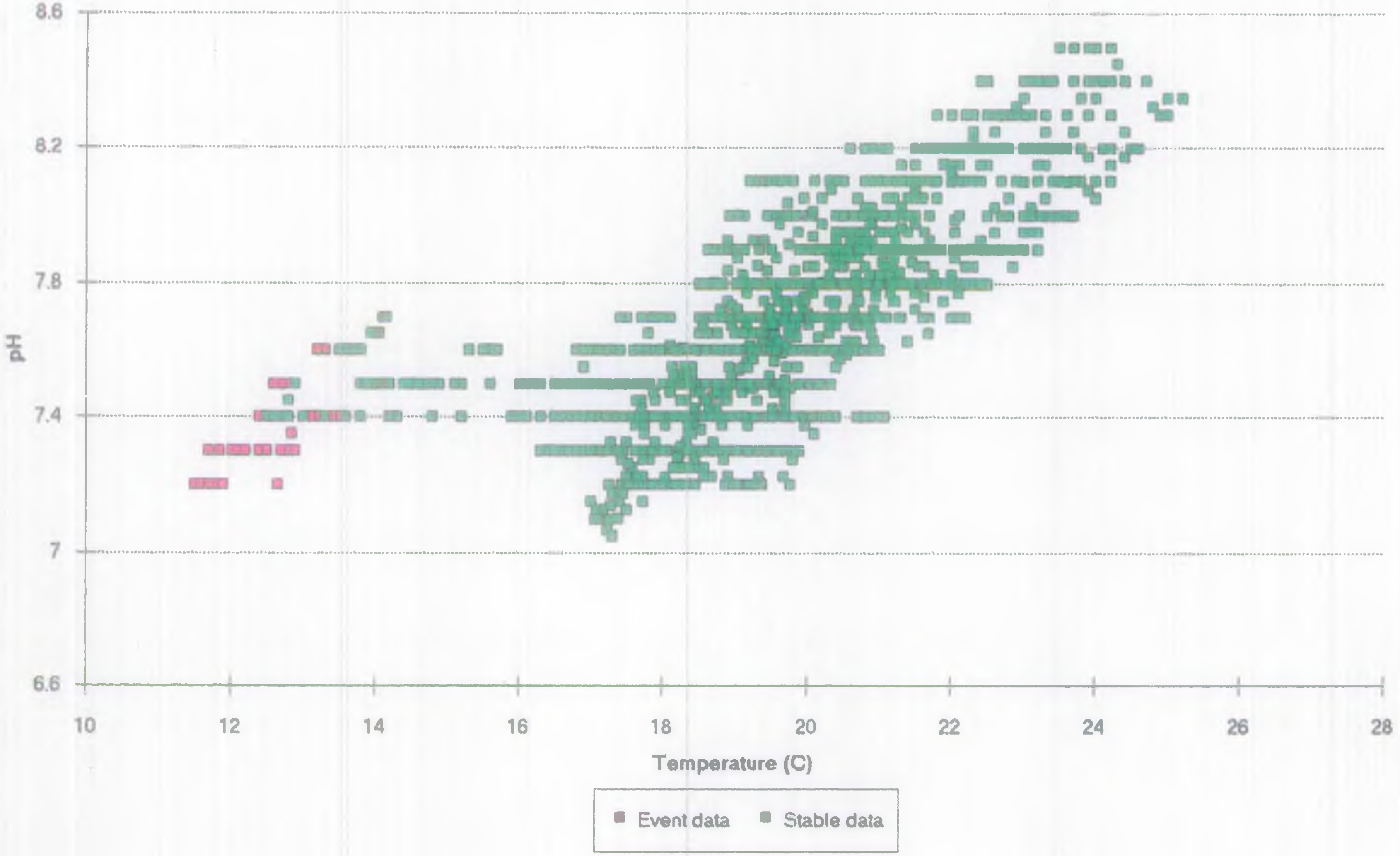


Figure B3 Graph of pH vs Temperature for Stable Data and an Event on the 7-8th June, 1989

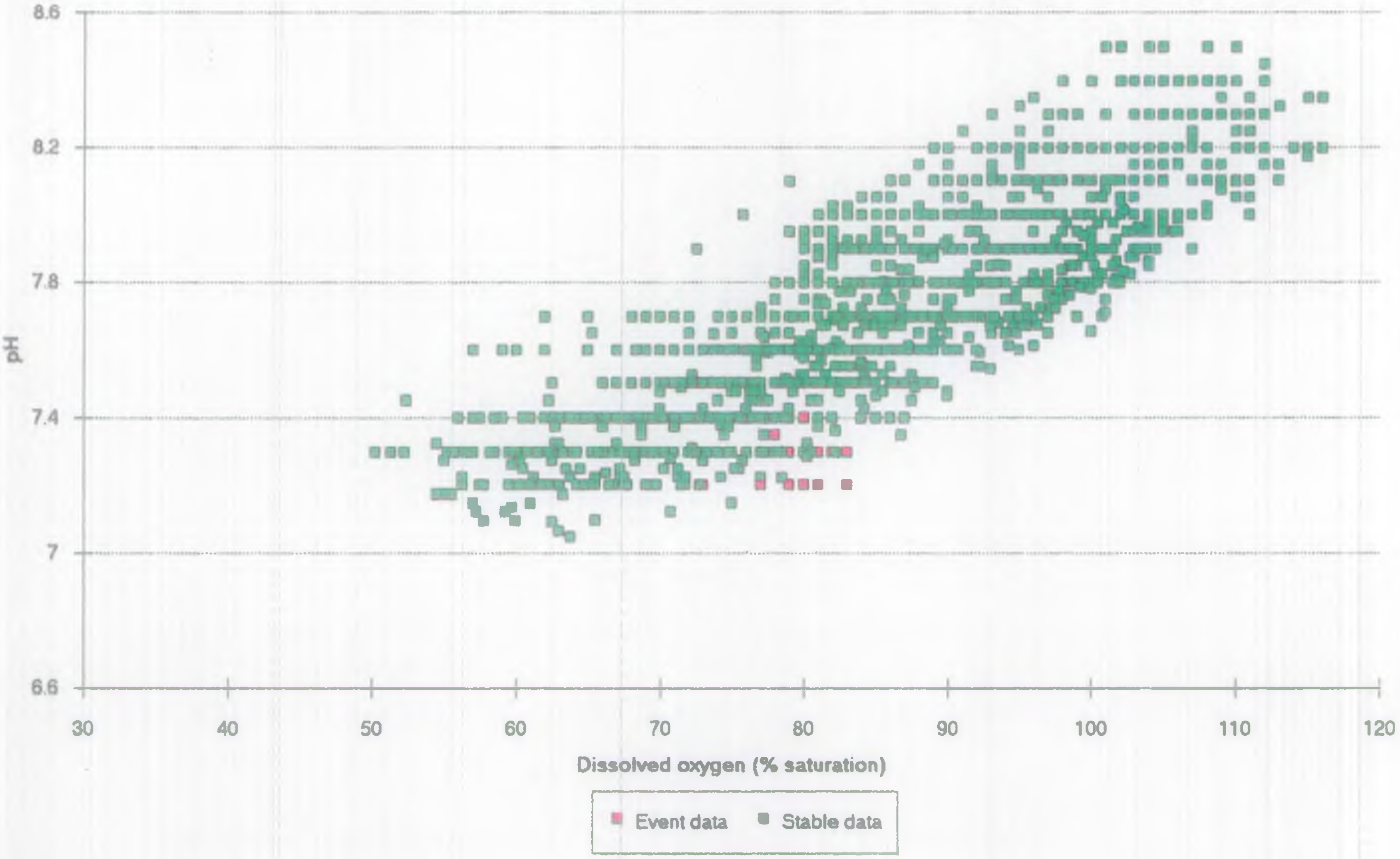


Figure B4 Graph of pH vs Dissolved Oxygen for Stable Data and an Event on the 7-8th June, 1989

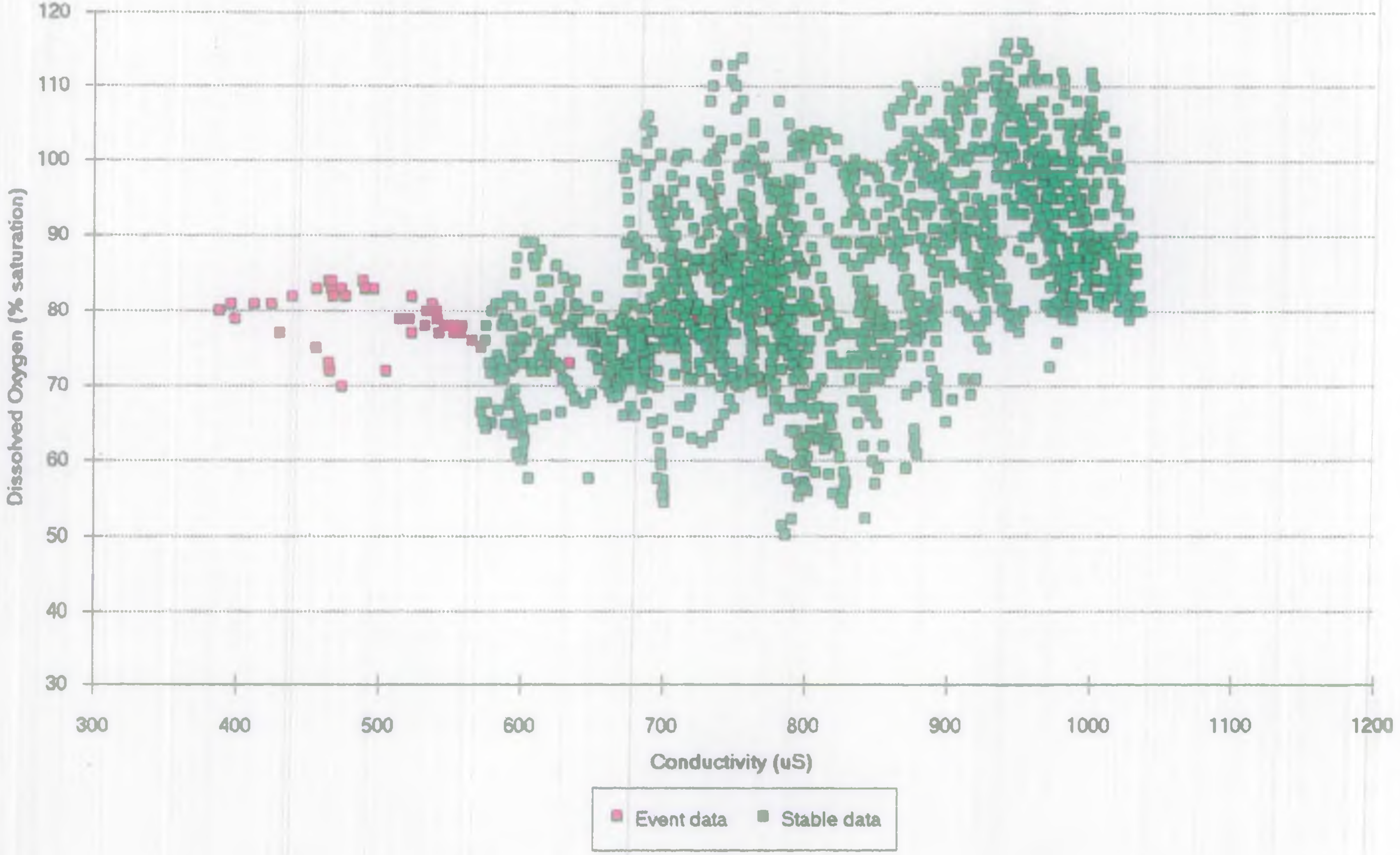


Figure B5 Graph of Dissolved Oxygen vs Conductivity for Stable Data and an Event on the 7-8th June, 1989

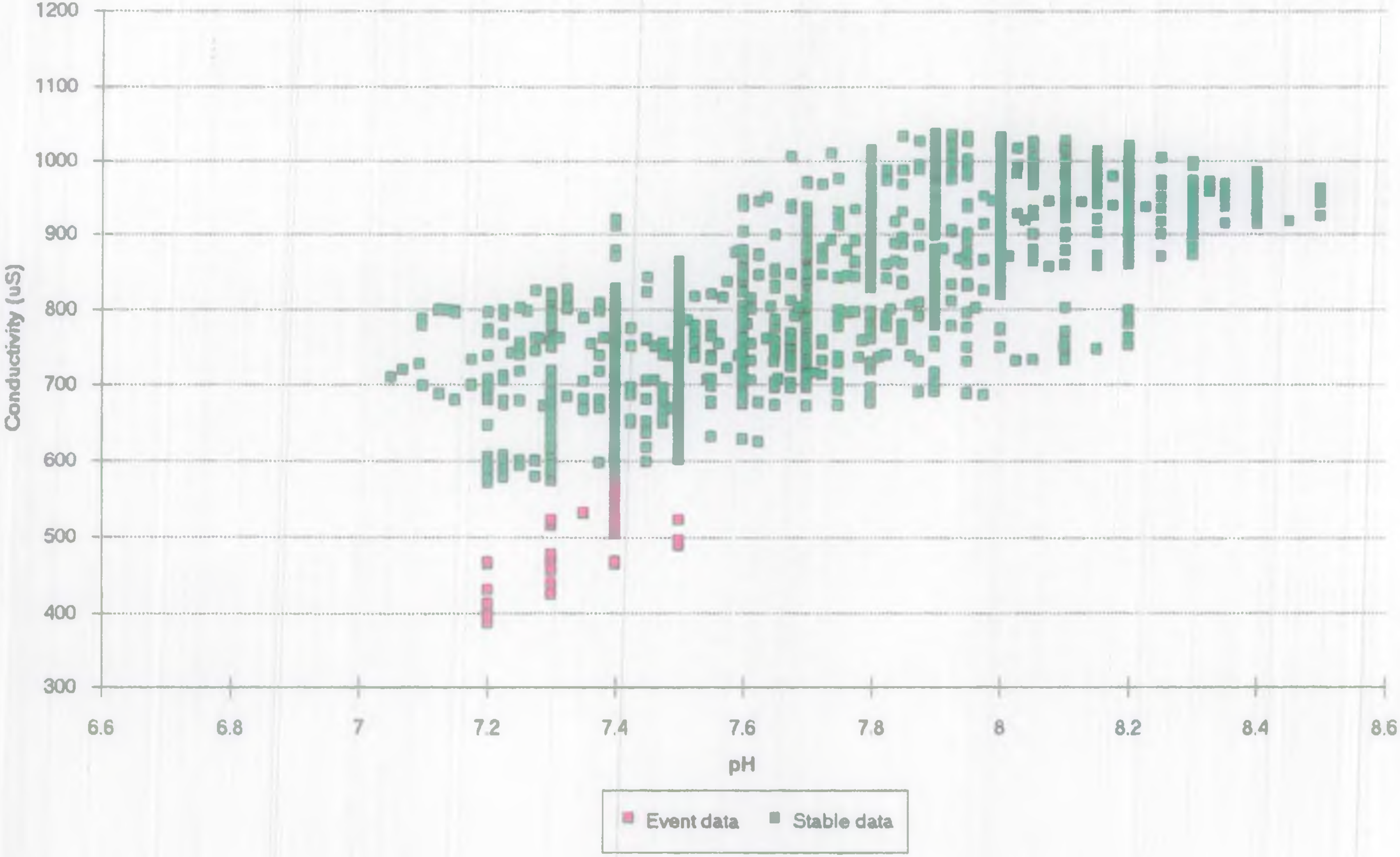


Figure B6 Graph of Conductivity vs pH for Stable Data and an Event on the 7-8th June, 1989

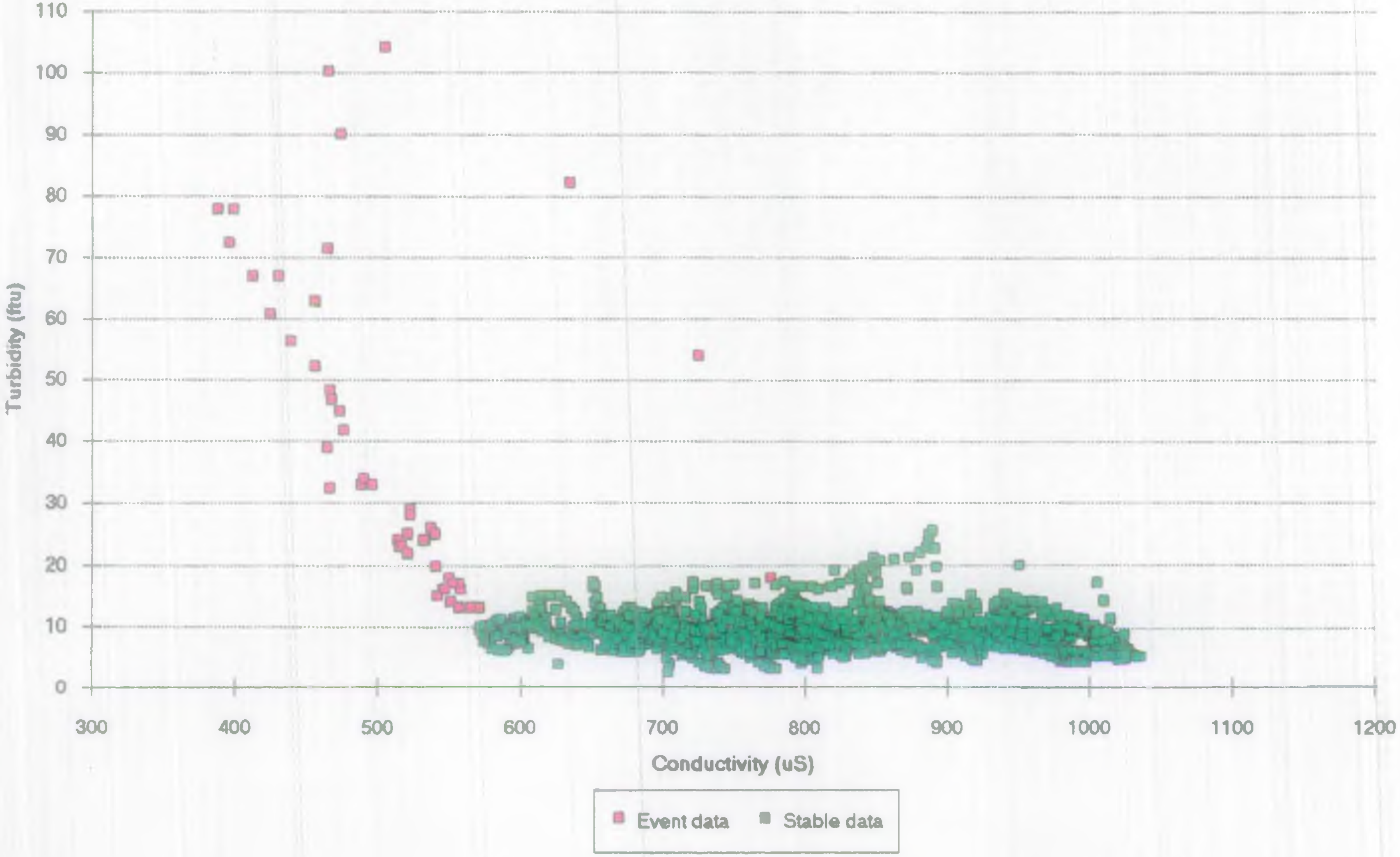


Figure B7 Graph of Turbidity vs Conductivity for Stable Data and an Event on the 7-8th June, 1989

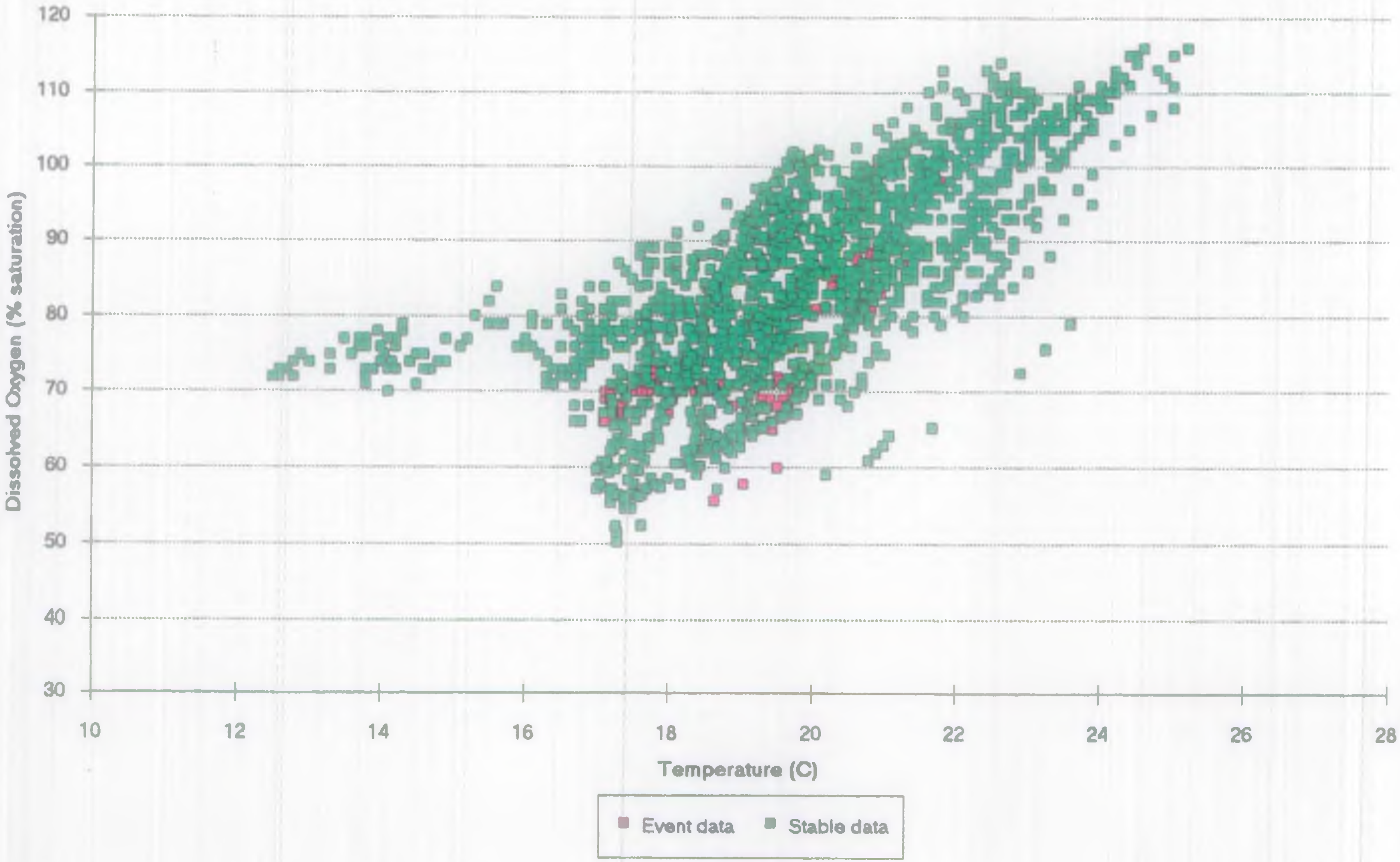


Figure B8 Graph of Dissolved Oxygen vs Temperature for Stable Data and an Event on the 7-11th July, 1989

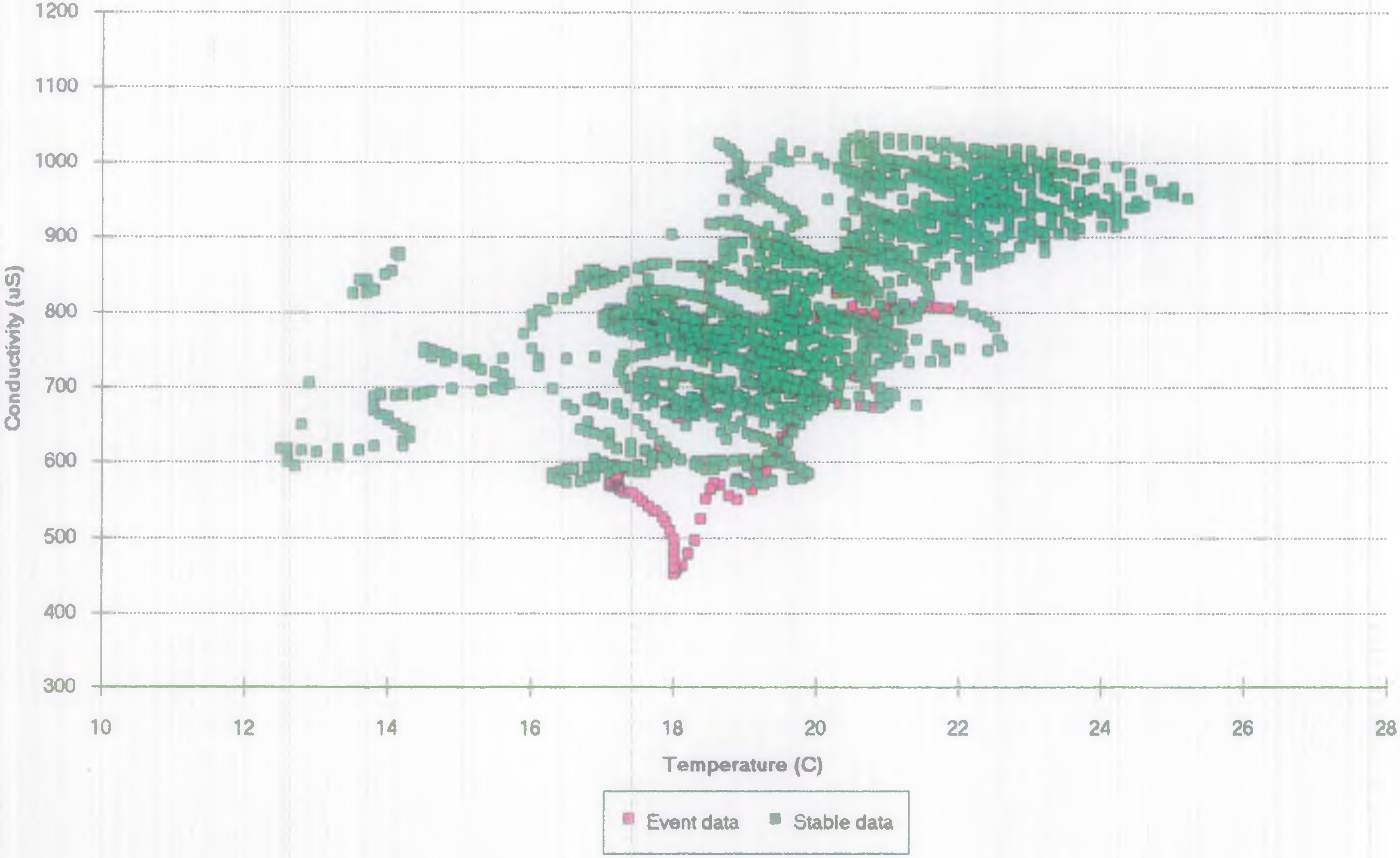


Figure B9 Graph of Conductivity vs Temperature for Stable Data and an Event on the 7-11th July, 1989

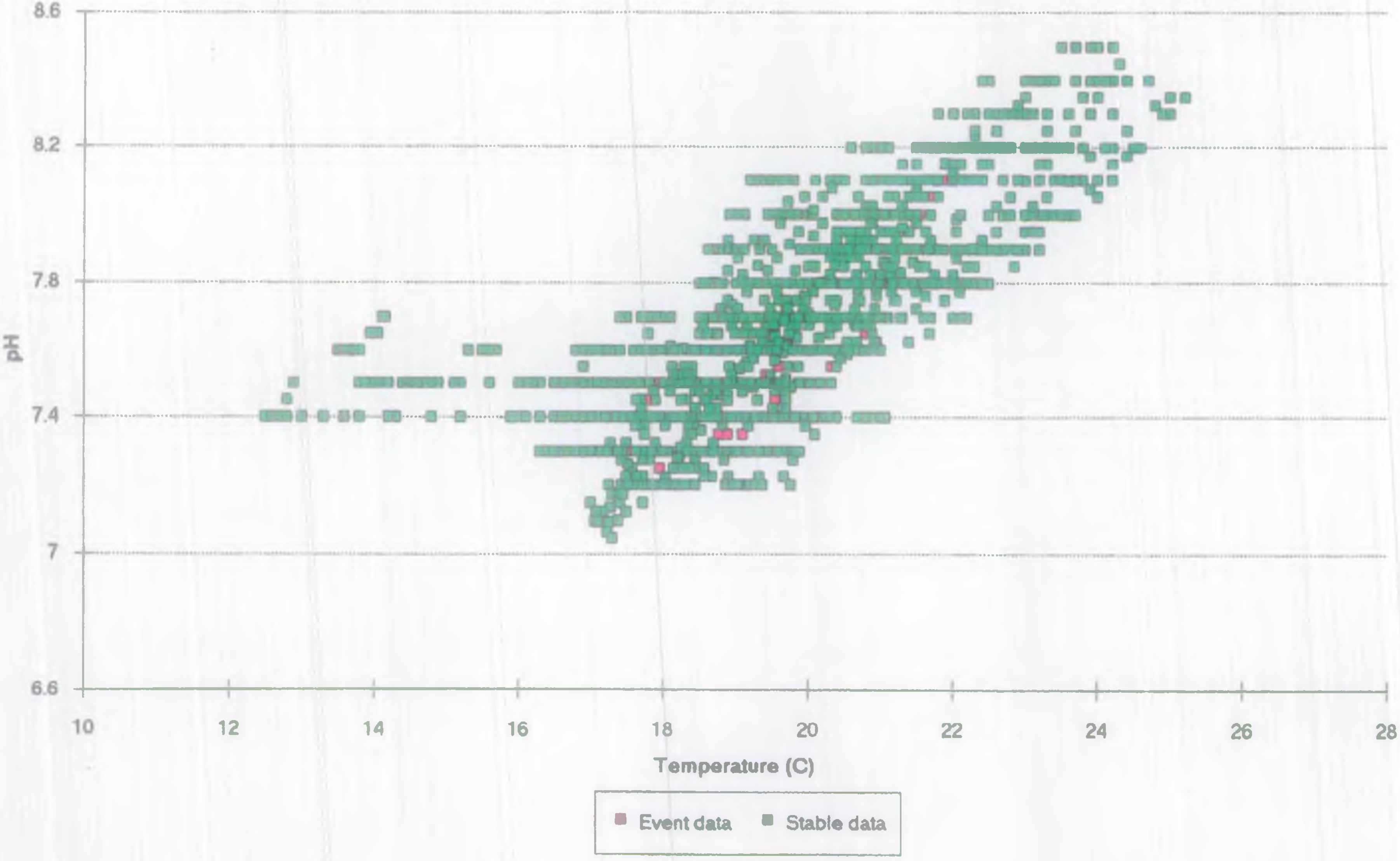


Figure B10 Graph of pH vs Temperature for Stable Data and an Event on the 7-11th July, 1989

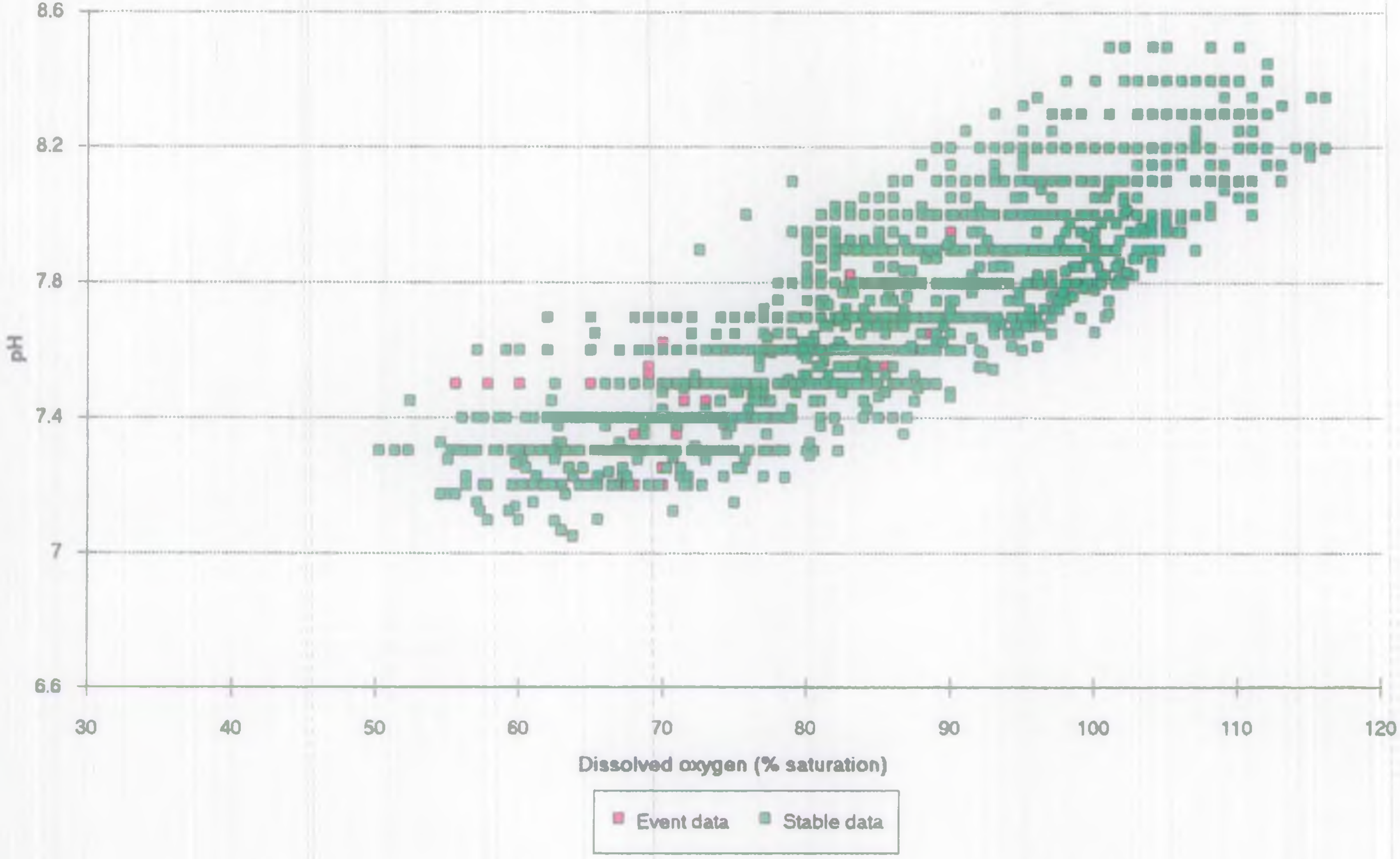


Figure B11 Graph of pH vs Dissolved Oxygen for Stable Data and an Event on the 7-11th July, 1989

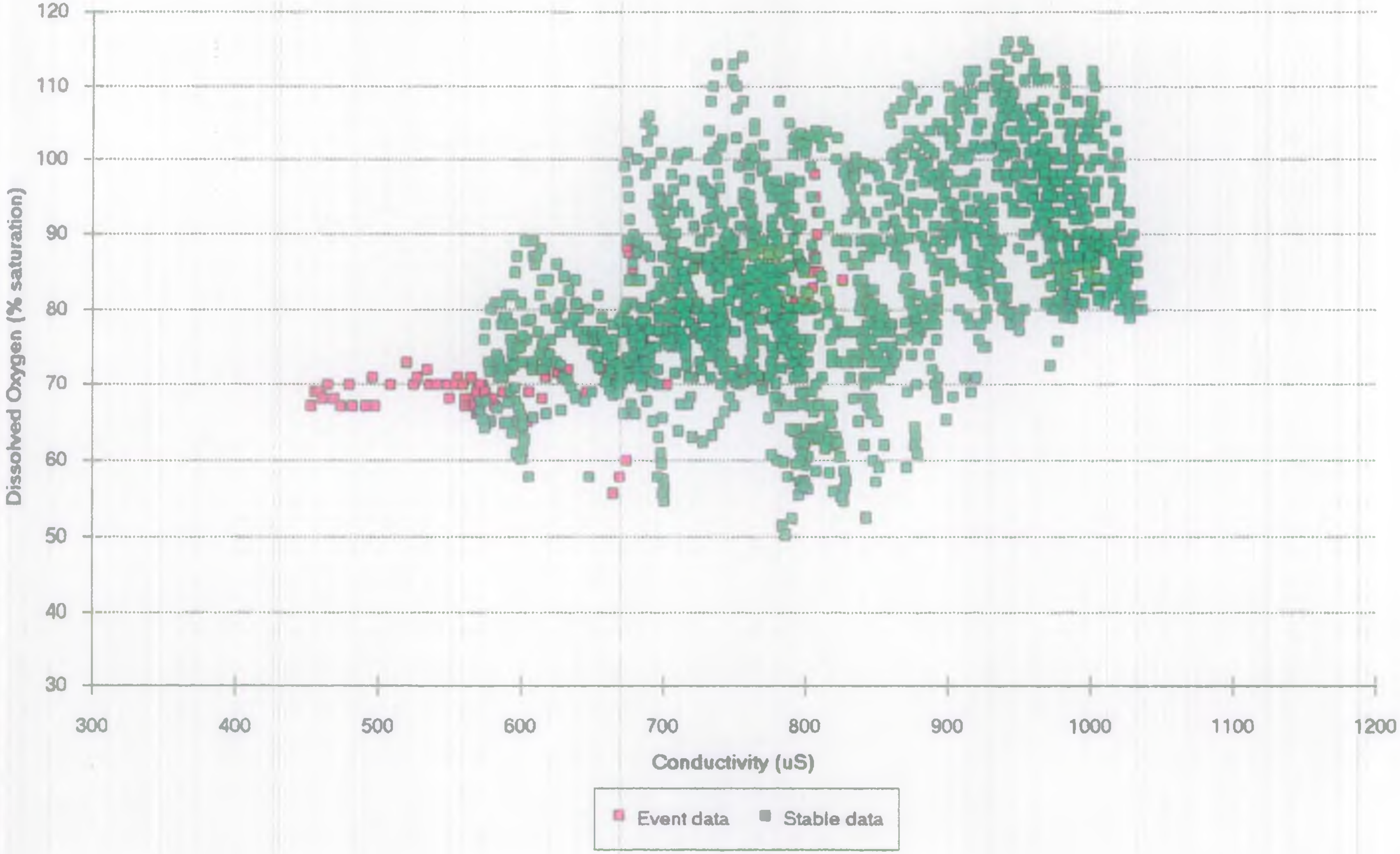


Figure B12 Graph of Dissolved Oxygen vs Conductivity for Stable Data and an Event on the 7-11th July, 1989

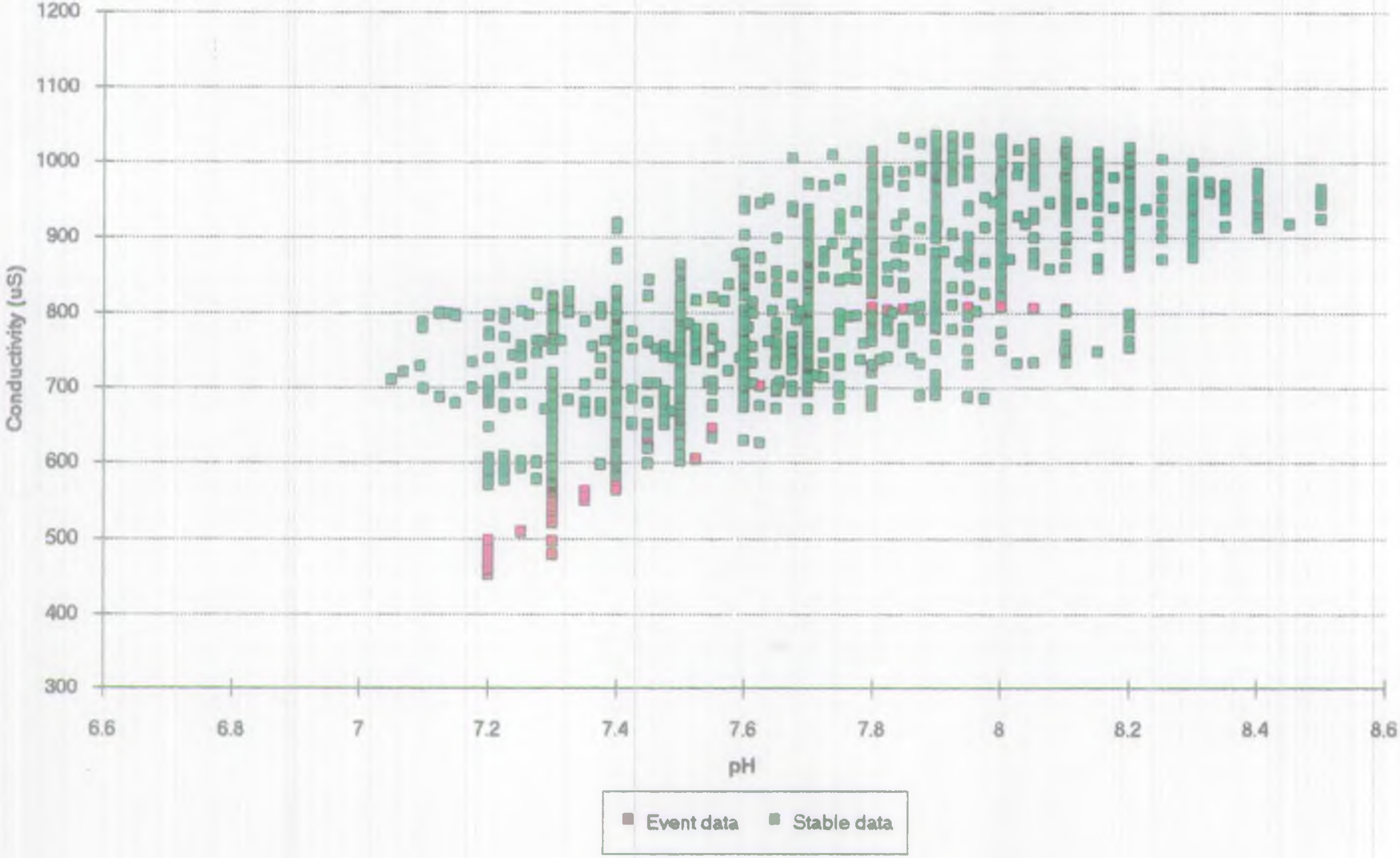


Figure B13 Graph of Conductivity vs pH for Stable Data and an Event on the 7-11th July, 1989

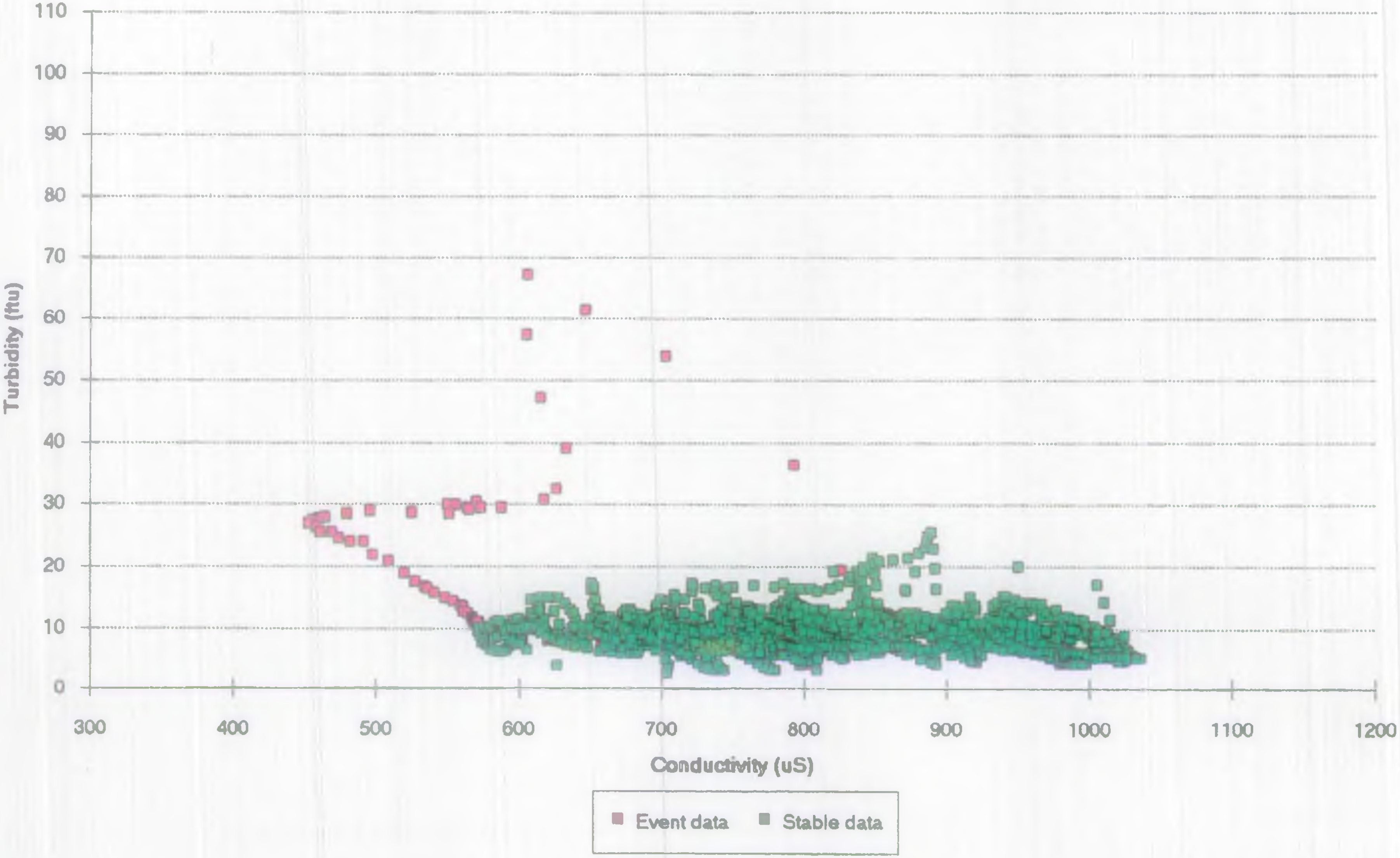


Figure B14 Graph of Turbidity vs Conductivity for Stable Data and an Event on the 7-11th July, 1989

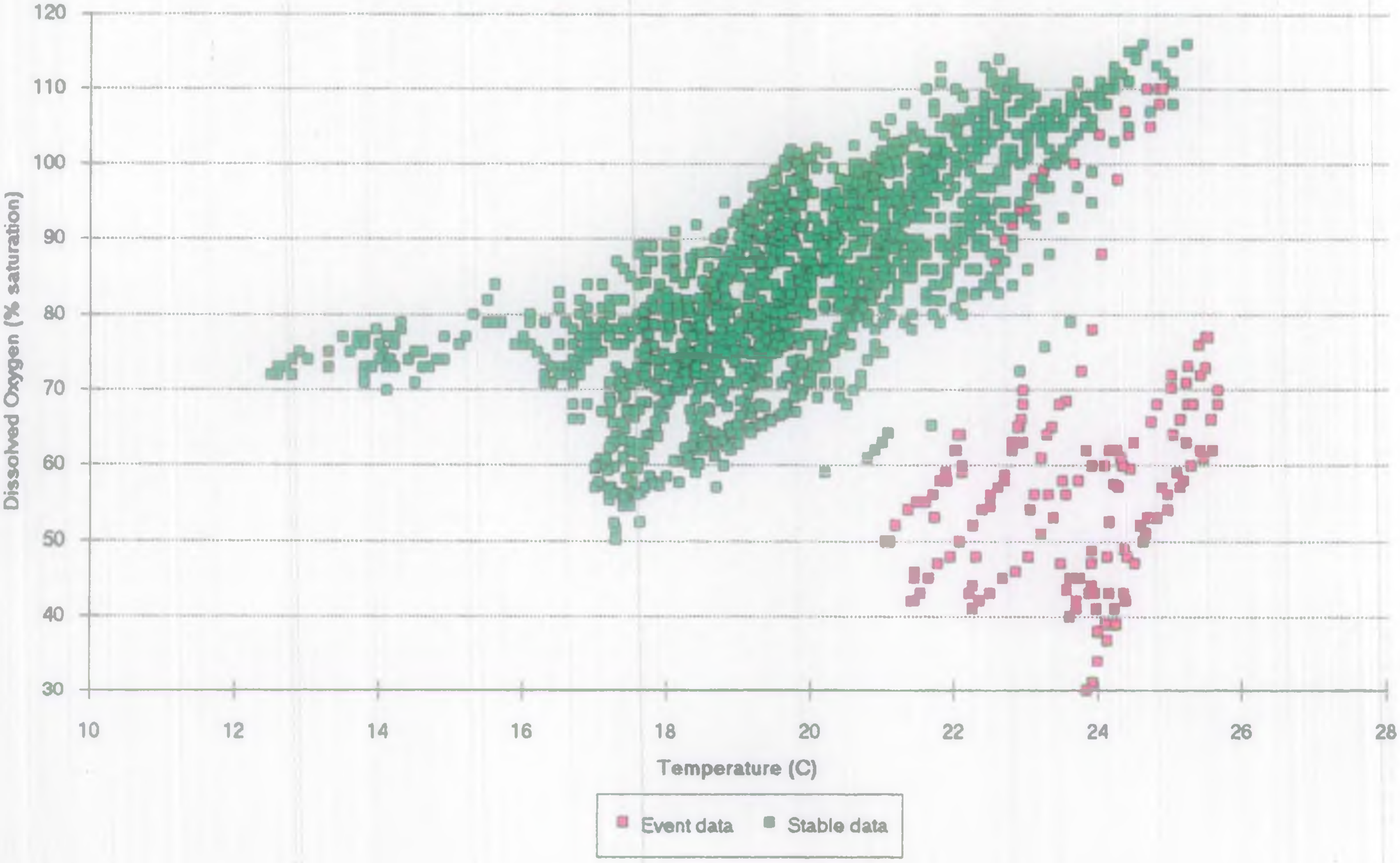


Figure B15 Graph of Dissolved Oxygen vs Temperature for Stable Data and an Event on the 22-29th July, 1989

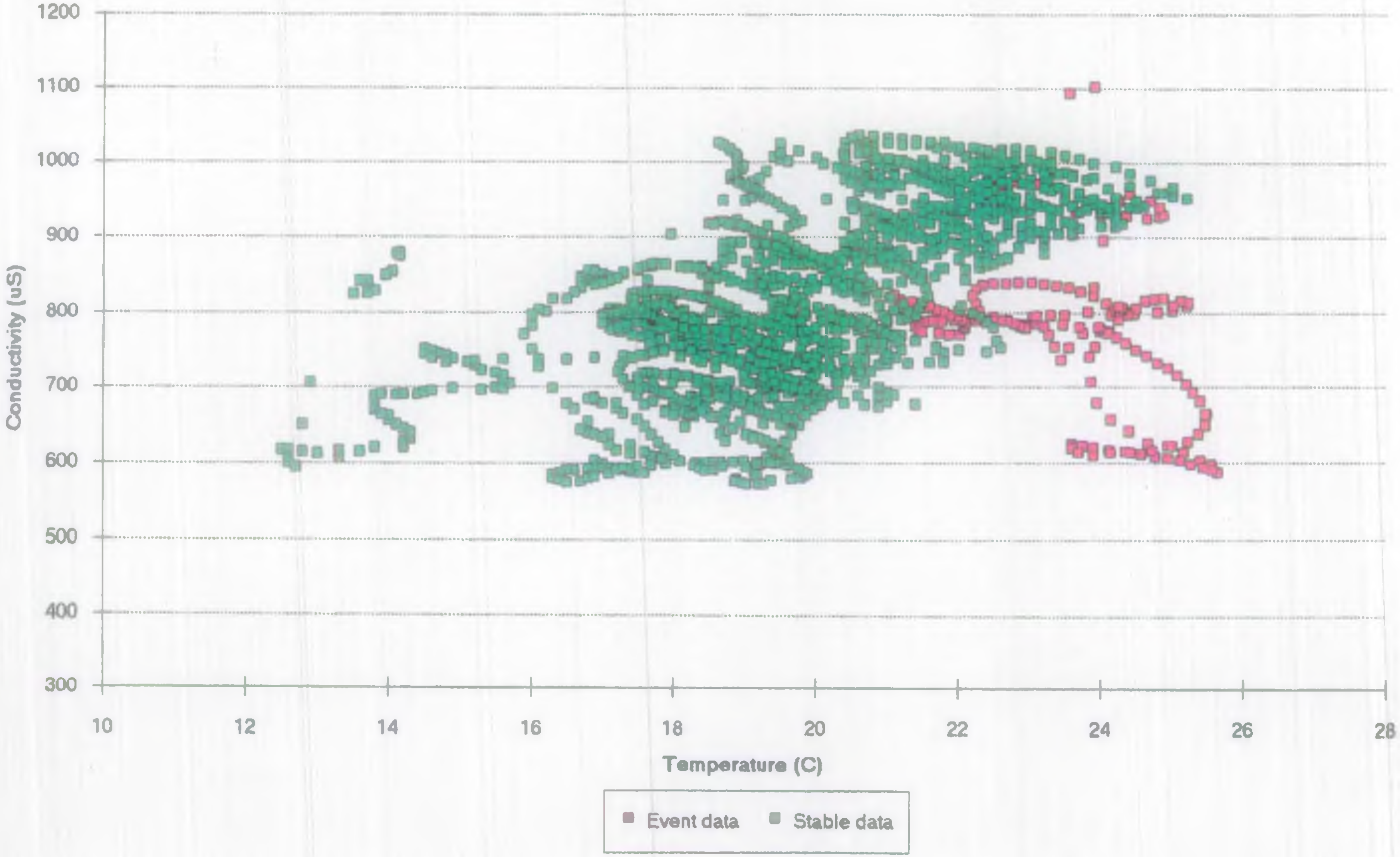


Figure B16 Graph of Conductivity vs Temperature for Stable Data and an Event on the 22-29th July, 1989

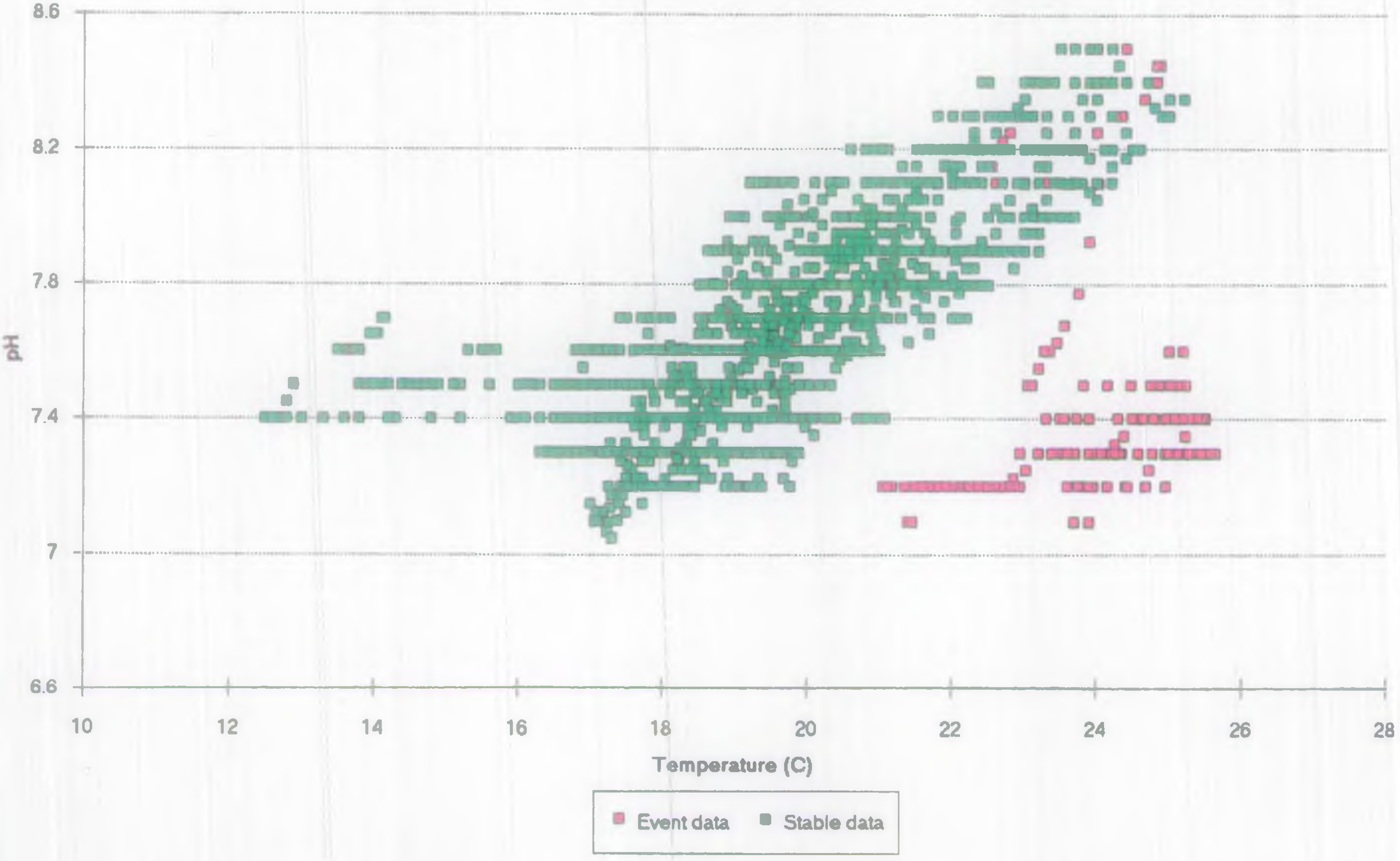


Figure B17 Graph of pH vs Temperature for Stable Data and an Event on the 22-29th July, 1989

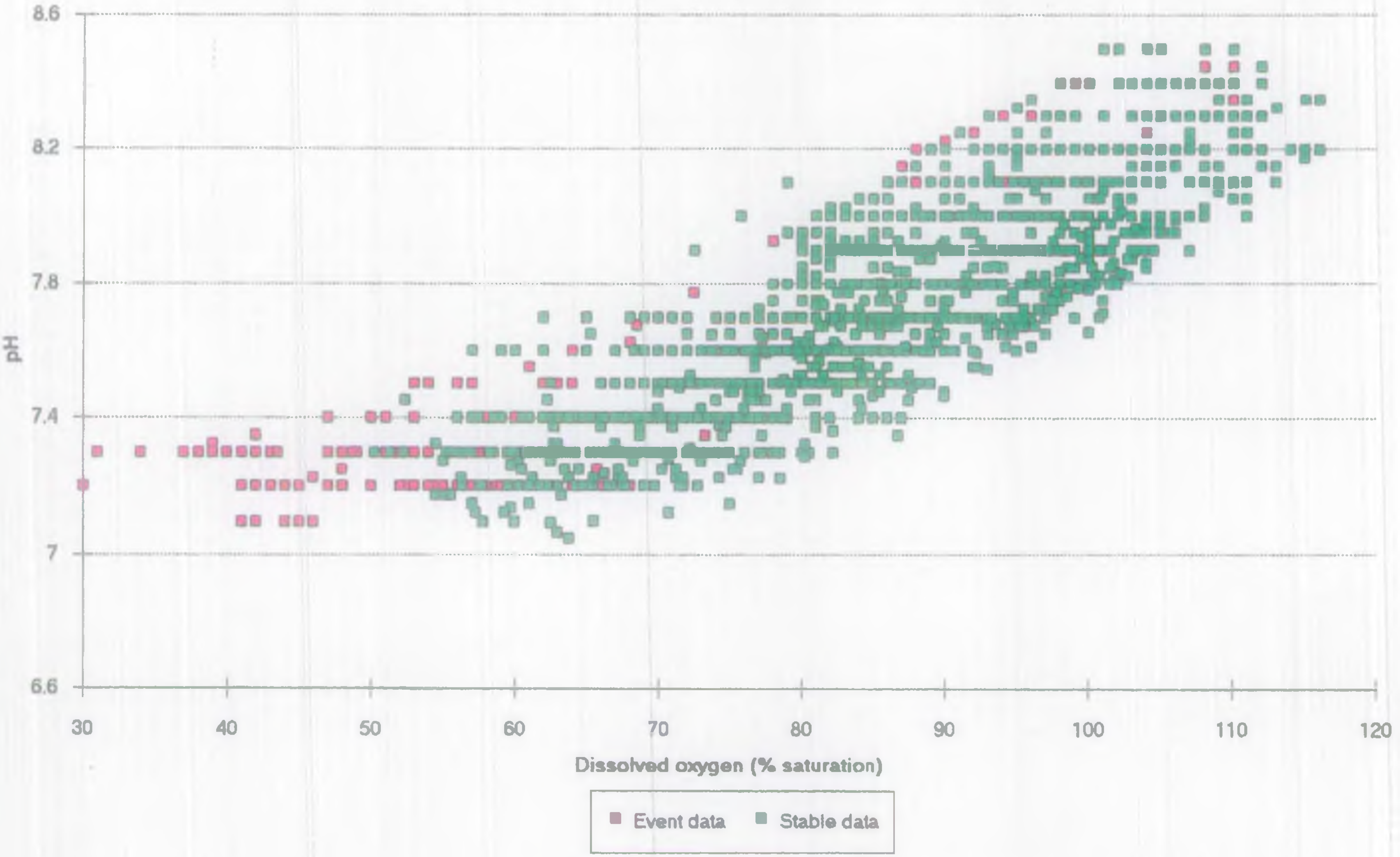


Figure B18 Graph of pH vs Dissolved Oxygen for Stable Data and an Event on the 22-29th July, 1989

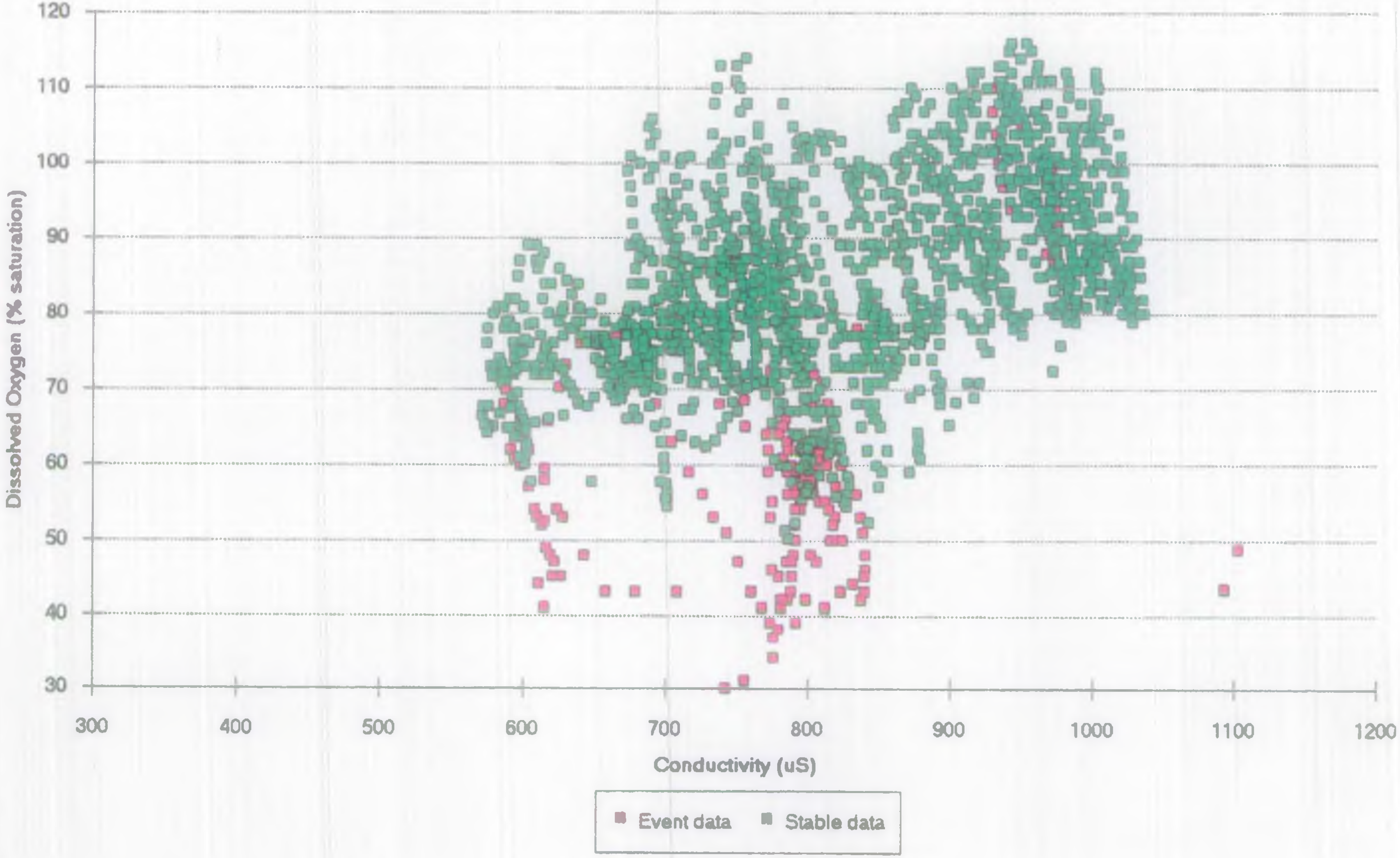


Figure B19 Graph of Dissolved Oxygen vs Conductivity for Stable Data and an Event on the 22-29th July, 1989

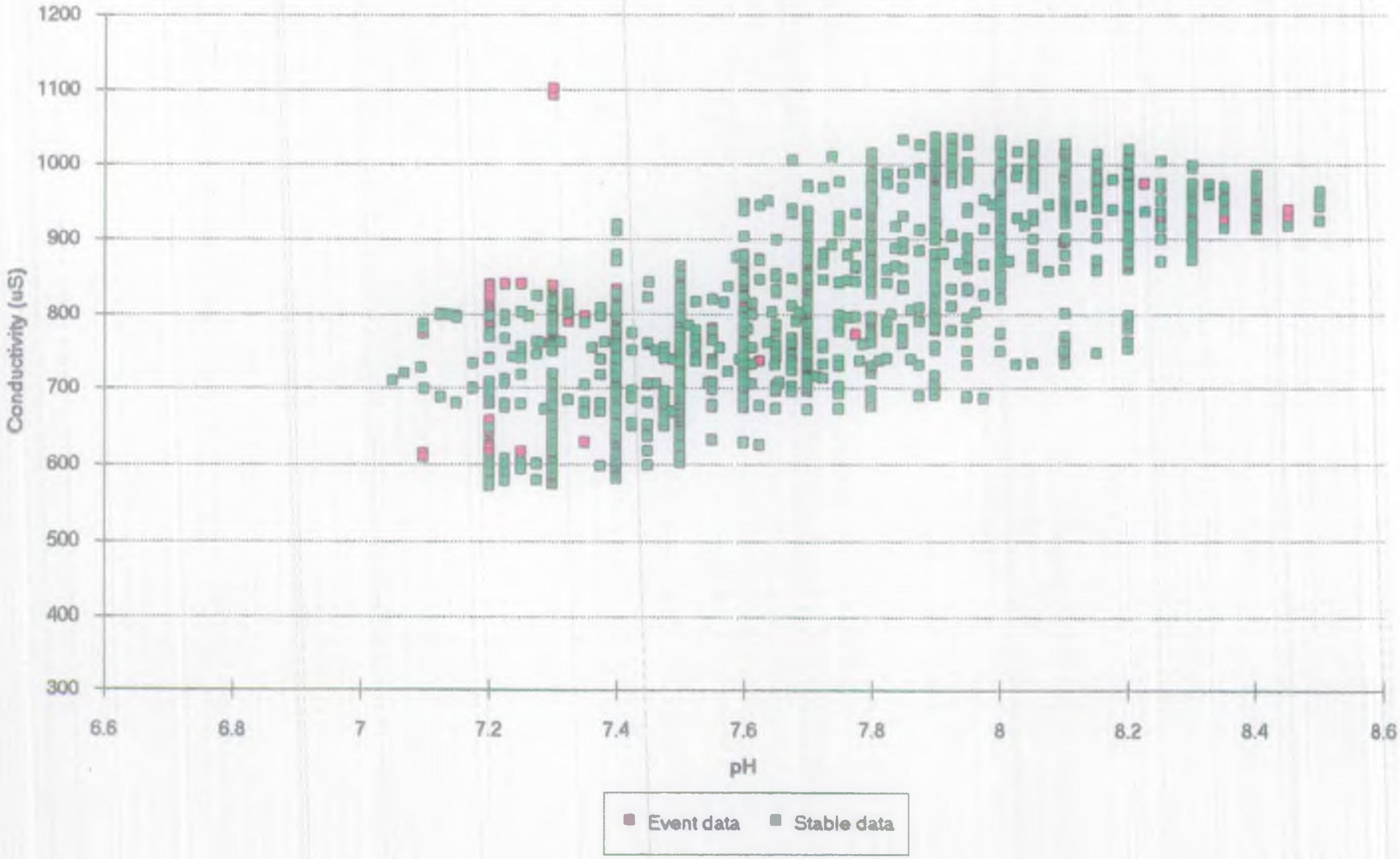


Figure B20 Graph of Conductivity vs pH for Stable Data and an Event on the 22-29th July, 1989

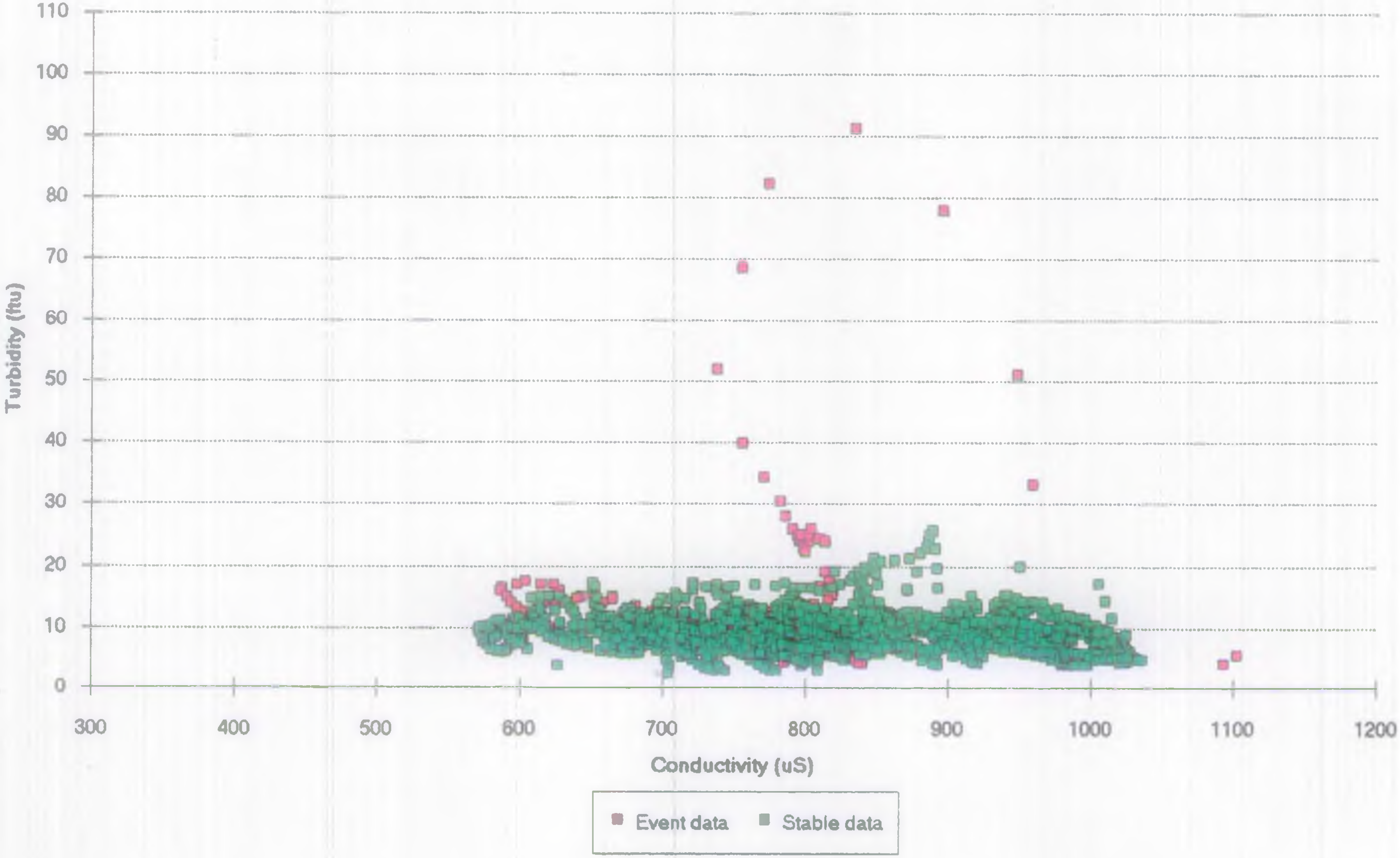


Figure B21 Graph of Turbidity vs Conductivity for Stable Data and an Event on the 22-29th July, 1989

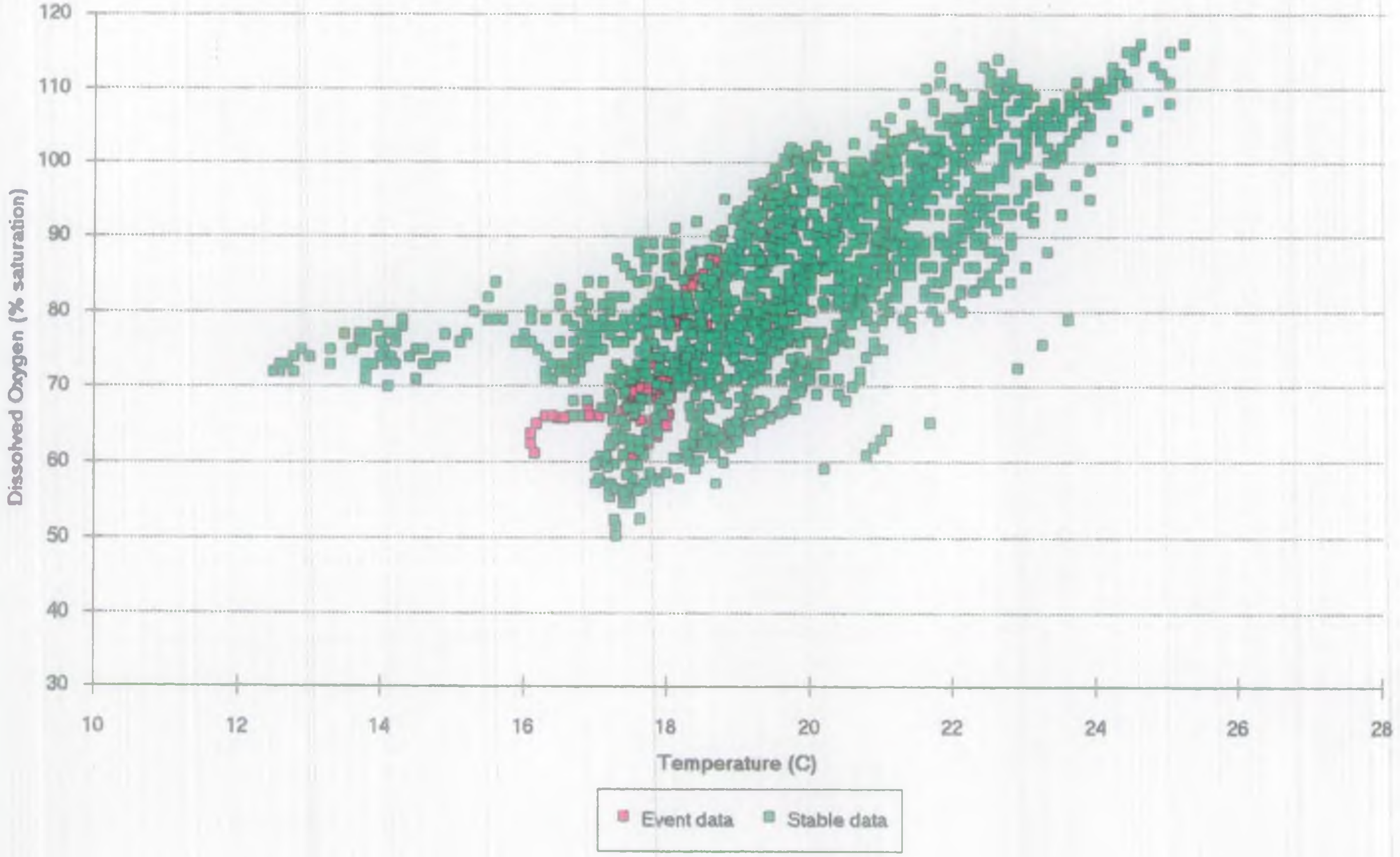


Figure B22 Graph of Dissolved Oxygen vs Temperature for Stable Data and an Event on the 26-28th August, 1989

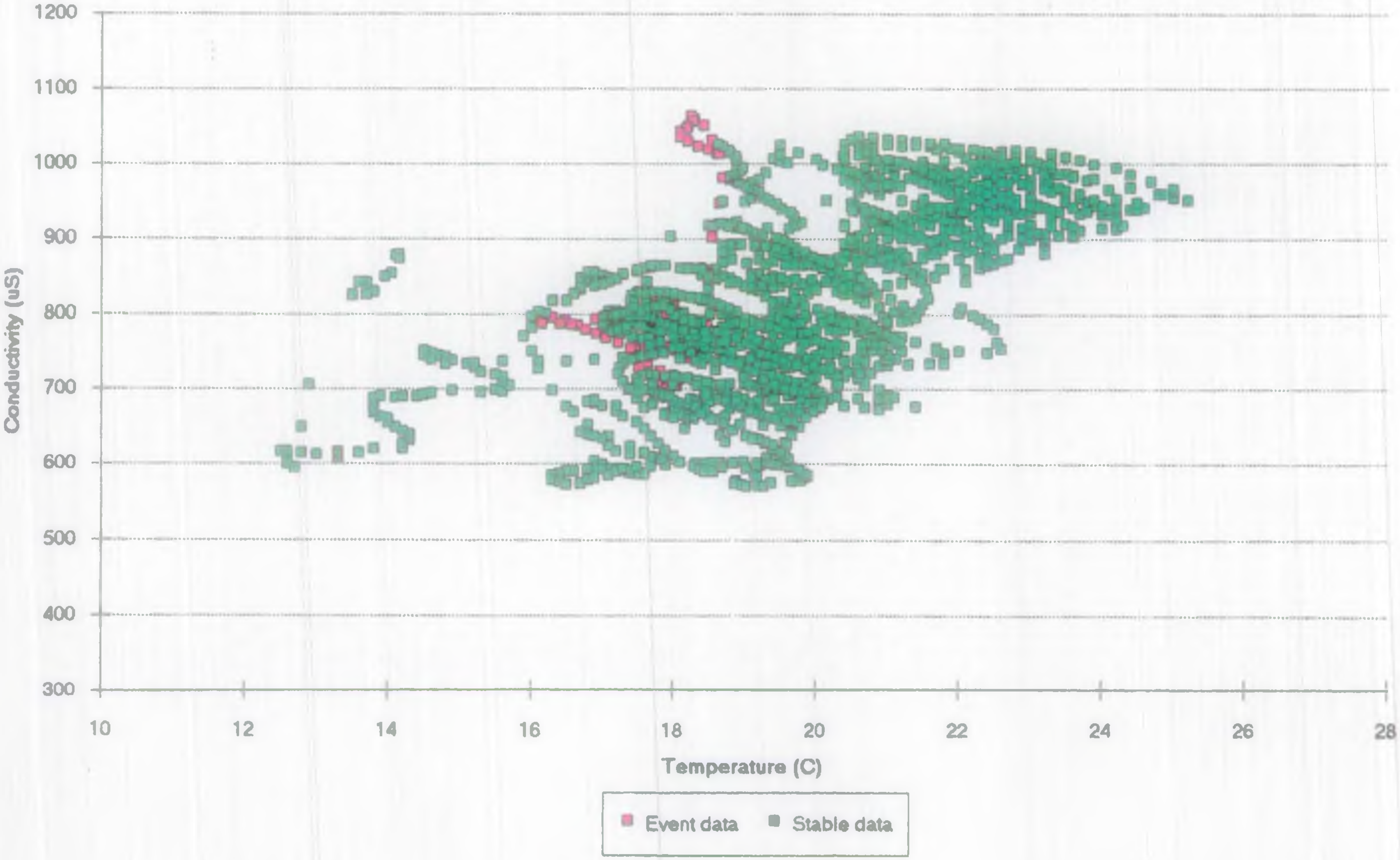


Figure B23 Graph of Conductivity vs Temperature for Stable Data and an Event on the 26-28th August 1989

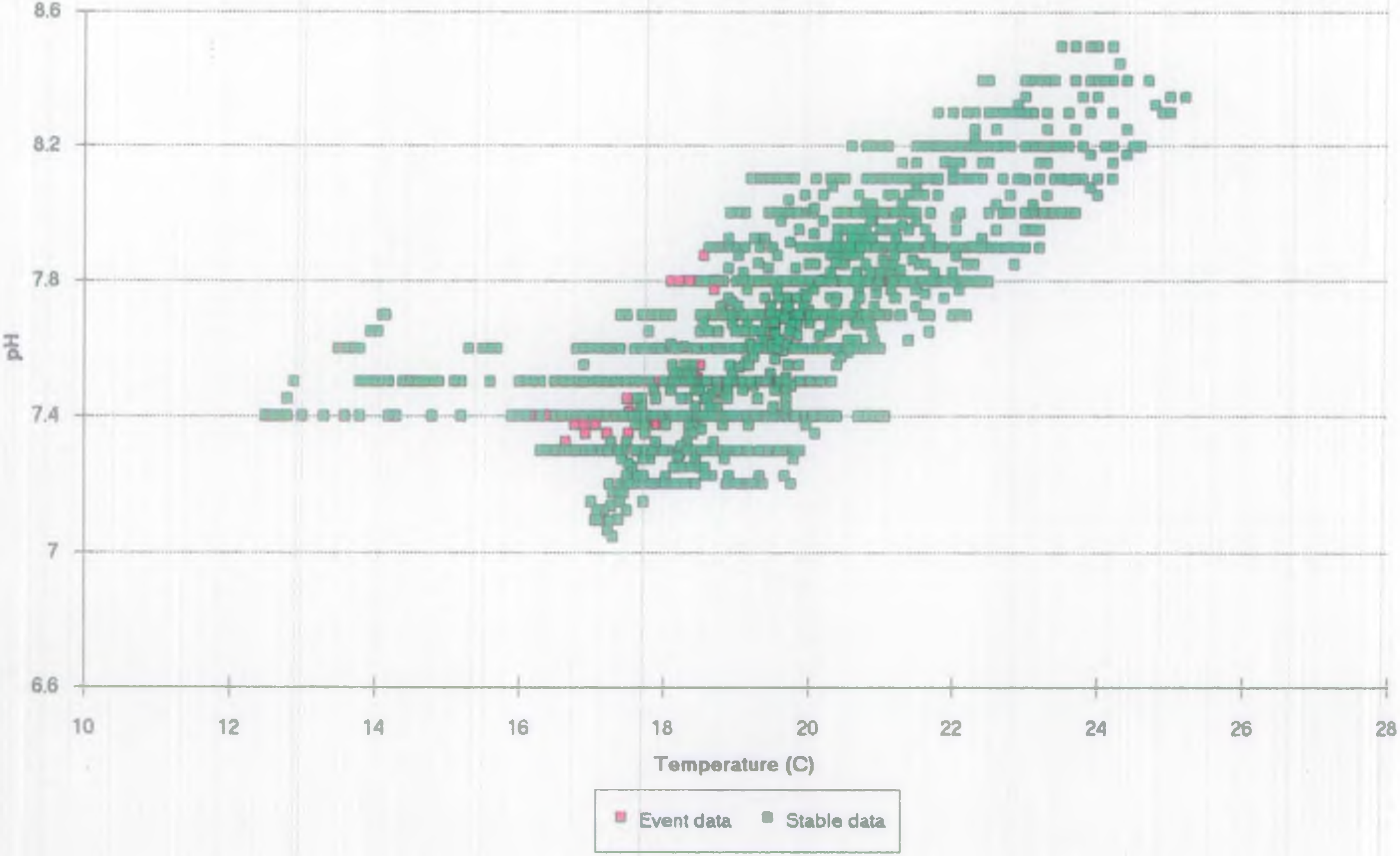


Figure B24 Graph of pH vs Temperature for Stable Data and an Event on the 26-28th August, 1989

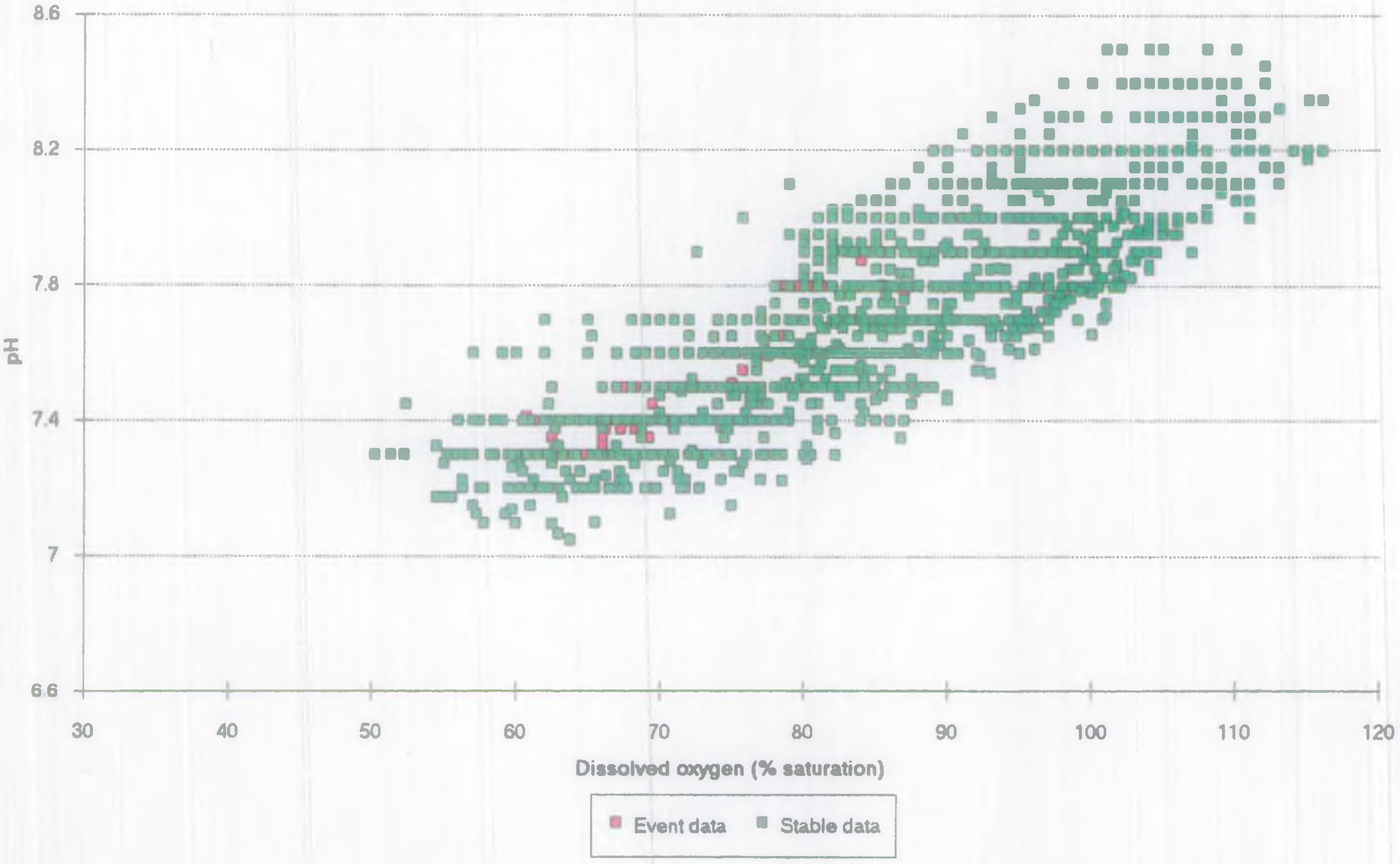


Figure B25 Graph of pH vs Dissolved Oxygen for Stable Data and an Event on the 26-28th August 1989

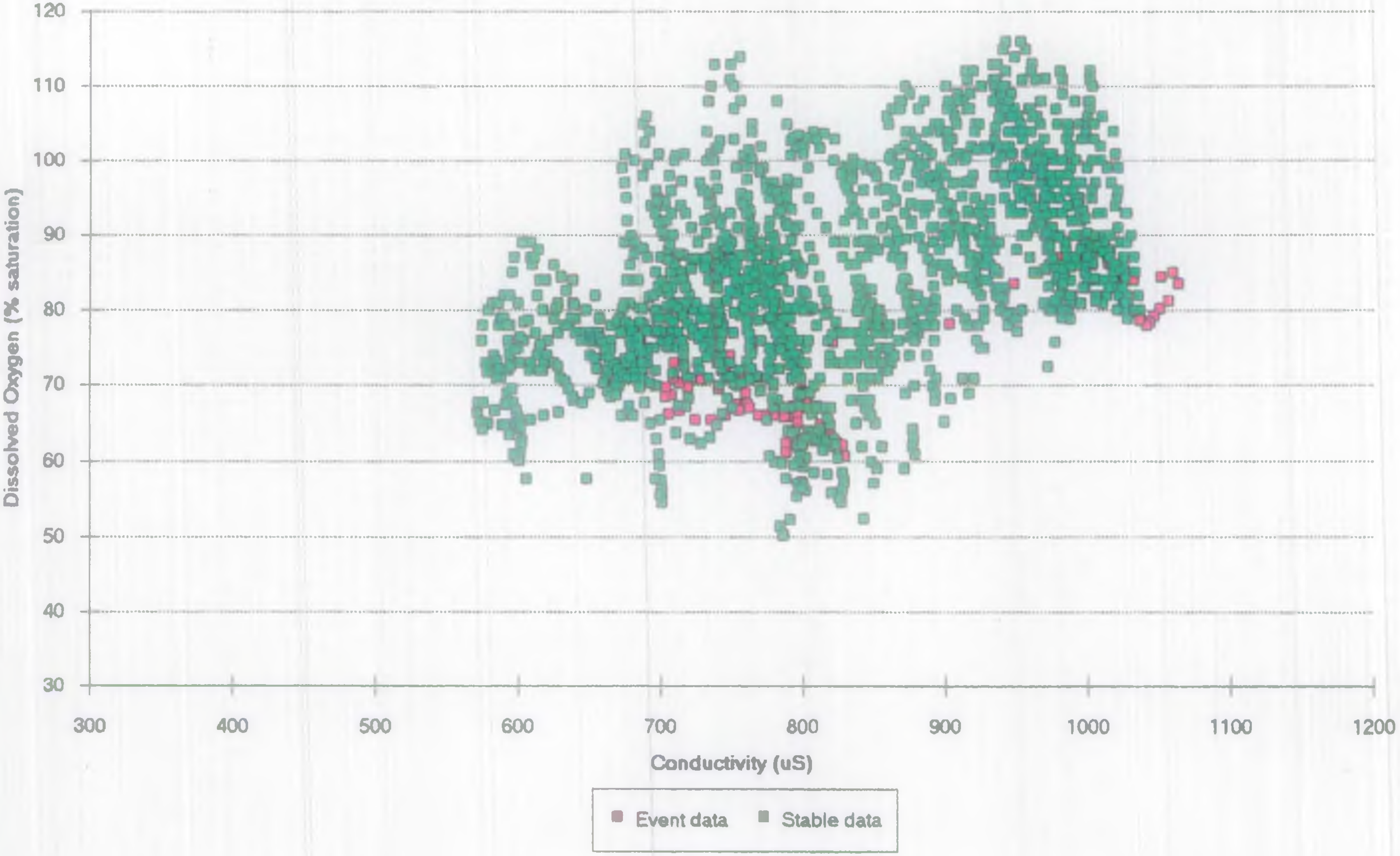


Figure B26 Graph of Dissolved Oxygen vs Conductivity for Stable Data and an Event on the 26-28th August, 1989

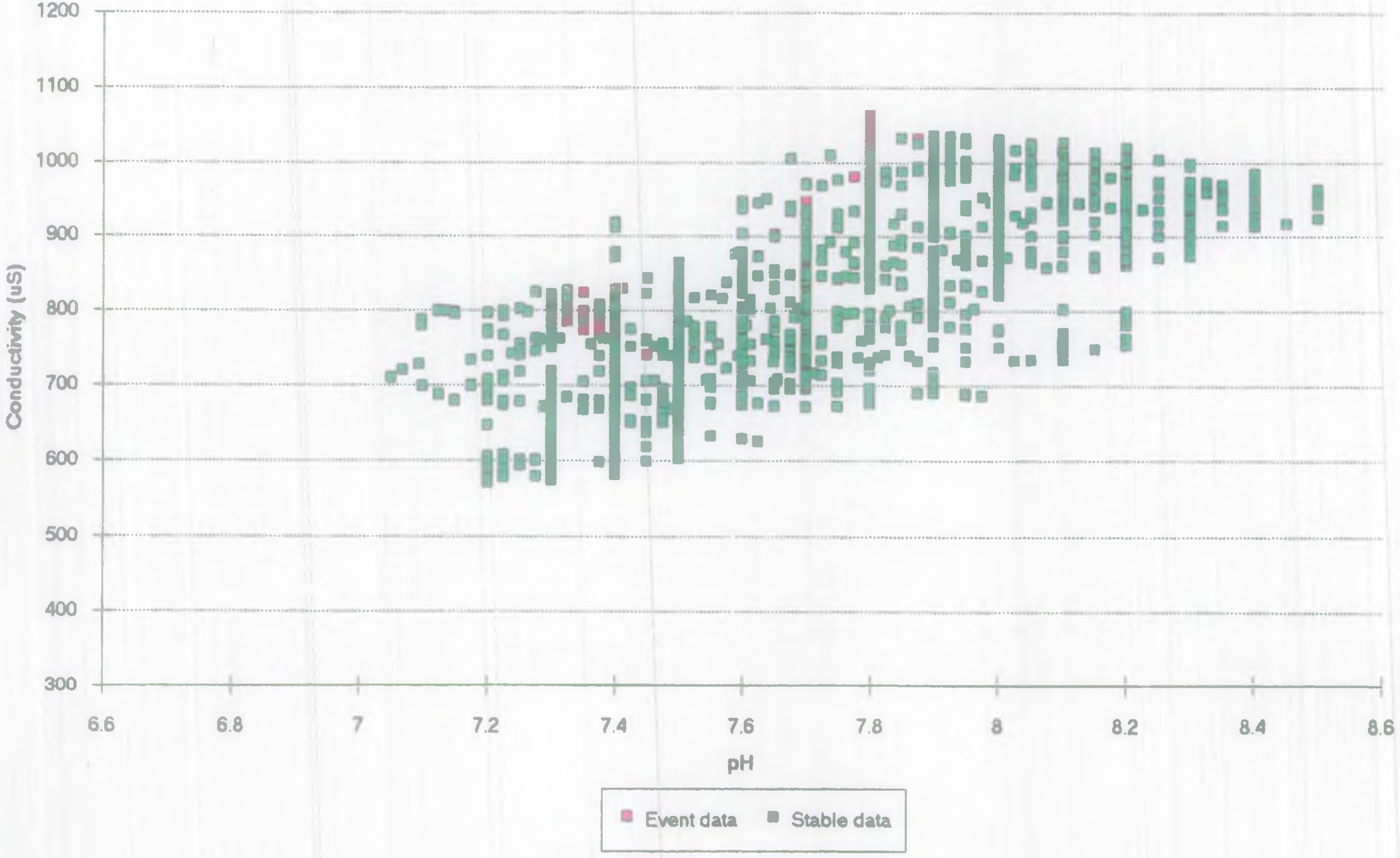


Figure B27 Graph of Conductivity vs pH for Stable Data and an Event on the 26-28th August 1989

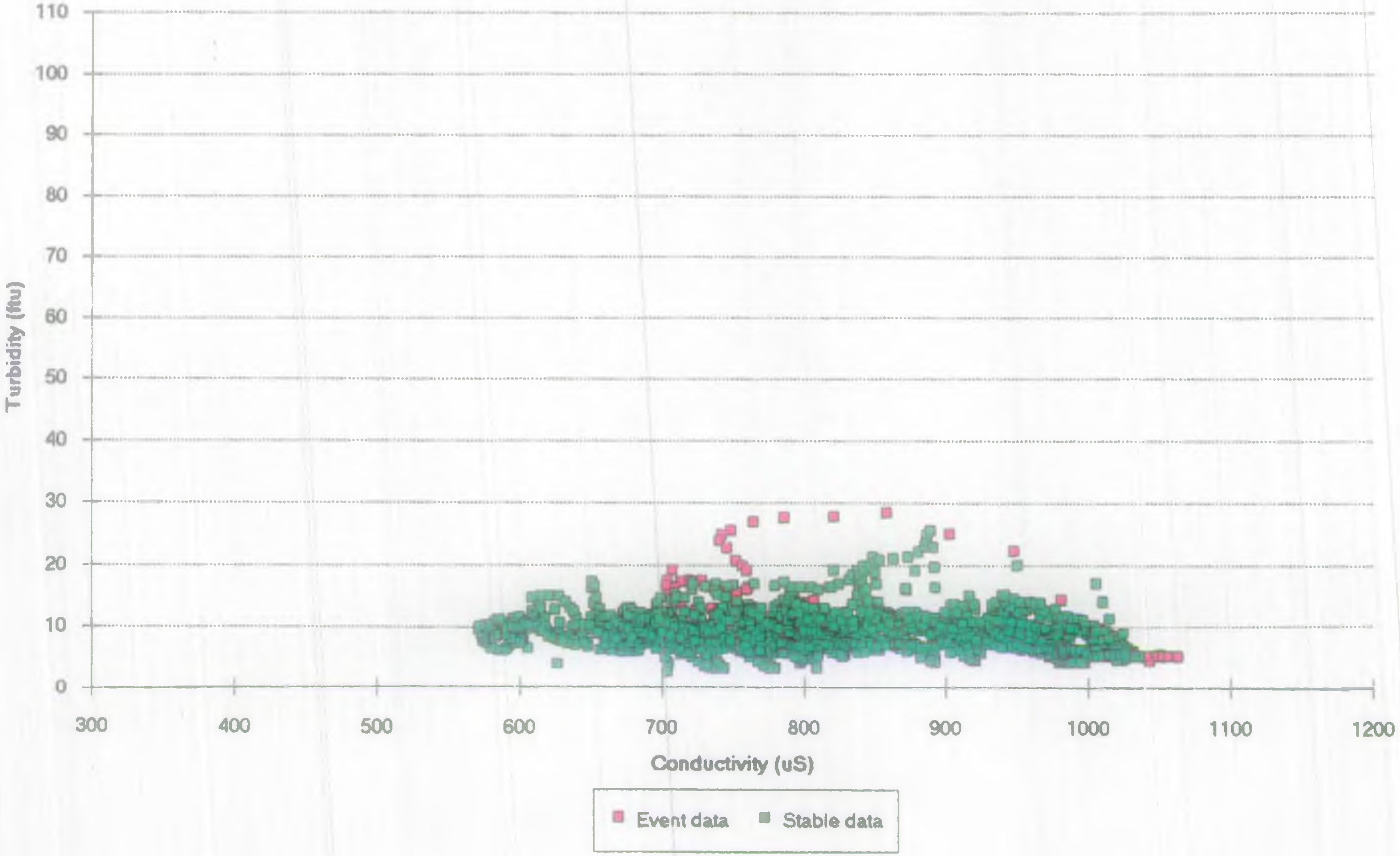


Figure B28 Graph of Turbidity vs Conductivity for Stable Data and an Event on the 26-28th August 1989

APPENDIX C

TIME SERIES PLOTS FOR THURMASTON AND JAMES BRIDGE

Thurmaston

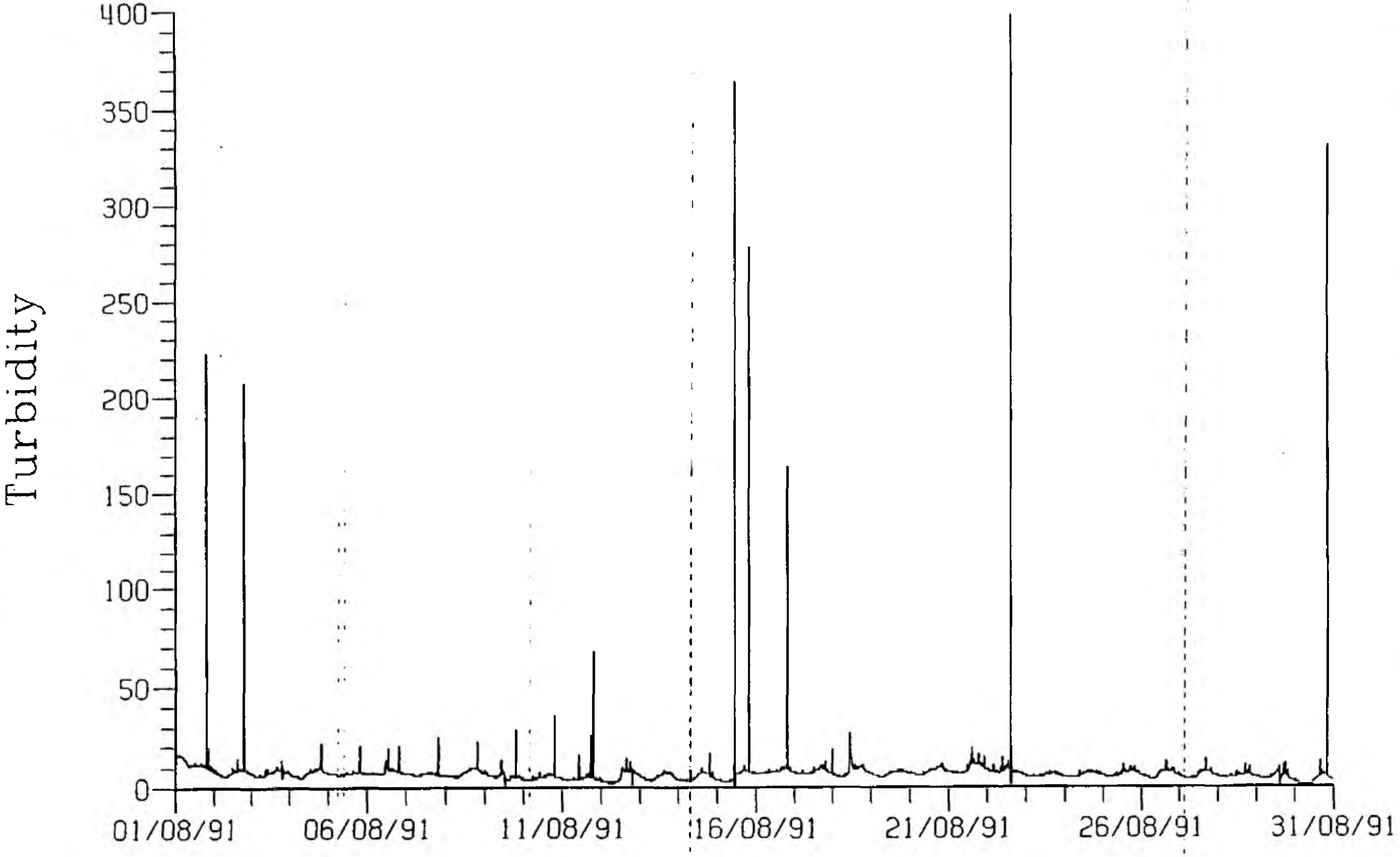


FIGURE C1

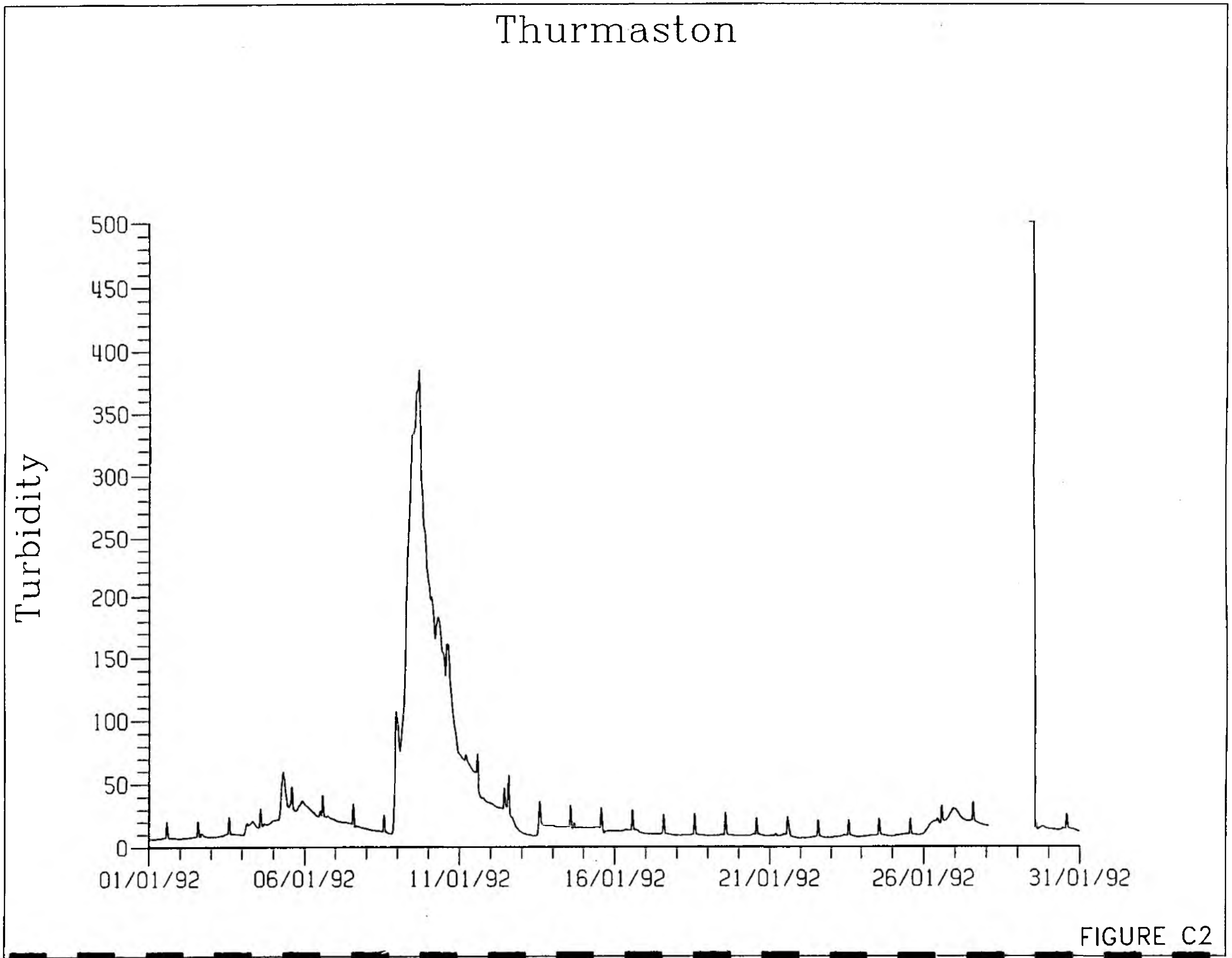

Thurmaston
Turbidity
500
450
400
350
300
250
200
150
100
50
0
01/01/92
06/01/92
11/01/92
16/01/92
21/01/92
26/01/92
31/01/92

FIGURE C2

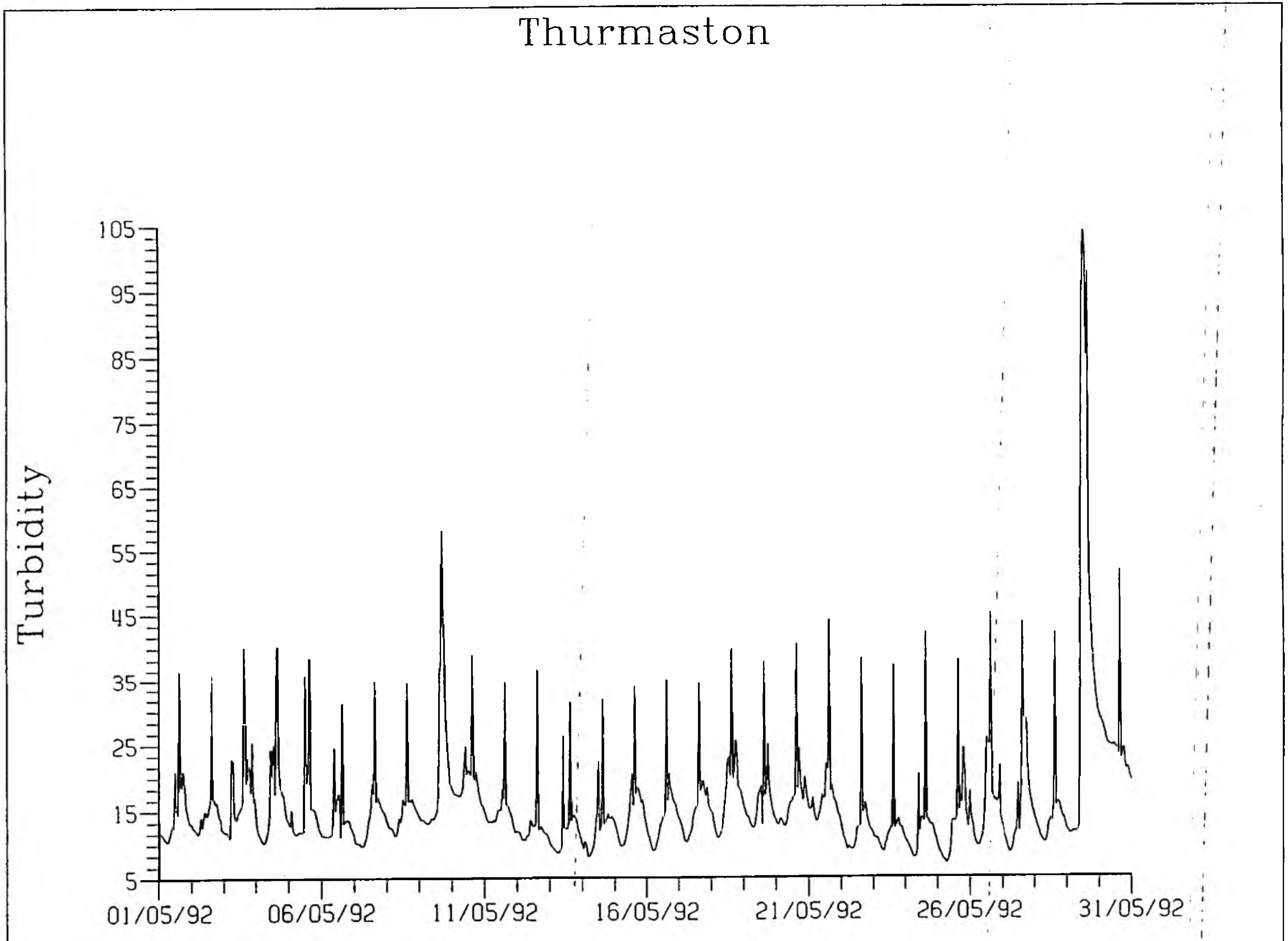
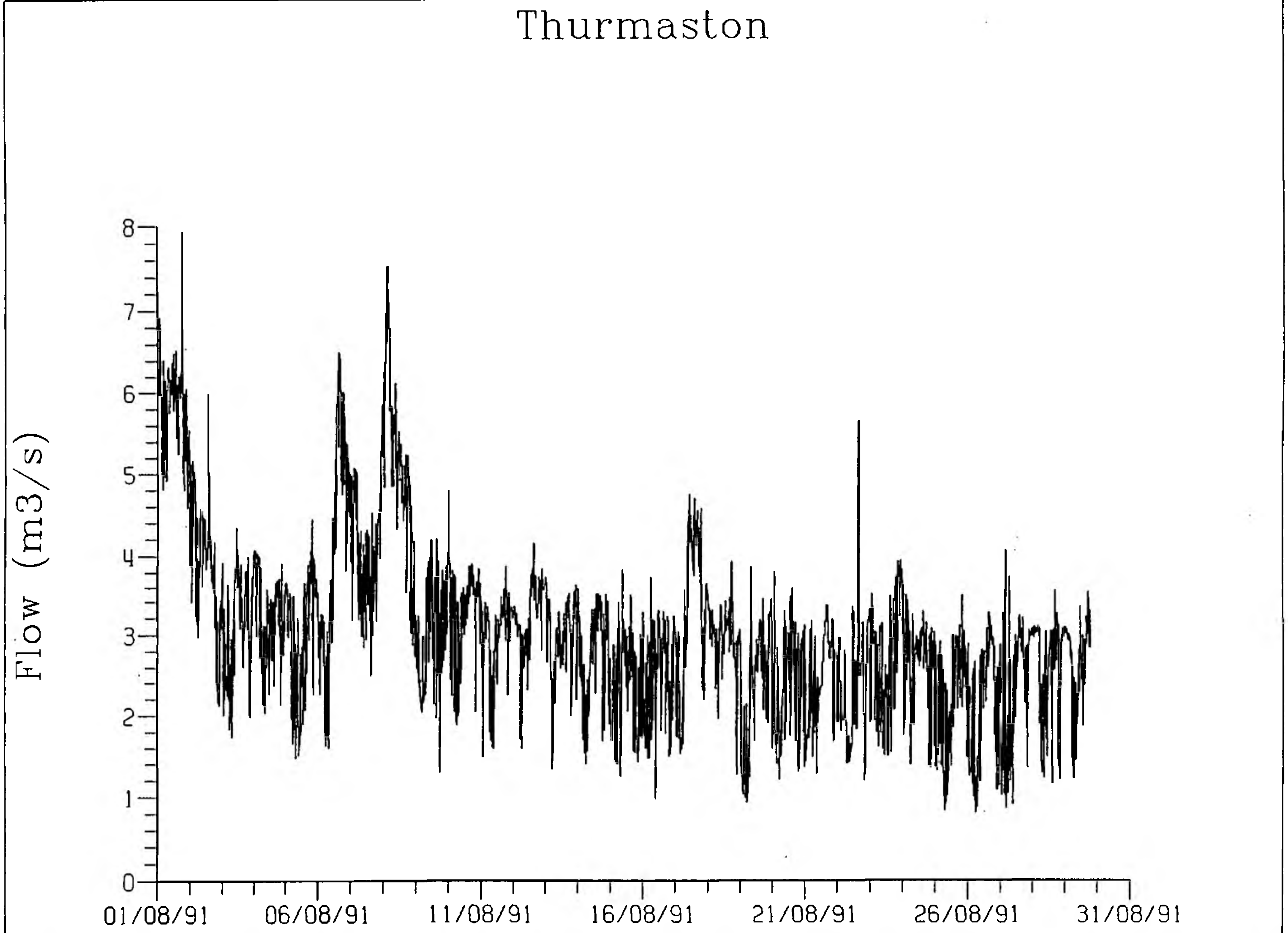

Thurmaston
Turbidity
105
95
85
75
65
55
45
35
25
15
5
01/05/92
06/05/92
11/05/92
16/05/92
21/05/92
26/05/92
31/05/92

FIGURE C3



Thurmaston

FIGURE C4

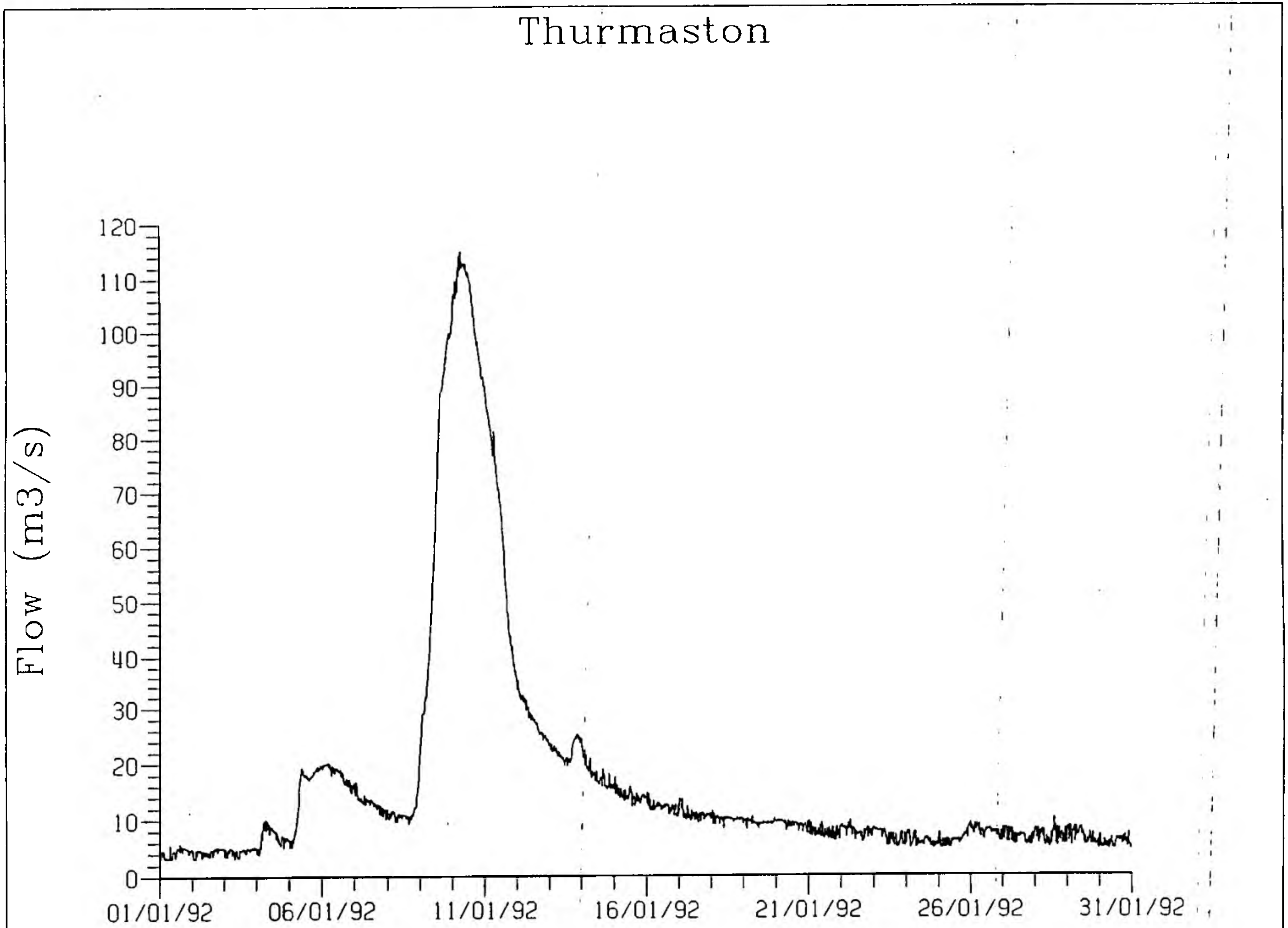

Thurmaston
Flow (m3/s)
120
110
100
90
80
70
60
50
40
30
20
10
0
01/01/92
06/01/92
11/01/92
16/01/92
21/01/92
26/01/92
31/01/92

FIGURE C5

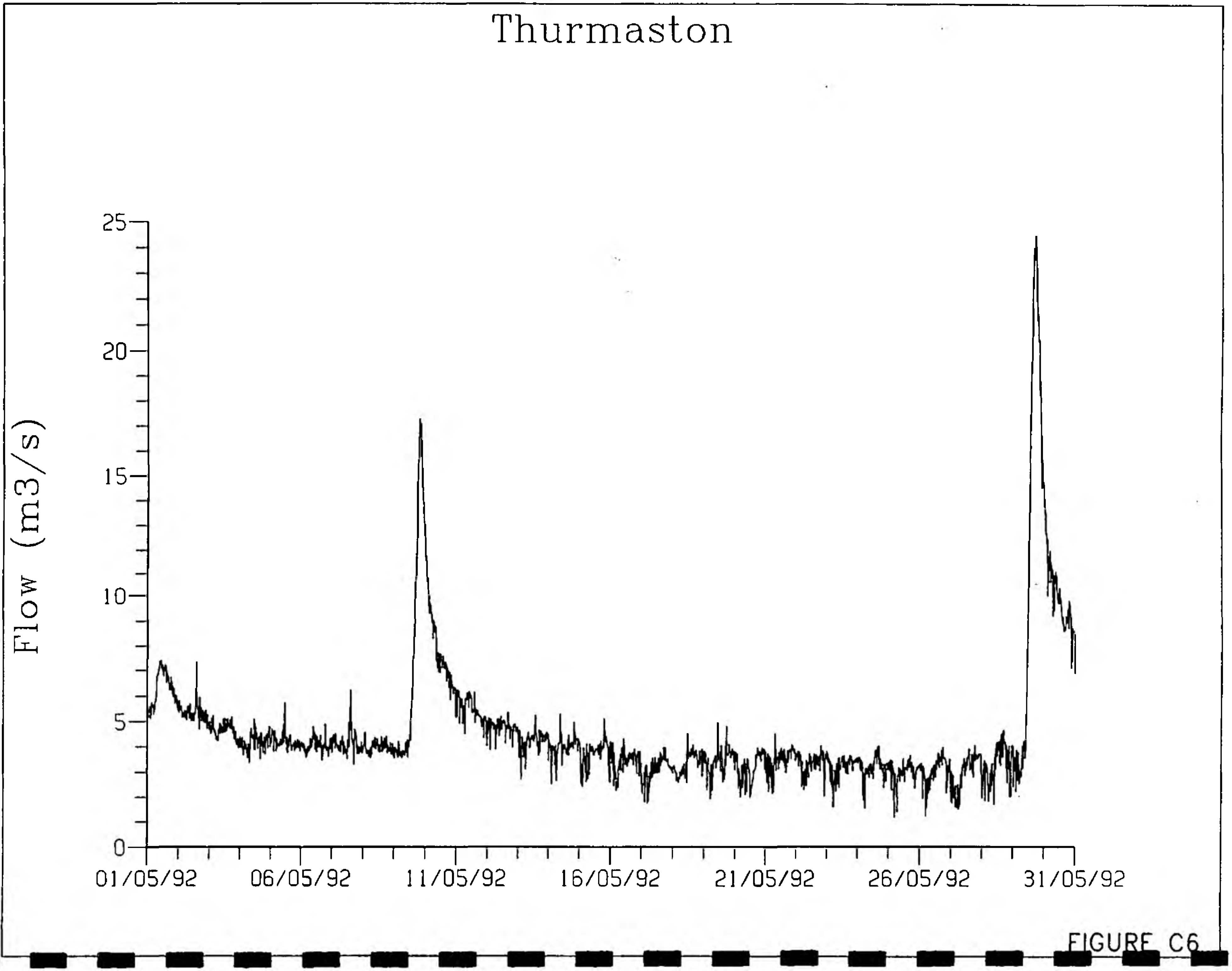

Thurmaston
Flow (m3/s)
0
5
10
15
20
25
01/05/92
06/05/92
11/05/92
16/05/92
21/05/92
26/05/92
31/05/92

FIGURE C6

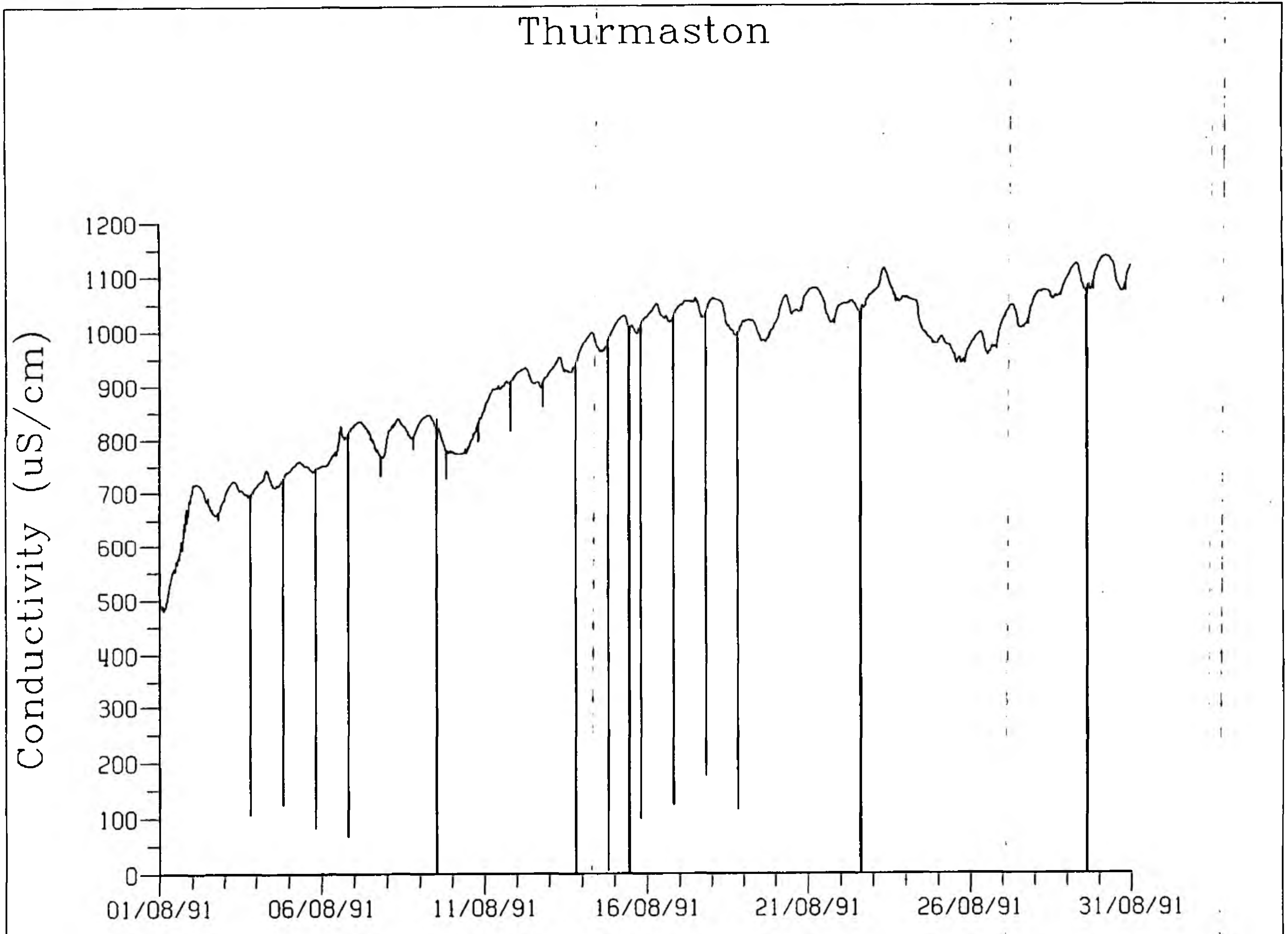

Thurmaston
Conductivity (uS/cm)
1200
1100
1000
900
800
700
600
500
400
300
200
100
0
01/08/91
06/08/91
11/08/91
16/08/91
21/08/91
26/08/91
31/08/91

FIGURE C7

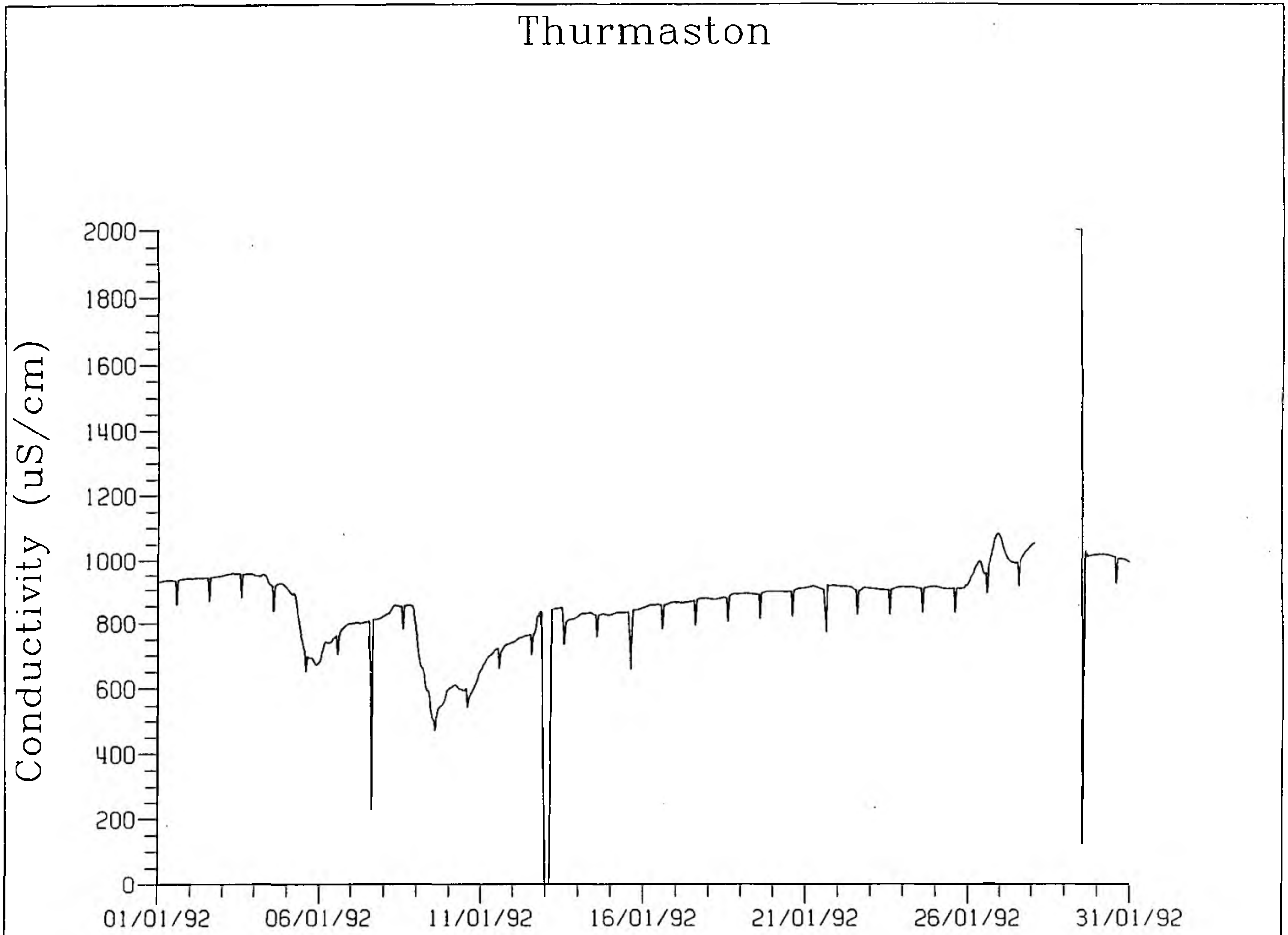

Thurmaston
Conductivity (uS/cm)
2000
1800
1600
1400
1200
1000
800
600
400
200
0
01/01/92
06/01/92
11/01/92
16/01/92
21/01/92
26/01/92
31/01/92

FIGURE C8

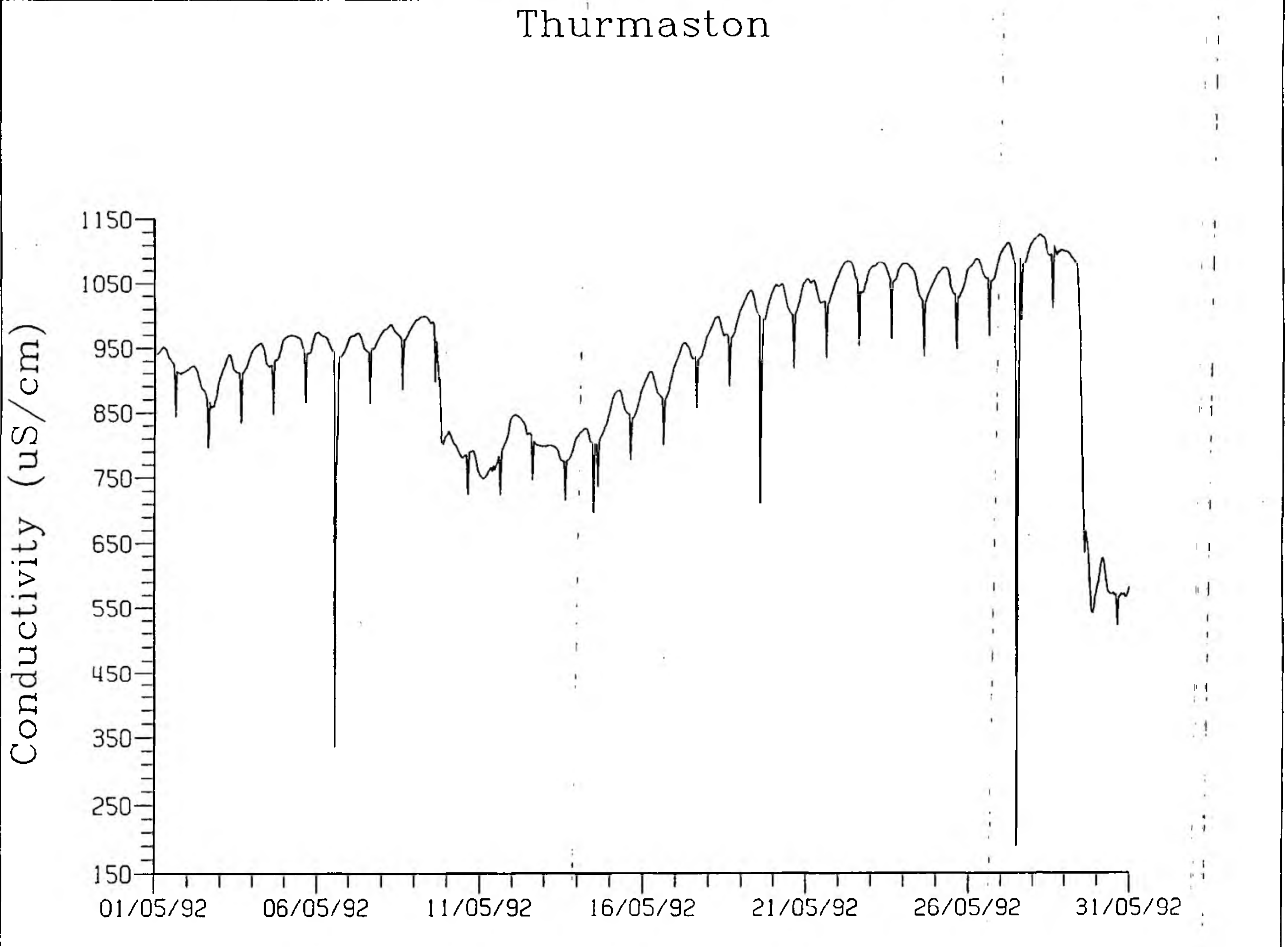

Thurmaston
Conductivity (uS/cm)
1150
1050
950
850
750
650
550
450
350
250
150
01/05/92
06/05/92
11/05/92
16/05/92
21/05/92
26/05/92
31/05/92

FIGURE C9

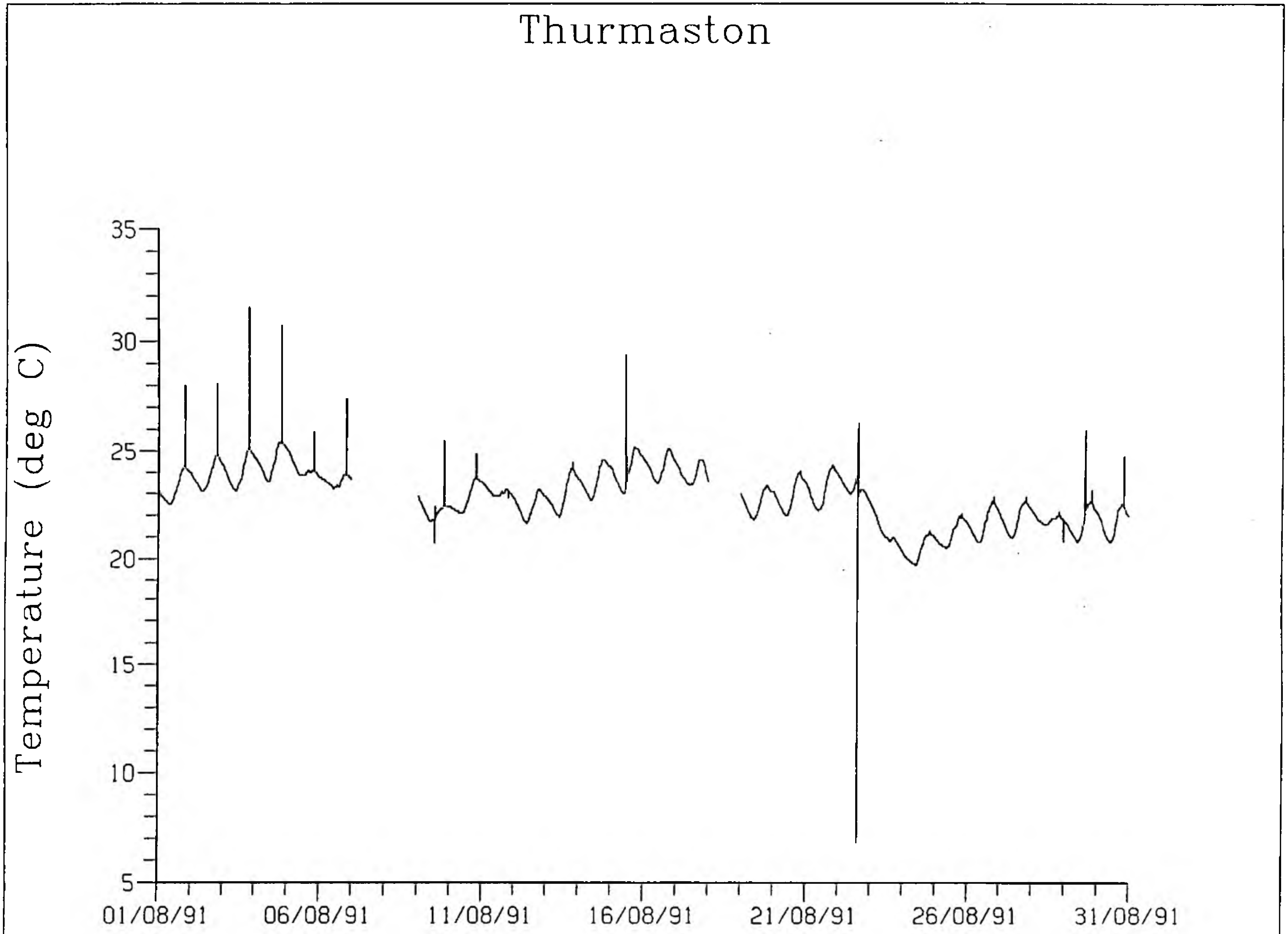

Thurmaston
Temperature (deg C)
35
30
25
20
15
10
5
01/08/91
06/08/91
11/08/91
16/08/91
21/08/91
26/08/91
31/08/91

FIGURE C10

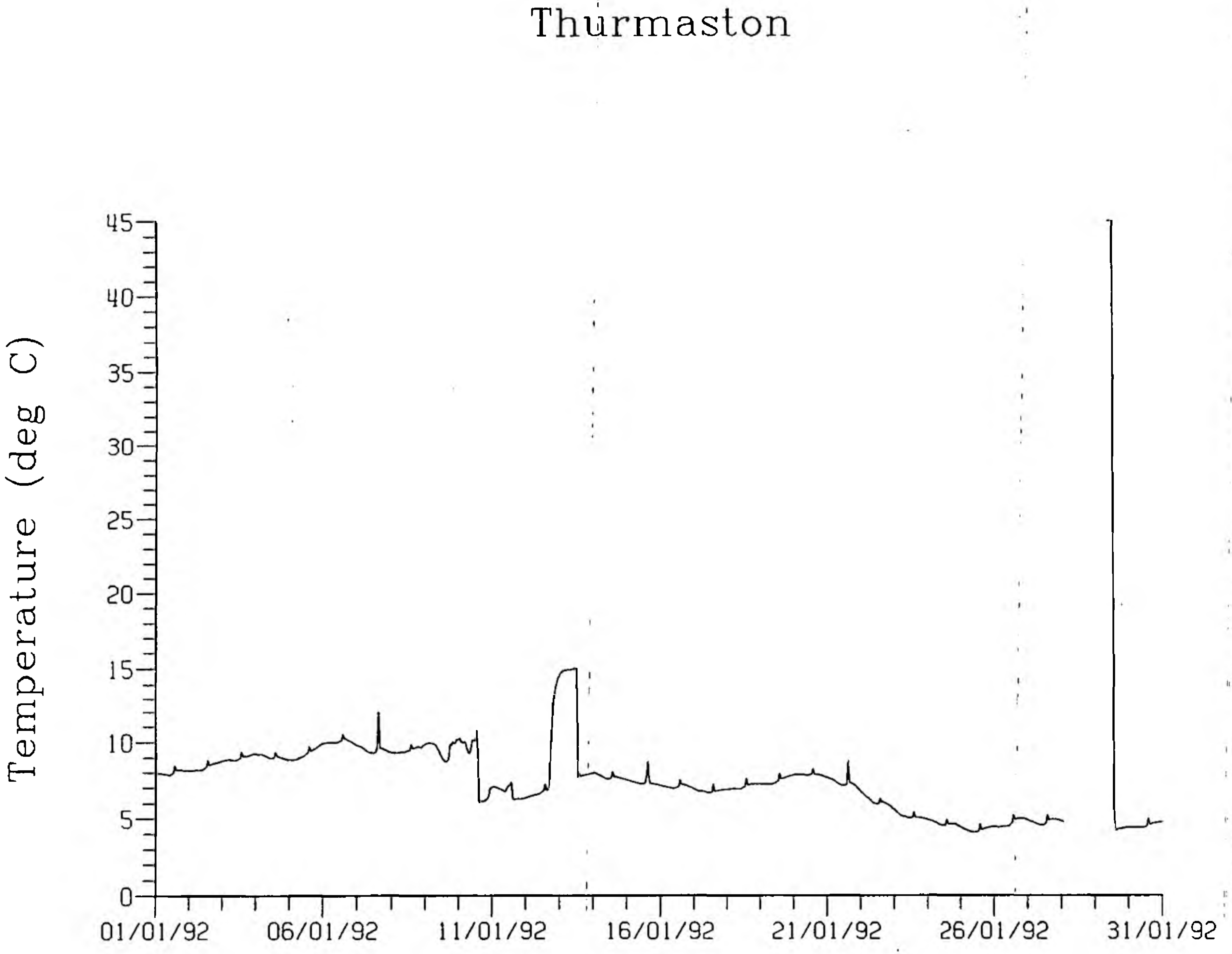

Thurmaston
Temperature (deg C)
0
5
10
15
20
25
30
35
40
45
01/01/92
06/01/92
11/01/92
16/01/92
21/01/92
26/01/92
31/01/92

FIGURE C11

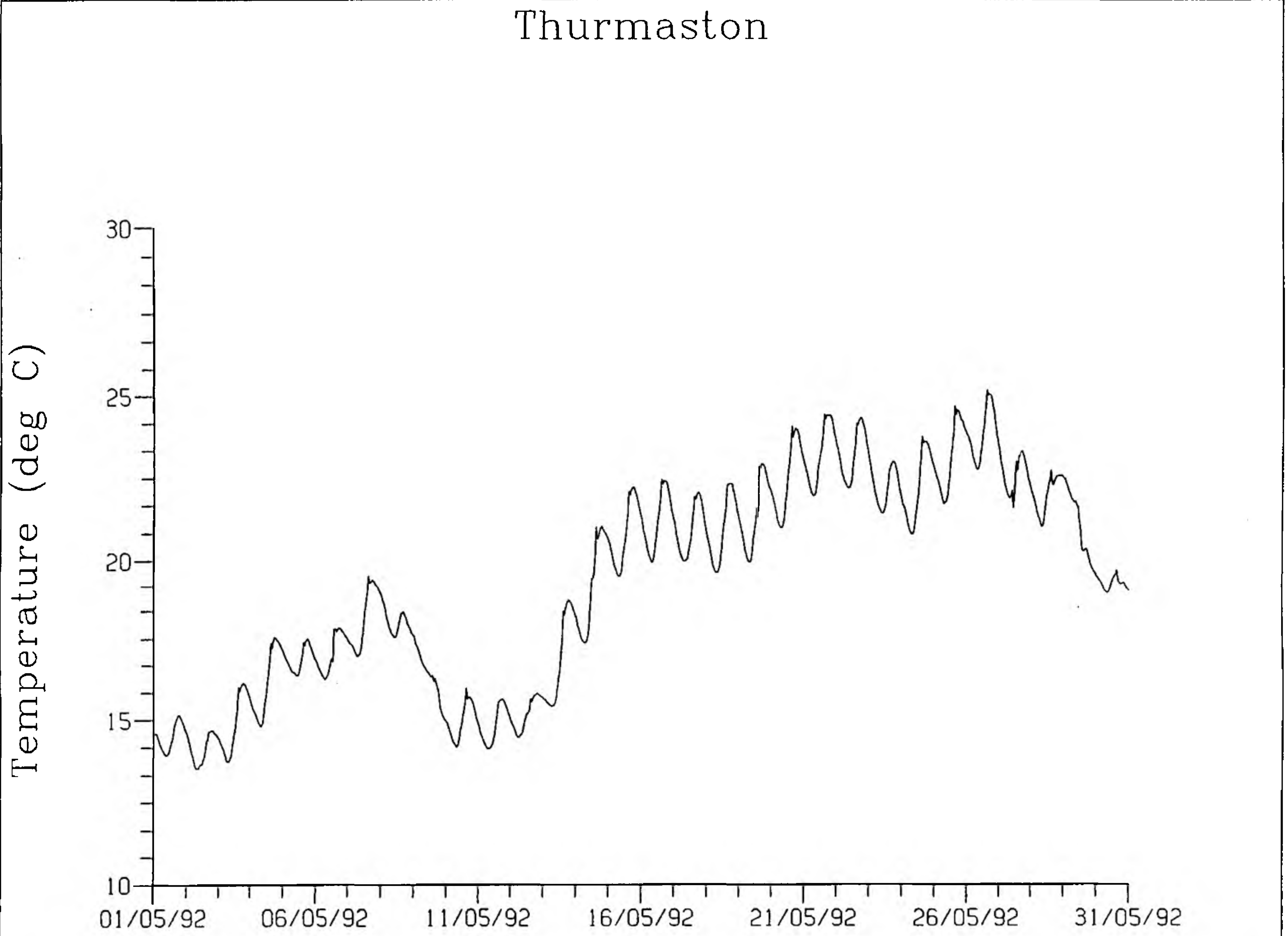
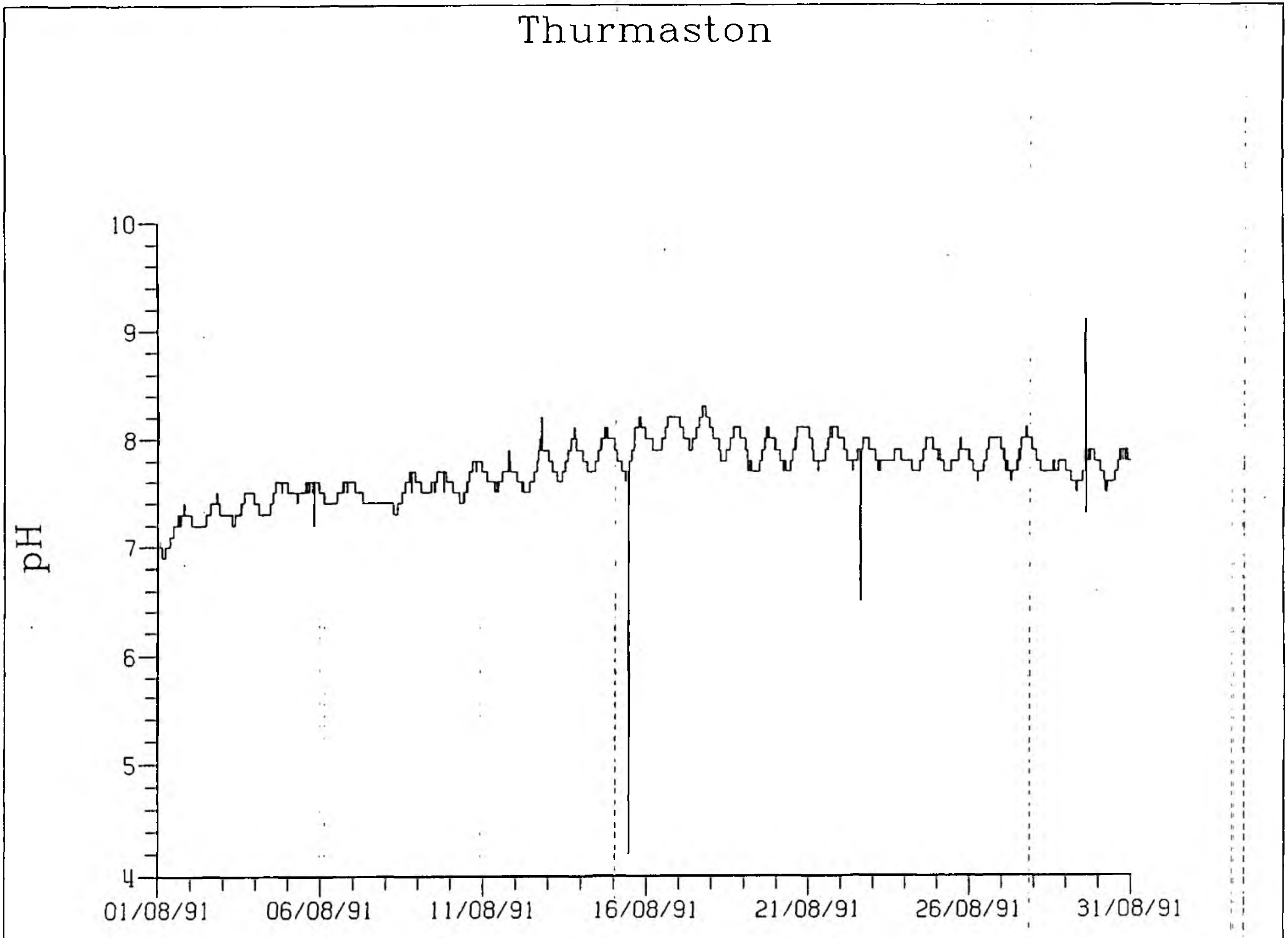

Thurmaston
Temperature (deg C)
30
25
20
15
10
01/05/92
06/05/92
11/05/92
16/05/92
21/05/92
26/05/92
31/05/92

FIGURE C12



Thurmaston
pH
10
9
8
7
6
5
4
01/08/91
06/08/91
11/08/91
16/08/91
21/08/91
26/08/91
31/08/91

FIGURE C13

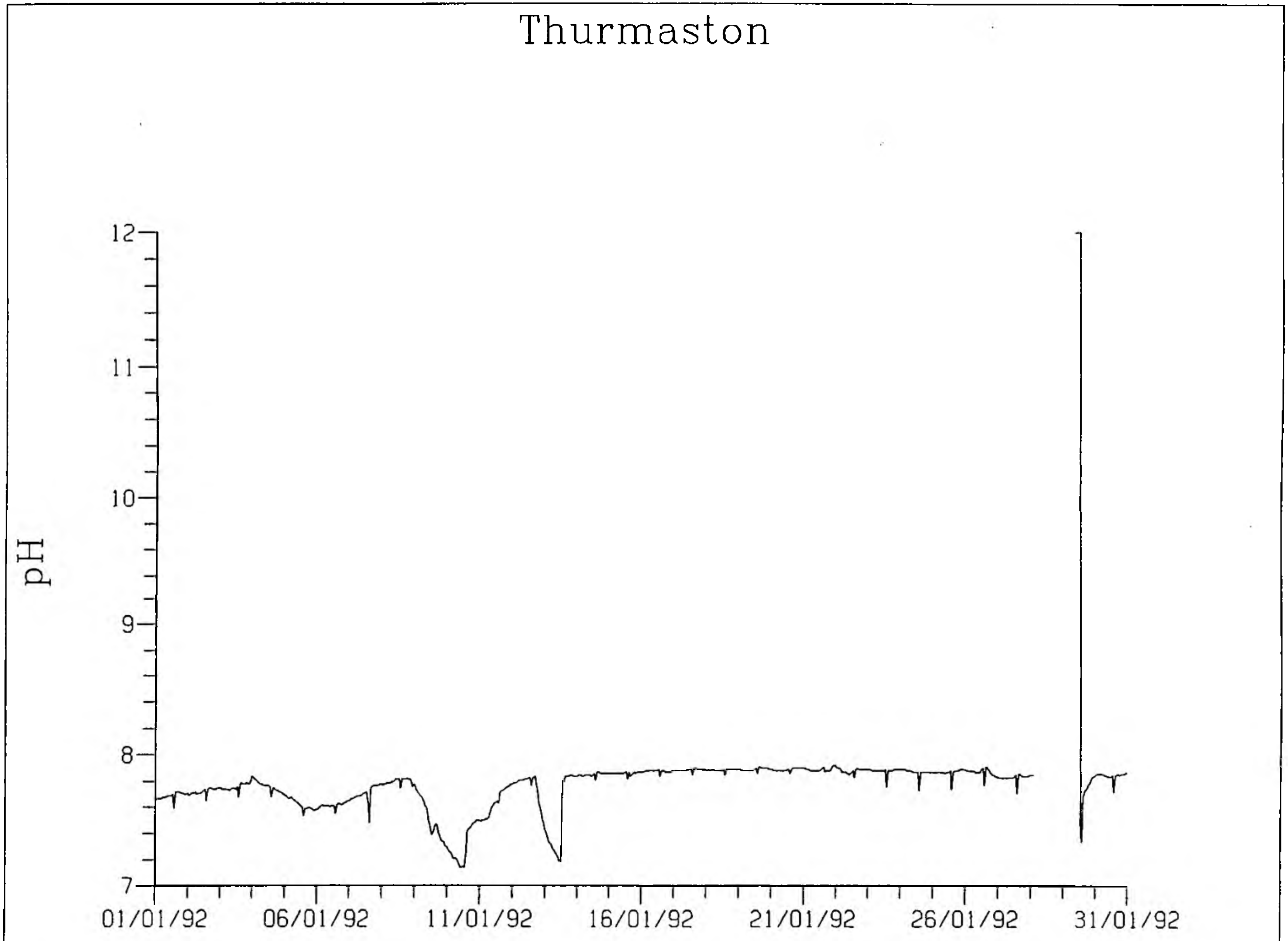

Thurmaston
pH
12
11
10
9
8
7
01/01/92
06/01/92
11/01/92
16/01/92
21/01/92
26/01/92
31/01/92

FIGURE C14

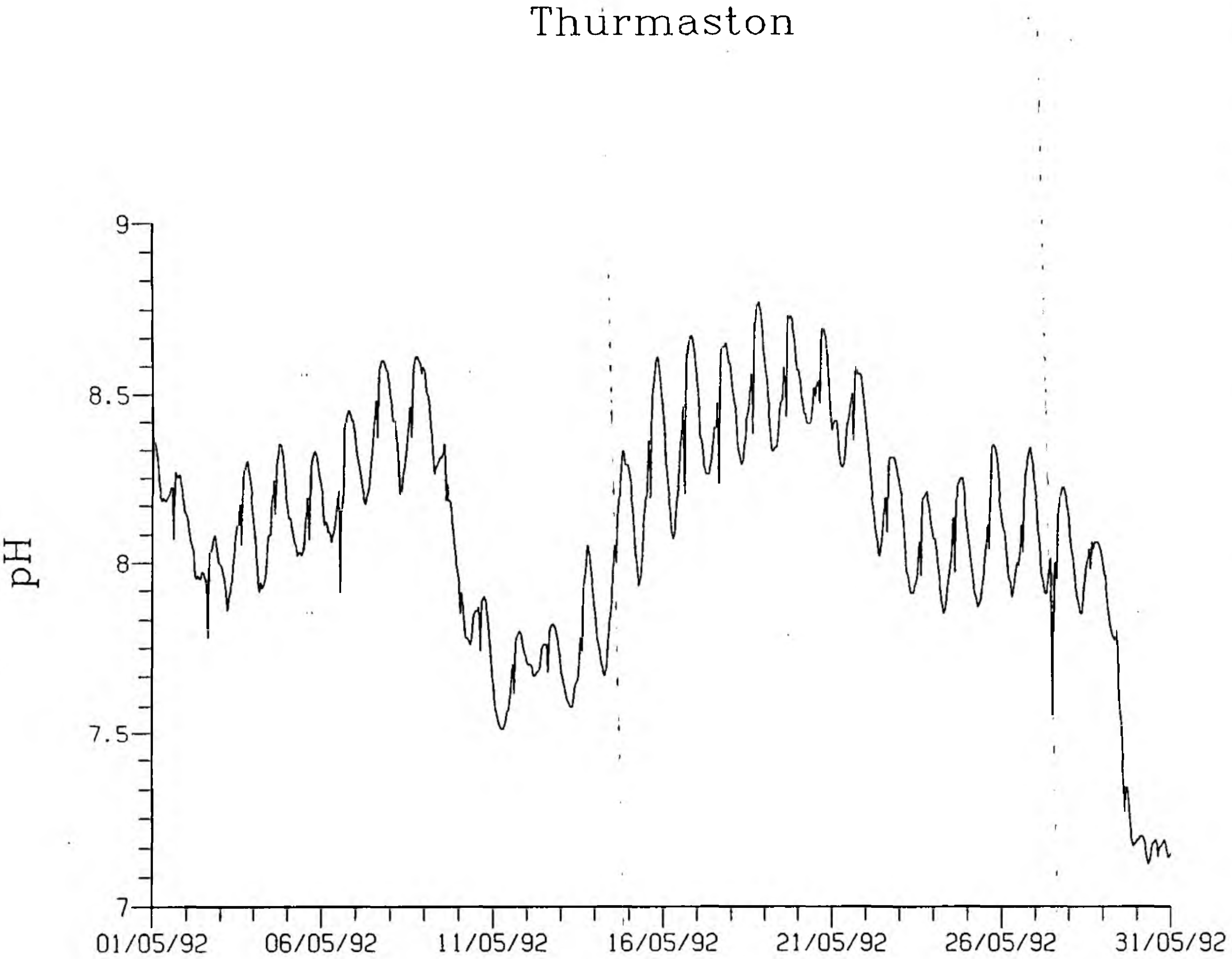

Thurmaston
pH
9
8.5
8
7.5
7
01/05/92
06/05/92
11/05/92
16/05/92
21/05/92
26/05/92
31/05/92

FIGURE C15

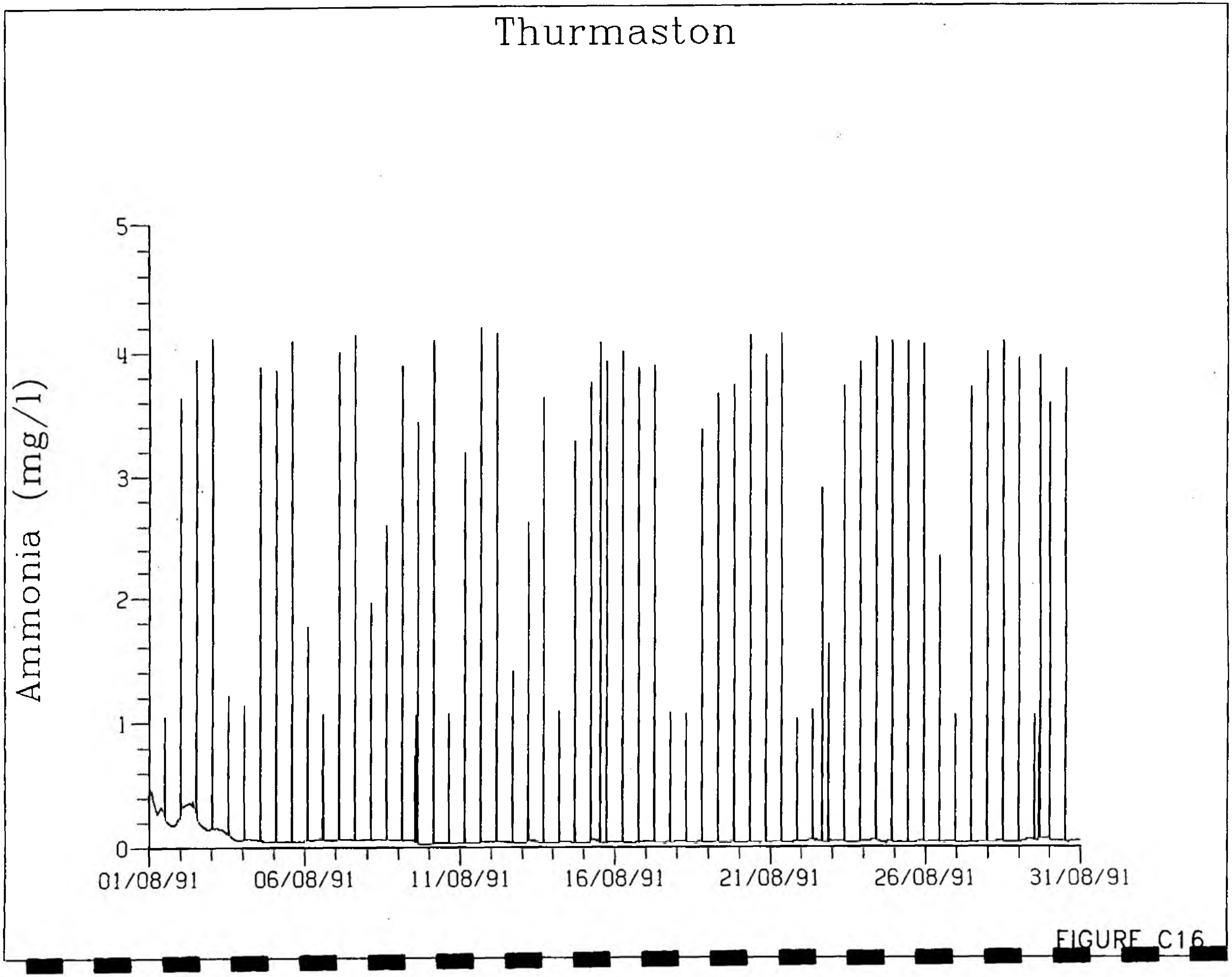

Thurmaston
Ammonia (mg/l)
0
1
2
3
4
5
01/08/91
06/08/91
11/08/91
16/08/91
21/08/91
26/08/91
31/08/91

FIGURE C16

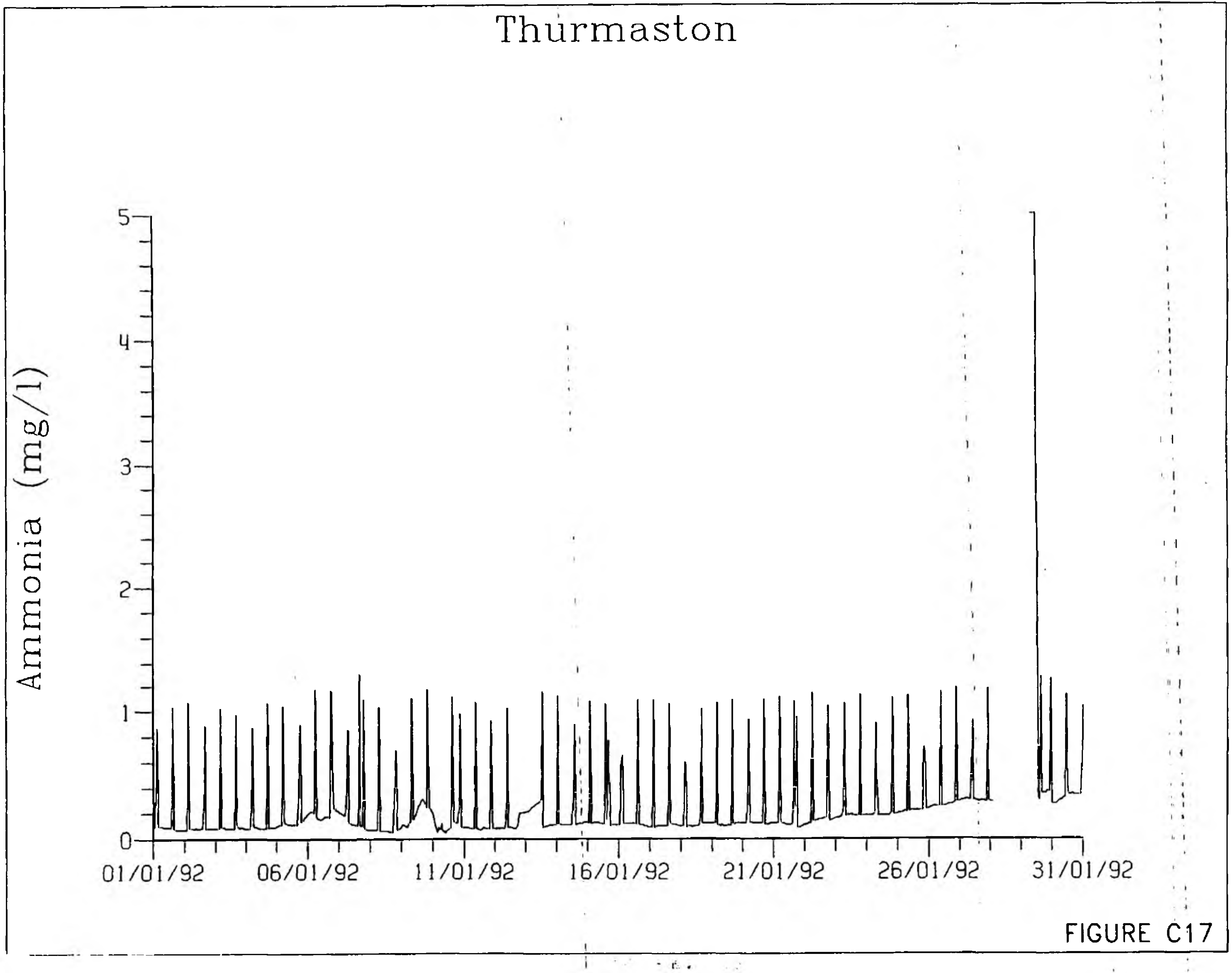

Thurmaston
Ammonia (mg/l)
5
4
3
2
1
0
01/01/92
06/01/92
11/01/92
16/01/92
21/01/92
26/01/92
31/01/92

FIGURE C17

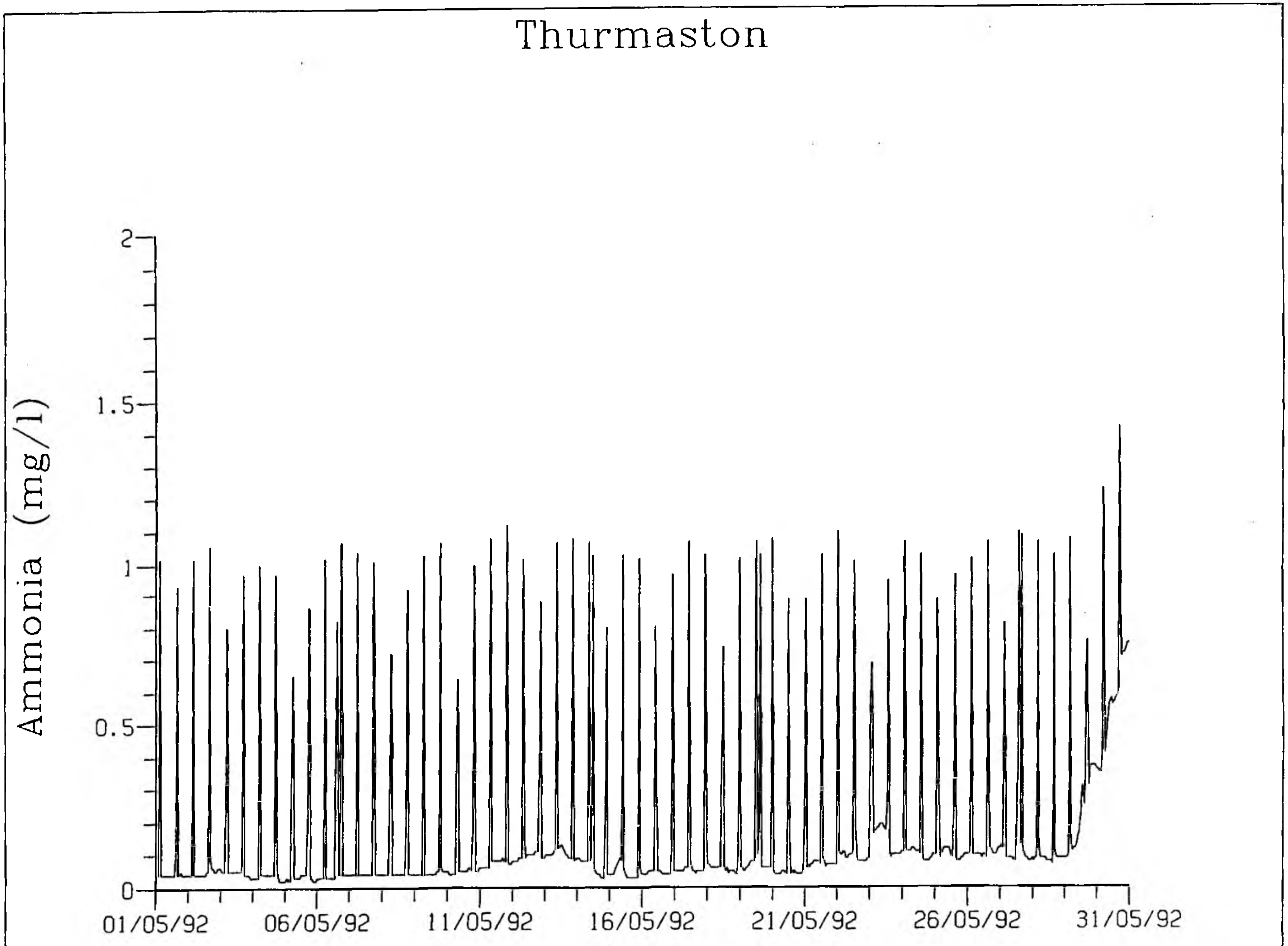

Thurmaston
Ammonia (mg/l)
2
1.5
1
0.5
0
01/05/92
06/05/92
11/05/92
16/05/92
21/05/92
26/05/92
31/05/92

FIGURE C18

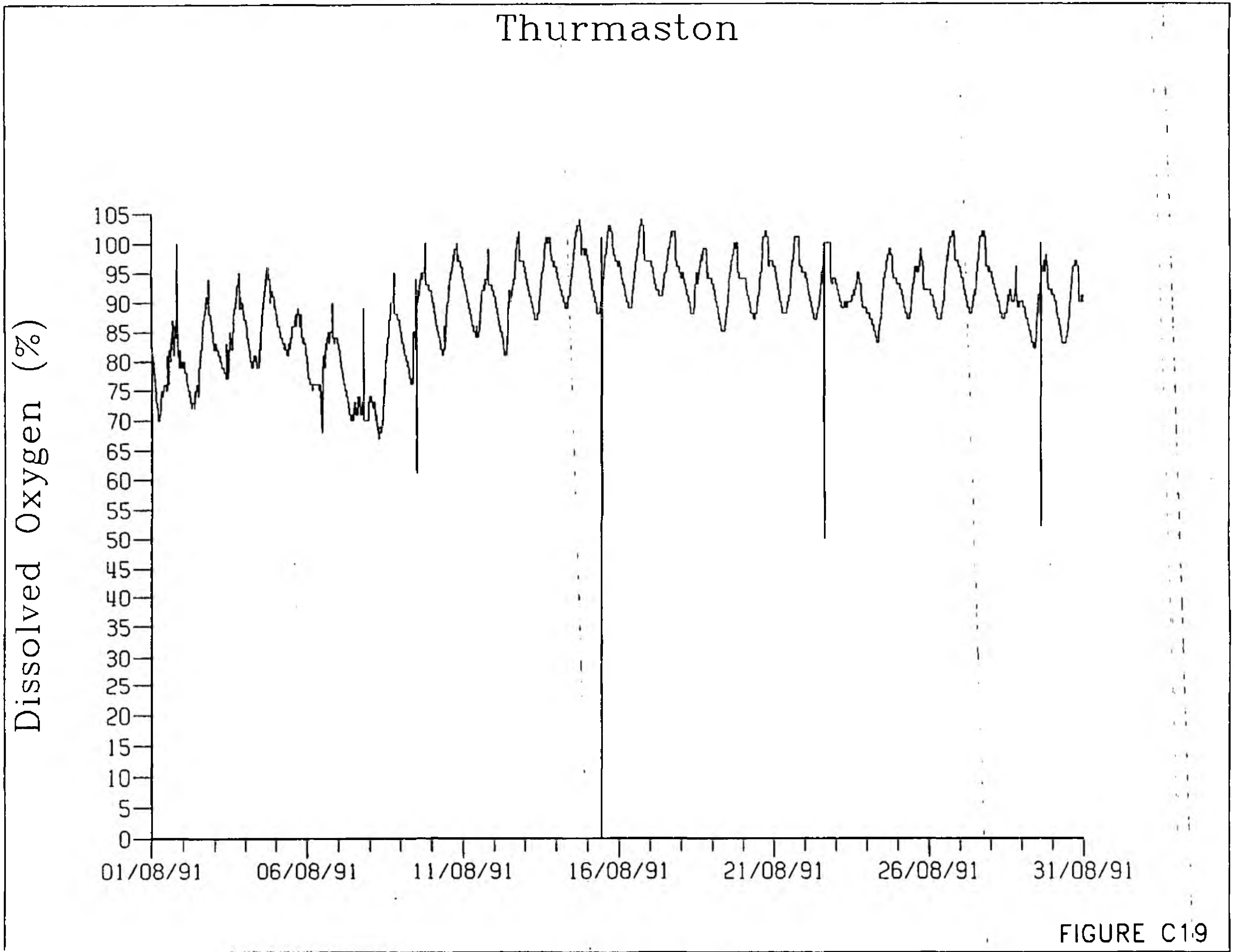

Thurmaston
Dissolved Oxygen (%)
105
100
95
90
85
80
75
70
65
60
55
50
45
40
35
30
25
20
15
10
5
0
01/08/91
06/08/91
11/08/91
16/08/91
21/08/91
26/08/91
31/08/91

FIGURE C19

Thurmaston

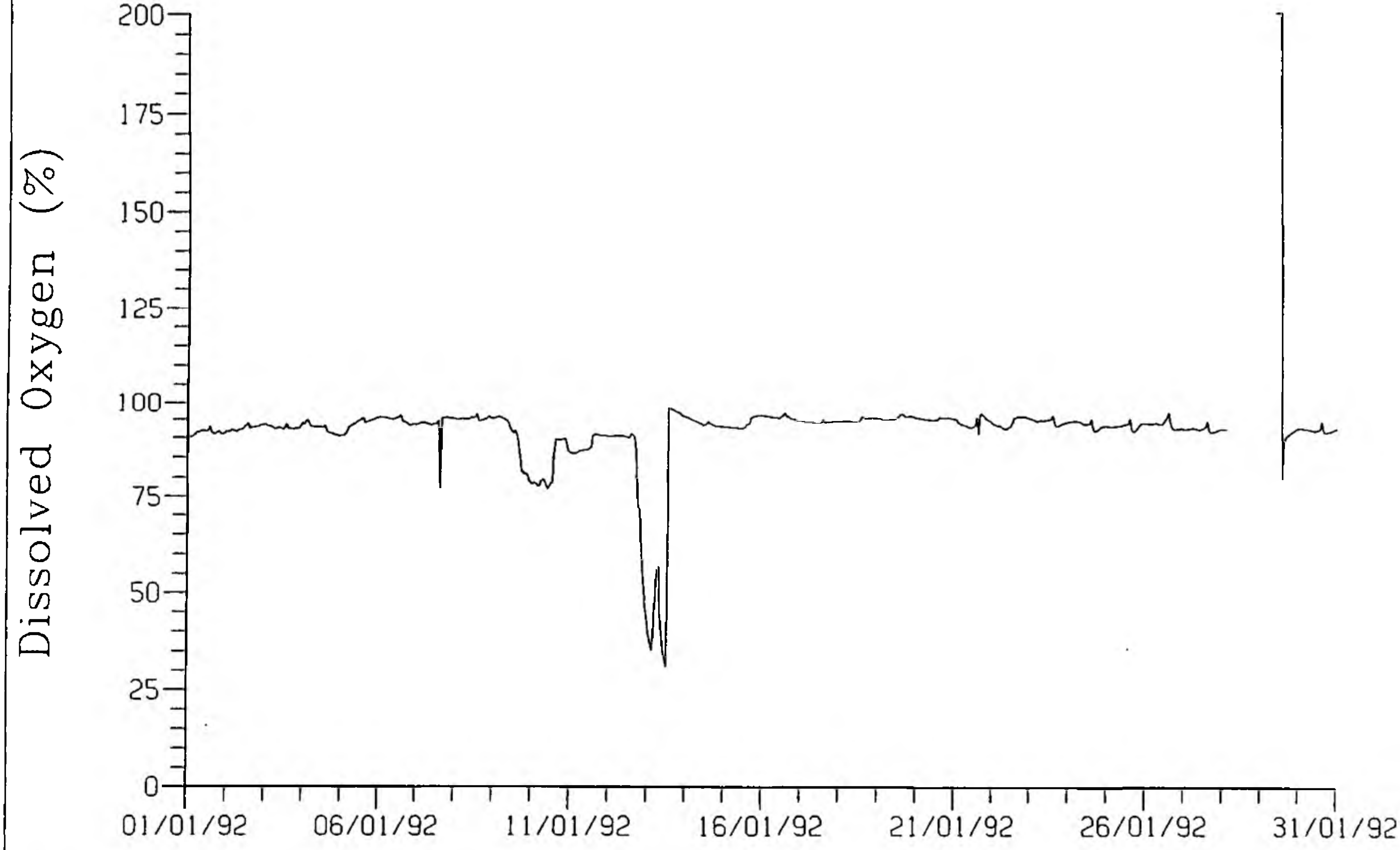


FIGURE C20

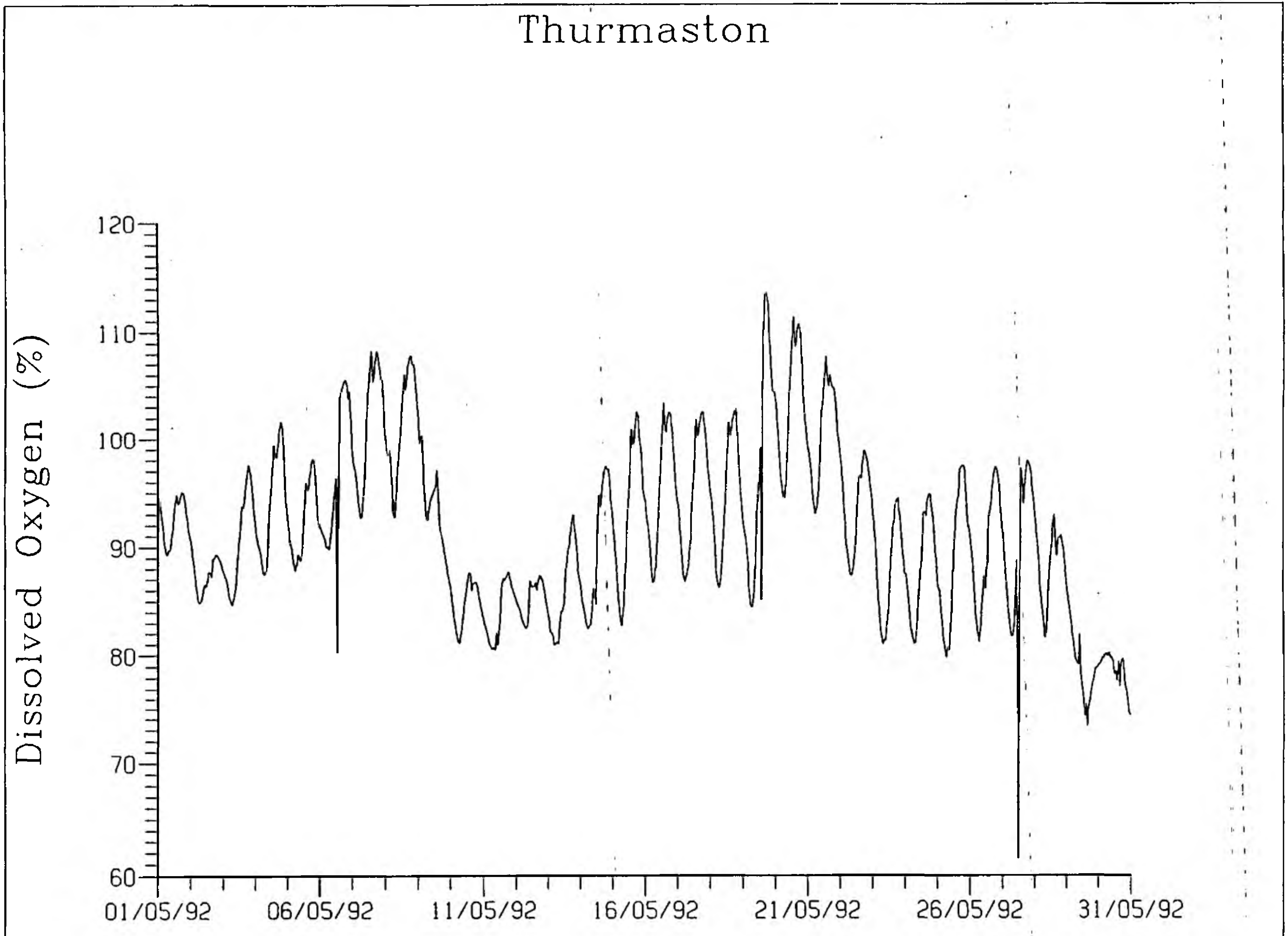

Thurmaston
Dissolved Oxygen (%)
120
110
100
90
80
70
60
01/05/92
06/05/92
11/05/92
16/05/92
21/05/92
26/05/92
31/05/92

FIGURE C21

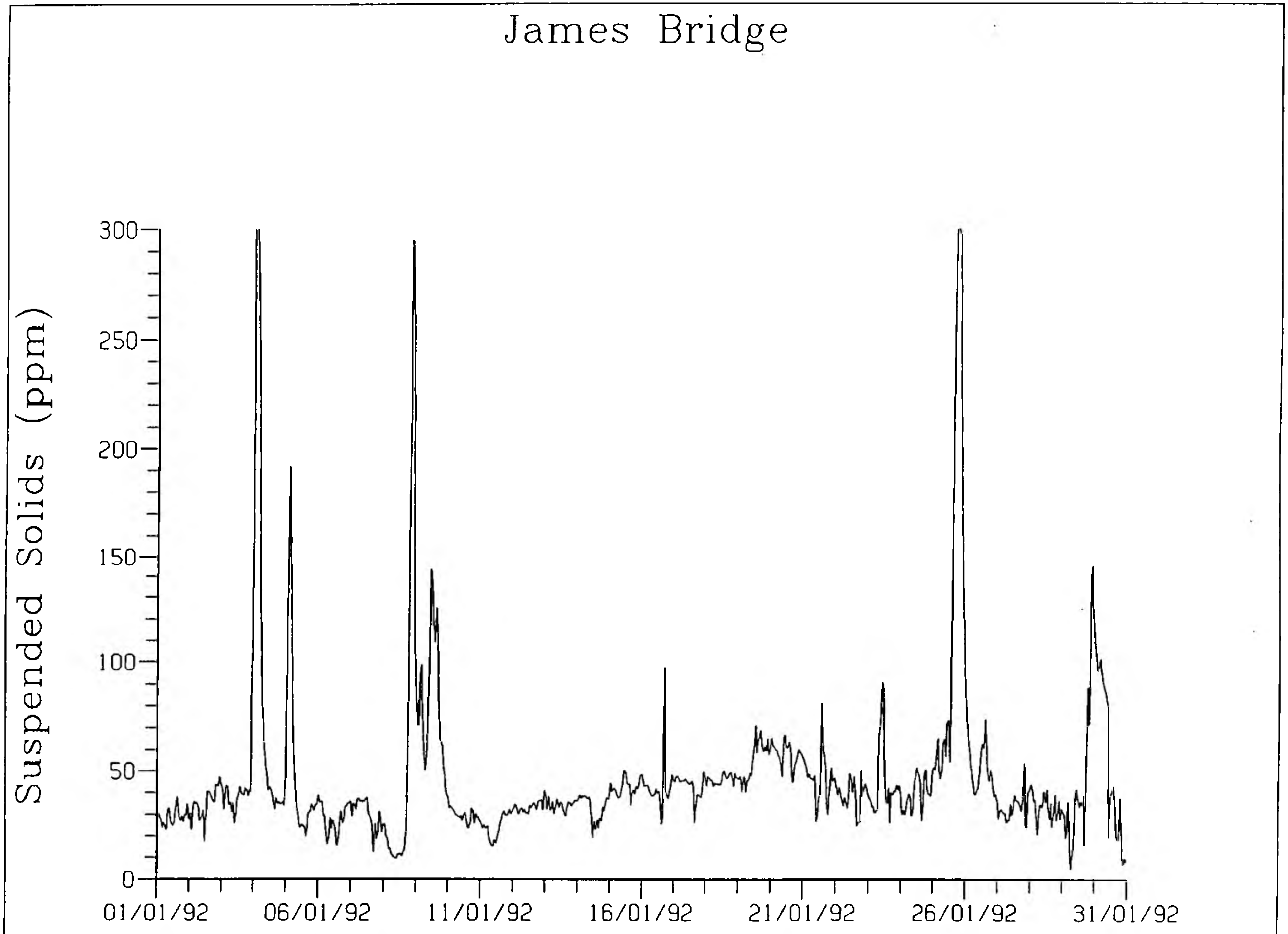

James Bridge
Suspended Solids (ppm)
300
250
200
150
100
50
0
01/01/92
06/01/92
11/01/92
16/01/92
21/01/92
26/01/92
31/01/92

FIGURE C22

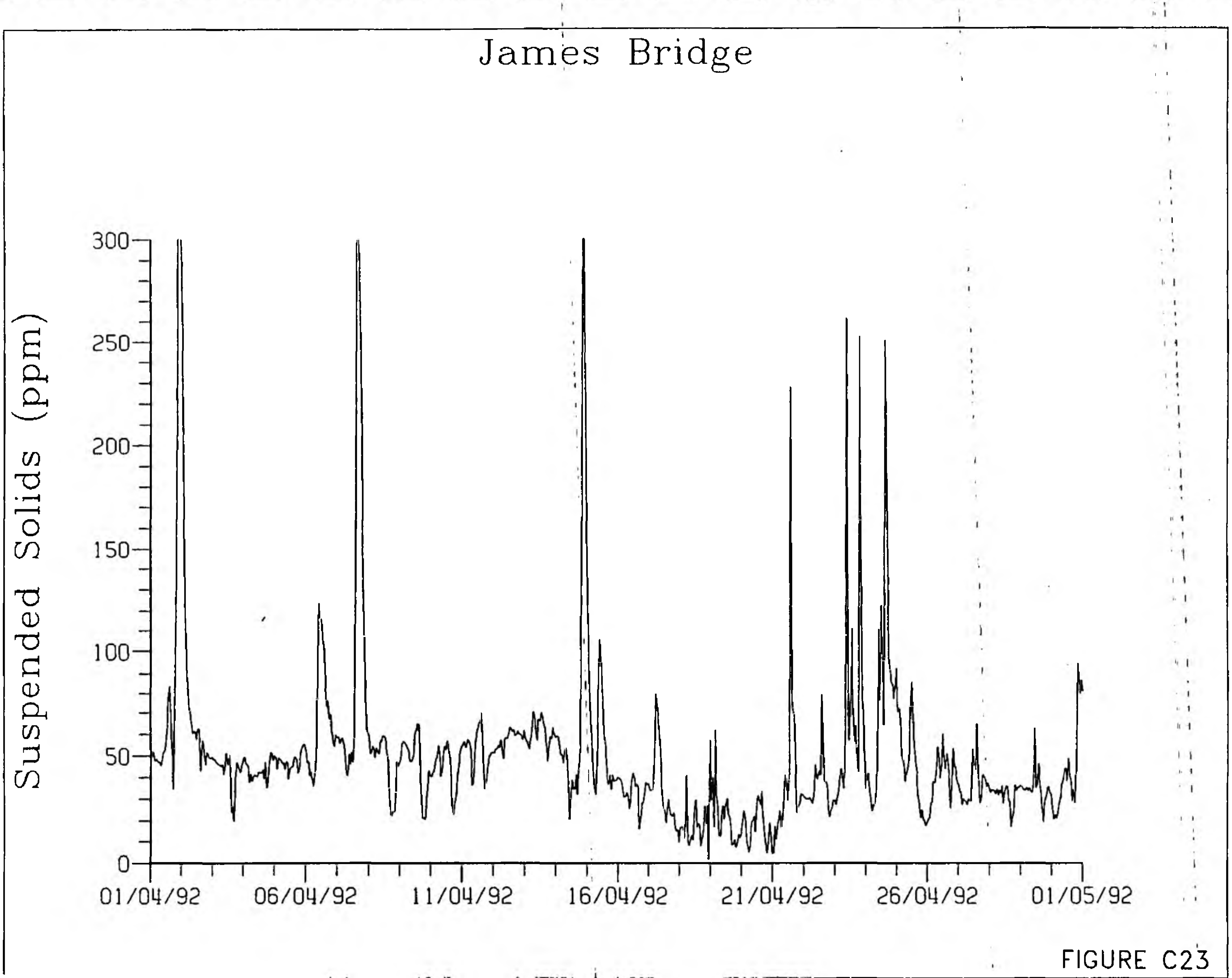

James Bridge
Suspended Solids (ppm)
300
250
200
150
100
50
0
01/04/92
06/04/92
11/04/92
16/04/92
21/04/92
26/04/92
01/05/92

FIGURE C23

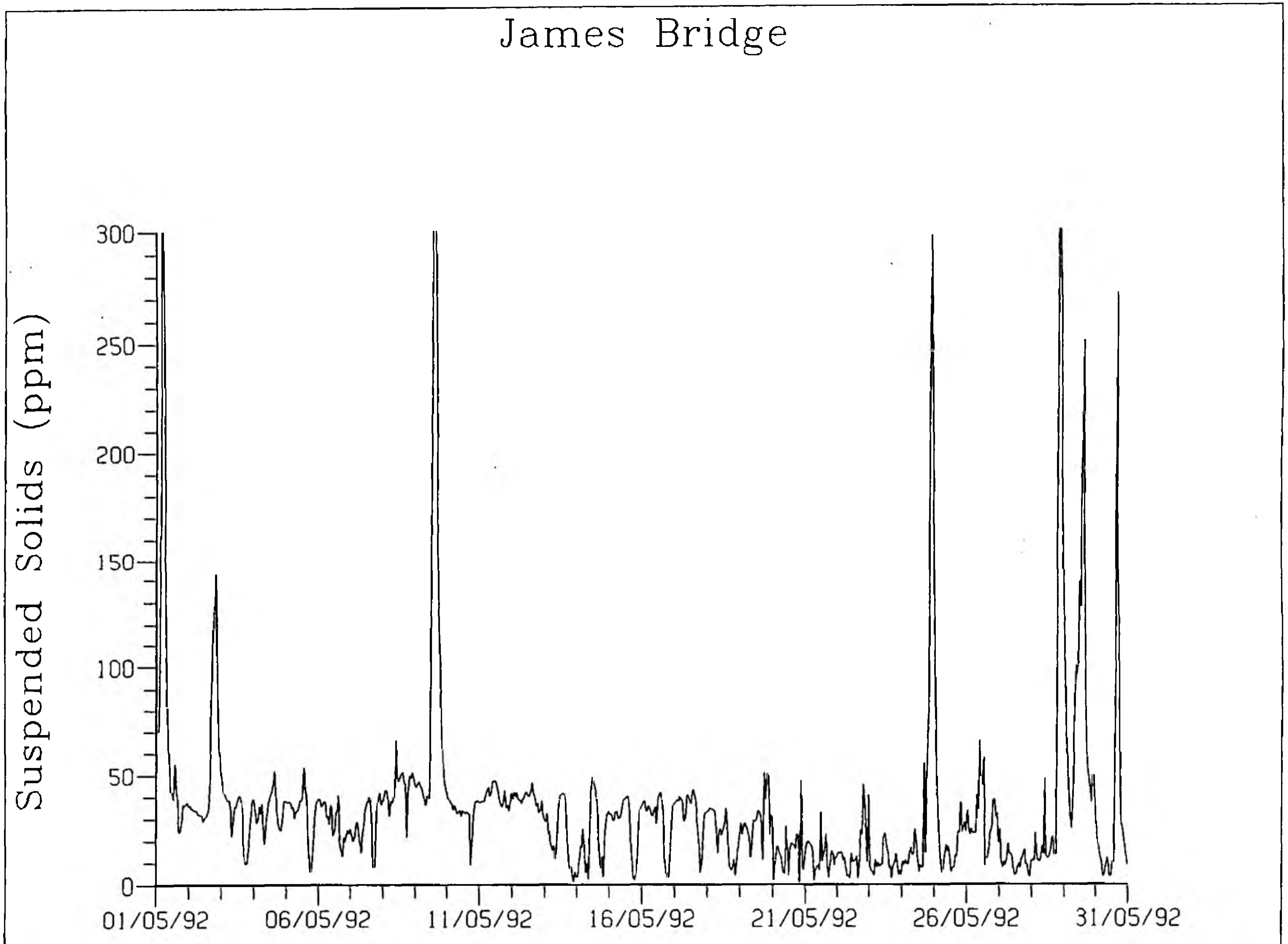

James Bridge
Suspended Solids (ppm)
300
250
200
150
100
50
0
01/05/92
06/05/92
11/05/92
16/05/92
21/05/92
26/05/92
31/05/92

FIGURE C24

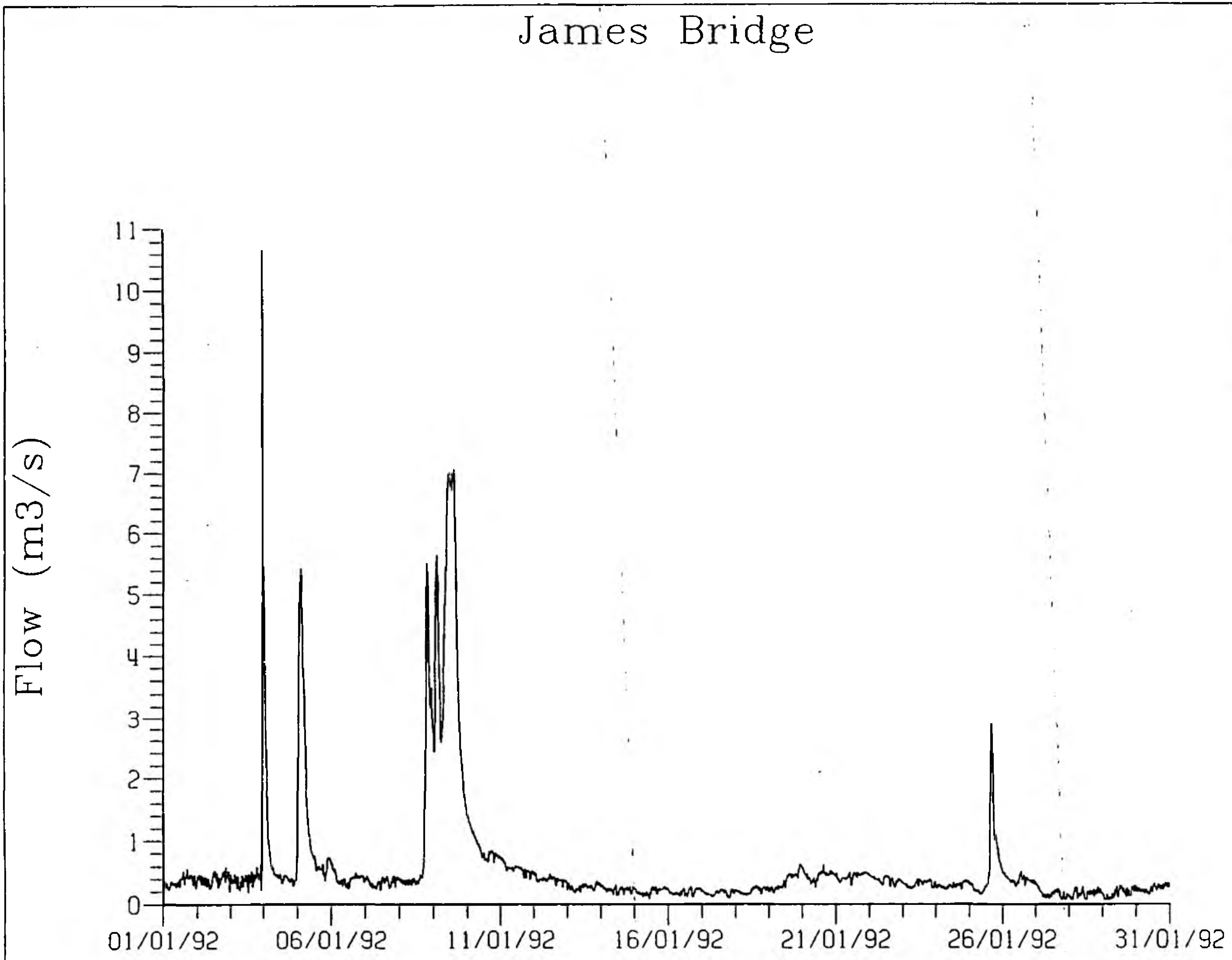

James Bridge
Flow (m3/s)
0
1
2
3
4
5
6
7
8
9
10
11
01/01/92
06/01/92
11/01/92
16/01/92
21/01/92
26/01/92
31/01/92

FIGURE C25

James Bridge

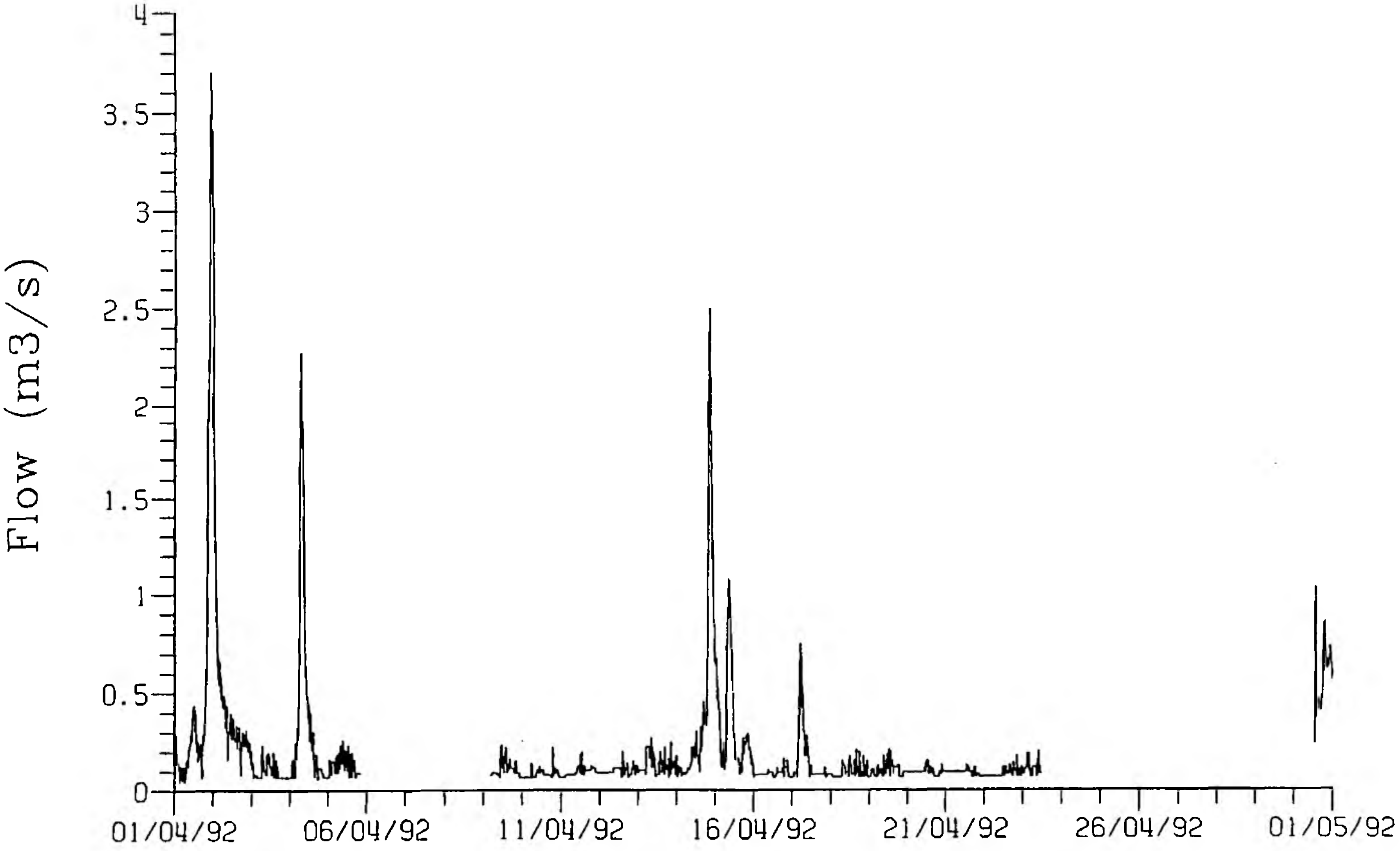


FIGURE C26

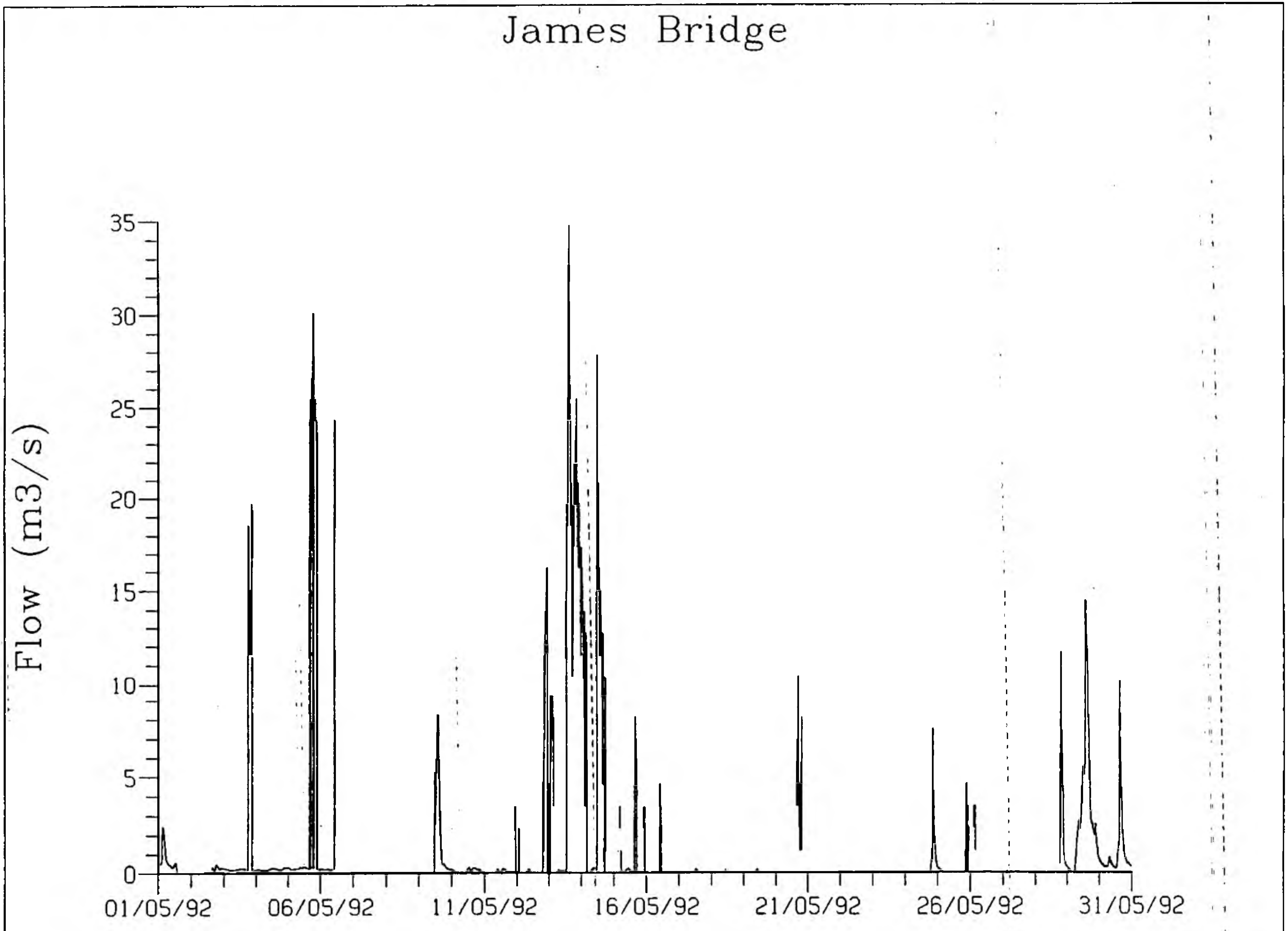

James Bridge
Flow (m3/s)
35
30
25
20
15
10
5
0
01/05/92
06/05/92
11/05/92
16/05/92
21/05/92
26/05/92
31/05/92

FIGURE C27

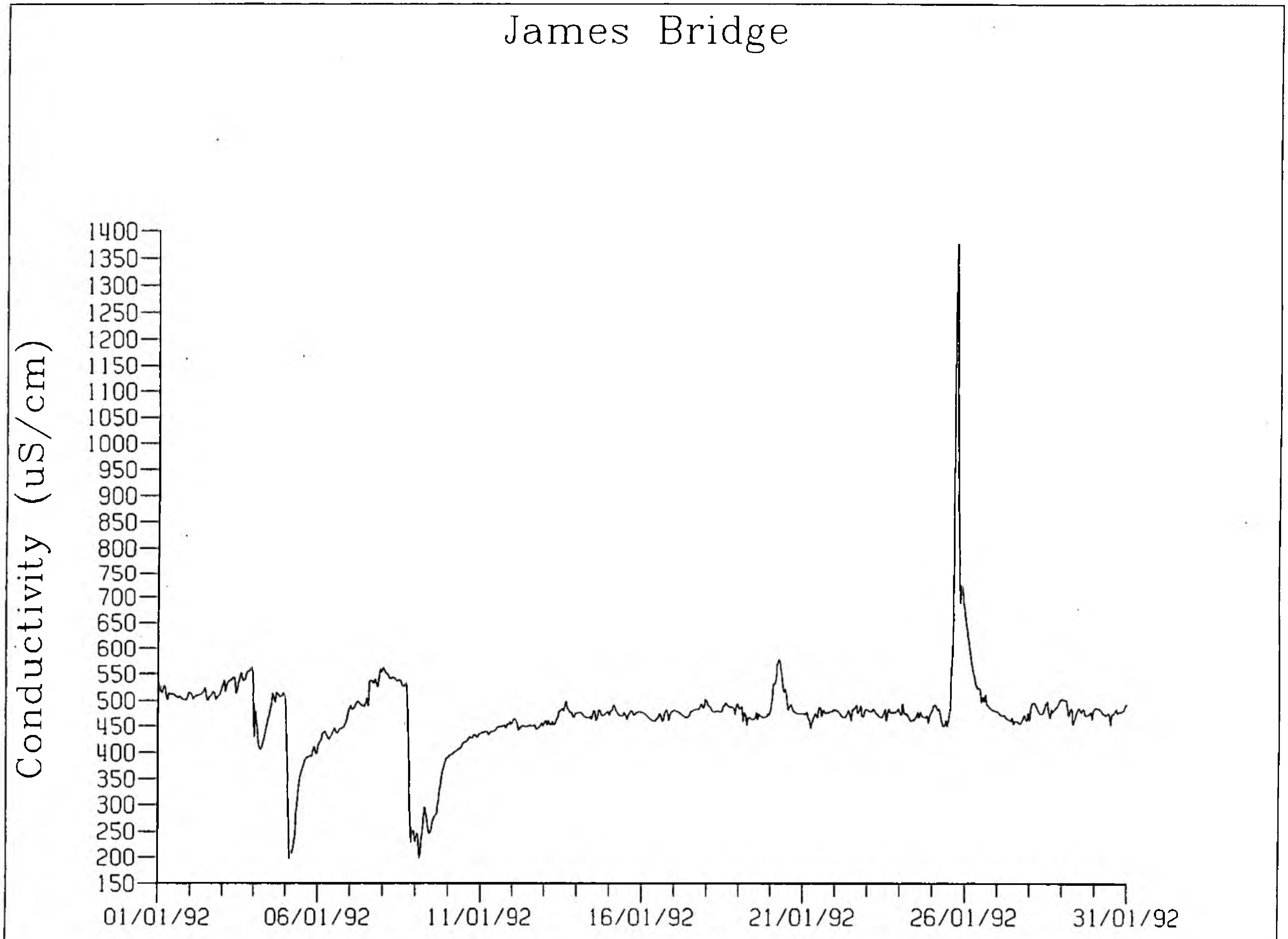

James Bridge
Conductivity (uS/cm)
1400
1350
1300
1250
1200
1150
1100
1050
1000
950
900
850
800
750
700
650
600
550
500
450
400
350
300
250
200
150
01/01/92
06/01/92
11/01/92
16/01/92
21/01/92
26/01/92
31/01/92

FIGURE C28

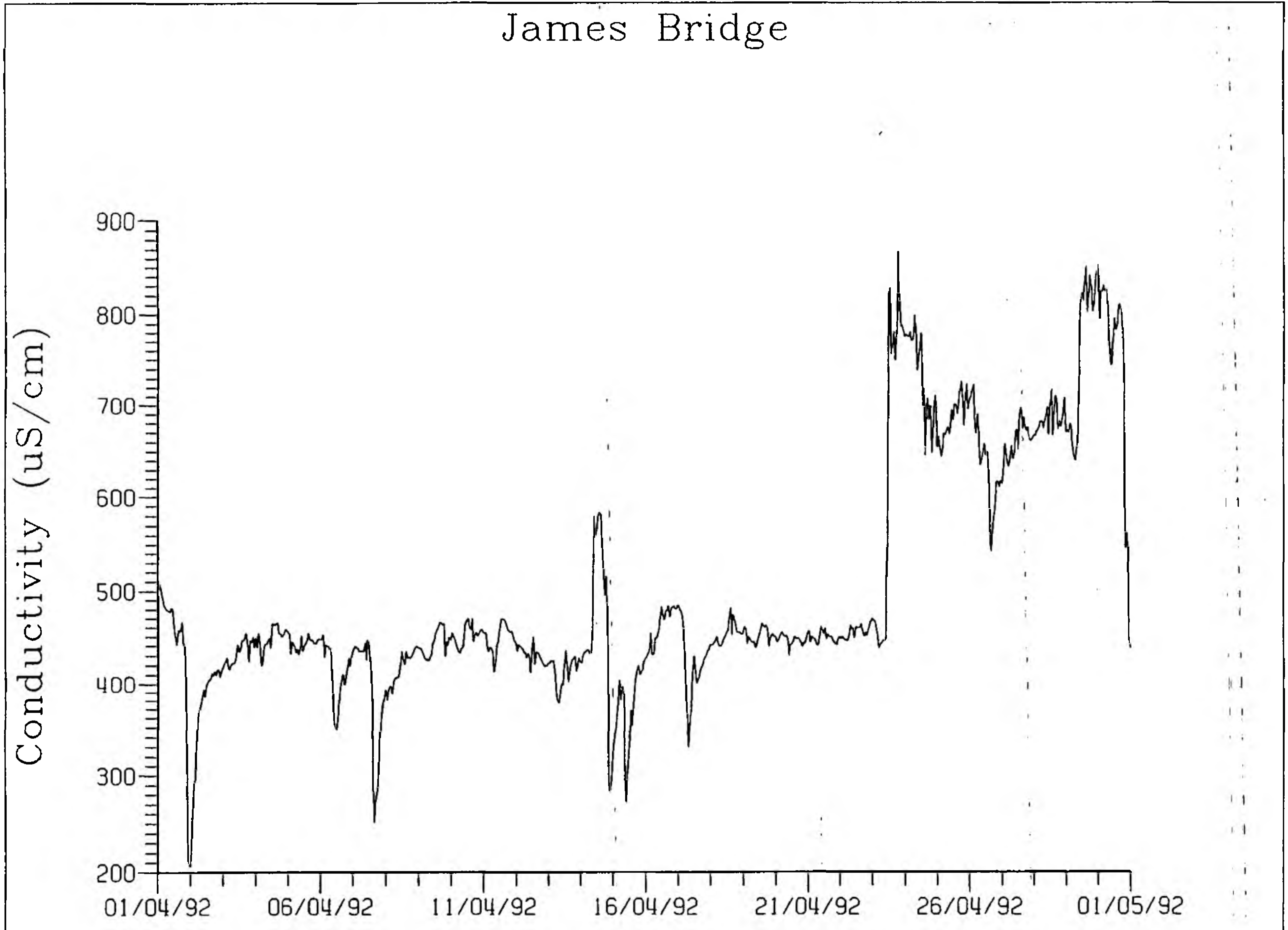
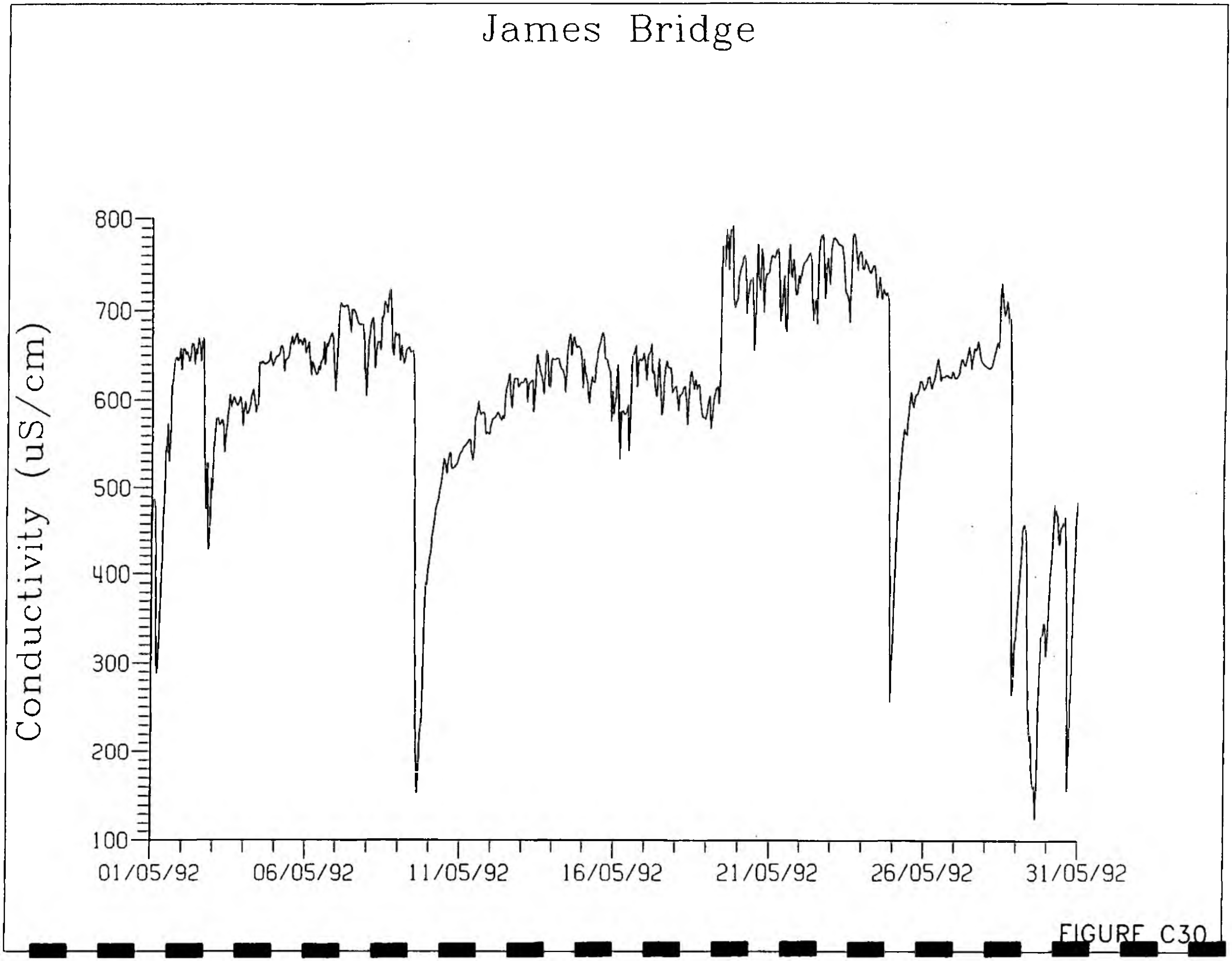

James Bridge
Conductivity (uS/cm)
900
800
700
600
500
400
300
200
01/04/92
06/04/92
11/04/92
16/04/92
21/04/92
26/04/92
01/05/92

FIGURE C29



James Bridge
Conductivity (uS/cm)
800
700
600
500
400
300
200
100
01/05/92
06/05/92
11/05/92
16/05/92
21/05/92
26/05/92
31/05/92

FIGURE C30

James Bridge

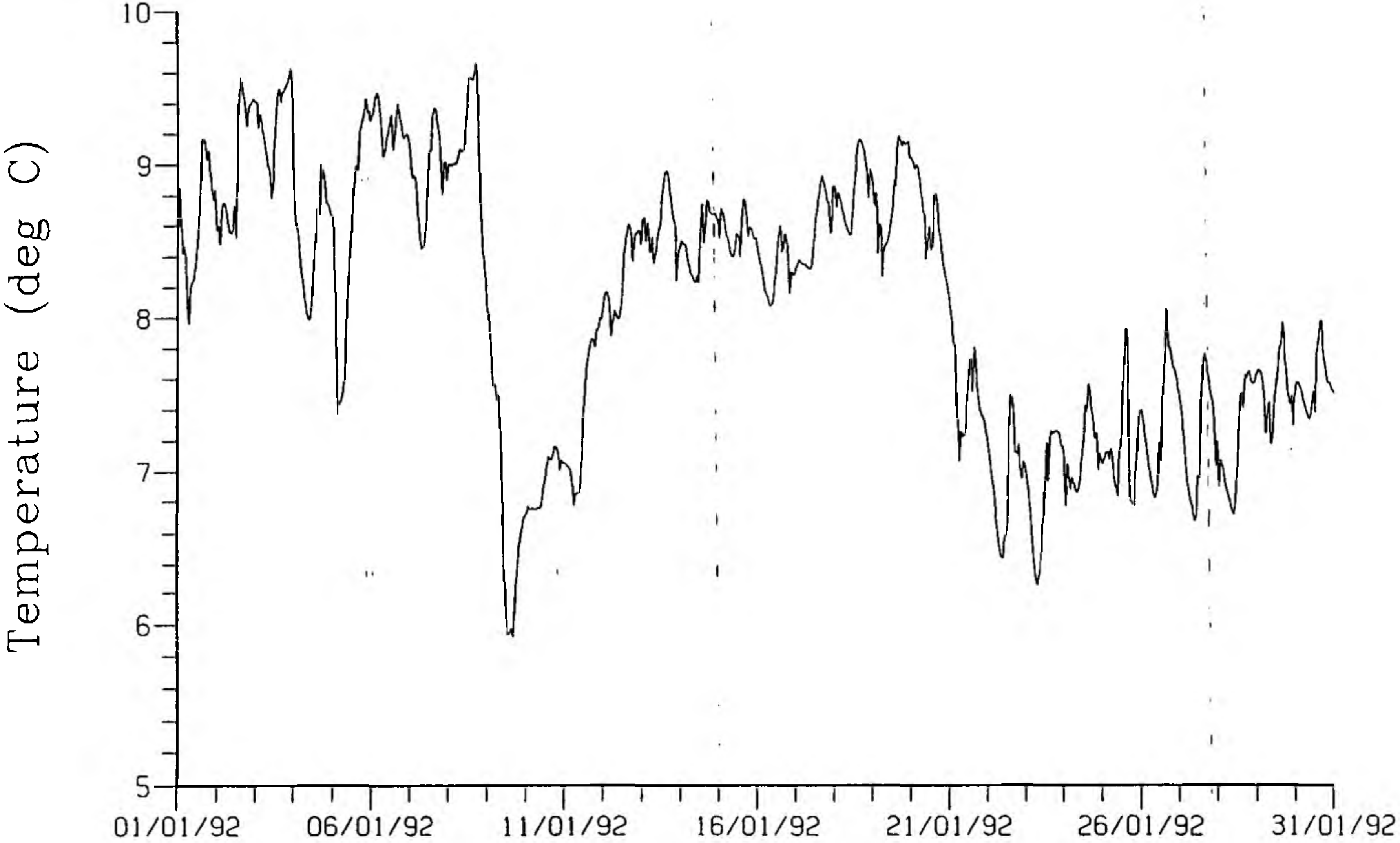


FIGURE C31

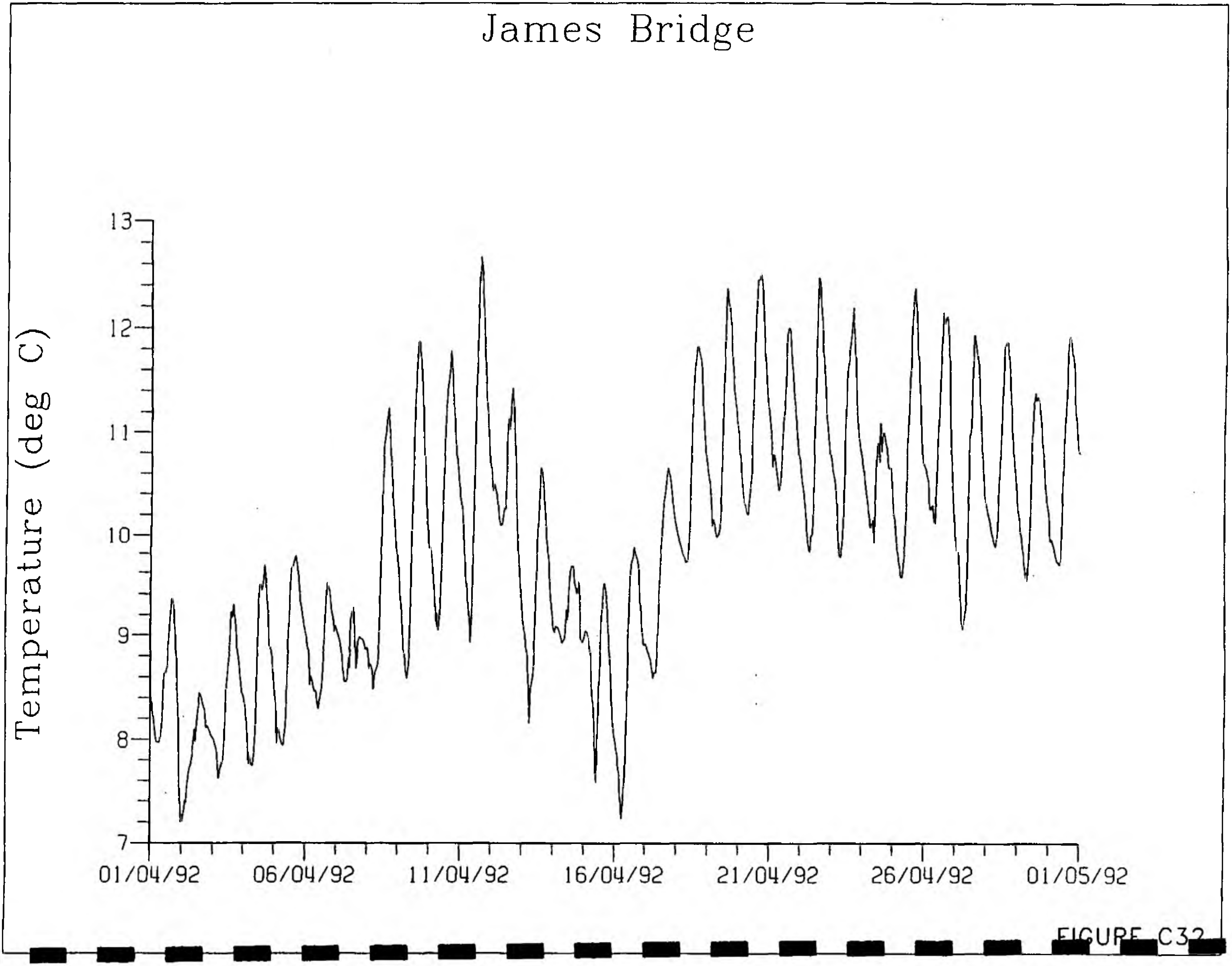

James Bridge
Temperature (deg C)
13
12
11
10
9
8
7
01/04/92
06/04/92
11/04/92
16/04/92
21/04/92
26/04/92
01/05/92

FIGURE C32

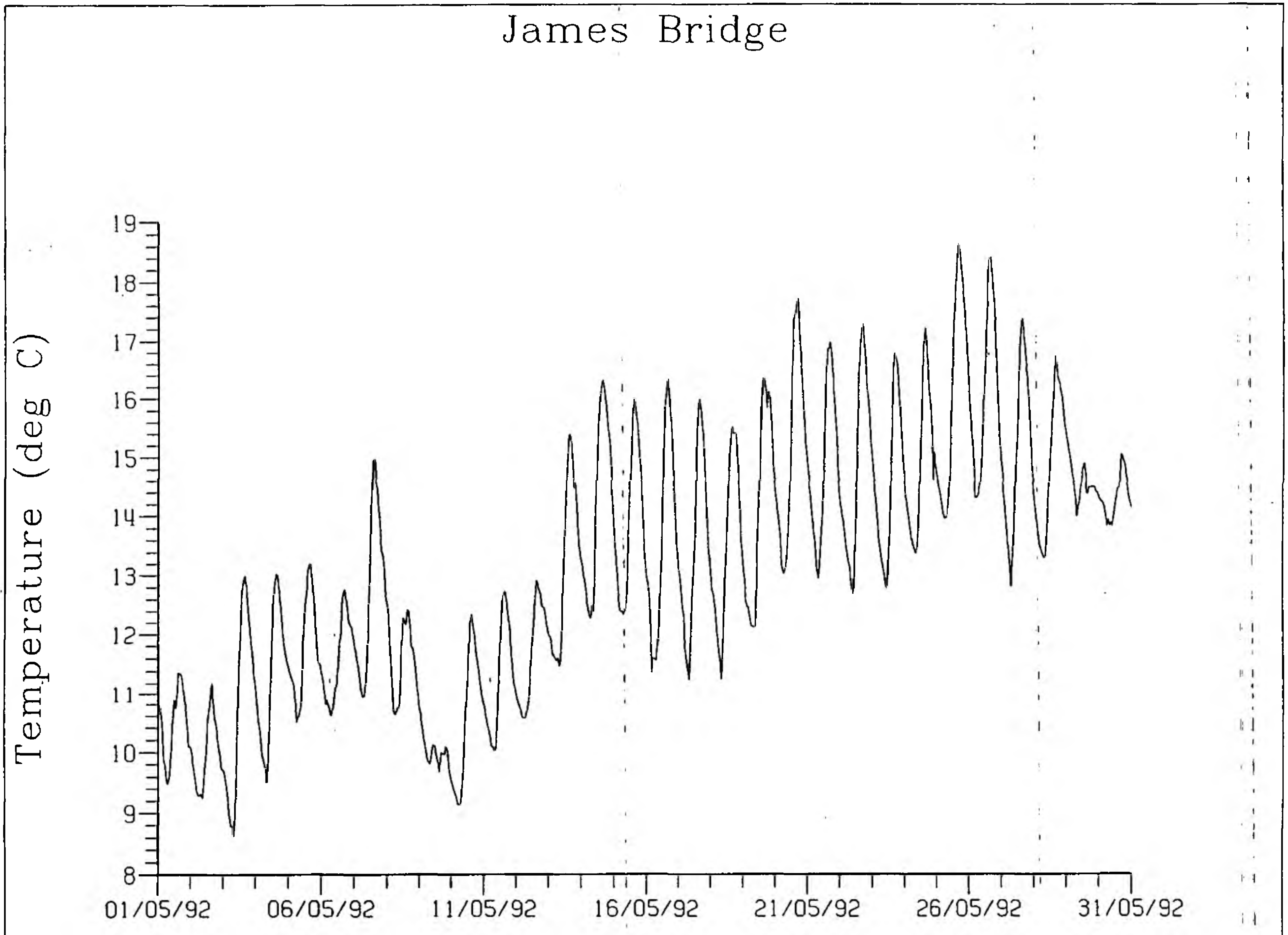

James Bridge
Temperature (deg C)
19
18
17
16
15
14
13
12
11
10
9
8
01/05/92
06/05/92
11/05/92
16/05/92
21/05/92
26/05/92
31/05/92

FIGURE C33

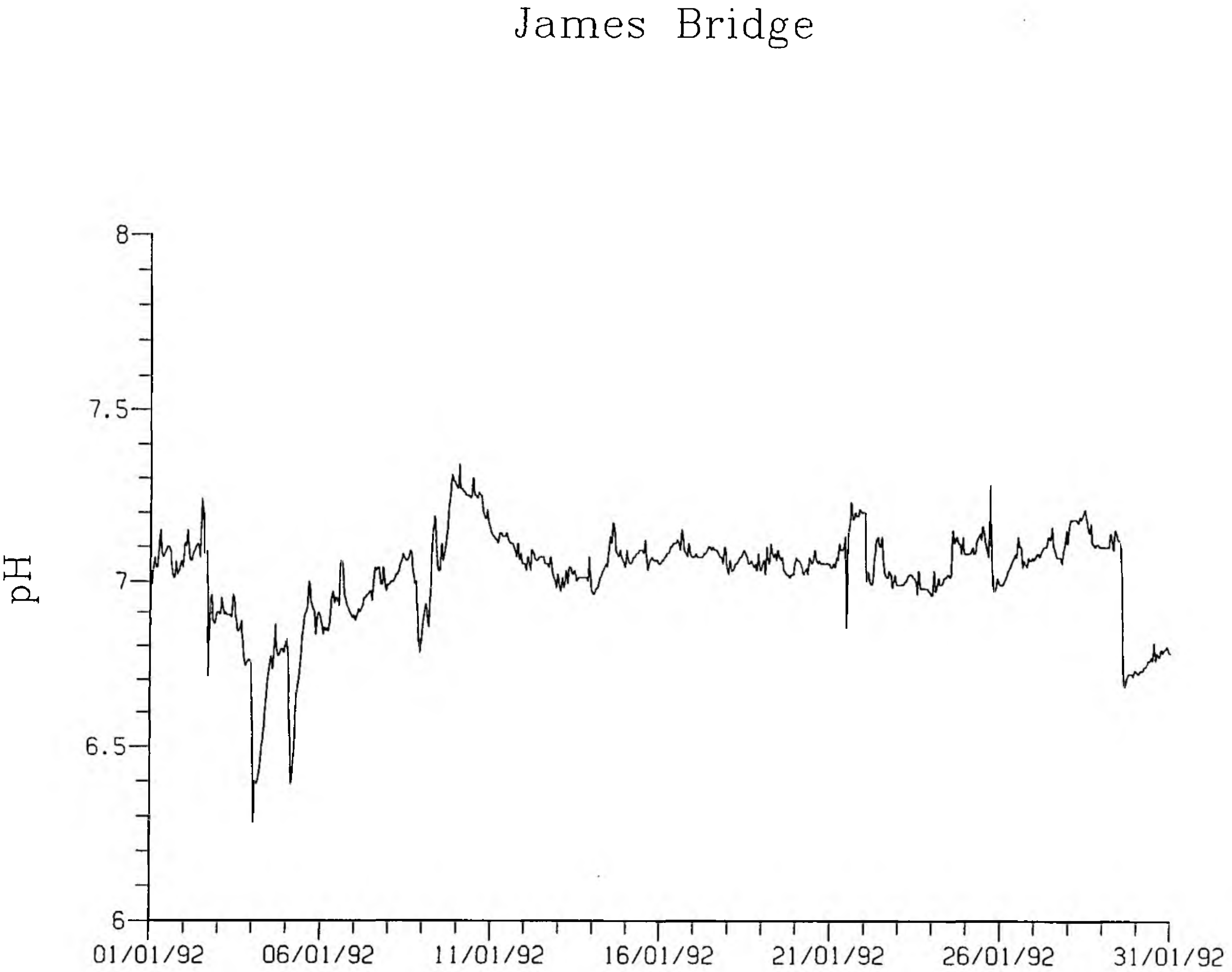

James Bridge
pH
8
7.5
7
6.5
6
01/01/92
06/01/92
11/01/92
16/01/92
21/01/92
26/01/92
31/01/92

FIGURE C34

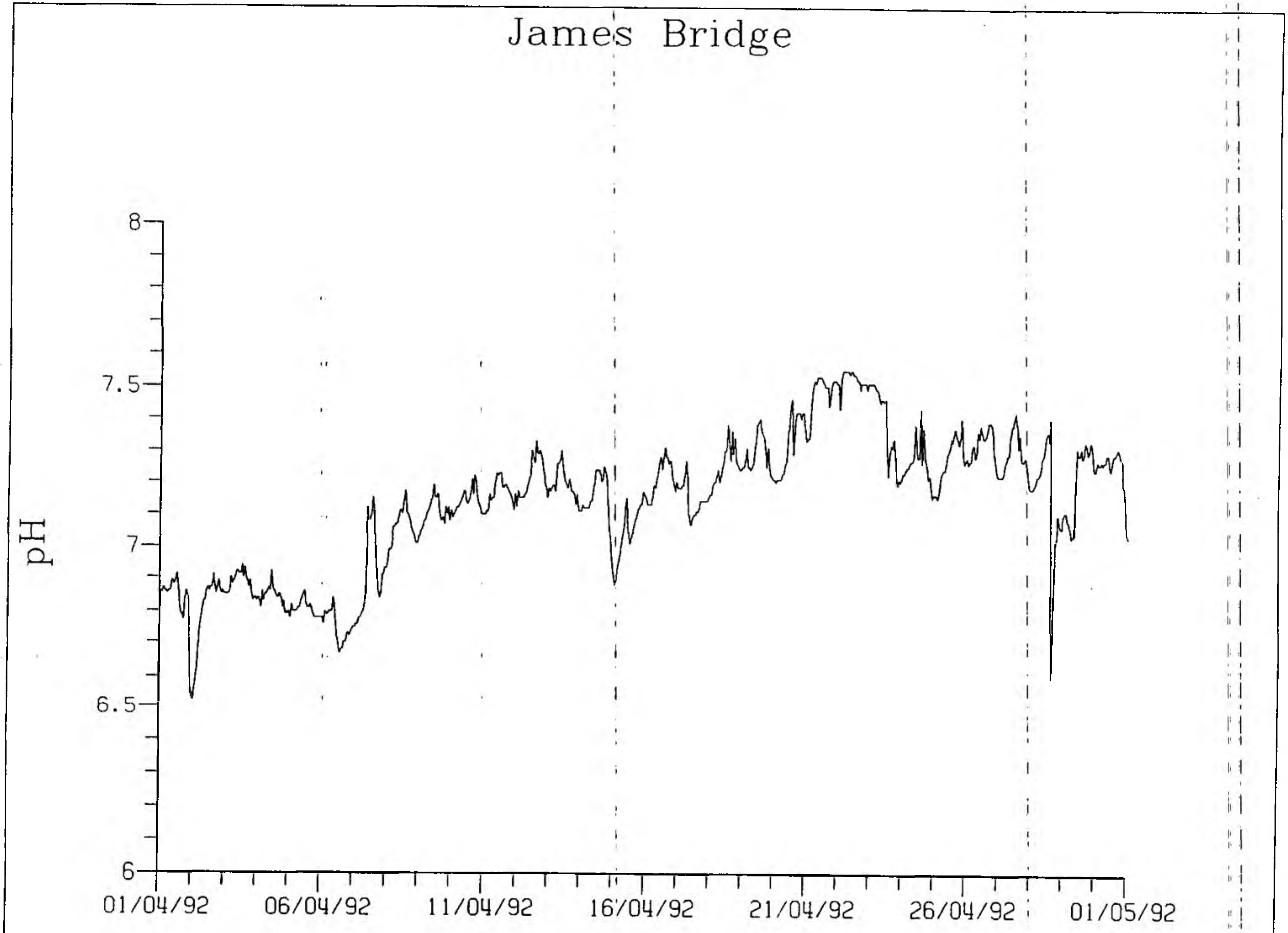

James Bridge
pH
8
7.5
7
6.5
6
01/04/92
06/04/92
11/04/92
16/04/92
21/04/92
26/04/92
01/05/92

FIGURE C35

James Bridge

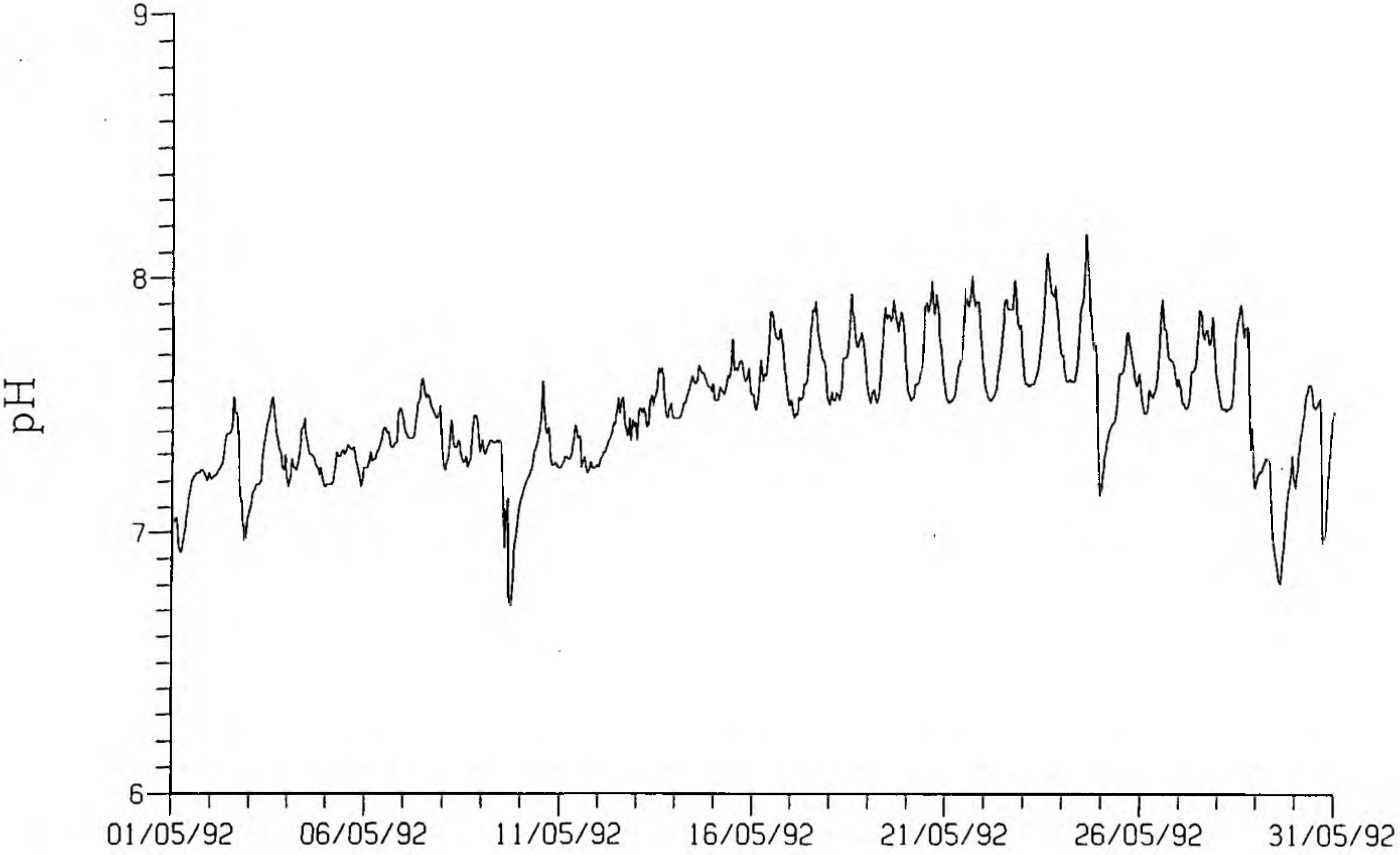


FIGURE C36

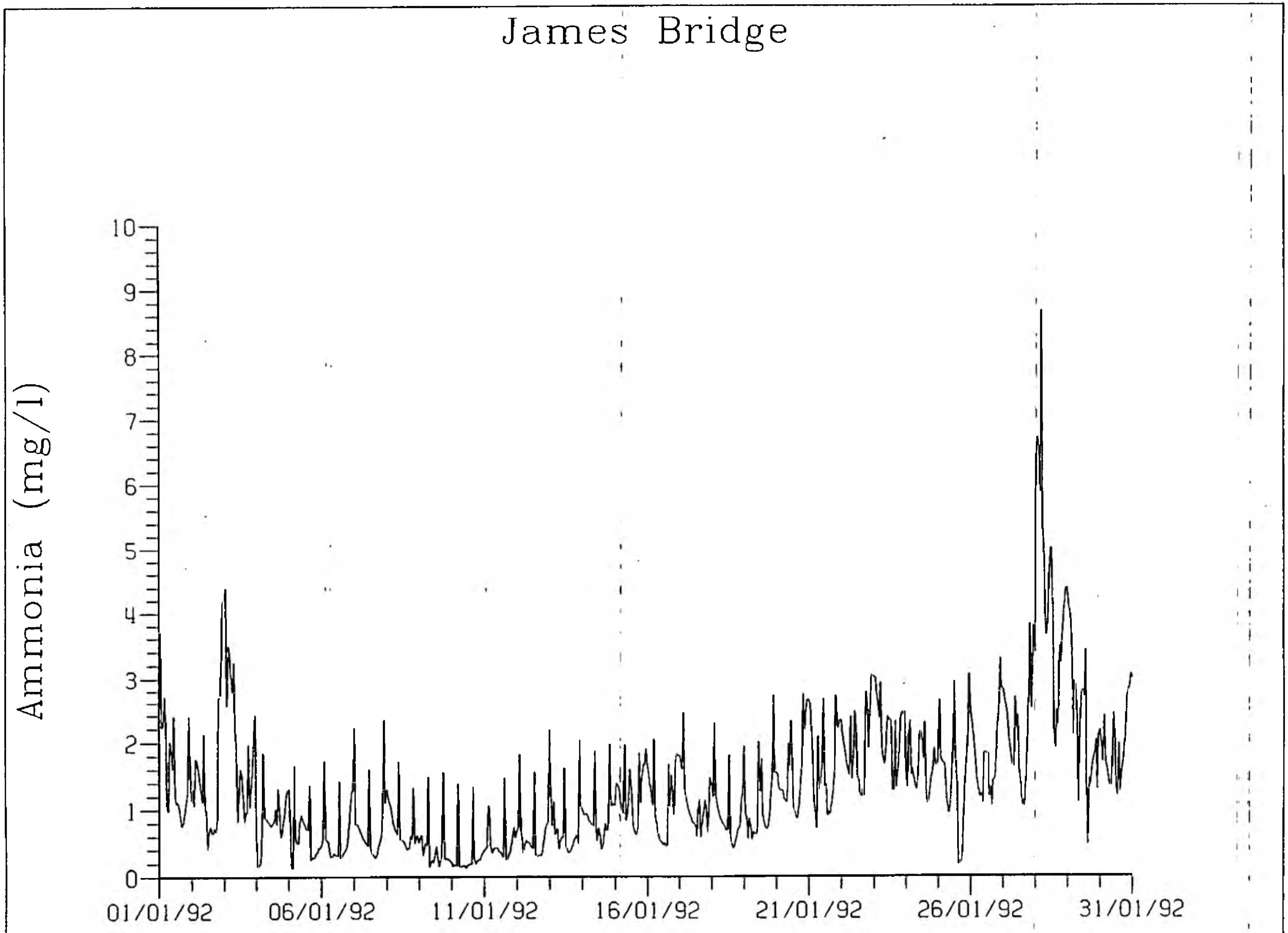

James Bridge
Ammonia (mg/l)
0
1
2
3
4
5
6
7
8
9
10
01/01/92
06/01/92
11/01/92
16/01/92
21/01/92
26/01/92
31/01/92

FIGURE C37

James Bridge

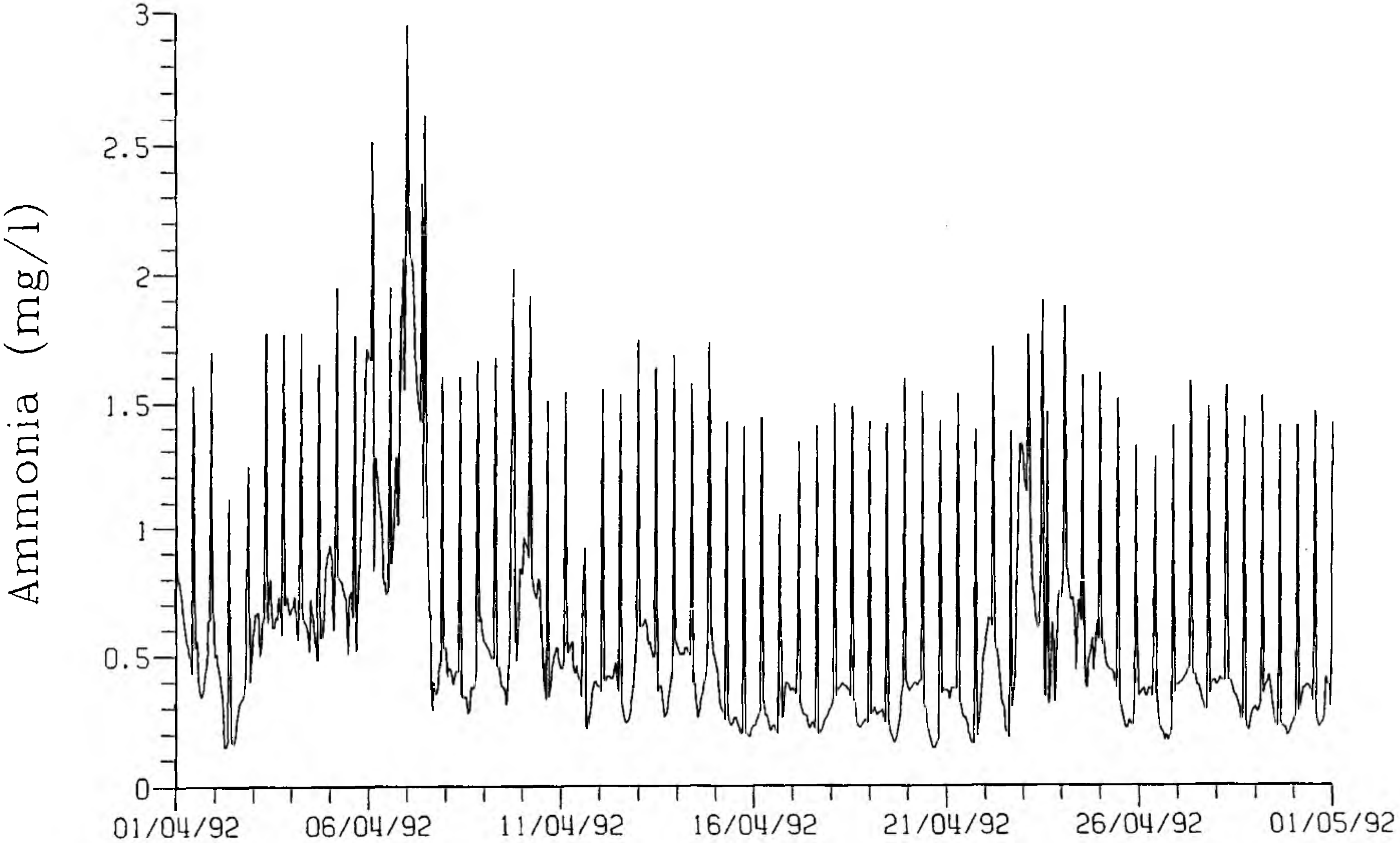


FIGURE C38

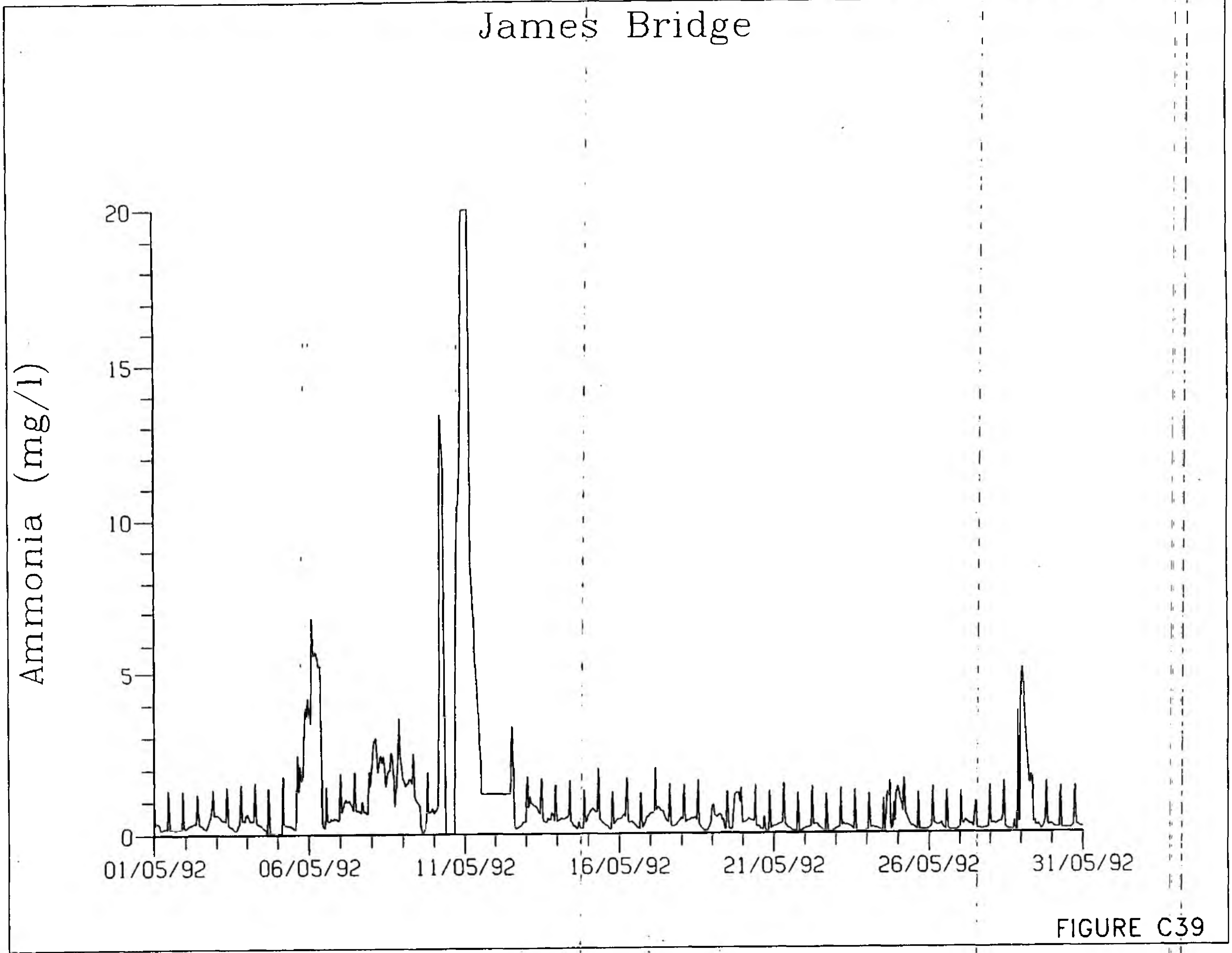

James Bridge
Ammonia (mg/l)
20
15
10
5
0
01/05/92
06/05/92
11/05/92
16/05/92
21/05/92
26/05/92
31/05/92

FIGURE C39

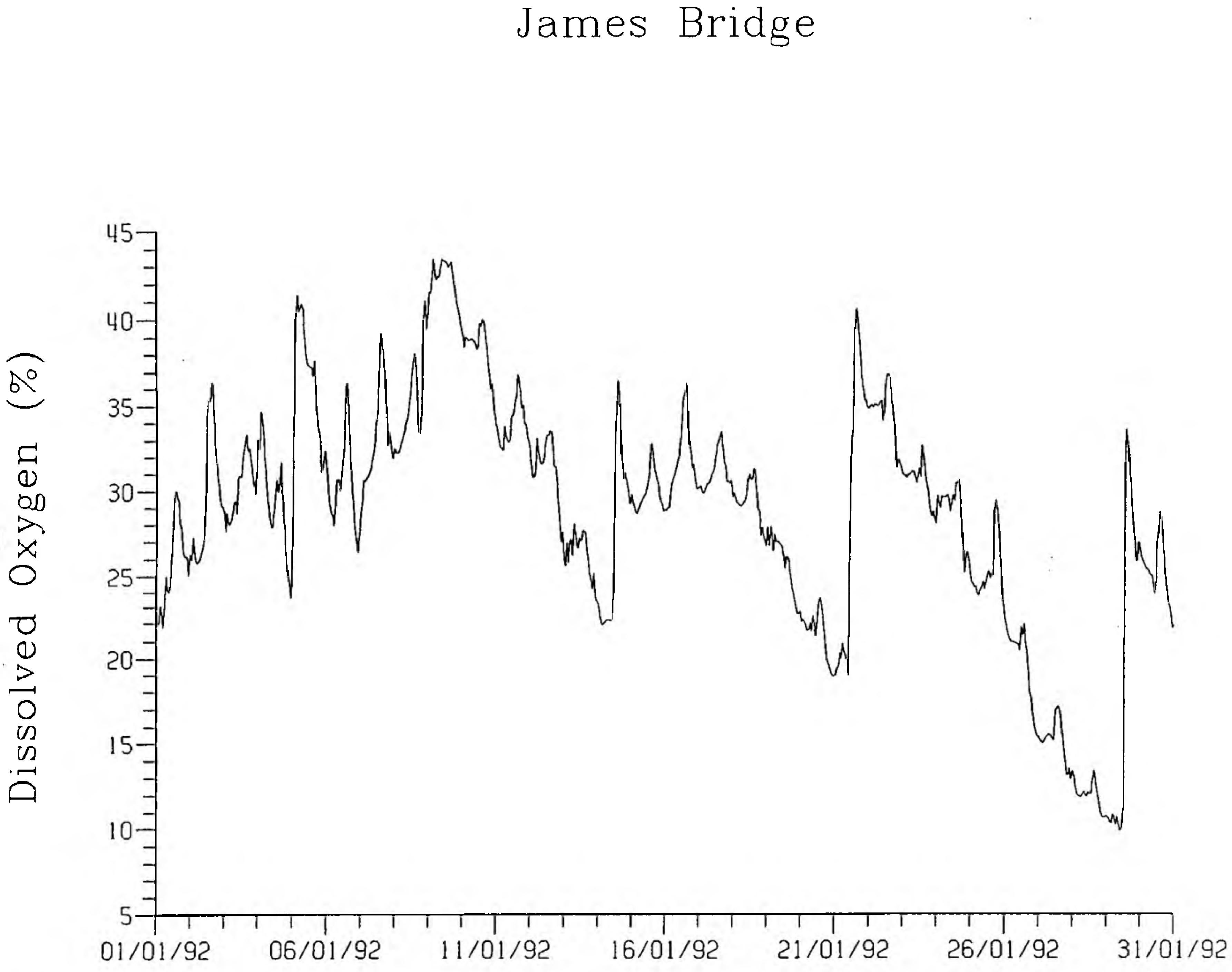

James Bridge
Dissolved Oxygen (%)
45
40
35
30
25
20
15
10
5
01/01/92
06/01/92
11/01/92
16/01/92
21/01/92
26/01/92
31/01/92

FIGURE C40

James Bridge

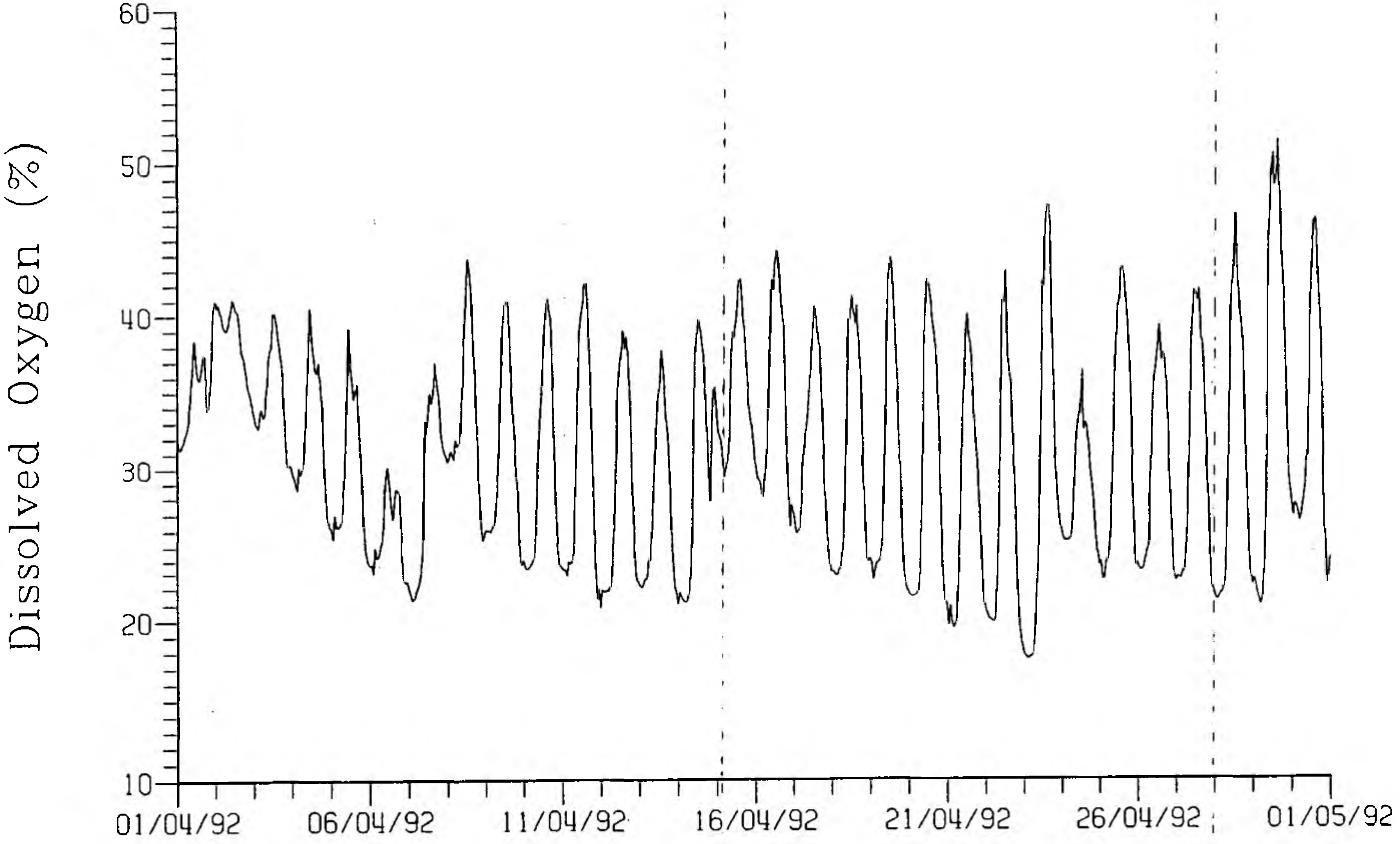


FIGURE C41

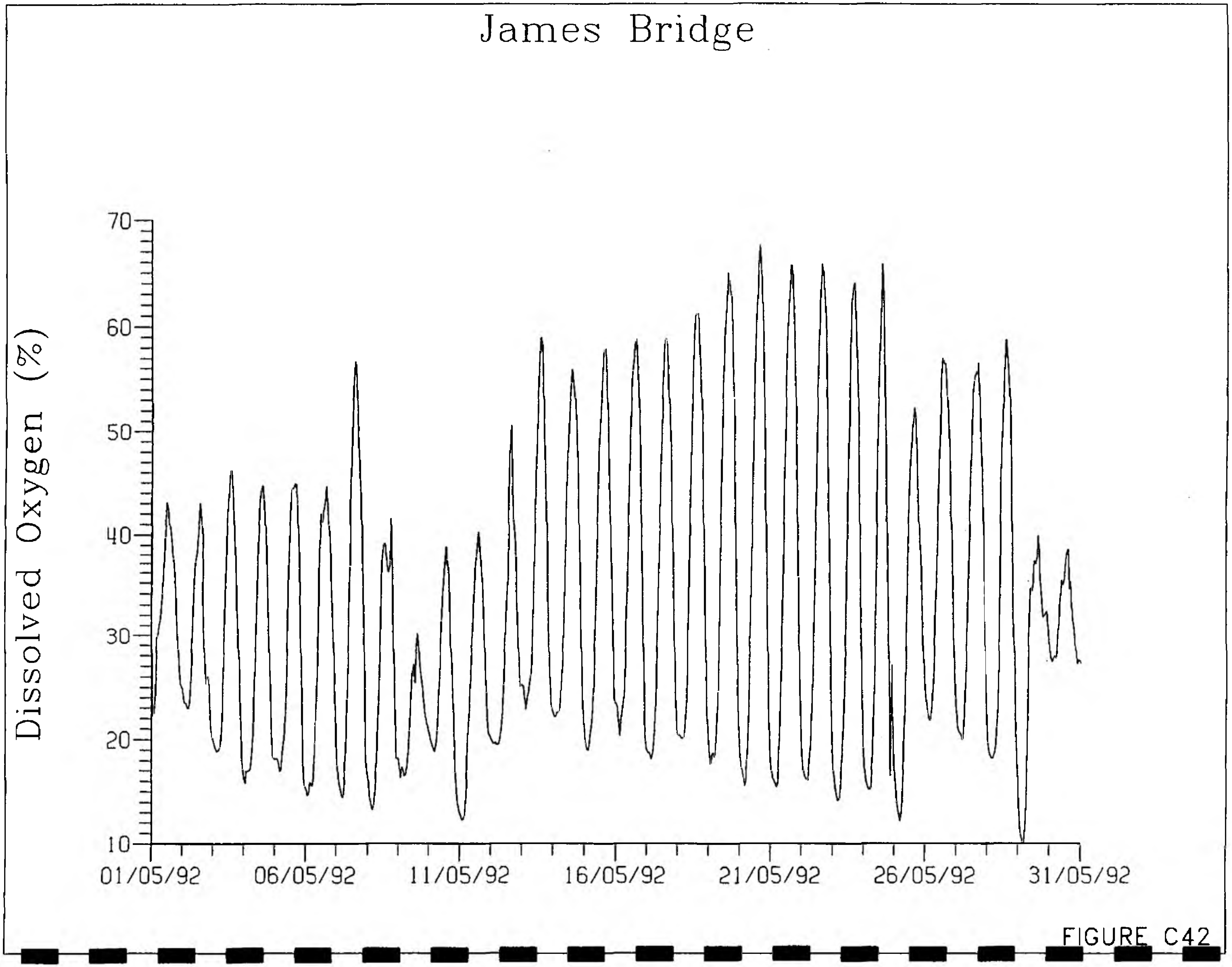

James Bridge
Dissolved Oxygen (%)
70
60
50
40
30
20
10
01/05/92
06/05/92
11/05/92
16/05/92
21/05/92
26/05/92
31/05/92

FIGURE C42

APPENDIX D

REVIEW

REVIEW

Introduction

An independent review of the analysis conducted under Phase II of the project was conducted by Dr. G. Woo, a competent mathematician. Dr. Woo's report and recommendations are presented below and were taken into account when formulating our recommendations for the completion of Phase II. They pre-date the use of Kalman filtering in this study.

Framework for Using Statistical Methods

An important objective in water quality monitoring is the capability of detecting real pollution incidents, while minimising the number of false alarms. This dual concern is well suited to statistical treatment, and may be addressed using a variety of statistical methods. The range of applicable statistical methods is broad in terms of complexity and degree of resolution, and at the outset, a strategy for the use of such methods needs to be framed.

A practical consideration in the choice of statistical methods is their feasibility of implementation, and ease of regular operation. Given the occasional nature of the occurrence of pollution events, a two-level system of statistical analysis would seem to be appropriate. At the first level, a basic statistical alarm system could be introduced, perhaps somewhat more sophisticated than that currently employed, which would have a fairly low threshold for warning of anomalous data, set to ensure that genuine pollution events are very unlikely to be missed. Of those events picked out by this first level alarm, some may be genuine, but many are likely to be represent innocuous excursions from the norm; attributable to a variety of causes, including instrument calibration, sensor malfunction, and stormy weather.

To filter out the false alarms, a secondary level of statistical analysis is envisaged, which would discriminate between the real and false events, with a high degree of rigour commensurate with the success rate demanded by a given criterion of acceptability.

First Level Statistical Methods

At the first level of statistical screening, the methods adopted should be simple, fast, and readily comprehensible. In the context of air pollution monitoring, Nelson et al. (1980) have suggested the following statistical procedures:

(i) Tests for the internal consistency of the data. These include plotting data for visual inspection and testing of outliers.

(ii) Comparing the current data set with historical data to check consistency over time. Examples are comparing data against upper limits obtained from historical data sets, or testing for historical consistency using control charts.

With respect to (i), outliers may be defined as observations which do not conform with the pattern established by other observations, and might result from instrument breakdowns, calibration problems, data processing and transmission errors, as well as from fluctuations in pollution levels. The simplest procedure for screening outliers is to introduce a cut-off at some definite extreme value for each monitored parameter: if the cut-off is exceeded, the first level alarm is raised. Alternatively, a more formal statistical test, e.g. Mann-Kendall nonparametric test, could be undertaken to decide, at a particular significance level, that an observation is an outlier.

With respect to (ii), the essential concept of the the control chart is to use historical datasets as a guide to the range of fluctuations to be expected while the river flow system is in control: i.e. pollution-free. In order for these datasets to be meaningful, care has to be taken in their selection so that they may be regarded as homogeneous. Because daily and seasonal fluctuations render data inhomogeneous, datasets from equivalent recording times should be used.

Rational subgroups is the terminology often reserved to describe these carefully chosen comparison datasets. Once each is chosen, various basic monitoring statistics are evaluated. For every monitored parameter, (e.g. temperature, pH, ammonia etc..), each dataset yields a mean value, a standard deviation, and a range. A chart of the means for each dataset might look as shown in Figure 6.1. The upper and lower control limits may be set to trigger a first level alarm, to indicate that a monitored variable deviates substantially from the expectations inferred from the collection of reference datasets.

Clearly, the construction of control provides a more refined procedure for flagging anomalous monitoring data than simply testing for outliers or setting threshold cut-off values. There is some flexibility in the statistical confidence level associated with the control limits, which will depend on the resolution of the second level statistical methods: the more sophisticated these are, the more relaxed these first level criteria can be.

Second Level Statistical Methods

The rationale for a second tier of statistical analysis is to reduce the number of false alarms, which otherwise might be triggered by a first level statistical alarm. A second level method can afford to be more sophisticated and computer-intensive, because it is only called upon once the first level alarm has been raised, which will be on particular occasions rather than continuously. The introduction of further analytical screening techniques will involve some extra cost in implementation and operation, but this should be more than offset by the savings in the avoidance of unnecessary alarms.

The guiding principle of a second level system is that it should make the most of past empirical observations of data anomalies and instances of pollution incidents. In an automated manner, such a statistical screening system should aim to replicate the trained eye of an experienced human monitor. A human inspector of a recorded times series can often pass a judgement on the existence and cause of an anomaly, just by recollecting precedents. A computerised inspection of a recorded time series should involve comparison with a library of past time series, which would be stored in some fashion in the computer's memory. The aforemention control chart approach is a simple realisation of this technique, which can be made rigorous if a comprehensive set of past time series is made available for comparison, and powerful statistical methods are adopted to identify similar precedents for a given observed time series.

One statistical method which has been quite widely used to detect changes or trends in pollution levels involves the theory of discrete linear random processes. This theory is mathematically tractable, as well as being flexible enough to represent both stationary and non-stationary time series.

A general discrete linear random process has the form:

$$Z_t - \phi 1 Z_{t-1} - \ldots\ldots - \phi p Z_{t-p} = A_t - \theta_q A_{t-1} - \ldots\ldots\ldots - \theta_q A_{t-q}$$

where $\phi 1$, ..., ϕp, θ_1, ... θ_q are parameters, and Z_t describes the behaviour of the time series at time t. The A_t's are uncorrelated input variables. Such a process is called a mixed autoregressive moving average process of order (p,q). The simplest case is that of white noise, where $Z_t = A_t$. The next simplest random process is the moving average process of order q, given by:

$$Z_t = A_t - \theta_1 A_{t-1} - \ldots\ldots\ldots - \theta_q A_{t-q}$$

In this case, Z_t involves a weighted average of the previous q white noise inputs. The last special case is the autoregressive process of order p, in which Z_t is a weighted average of the previous p outputs, plus an independent input A.

$$Z_t = \phi 1 Z_{t-1} + \phi 2 Z_{t-2} + \ldots \phi p Z_{t-p} + A_t$$

Time series are shown in Figure 6.2 for the three cases:

(a) (p=0, q=0)
(b) (p=0, q=1, ϕ=-0.7)
(c) (p=1, q=0, ϕ=0.7)

Multi-parameter Alarms

Each monitored parameter may be associated with an alarm, which may be raised if its characteristics, as gauged by control chart, random process time series modelling, or direct library comparison, resemble those of a pollution incident. Correlations between variables provide additional criteria upon which to gauge the legitimacy of a pollution alert.

To make use of all available methods for discriminating between real and false alarms, and to allow for uncertainty in each, the most robust approach is to assign a weight to the conviction associated with each separate alarm criterion, trigger a pollution alert if the cumulative weight excess some specified threshold figure. The choice of weights for the various alarm criteria would need to be decided on the basis of judgement and experience; the threshold figure for ultimately raising a pollution alert would depend on the relative cost of false alarms as opposed to the consequences of missing an actual pollution incident.

This approach is akin to a voting system, in which each discrimination method is awarded a certain number of votes, according to its aptitude at identifying genuine pollution incidents. This approach is favoured in signal-processing situations, such as prevail in water pollution detection, where the signal (i.e. pollution indicators) may be swamped by general system noise. With such a system for the ultimate screening of false alarms if insufficient votes are cast, the efficiency of the system in rejecting false alarms, but signalling real pollution incidents, can be made very high.