The Influence of Variations in Flow on General Quality Assessment of Rivers

July 2002

The Influence of Variations in Flow on General Quality Assessment of Rivers

R&D Technical Report E1-112/TR

Julian Ellis and David Hunt

Research Contractor:

WRc plc

Publishing Organisation
Environment Agency, Rio House, Waterside Drive, Aztec West, Almondsbury, BRISTOL, BS32 4UD.
Website: www.environment-agency.gov.uk

Tel: 01454 624400 Fax: 01454 624409

ISBN: 1 85705 926 3

Dissemination Status
Internal: Released to Regions
External: Released to Public Domain

Statement of Use
This technical report describes a study to quantify the statistical relationship between river quality and river flows. The document is for use by Environment Agency staff and others interested in trends in the General Quality Assessment for water quality in England and Wales.

Key Words
River Quality, General Quality Assessment, GQA, River Flow, B OD, DO% Ammonia.

Research Contractor
This document was produced under R&D Project E1-112 by:

WRc plc, Frankland Road, Blagrove, Swindon, Wiltshire, SN5 8YF.

Tel: 01793 865000 Fax: 01793 865001 Website: www.wrcplc.co.uk

Environment Agency Project Manager
The Environment Agency's Project Manager for R&D Project E1-112 was:
Julian Struve

Further copies of this report are available from:
Environment Agency R&D Dissemination Centre, c/o
WRc, Frankland Road, Swindon, Wilts SN5 8YF

tel: 01793-865138 fax: 01793-514562 e-mail: publications@wrcplc.co.uk

CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

EXECUTIVE SUMMARY

Background

The Environment Agency monitors river quality mainly through spot sample analyses of Biochemical Oxygen Demand (BOD), ammonia and Dissolved Oxygen (DO%). The resulting data is used for chemical classification according to the General Quality Assessment (GQA). The Agency needs to comment on changes in GQA results: these may occur for many reasons, but flow variations were thought to have a large impact, and Agency press releases have attributed apparent deterioration in GQA results to low flows. However, with no formal method to quantify the effect, the Agency has had to rely on the observations and experience of its regional staff. It therefore commissioned WRc plc to undertake a statistical study of the relationship between flow and various measures of water quality, including GQA class in particular. This report describes the methodology adopted and the results obtained from that study.

Objectives

The broad aims of the project were to (a) build up an understanding of the relationships between quality and flow for the three GQA determinands, and (b) clarify how any relationships discerned are influenced by the nature of the river. The project also aimed to assess the effect on quality of factors other than flow, such as AMP-related improvements.

Methods

A fundamental element of the approach was the use of several different but complementary approaches, defined primarily by the degree of data aggregation involved:

- **Low-level** - no aggregation, data being used at the level of individual samples.
- **High-level** - data aggregated over time and determinands, but not over space (i.e. sites).
- **Regional-level** - data aggregated over all three dimensions.

At each level, a variety of statistical techniques were applied to flow and quality data. For the low- and high-level approaches, this consisted principally of flow and chemical quality data for all 565 GQA sites in Thames Region, supplemented by selected sites in other Regions. For the Regional-level approach, several aggregated data sets provided by the Agency were analysed.

Findings and Conclusions

- Statistically significant correlations between individual GQA determinands (DO%, BOD and ammonia) and flow are found at only about one-half of Thames Region sites, and are weak, providing no convincing evidence of an effect of flow on GQA results.
- GQA monitoring data gives only suggestive evidence of a greater number of BOD improvements - or smaller number of deteriorations - during the AMP years, with even weaker evidence for a positive effect of AMP activity on ammonia and DO%. However, there may be reasons why the analysis might not have been expected to detect an effect of AMP improvements.

- GQA class changed markedly at many Thames Region sites over the past 20 years, and a widespread improvement in river quality is evident over the period. Flow was much more variable in the 1990s than in the 1980s; mean summer and annual flows fell over each of the last three 5-year periods; and flow was more persistent in the 1990s than in the 1980s.
- The extent to which GQA class varies appears to be largely unrelated to variation in site mean flow, and sites with more variable flow do not have a greater proportion of significant quality v. flow models. Only 1 in 5 sites show significant associations between 1-year GQA class and log mean flow. At virtually all of these, flow increases are associated with GQA improvement. However, the relationships are too weak to account for more than a small proportion of the observed improvements in GQA class.
- Previous Agency examinations of (a) aggregated data for each of the eight Regions, and (b) aggregated data for the North-East Area of Thames Region appeared to show quite strong associations between GQA change and flow, but were compromised by the presence of strong autocorrelation arising from the use of 3-year rolling GQA results. When analysed with correction for autocorrelation:
  - Paper (a) shows a statistically significant GQA v. flow association for only Thames Region individually, but also a highly significant association when data is pooled across all Regions.
  - Paper (b) shows no significant GQA v. flow associations for two sub-groups of the data, but a highly significant correlation for another (mainly because of good GQA performance and high flows in 1993-95, and poor GQA results and low flows in 1996-98).
- The absence of clear evidence of relationships at the level of individual data points is not inconsistent with a strong high-level GQA v. flow association, but it severely hampers attempts to explain, and justify any projected use of, such an association.
- No consistent pattern was seen in the data for the nine Thames Region sites showing the greatest recent GQA improvement. However, in 3 cases there was evidence of a threshold-type relationship between DO% and flow, whereby the risk of obtaining low DOs increases sharply once flows have dropped below some critical value.
- It is concluded that an association between GQA and flow does exist at some sites, but is not readily discernible at the individual determinand and site levels for several reasons, including (a) the limitations of monthly GQA data in relation to infrequent and transient events, and (b) the complexities of behaviour of individual GQA determinands, especially dissolved oxygen, in response to low flows.
- In particular, the project has found no convincing evidence to support previous suggestions by the Agency that a deterioration in national GQA results could mainly be attributed to low flows.

Recommendations

- Analysis of aggregated data should be repeated for smaller groups of sites than whole Regions: this would allow more representative flow data to be used with the aggregated GQA data, and would also give the opportunity for differences to be seen in the strength of association within Regions as well as between Regions.
- A detailed search, automated using suitable software and statistical tools, should be conducted for patterns of behaviour such as the low flow/low DO% 'threshold' effect

postulated here. This should focus primarily on DO, but also include exploratory searches for potentially relevant patterns in BOD and ammonia data.

- Consideration could also be given to examination of river quality records involving more frequent sampling - especially in relation to DO.

1. INTRODUCTION

The Environment Agency ('the Agency') monitors river quality in England and Wales mainly through spot samples analysed for Biochemical Oxygen Demand (BOD), ammonia and Dissolved Oxygen (DO%). The resulting data is used for chemical classification of rivers according to the General Quality Assessment (GQA).

The Agency has a duty to comment on possible reasons for any changes in overall and regional GQA results. Such changes may have a variety of causes, including variations in river flow, variability in the degree of algal growth, sampling error (especially systematic error), and changes in the flow and quality of sewage treatment works effluents. Variations in river flow in particular are thought to have a large impact, and Agency press releases have attributed apparent deterioration in GQA results to low river flows. However, no formal method exists to quantify the effect of a change in flow on quality, and so the Agency has had to rely on observations by, and the experience of, its regional staff.

Recognising the limitations of this situation, the Agency commissioned WRc plc ('WRc') to undertake a study to clarify the relationship between flow and various measures of water quality variables, and specifically between flow and GQA class. The study took place in 2001 and 2002, and this Technical Report describes its objectives, detailed methodology and results.

2. BACKGROUND AND OBJECTIVES

2.1 Background

2.1.1 Causes of variation in river quality

Over periods of a year and longer, river quality at a site varies for a variety of reasons - some readily explicable, others less so. One of the known explanatory factors is flow, with two common (but potentially opposing) mechanisms being:

1. high rainfall causing increasing run-off of agricultural or other substances, leading to an increase in substance concentrations with flow; and
2. high rainfall and flows diluting a fairly constant load into a river, leading to a decrease in substance concentrations with flow.

All but the most highly regulated rivers show substantial variations in flow over a typical year, so some degree of association between quality and flow can therefore be expected.

Changes in the flow and strength of upstream discharges can have a major influence on river quality, with improvements resulting from Asset Management Plan (AMP) investment being of particular interest. Quality can also be affected by a variety of weather-related influences conveniently labelled 'seasonal' – e.g. the effect of sunshine/photosynthesis on dissolved oxygen, or the tendency for sewage treatment works to perform better in summer because of the increased biological activity at higher temperatures. It should also be noted that, because flow itself has an element of seasonality (tending to be low in summer and high in winter), it can be difficult to distinguish seasonal quality effects from those due specifically to flow.

2.1.2 Summarising river quality

The Agency uses two methods of summarising chemical river quality: the GQA classification system and the Rivers Ecosystem (RE) method. For GQA, river quality data for a rolling three-year period is summarised in the form of 10%ile (DO%) or 90%ile (BOD and ammonia) values; and the performance of the poorest determinand in relation to the class boundaries dictates the GQA class. The RE system uses essentially the same approach, except that 90% confidence intervals are calculated for the three percentiles, and the optimistic confidence limits then used for comparison against the class limits for the required class. (The RE system also incorporates a number of other determinands which are handled in a similar way.)

This work has intentionally addressed the impact of flow upon GQA classification specifically, this being the principal summary measure of river quality.

2.2 Objectives

2.2.1 General

The Agency wished to be able to explain why GQA results vary from one assessment period to the next – particularly in relation to the quinquennial surveys of river quality, the results of which receive wide coverage. Variations in the behaviour of river flow from one assessment period to the next were thought likely to have a substantial impact on the results, and the Agency had attributed the deterioration in national GQA results noted just before the project commenced to low flows. However, this belief was based largely on the informal observation and experience of local staff, with no more formal body of knowledge to support it.

The broad aims of the project, therefore, were to:

- build up an authoritative understanding of the relationship between river quality and flow for each of the three main GQA determinands - BOD, dissolved oxygen (DO%) and ammoniacal nitrogen (ammonia); and

- clarify how any quality v. flow relationships discerned are influenced by the nature of the river - such as its hydromorphological characteristics, and the degree of impact of effluent discharges upstream.

The project also aimed to determine the effect on quality of explanatory factors other than flow, such as AMP-related improvements.

The ultimate aim was to establish a quantitative link between (a) changes in the pattern of river flow across a collection of sites, and (b) the resulting changes in their GQA assessment. This would provide an objective basis, given year-to-year variations in regional and national GQA results, for judging how much of these are attributable to changes in the flow regime.

Detailed objectives

The Agency's detailed objectives, as stated in the project specification, were as follows:

1. To investigate the relationship between river flow and individual water quality determinands DO%, BOD and Ammonia. To investigate which river statistic, e.g. mean flow, 5%ile or 95%ile, plays the most important role. To test if the resulting relationships are predictive. To develop a method to correct for the effects of flow variation from sample data at the time series or summary statistic level.
2. To investigate if the sensitivity of rivers or reaches against river flow variation can be predicted from river characteristics. A possible predictor of sensitivity is the ratio of effluent flow to river flow. To investigate if any predictive approaches that relate the response of river quality to fluctuations in flow can be applied with data from another region.
3. To investigate other potential factors that could influence the sensitivity of river quality against flow variations, i.e. hydromorphological factors such as river gradient and degree of eutrophication.

4. To investigate if a systematic relationship exists between river flow and summary GQA class, and if it could be used to quantify the effects of flow variations on these results. To investigate how this approach would compare with the results obtained from correcting the original data (see Objective 1).
5. To investigate the relative importance of river flow variations compared to other factors influencing water quality in Thames Region, especially the impact of AMP improvements.
6. To investigate the problem of scale in determining the influence of river flow variations on Agency water quality results. To investigate if relationships developed for individual catchments, areas, or regions are predictive for other catchments, areas and regions.
7. To quantify the role of river flow in the national water quality results. Rainfall and river flow vary in time and magnitude among the EA regions and it is also possible that the response of rivers to flow changes is different in different regions. Investigate how these differences impact on the national GQA results.

It was recognised at the outset that the direction taken by the project would to some extent be shaped by the degree of success with which the first few objectives were met. In particular, some of the later objectives might cease to be relevant (or indeed achievable) if the established quality v. flow models were not strong enough or applied to an insufficiently large number of sites.

In the event, it did prove necessary as the project progressed to modify the detailed scope, as we indicate where relevant in subsequent chapters.

2.3 The Steering Group

The project Steering Group consisted of Juliane Struve (Project Manager) and Paul Davidson from Thames Region, plus Simon Bingham from Head Office. Progress meetings were held at roughly quarterly intervals at which interim results were presented by the WRc project team, Julian Ellis and David Hunt, and planned future work reviewed. These meetings, together with frequent email communication between WRc and members of the Group, proved extremely useful in helping to shape the direction taken by the project - especially (as noted above) in its later stages.

3. DATA SOURCES

3.1 GQA data

At the start of the project, Paul Davidson (PD) carried out WIMS retrievals for all 565 GQA sites in the Region. For each of these sites he identified the nearest relevant flow gauging station, and then used the Agency's LOUISE utility to match each GQA sample with the mean daily flow for the corresponding date. Finally he used LOUISE to create an Aardvark-format data file for each GQA site containing the following determinands:

- Flow (Ml/d)
- BOD (5 day using ATU), (mg/l)
- Ammonia as N, (mg/l)
- Nitrate as N, (mg/l)
- Orthophosphate as P, (mg/l)
- Chlorophyll A Meth, (ug/l)
- Dissolved Oxygen, (%Sat)
- Dissolved Oxygen, (mg/l).

The data spanned a nominal 21-year date window from 1980 to 2000 (although there were gaps in the record for some of the sites during the 1980s). The 565 pairs of .DAT and .CTL files were supplied to WRc by email.

3.2 Flow data

During the process of linking GQA sites to flow sites, PD identified a total of 116 relevant gauging stations across Thames Region. For each of these stations, he provided WRc with a file in standard Hydrolog 'SCF' format containing daily mean flow data for the nominal 21-year date window from 1980 to 2000. (As with the quality data there were occasional gaps; also, for some stations the data was not yet available for 2000.)

3.3 AMP and other site-related data

In the first progress meeting, various site-related factors were discussed which might influence the way in which flow affected quality – i.e. affect the slope of any mathematical function relating quality to flow. Arising out of this, agreement was reached on a list of potentially useful explanatory variables, such as the proximity of an upstream sewage treatment works (STW) discharge, and the typical dilution provided by the river.

Data was also needed for the proposed investigation into the impact of AMP improvements. Specifically, it was agreed that two representative groups of sites would be identified: one for rivers where there had been no AMP-related development work, and the other where there had been AMP improvements with completion dates known (at least approximately).

Finally, data was needed on rolling three-year GQA Class for each of the 565 sites, to serve as a check on the values calculated by WRc.

Data to meet these various requirements was assembled early in the project by PD, and provided to WRc as Access files.

3.4 Data from Regions other than Thames

In view of how the project evolved, there was less emphasis than had originally been envisaged on repeating the modelling for data from other Regions. Nevertheless, some GQA quality and flow data sets were provided by three other Regions - Anglian, Midlands and Southern - as shown in Table 3.1. These sites were put forward by the Regions concerned as being good examples of where flow was thought to have a marked influence on water quality.

Table 3.1 Data provided by Regions other than Thames

Region	GQA site	Corresponding flow site
Anglian	Dernford on R. Cam	Dernford
Anglian	Billingford Bridge on R. Waveney	Billingford
Anglian	New Mills on R. Wensum	New Mills (but too little data available)
Midlands	Clifton on R. Avon	Stareton
Midlands	Stanbridge Farm on Glynch Brook	Wedderburn
Midlands	Water Lane on Oxton Dumble	Lowndham
Southern	E0001453	Crabble Mill
Southern	E0001456	Crabble Mill
Southern	E0001462	Crabble Mill

In most cases the river quality data was provided in Aardvark format, and so relatively little effort was needed to modify the software analysis tools that had been developed specifically for the Thames data (see Chapter 4). One region, however, provided its river quality data in an unconventional Hydrolog-type Excel format, and quite a lot of effort was needed to re-format this into a more conventional 'dates × determinands' layout.

3.5 Regional-level data

Several ad hoc data sets containing GQA and flow information aggregated to the Regional level (see Section 4.1.1) were provided by Steering Group members during the course of the project.

4. METHODOLOGY

4.1 General approach

4.1.1 Levels of data aggregation

In investigating the impact of flow on quality, a fundamental element of the original project specification - as indicated in Chapter 2 - was the use of several different but complementary approaches, defined primarily by the degree of data aggregation that each approach calls for. This has accordingly been a key theme running through the project, and so it is useful to start with a brief discussion of the types of aggregation that are possible, and why some are more useful than others.

Table 4.1 below shows the types of data aggregation that can be adopted when looking for relationships between river quality and flow. For example, the aggregation can be:

- over time - for example, when looking at the 3-year 90%ile BOD concentration for a particular site rather than the individual data values;
- over determinands and time - for example, when looking at GQA class at a particular site;
- over space and time - for example, when looking at the proportion of river sites in a Region for which the 3-year 90%ile BOD concentration is 4 mg/l or less;
- over determinands, space and time - for example, when looking at the proportion of river sites in a Region falling into class A or B.

There are three dimensions over which data can in principle be aggregated - time, space, and determinand. As any of these may or may not be aggregated, there are $2^3 = 8$ possible combinations. However, not all of these are sensible options. For example, it would not be useful to aggregate over space alone. This would imply pooling individual data values for a number of sites (and losing track of which values were associated with which rivers). It would be similarly unhelpful to aggregate over determinands and space without also aggregating over time. The unsuitability of these options is indicated by the dark shading of the corresponding cells in the table.

For the present exercise, we have used three types of aggregation, as indicated by the unshaded cells of the table, namely:

- **Low level** - where there is no aggregation, and data is used at the level of individual samples;
- **High level** - where data is aggregated over time and determinands but not over space; and
- **Regional level** - where data is aggregated over all three dimensions.

Table 4.1 Types of data aggregation

Temporal aggregation	Aggregation over determinands	Spatial aggregation	
		No - use individual sites	Yes - use grouped sites
No - use data for individual samples	No - use individual determinands	**'Low level'**	**Unhelpful**
	Yes - use combined determinands	*e.g. Water Quality Index on a particular date* **Not relevant**	**Unhelpful**
Yes - use summary statistics over 3 years (or any other whole no of years)	No - use individual determinands	*e.g. 90%ile BOD at a site* **Not considered**	*e.g. % of sites with 90%ile BOD ≤4 mg/l* **Not considered**
	Yes - use combined determinands	*e.g. GQA Class at a site* **'High level'**	*e.g. % of sites with GQA class of B or better* **'Regional level'**

A fuller discussion of the data aggregation concept can be found in Chapter 11 of the Sampling Handbook (WRc report NS29).

4.1.2 Relating the different levels of analysis

With investigations being pursued at different levels of data aggregation, it is important that the various sets of results can be set into a coherent overall framework so that it is clear how the findings obtained from one level of analysis reinforce or relate to those gained from another level. We address this issue in Chapter 8.

4.2 Low-Level analyses

4.2.1 Summarising quality determinands

With the quality data having been provided in Aardvark-format files, it was a straightforward matter to carry out an initial appraisal of the data using Test Data Facility (TDF) routines. The TDF is a flexible batch-based system for carrying out a variety of statistical analyses on any specified collection of data sets, and was developed by WRc in the early 1990s during the 'Codes of Practice for Data Handling' project undertaken for the NRA.

At the start of the project we applied the TDF routine MOT (Multiple Outlier Test) to all quality determinands in each of the 565 data sets. This served two purposes: it provided a check that there had been no problems in extracting or transmitting the data files, and it also

gave an indication of the severity and rate of occurrence of potential outliers. (The latter were not removed, however, as to do so could have distorted the calculated GQA classes.)

4.2.2 Measures of antecedent flow

In looking for a relationship between quality and flow, the simplest approach is to relate each water quality sample to the corresponding same-day mean flow. However, the behaviour of water quality on a particular day may well be influenced not only by same-day flow but also by the behaviour of flow over preceding days. For the low-level modelling work, therefore, we devised the following three groups of potentially useful antecedent flow measures:

Group 1 - mean flows over various numbers of days counting backwards from 'today'…

- flow(d) - i.e. the 'same-day' flow as used in the preliminary analyses;
- mean flow over d to d-1;
- mean flow over d to d-2;
- etc, etc
- mean flow over d to d-14;

Group 2 - mean flows over various numbers of days counting backwards from 'yesterday'…

- flow(d-1);
- mean flow over d-1 to d-2;
- mean flow over d-1 to d-3;
- etc, etc
- mean flow over d-1 to d-14;

Group 3 - weighted mean flows over various numbers of days, with linearly declining weights…

- [2×flow(d) + flow(d-1)]/3;
- [3×flow(d) + 2×flow(d-1) + flow(d-2)]/6;
- etc, etc
- [15×flow(d) + 14×flow(d-1) +… + flow(d-14)]/105.

4.2.3 Relating quality to flow

For the first stage of the quality v. flow modelling work, we exploited the fact that the data sets provided by Thames had very helpfully included same-day flow as another 'determinand' along with the various quality determinands. Started with an existing TDF routine CFC (Concentration-Flow Correlations), we developed this in various ways, as described below, to produce the new routine **CAFE** (**C**oncentration **A**gainst **F**low **E**valuation). For any data set, this showed the strength of association between (a) quality and flow, and (b) log(quality) and log(flow) for all specified determinands.

Subsequently we extended CAFE in the following ways:

1. We developed a system for linking any given GQA data file to any desired SCF flow file. (We did continue using the quality-flow links that had initially been defined by Thames, but the new system gave the capability to use other flow files later in the project should this be needed.)
2. We added a date-matching routine which located the same-day flow corresponding to each record in the quality file, and then calculated the various antecedent flow measures as described in the previous section.
3. We arranged for CAFE to loop through all the candidate flow fingerprint variables in turn to see how much improvement could be gained in the association between quality and flow if we give the modelling process free rein over the choice of explanatory variable (i.e. the freedom to try each antecedent flow in turn). As with the earlier version, CAFE did the analyses both in the unlogged world and the logged world.

4.2.4 Testing the adequacy and temporal stability of fitted models

A statistical model may have a high correlation coefficient and yet not pass very well through the data. Similarly, the model may fit the data better in some years than in others - which would be a limitation if the model were to be used for prediction. When we were developing CAFE from the earlier CFC routine, accordingly, we thought it important to introduce a number of statistical enhancements which explored the structural 'adequacy' and temporal stability of each fitted model.

The tests are all based on calculating the 'residuals' from each model - that is, the differences between the actual and fitted values - and then subjecting these to various tests for randomness. A detailed account of the methodology is provided in Appendix A.

4.2.5 Assessing the effect of AMP changes on quality

We used a two-stage approach in assessing the strength of evidence for improvements in quality that might reasonably be attributed to AMP investment. As noted earlier, 50 sites had been identified where associated AMP schemes with (approximately) known completion dates had been put in place during the 1990s. First, using the TDF routine SAD, we ran cusum analyses on DO%, BOD and ammonia for each of these 'AMP' sites. We made the assumption that a change occurring within a year of the completion date could be termed 'AMP-related'. Thus, by using the date window 1990 -2000 for the cusum analyses we identified 10 years of non-AMP-related changes, and one year of AMP-related changes. Under the null hypothesis that the frequencies of improvements and deteriorations in quality are unrelated to whether or not an AMP scheme had just come on-stream, the expected frequency in the 1-year AMP window is one tenth of the observed frequency in the other 10 years. Thus we can compare observed and expected frequencies in the AMP window and use the Poisson distribution to assess whether any discrepancies are statistically significant.

This approach breaks down if the background rates of improvements and deteriorations in quality *themselves vary through time*. For example, if conditions happened to be particularly favourable in the late 1990s, this might be the reason for a sudden increase in the numbers of improvements in quality rather than AMP investment. To protect against this possibility, we repeated the above analysis on a set of 50 'Control' sites, these being sites with sewage treatment works upstream but with no history of AMP investment over the period. We could

then compare any actual:expected ratio for the AMP sites with the corresponding ratio for the Control sites, using Fisher's test for 2×2 contingency tables, and hence determine whether the effect was genuine. (This is an example of a Before/After Control/Impact - commonly used by biologists.)

4.3 High-Level analyses

4.3.1 Calculating GQA class

To calculate GQA class for any data set we developed the TDF program **GCSE** (**G**QA **C**lass **S**equentially **E**valuated). This carried out the required GQA percentile calculations for ammonia, BOD and DO% for each consecutive 3-year period in the data set to generate rolling 3-year GQA class. We first confirmed the accuracy of the statistical calculations in GCSE by checking that they agreed with the corresponding calculations in Aardvark. We then extracted a table of historical GQA values from one of the Access files previously provided by the Agency, and did some manipulation in Excel to compare our results from GCSE with the Thames values. There were a few disagreements in the early years, due to the Aardvark file not containing the full 3-year data set. But once we moved into the 1990s there was virtually 100% agreement between the two sets of results.

Towards the end of the project it was agreed that we would extend the high-level analysis to look at the effect of flow on '1-year GQA class' - that is, class as calculated from the data for individual years rather than 3-year blocks. We added an additional routine into GCSE to produce these new statistics. (We did, however, impose a minimum requirement of 12 sample values, in view of the substantial statistical uncertainty associated with 10%ile and 90%ile estimates from such limited data sets.)

4.3.2 Analysing trends in GQA class

In view of the general and widespread improvement in river quality seen over the last 20 years, it was thought useful to quantify the strength of this overall time trend as a preliminary to starting on the modelling proper. (Note: for the purpose of this and subsequent analyses, we converted GQA classes A to F into the numbers 1 to 6.)

Using the GenStat package we carried out an analysis of variance (ANoVA) on GQA class. There were three terms in the model:

- A **site factor -** at 565 levels. This was included merely as a nuisance term to absorb the variation due to the fact that some sites are generally better than others.
- A **time factor** (at the four levels 89-91, 92-94, 95-97 and 98-00. (Note that there were too many gaps in earlier years for it to be worthwhile extending the analysis pre-1989.) This factor was used to determine the strength of evidence for an overall time trend in GQA class across all sites.
- A **time covariate** This was included in order to test for a specifically *linear* time trend.

4.3.3 Developing 'Flow Fingerprint' measures

One of the requirements of the project was to develop a small number of flow summary statistics which jointly provided a concise but comprehensive description of flow in any given river. We accordingly spent some time visualising the various characteristics of flow which help to make one river different from another; and to assist in this process we wrote the Fortran program **FFION** (**F**low **F**ingerprint **I**nformation **O**f **N**ote), which calculated a great variety of flow statistics from any given daily flow series.

Because of the varying uses to which these statistics were going to be put, we built into FFION the flexibility to operate on the data in consecutive blocks of either 2, 3, 4 or 5 years.

Following exploratory investigations with FFION, we defined a selection of summary statistics or measures which we termed 'flow fingerprint candidates' (FFCs). These formed four groups, as detailed in Table 4.2. The first, main group consists of summary statistics relating to flow itself - such as log10(mean) and relative standard deviation. Two of these - log10(summer mean) and RSD(annual mean flow) - were suggested later in the project by the Steering Group, which explains why their numbers are out of sequence.

Table 4.2 Flow fingerprint candidates (FFCs)

No	Category	Name	Description
1	Flow	N	No of valid daily flows
2		Log10Mean	Log of mean flow
25		Log10SummerM	Log of summer mean flow
3		CoV	Relative st.deviation
24		CoV(AnnAv)	Rel.st.dev. of annual mean flow
4		Mean/P50	Mean:50%ile ratio
5		P05/P50	5%ile:50%ile ratio
6		P95/P50	95%ile:50%ile ratio
7		P95/P05	95%ile:5%ile ratio
8		ACC1	Lag-1 autocorr. coefficient
9		ACC15	Lag-15 autocorr. coefficient
10		ACC30	Lag-30 autocorr. coefficient
11	F(i)/F(i-1)	N	No of valid consec.day ratios
12		Mean	Mean ratio
13		CoV	Relative st.deviation of ratio
14		P75	75%ile of ratio
15		ACC1	Lag-1 autocorr. coeff. for ratio
16	UpRuns	N	No of upward runs
17		P50	50%ile no of days in upward run
18		P75	75%ile no of days in upward run
19		P95	95%ile no of days in upward run
20	DownRuns	N	No of downward runs
21		P50	50%ile no of days in downward run
22		P75	75%ile no of days in downward run
23		P95	95%ile no of days in downward run

The second group of FFCs relates to the ratio of flows on consecutive days. The other two groups relate respectively to 'up runs' and 'down runs'. A 'run' in this context is any consecutive period of days over which daily flow changes in the same direction (or perhaps stays unchanged).

Note that, in nearly all cases, the FFCs have been constructed in such a way that they are scale-free. This is so that they can be compared more readily between river sites.

4.3.4 Analysing trends in flow

For each of the 25 FFC measures we carried out a two-way ANoVA with the factors **site** (at 565 levels) and **time** (at the four levels 80-84, 85-89, 90-94 and 95-99). We used 5-year rather than 3-year blocks, as the flow data went back to 1980 with relatively few gaps. We used the analysis to answer two main questions:

1. How strong are the differences *between* sites in relation to the *within*-site variability?

2. Is there any evidence of a time trend that is common to all sites?

We also undertook a brief multivariate analysis exercise to provide additional statistical insight into the similarities and differences between the various FFCs. The broad aim was to identify any patterns and similarities existing amongst the FFCs, and thereby assist in the process of deciding which small subset of the FFCs might be most useful for the purposes of the project. First, we used *principal component analysis* to analyse the inter-correlation matrix of the FFCs. If this were to show that the FFCs fell into several distinct, strongly correlated groups, this would indicate that FFCs falling within any one group were broadly interchangeable, whilst FFCs falling in different groups would be potentially more useful in that they tended to measure different aspects of flow. Secondly we used *cluster analysis* to see if the FFC measures were instrumental in grouping the flow sites into geographically distinct river types. (The search for such a categorisation was one of the tasks specified in the original programme specification.)

4.3.5 Relating GQA to flow

As noted earlier, rolling 3-year GQA class was calculated for each site. This produced 19 values, running from 1980-82 to 1998-2000. However, for the purpose of high-level modelling it would be invalid to use all 19 values because they are autocorrelated - they do not provide statistically independent information. Specifically, two of the three years' data used in calculating 3-year class for any two consecutive years are common to both calculations. The problems caused by autocorrelation - and possible solutions - are discussed in some detail later in Appendix B and Chapter 7.

For each site analysis, therefore, we selected the six *non-overlapping* class values for the periods 1982-84 to 1997-99. (We had to choose 1999 rather than 2000 as the final year because at the time of the analysis some of the flow data was not yet available for 2000.) From an output file produced by FFION we also extracted the values of the 25 flow fingerprint candidates (FFCs) for the corresponding 3-year periods. We then carried out linear regressions of GQA class against each of the FFC measures - thus giving a grand total of 25 regressions for each of 565 sites.

One unavoidable problem with this approach was the small number of data points in each regression. Even if all six pairs of values are available, the correlation coefficient R needs to be numerically greater than 0.81 for it to be statistically significant at the $P<0.05$ level. The problem was exacerbated by missing values; with only four points, for example, R must be as high as 0.95 before a regression provides useful evidence of an association. (With fewer than four points the data was too limited to be trustworthy and so we discarded the analysis.)

Largely because of this inherent data limitation, it was decided in the later stages of the project to repeat the regression analysis using 1-year rather than 3-year GQA class. This was certainly a more attractive proposition on statistical grounds: there were generally 15 or more data points available per site in comparison with 6 previously, and this more than compensated for the greater uncertainty in each individual GQA class value.

4.4 Regional-level analyses

Quality data aggregated to the Regional level (see Section 4.1.1) typically consists of measures such as:

- % of sites showing an improvement in class from one survey to the next;
- net % improvement in class by river length; or
- % of sites in classes A-C.

For the modelling work that we undertook at the Regional level, a useful focus was provided by two discussion papers supplied by Steering Group members. These brought to the forefront the issue of autocorrelation, and this prompted a short exercise using computer simulation to demonstrate the bias that can be introduced when using regression analysis on autocorrelated data. (As noted earlier, this exercise is summarised in Appendix B.) Following on from this, we carried out some additional analysis on the data presented in the above papers to assess the extent to which autocorrelation was masking or distorting the real underlying influence of flow on quality.

Finally, using the new insights provided by these papers we examined the detailed results for a representative selection of Thames Region sites in an attempt to reconcile the apparent conflict between the generally weak low-level findings and the stronger effect seen at the Regional level.

5. LOW-LEVEL RESULTS

5.1 Quality v. flow relationships for DO%

5.1.1 Models using same-day flow

Figure 5.1 summarises the performance of the DO% models obtained from the unlogged data using same-day flow as the explanatory variable. Each point represents a GQA site, and it plots the correlation coefficient ('R') for the DO% v. flow regression for the site against the number of samples available for that site. With monthly sampling over 21 years, we would expect around 250 samples in all. However, the majority of sites actually fall in the 100-200 samples range. This shortfall is a consequence of the lower frequencies often seen in the early 1980s, coupled in some cases with gaps in the flow record.

The figure also plots a 95% probability region for R under the null hypothesis that R = 0. In other words, an R value falling anywhere within the two curves could well have arisen solely by chance rather than because of any underlying relationship between DO% and flow. With 100 points, for example, R must be numerically greater than about 0.2 for it to be statistically significantly different from zero at the P<0.05 level.

We see, therefore, that the majority of sites (actually 422 out of the 565, or 75%) fall within the not-significant region.

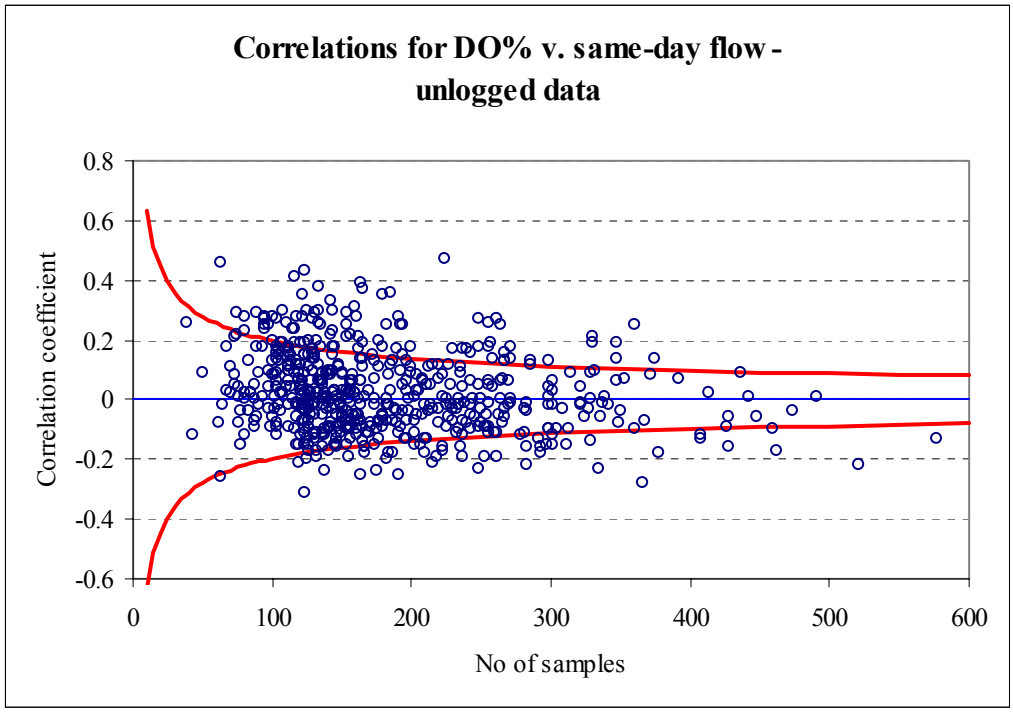


Figure 5.1 Performance of DO% models using same-day flow (unlogged data)

Even where the R value for a site is *statistically* significant, this does not necessarily mean that it is *practically* significant. In other words, it does not follow that variations in flow will explain a useful amount of the variations seen in DO%. For example, consider the case of a model with R = 0.4 (a value that is exceeded by only four sites in the figure). The proportion of the total variance accounted for by the model is given by R^2, which is 0.16. Thus the *unaccounted-for* variance proportion is 0.84; and we need to take the square root of this to get back to the original measurement units, giving 0.916. The uncomfortable message from this is that the standard deviation of the scatter around the model is barely 8% smaller than the original overall standard deviation of DO%. Such a model would clearly be of little use for predictive purposes. (We demonstrate this later in the chapter after presenting the results for BOD and ammonia.)

Figure 5.2 similarly summarises the regression results that we obtained when carrying out the analyses on log(DO%) and log(same-day flow). There is a noticeable improvement in the number of sites having statistically significant R values. Even so, these are still in a minority, with 58% of sites *not* showing a significant effect. Moreover, the R value for even the best model is only 0.6. For this model the residual standard deviation, expressed as a proportion of the overall standard deviation, is still as high as $\sqrt{(1-R^2)} = 0.8$. Thus, if such a model were used to predict DO%, the prediction uncertainty would be only 20% narrower than the uncertainty associated with simply using the overall mean as the prediction.

Notwithstanding the weakness of the identified effects, one interesting point to note is that the great majority (81%) of the statistically significant R values are positive. Thus an increase in flow tends to be associated with an increase in DO%.

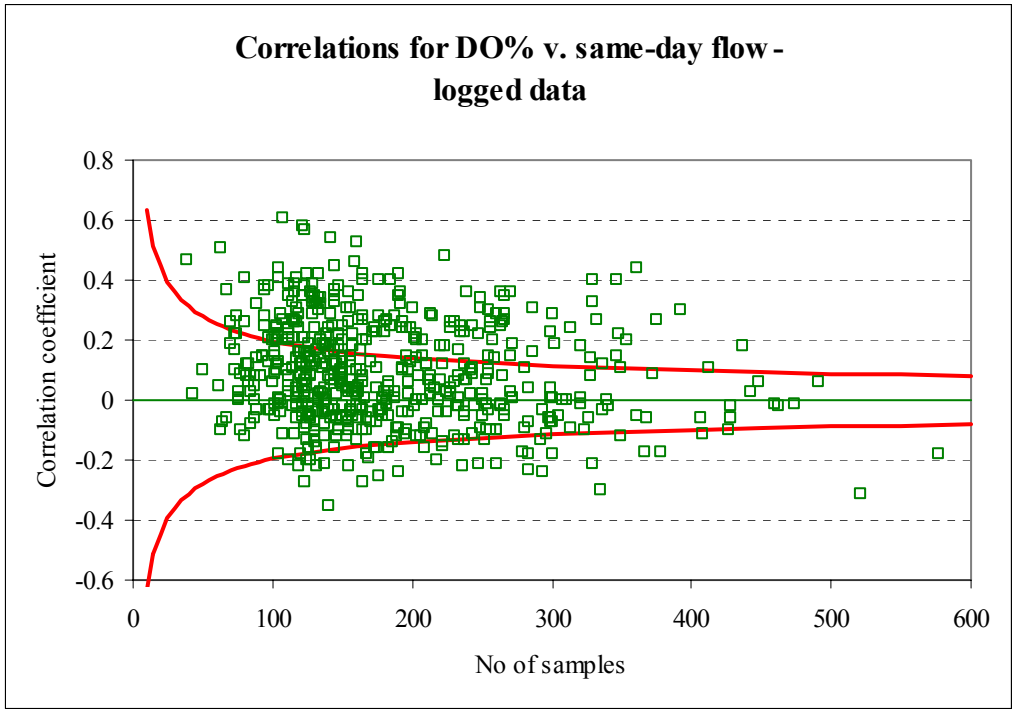


Figure 5.2 Performance of DO% models using same-day flow (logged data)

5.1.2 Models using other antecedent flow measures

The models summarised in the previous section were restricted to the use of same-day flow as the explanatory variable. Section 4.2.2 has described how we generalised program CAFE so that it searched through a variety of antecedent flow measures and picked out the one which produced the greatest improvement in R. We carried out these extended analyses in both the unlogged and logged modes. The results for the logged data case are shown in Figure 5.3. Each stalk represents a GQA site, and its height shows the numerical improvement in R obtained in going from same-day flow to the optimal antecedent flow measure for that site.

It is evident that in the great majority of cases the increase in R is trivially small (or even zero in some cases), and there are only a few sites for which R increases by more than 0.15. Thus there is no practical benefit to be gained in using anything more complicated than same-day flow.

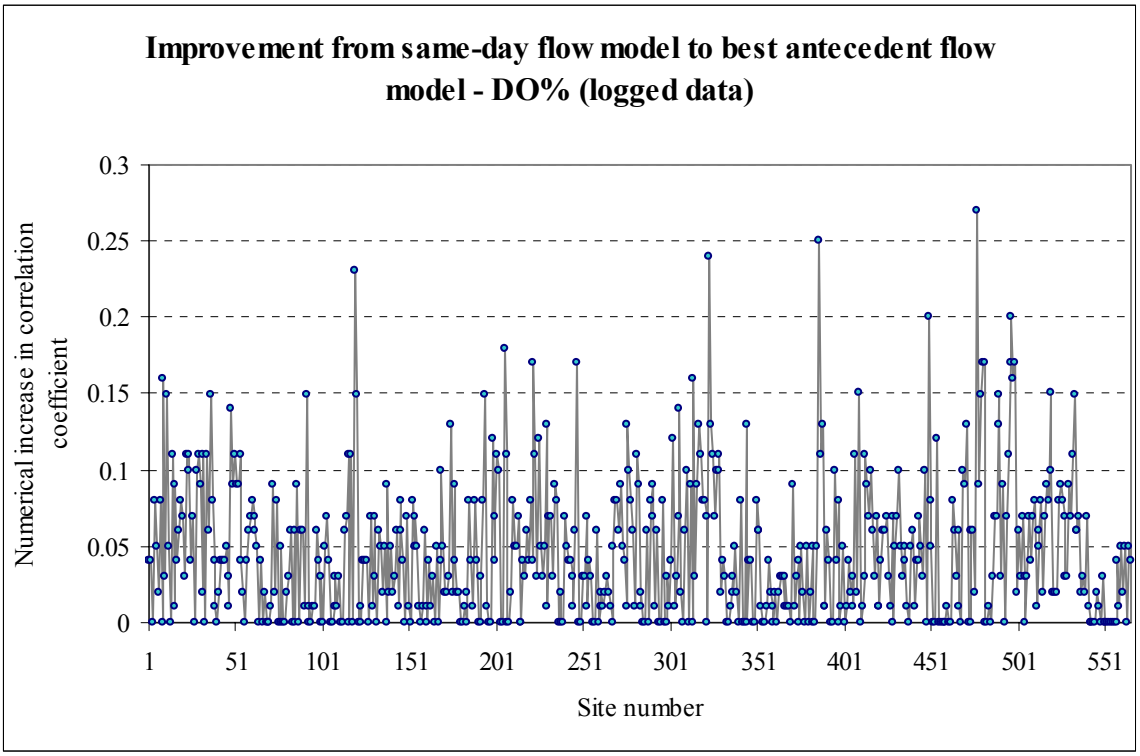


Figure 5.3 Improvements achieved in DO% models by using more general antecedent flow measures (logged data)

5.2 Quality v. flow relationships for BOD

5.2.1 Models using same-day flow

Figure 5.4 summarises the performance of the same-day flow models for BOD obtained using the unlogged data and using same-day flow as the explanatory variable. It follows exactly the same format as seen earlier in Figure 5.1. As before, the majority of sites (72%) fall within the

not-significant region. The position improves slightly for the logged data models (see Figure 5.5), with the proportion of sites in the not-significant region falling to 63%.

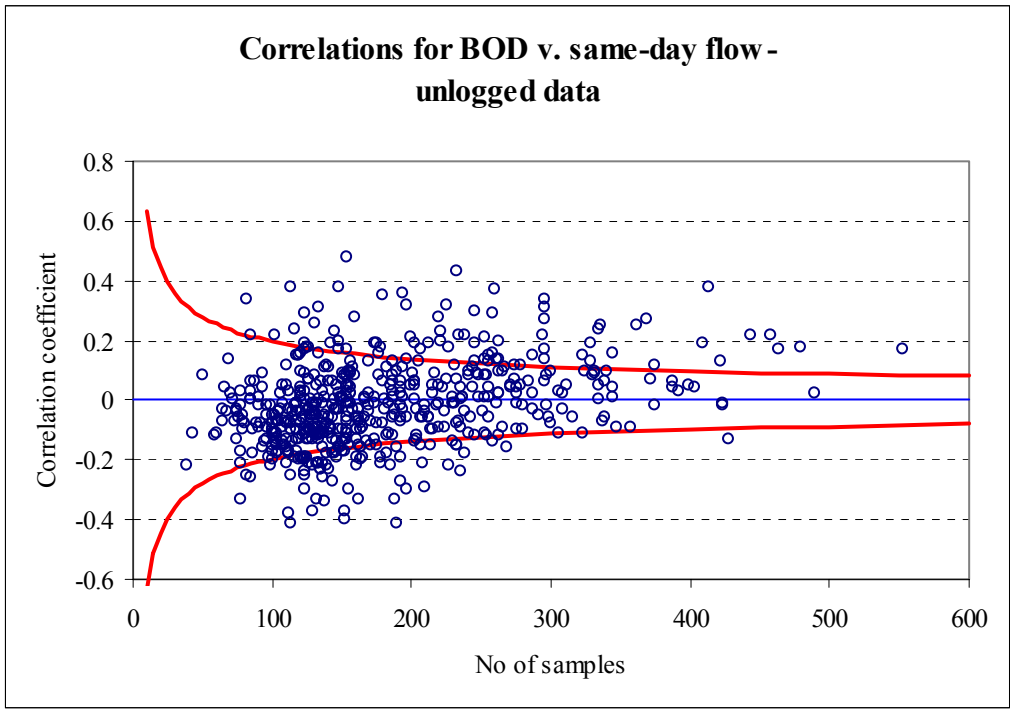


Figure 5.4 Performance of BOD models using same-day flow (unlogged data)

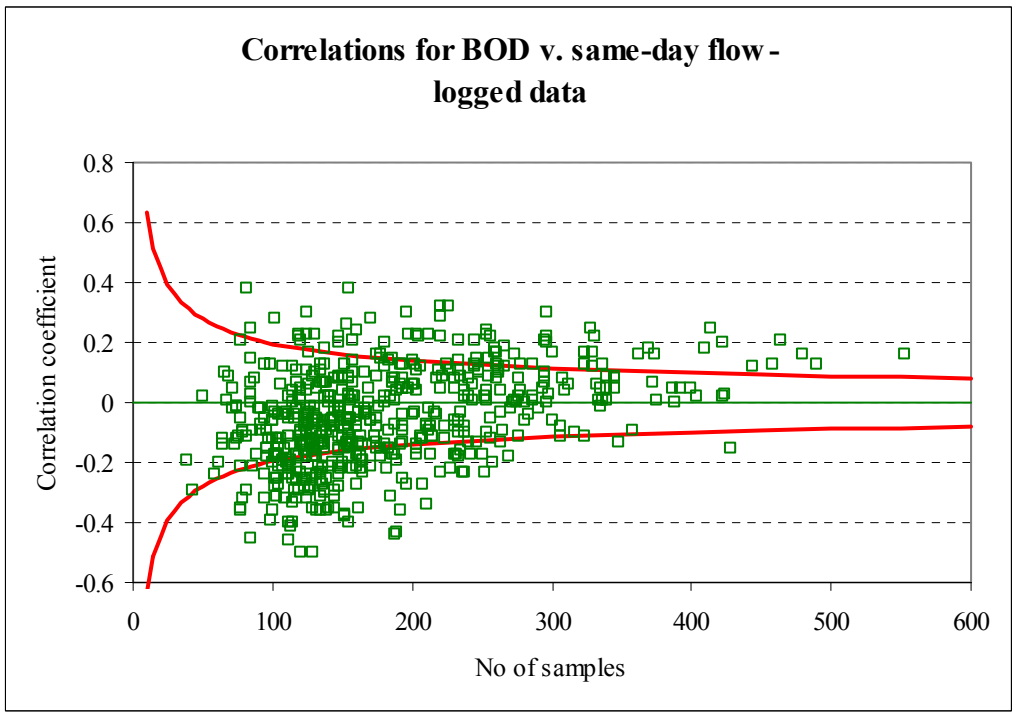


Figure 5.5 Performance of BOD models using same-day flow (logged data)

For those sites that do have a statistically significant R value, the form of the relationship is less clear-cut for BOD than it was with DO%. Although the relationship is predominantly negative - that is, increases in flow tend to be associated with decreases in BOD - the contrary effect is seen for about 1 site in 3. This clearly militates against there being an unequivocal improvement GQA class with an increase in mean flow - at least for those sites where BOD is the class-critical determinand.

5.2.2 Models using other antecedent flow measures

The effect of generalising the modelling to choose from a variety of possible antecedent flow measures is summarised in Figure 5.6 (for the logged data case). As with DO%, the improvement in performance over that achieved by the simple same-day flow model is, in almost all cases, extremely slight or non-existent.

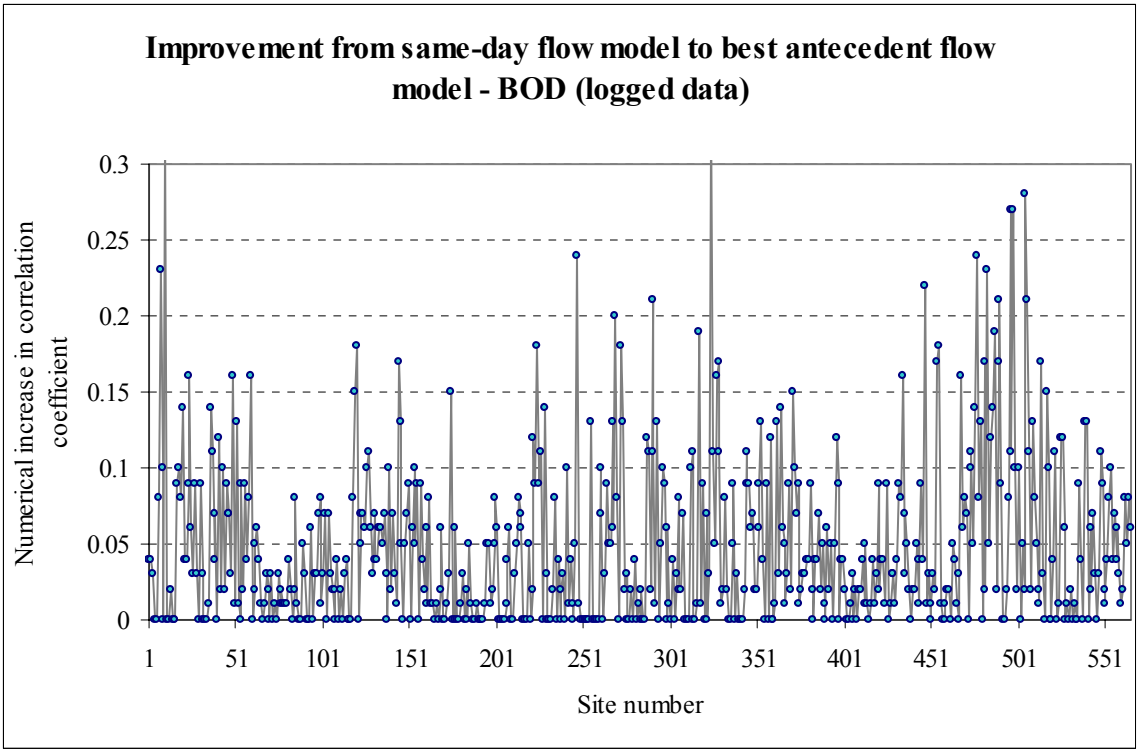


Figure 5.6 Improvements achieved in BOD models by using more general antecedent flow measures (logged data)

5.3 Quality v. flow relationships for ammonia

5.3.1 Models using same-day flow

Figure 5.7 summarises the performance of the same-day flow models for ammonia obtained using the unlogged data and with same-day flow as the explanatory variable. The picture is very similar to those shown by DO% and BOD, with 76% of sites falling within the not-significant region. Again the performance improves when the modelling is done on the logged data (see Figure 5.8): the proportion of sites in the not-significant region decreasing from 76% to 54%. As before, however, even for the statistically significant models the R values still remain below 0.6. It is particularly interesting to see that at 80% of those sites the relationship is positive - that is, increases in flow are associated with increases in ammonia. This means that, for sites where ammonia is the class-critical determinand, we would expect an increase in flow to have a potentially *harmful* effect on GQA class. (Presumably, the observed overall increase in ammonia concentrations with increasing flow is a consequence of enhanced leaching activity outweighing the tendency for increased flow to cause dilution of both effluent and river ammonia concentrations.)

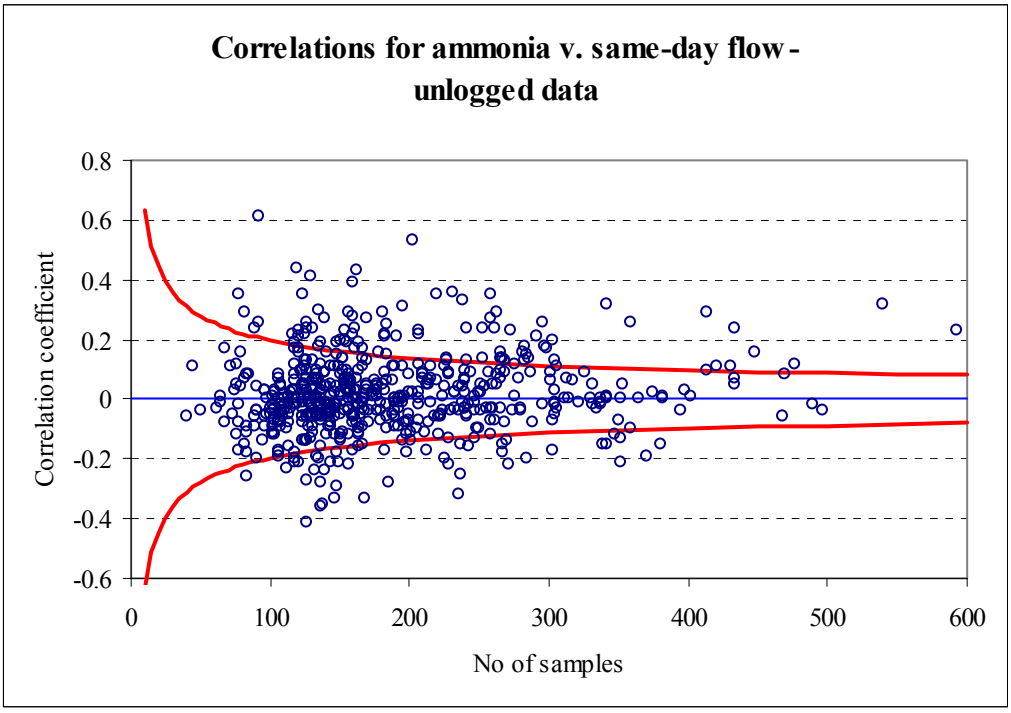


Figure 5.7 Performance of ammonia models using same-day flow (unlogged data)

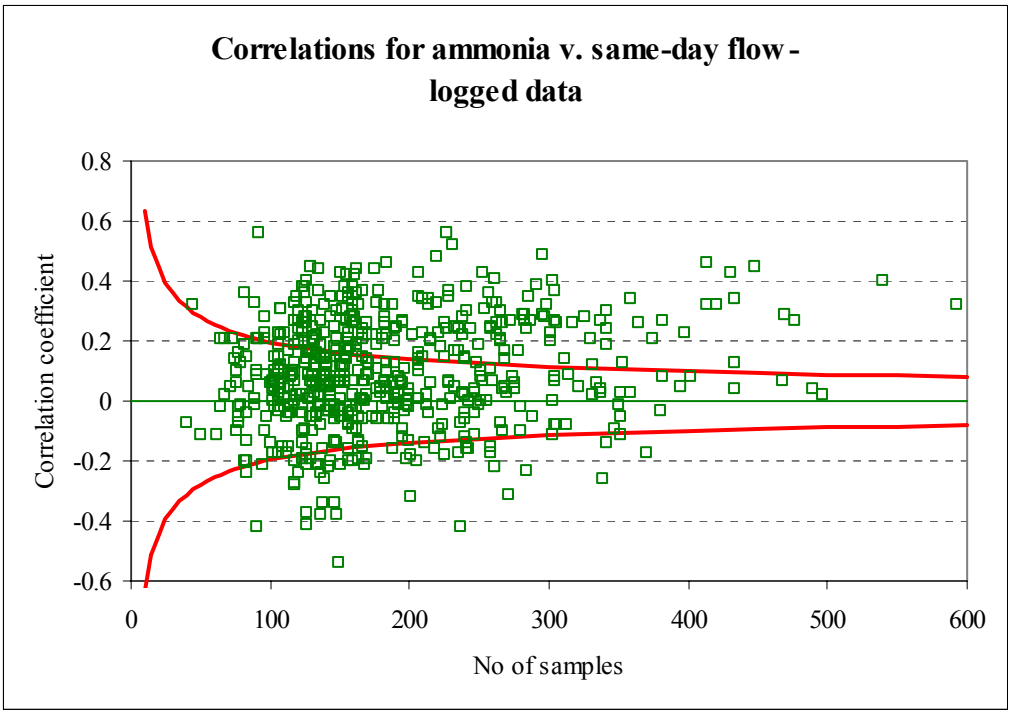


Figure 5.8 Performance of ammonia models using same-day flow (logged data)

5.3.2 Models using other antecedent flow measures

To complete the picture, Figure 5.9 shows the improvements achieved in moving from the simple same-day flow models to the more general range of antecedent flow measures. Once again the improvements are negligible.

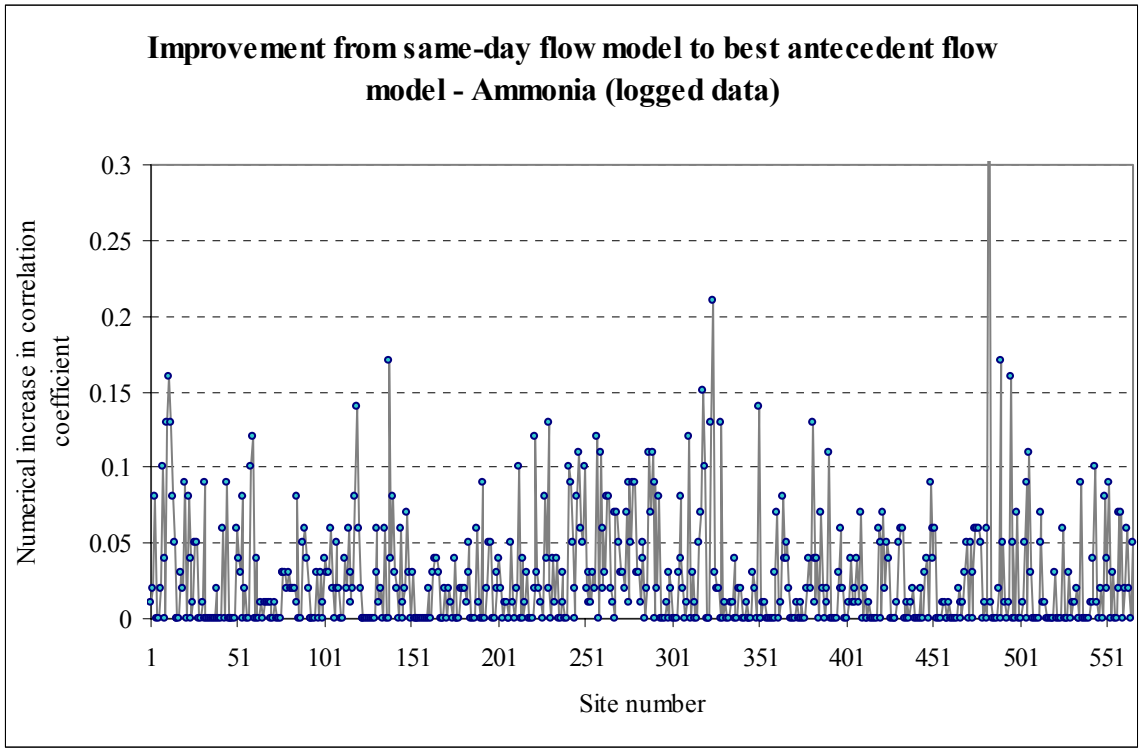


Figure 5.9 Improvements achieved in ammonia models by using more general antecedent flow measures (logged data)

5.4 Adequacy and stability of models

5.4.1 Adequacy of models

We testing the structural adequacy of each model by applying a cusum analysis to the *flow-sorted* residuals from the model, as described in Appendix A. For the same-day flow models (using logged data), the results are summarised in Figure 5.10.

The findings, which were consistent across all three determinands, provided good support for the assumed structure. For 500 or more of the 565 sites, no step changes were detected in the residuals. In other words, there was no evidence that the quality v. flow models fitted less well over some flow bands than others. This means that, whatever the shortcomings of the models may be in terms of their explaining power, the log-log formulation in the great majority of cases provides an adequate representation of the relationship over the whole flow range.

This conclusion was reinforced by the additional test that we built into CAFE whereby the lag-1 autocorrelation coefficient was calculated for the flow-sorted residuals from each of four 5-year blocks. As Table 5.1 demonstrates, only a handful of the 565 sites showed any evidence of a problem - a conclusion that was upheld for each of the three determinands.

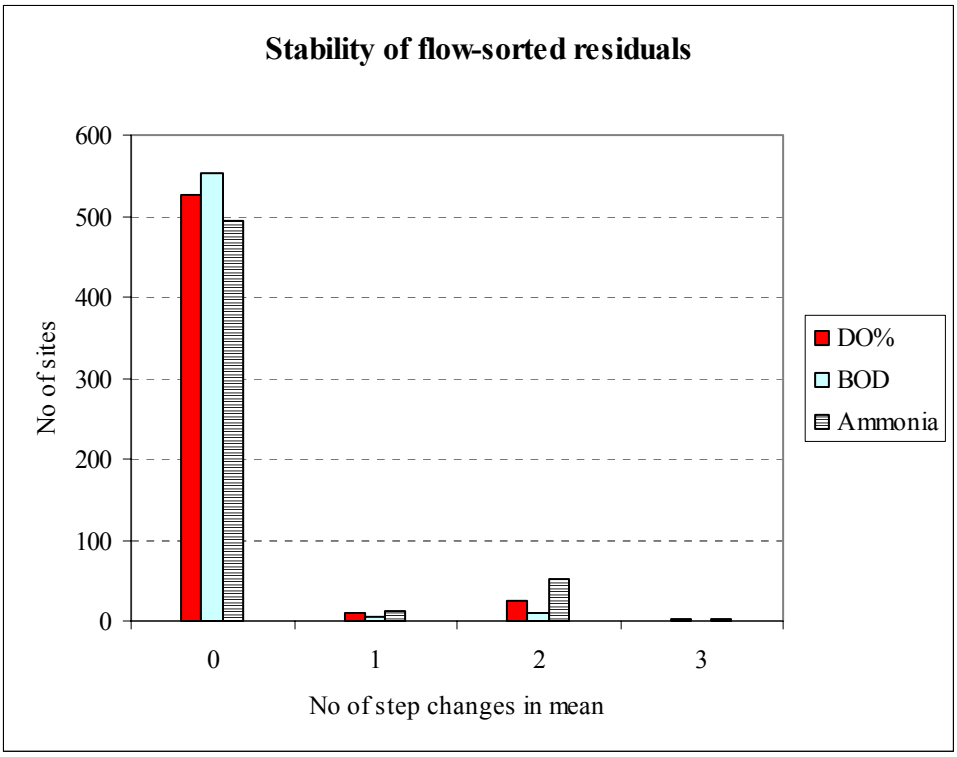


Figure 5.10 Structural adequacy of quality v. flow models

Table 5.1 Results of the N/4 autocorrelation test

No of 5-year blocks in which the flow-sorted residuals showed a stat. significant degree of autocorrelation	No of data sets		
	Amm.	BOD	DO%
0	526	532	539
1	35	32	25
2	4	1	1
3	0	0	0
4	0	0	0

(Note: see text for details)

5.4.2 Temporal stability

We tested the temporal stability of each model by applying a cusum analysis to the *date-sorted* residuals from the model, as described in detail in Appendix A. The purpose of the test is to see whether or not there are any time trends in the data *after allowing for the effect of flow*. If, on the one hand, the residuals are found to be randomly scattered through time, this indicates that the flow model successfully accounts for any time trends that may have been present in the original quality data. If, on the other hand, the residuals show one or more step changes through time, this provides evidence of temporal changes in quality that are *not* explained by variations in flow.

For the same-day flow models (using logged data), the results are summarised in Figure 5.11. The main message is that, for all three determinands, non-flow-related changes in mean quality do occur at the majority of sites. The most stable determinand in this regard is ammonia, with no changes detected at about 200 of the 565 sites, whilst BOD is the least stable, with one or more step changes found at all but about 120 sites.

At around 200-250 sites just one step change in the mean was detected, whilst at a further 100 sites two changes were detected (i.e. there were three different mean levels). The determinand showing the greatest change was DO%, with about 100 sites showing 3 or more step changes.

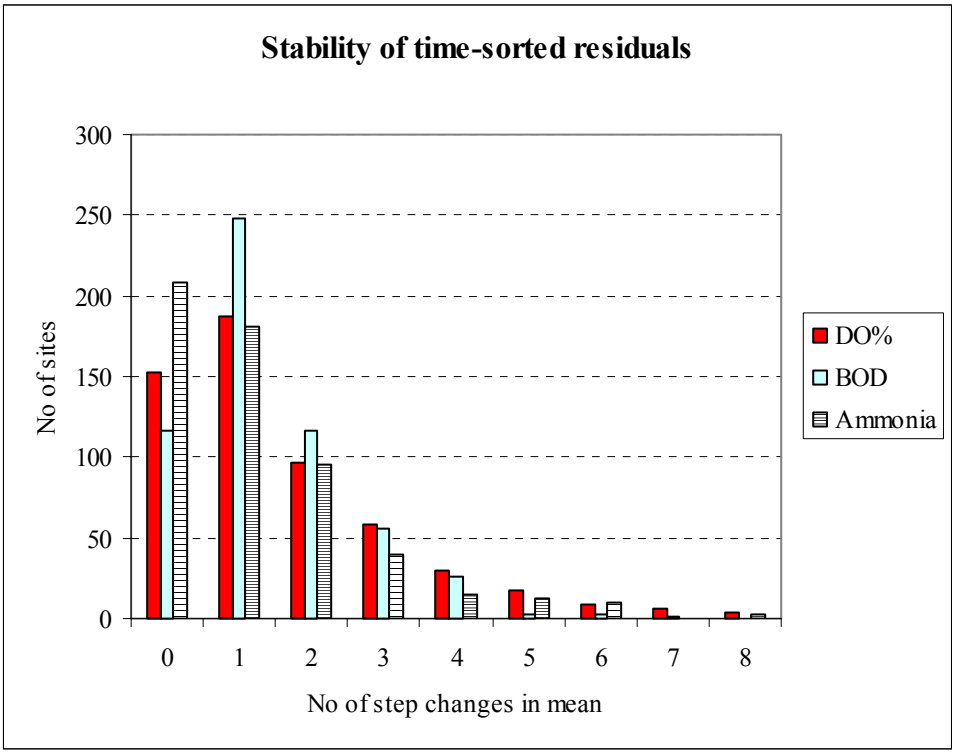


Figure 5.11 Temporal stability of quality v. flow models

Of course, as we are looking at time series spanning up to 20 years, it is hardly surprising that there are several different mean levels over the period. What is important for the present study, however, is that (as far as we can tell) these changes in quality are *unrelated to flow*.

5.5 Predictive capability of the identified models

We noted earlier in the chapter that models with low correlation coefficients are of little practical use for prediction. As a key feature of the results in Sections 5.1 to 5.3 has been the generally low R values obtained, the purpose of the present section is to provide a practical insight into the consequences of this.

In essence, the problem is that, for any given level of overall variability in quality and flow, the slope of the regression line is proportional to R. Thus the lower the value of R, the smaller is the predicted change in quality corresponding to a given change in mean flow.

To illustrate this, we have taken the 565 BOD v. flow models (logged data) and calculated for each the predicted change in 90%ile BOD associated with a *20% increase in mean flow*. Assuming for simplicity that each site was at the Class C/D border, we then converted these predicted changes in BOD *90%ile* into multiples of BOD *class-width*. (This was to make it easier subsequently to compare the BOD findings with those for the other two determinands.)

The results are summarised in Figure 5.12. First, the strong diagonal pattern confirms that the bigger the model's correlation coefficient, the bigger the predicted change in BOD tends to be. But more importantly the figure shows that, even for the most successful models, the predicted change in BOD GQA class is very slight. In particular, even for the models with the strongest negative correlation (around -0.5), the predicted improvement in BOD class associated with a 20% increase in mean flow is almost never more than a third of a class-width (i.e. -33%). In other words, if the association between BOD and flow truly behaved in the way implied by these models, even very marked changes in mean flow could rarely account for changes of more than one GQA class.

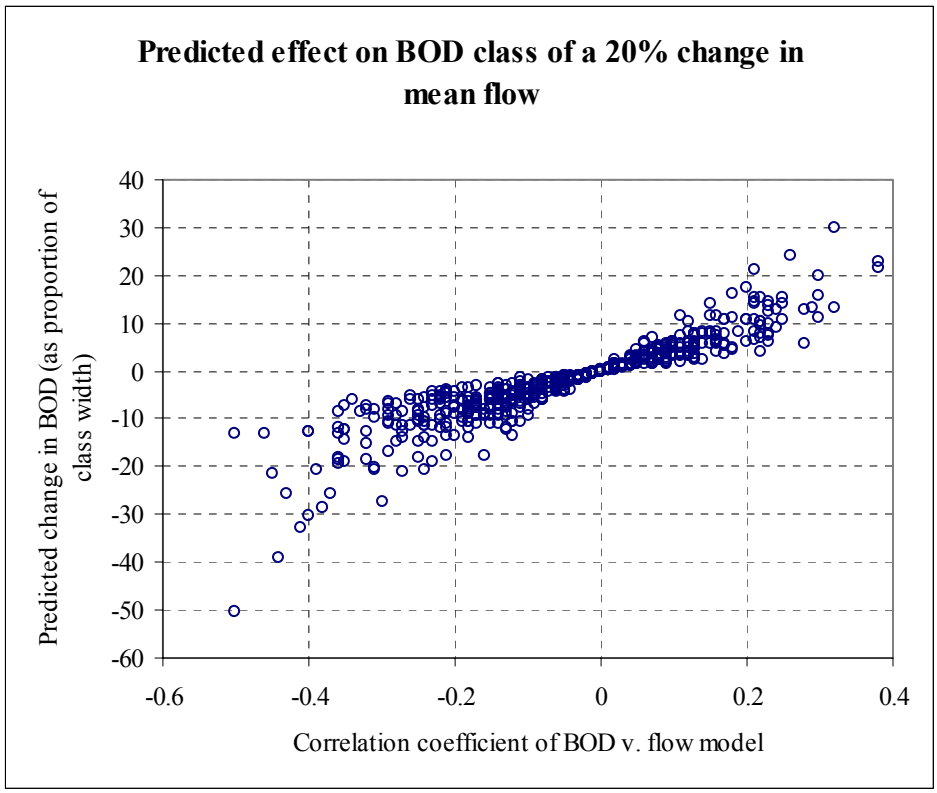


Figure 5.12 Example of predictions using the low-level BOD v. flow models

The assessment was repeated for ammonia and DO%; and the results for all three determinands are summarised in Figure 5.13. First look at the BOD results (the solid red towers). The four tallest towers, representing about 80% of sites, lie between -10% and +10%. This shows that the predicted effect of a 20% increase in mean flow on BOD quality amounts to less than a tenth of a class-width for the great majority of sites, whilst only about 1% of sites have a predicted improvement of more than a quarter of a class-width.

A broadly similar pattern is shown by DO%, with the predicted changes in quality falling within ±10% of a class-width for nearly 90% of sites. For ammonia there is a slight shift along the 'deterioration' axis. For about 70% of sites the effect of a 20% increase in mean flow is predicted to be less than 10% of a class-width, whilst for a further 25% of sites the predicted deterioration is 10-25% of a class-width.

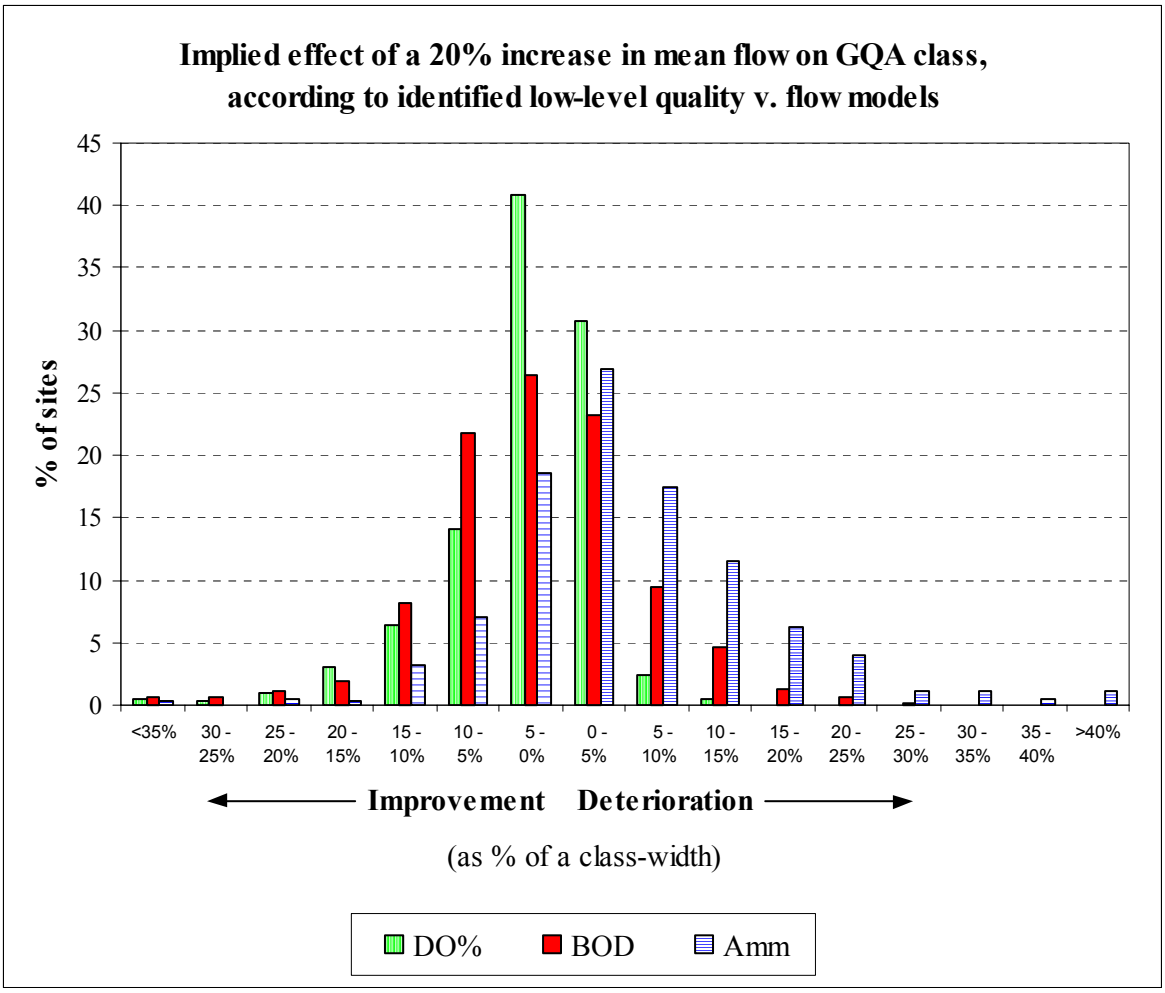


Figure 5.13 Example of predictions using the low-level quality v. flow models

These results demonstrate conclusively that the low-level quality v. flow models identified during this study are too weak to account for more than a very small proportion of observed class-changes. This conclusion is further strengthened when it is remembered that, at many sites, the predicted improvement would need to be shown for more than one determinand.

5.6 Quality v. AMP relationships

5.6.1 Results for the 50 'AMP' sites

Figure 5.14 shows the results of the cusum analysis for BOD at the 50 'AMP' sites. The green circles mark the dates at which a step-change improvement was detected; the red squares similarly mark the dates of step-change deteriorations. The solid black line for each site shows the 1-year date window within which any effect of the AMP scheme might be expected to be seen. We refer below to this window as 'the AMP scheme completion year'.

The numbers of improvements and deteriorations seen in Figure 5.14 are summarised in the top half of Table 5.2. This shows that there were 58 improvements and 19 deteriorations over the 10 years falling outside the AMP scheme completion year. Other things being equal, therefore, the expected number of quality improvements in a 1-year period *without* AMP scheme completions is 5.8. In contrast, we actually see 10 improvements in the AMP scheme completion year. However, this increase is not statistically significant: a calculation using the Poisson distribution (which governs the behaviour of independent 'rare' events) shows that the chance of getting 10 or more events, given a mean of 5.8, is about 7%. So the evidence of a greater number of improvements during AMP scheme completion years is no more than suggestive.

A similar conclusion is reached when we look at the number of deteriorations (i.e. zero) in AMP scheme completion years. Although this is lower than the expected number (1.9), it is not unusually so: the chance of getting zero is about 15%.

Note, incidentally, the striking number of improvements in mean seen in 1994-95. One interesting explanation put forward for this is that it is the consequence of AMP1 projects having been finished off before the deadline.

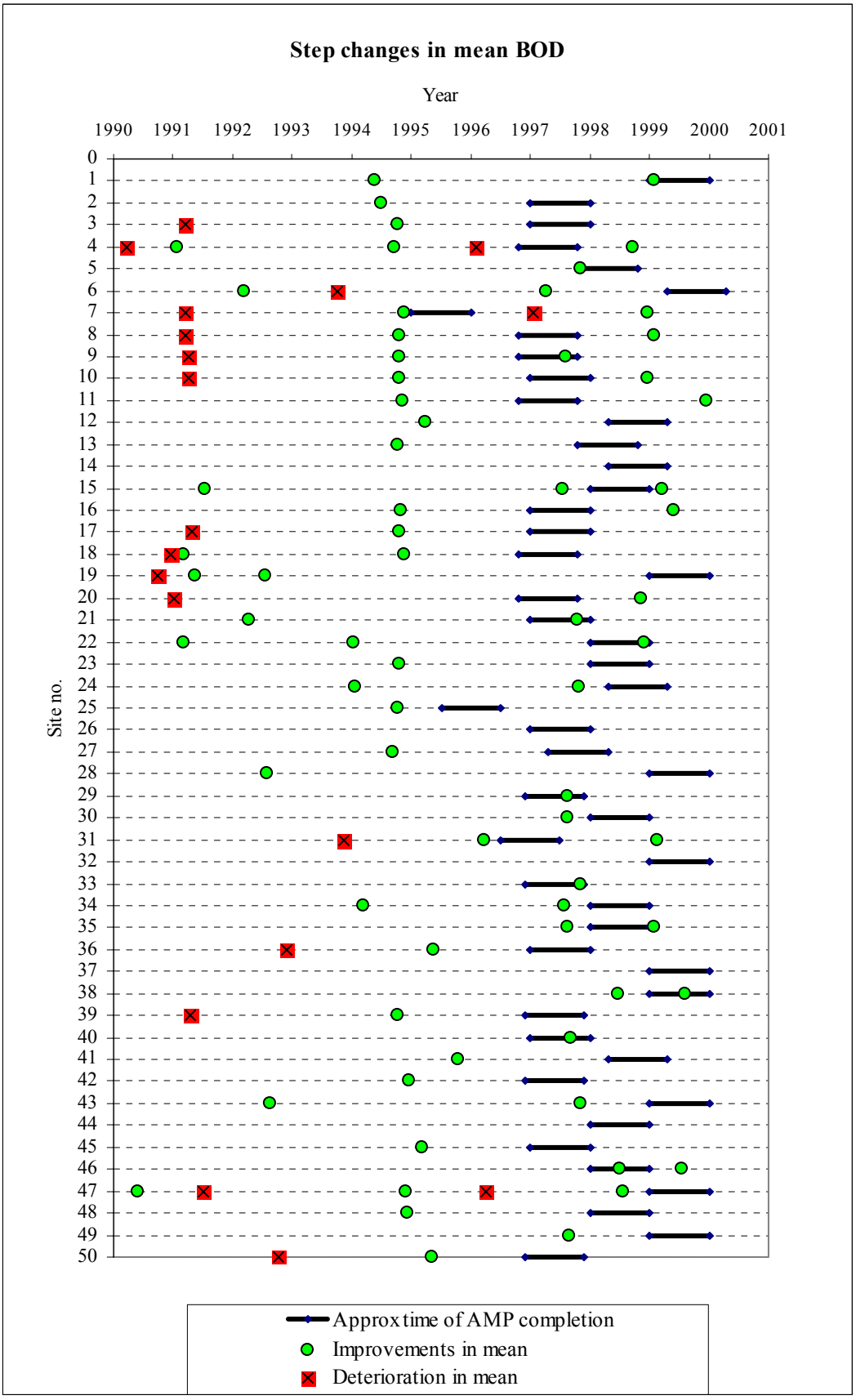


Figure 5.14 Time trends in BOD for the 50 'AMP' sites

Table 5.2 Summary of step changes in quality detected at the 50 'AMP' sites and the 50 'Control' sites

	DO%		BOD		Ammonia	
	Act.	*Exp.*	Act.	*Exp.*	Act.	*Exp.*
50 AMP sites						
10 yrs outside AMP window						
No of improvements	37		58		56	
No of deteriorations	19		19		32	
1 yr inside AMP window						
No of improvements	1	*3.7*	10	*5.8*	8	*5.6*
No of deteriorations	0	*1.9*	0	*1.9*	5	*3.2*
50 Control sites						
10 yrs outside Control window						
No of improvements	42		65		46	
No of deteriorations	24		25		11	
1 yr inside Control window						
No of improvements	6	*4.2*	9	*6.5*	6	*4.6*
No of deteriorations	2	*2.4*	0	*2.5*	1	*1.1*

Table 5.2 shows that, of the three determinands, BOD was in fact the one for which the apparent improvement during the 1-year AMP window was most marked. For ammonia the number of step-change improvements was 8 against an expected number of 5.6, whilst for DO% the number of improvements (1) was actually less than the expected number of 3.7. Overall, therefore, we can conclude that that the AMP schemes had no detectable effect over these 50 sites. (Of course, there may be reasons why the analysis failed to detect an effect of AMP improvements. For example, the date on which the new plant was commissioned might not be known well enough; the improvement might have been gradual rather than a step change; the investment might have been for screening or storm tanks; the investment might have been to protect against potential deterioration, or there might have been an improvement which the monitoring programme was unable to detect.)

5.6.2 Results for the 50 'Control' sites

Figure 5.15 similarly shows the results of the cusum analysis for BOD at the 50 'Control' sites. To mimic the AMP scheme completion year windows, we arbitrarily alternated between 1997 and 1998 for the 'Control windows'.

The numbers of improvements and deteriorations seen in the figure are summarised in the bottom half of Table 5.2. There were 65 improvements and 25 deteriorations over the 10 years falling outside the 1-year Control window - which means that the expected numbers of improvements and deteriorations in any 1-year period are 6.5 and 2.5. The actual numbers in

the 1-year Control window were 9 and 0. Thus, over the 1-year Control period there was a relative increase in the frequency of *improvements* (9 versus 6.5), and a relative decrease in the frequency of *deteriorations* (0 versus 2.5) - but neither effect is statistically significant. These results are in striking agreement with those for the 50 AMP sites. This suggests that the modest improvements seen for the AMP sites during the AMP scheme completion year could plausibly be attributed to a general improvement in river quality in the late 1990s. This reinforces the conclusion of the previous section that no AMP effect could be detected from the GQA data.

There is one other point of interest. This concerns the degree of comparability between the 50 AMP sites and the 50 Control sites during the 10 years outside the AMP scheme completion year (or the one-year Control window). Both for DO% and BOD, there is a broad level of agreement between the two groups of 50 sites in their relative proportions of improvements to deteriorations. For DO% the proportions are 37/19 (= 1.95) for the AMP sites, and 42/24 (= 1.75) for the Control sites. That is, for both types of site there were nearly twice as many improvements as deteriorations. For BOD the proportions are 58/19 (= 3.05) and 65/25 (= 2.60). For ammonia, however, the proportions are very different. At the Control sites there were over four times as many improvements as deteriorations (46/11 = 4.2), whereas at the AMP sites there were fewer than double (56/32 = 1.8). It should be noted that this finding has no bearing on the statistical assessment of AMP improvements described above. However, the increased propensity for ammonia deteriorations at the AMP sites (in the years prior to completion of the AMP schemes) does perhaps help to explain why those sites had been selected for AMP schemes in the first place.

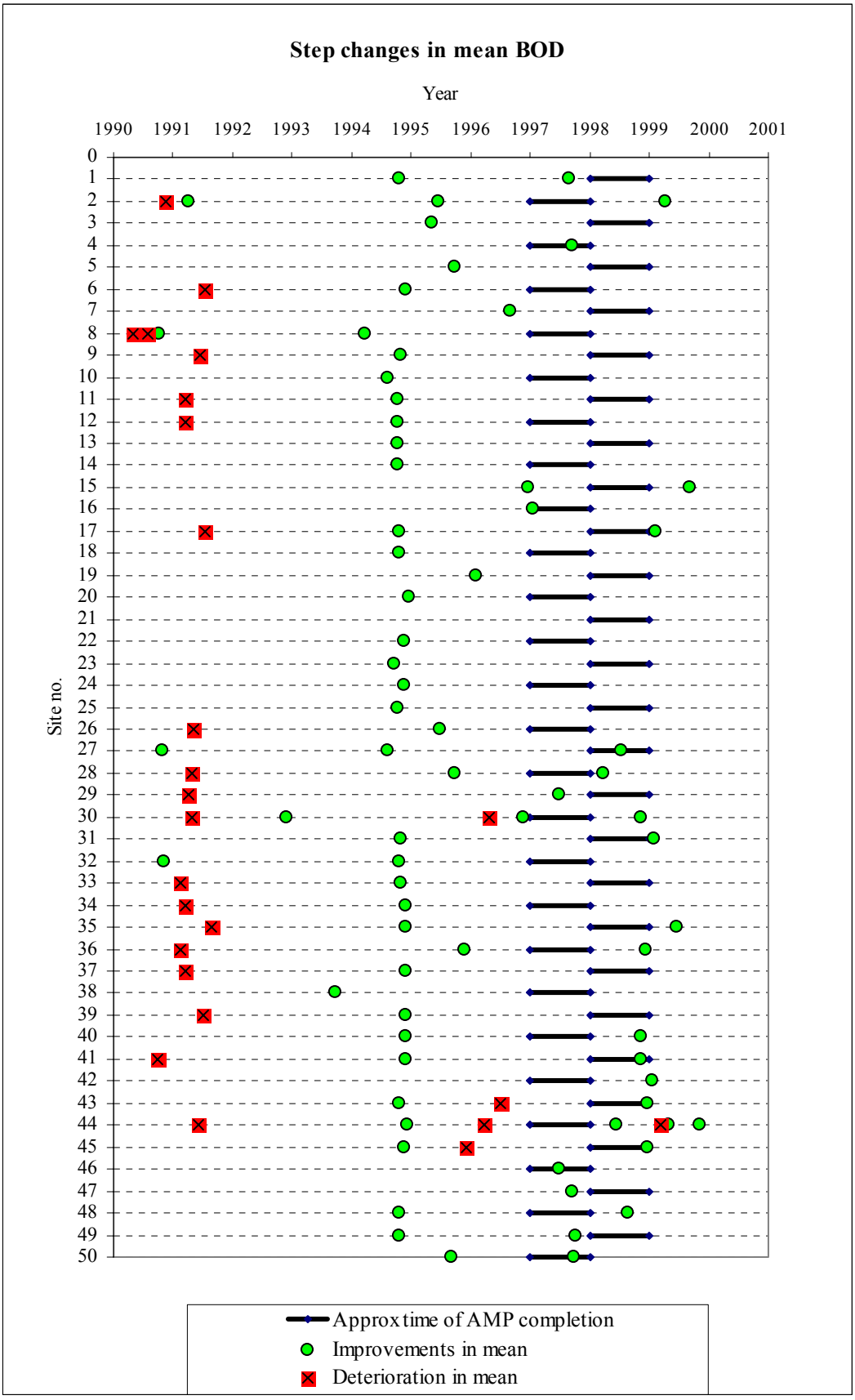


Figure 5.15 Time trends in BOD for the 50 'Control' sites

5.7 Results for Regions other than Thames

A limited amount of low-level modelling was undertaken using data from three other Regions. Useful data was available for eight sites - two from Anglian, three from Southern and three from Midlands - and we submitted the DO%, BOD and ammonia data for each of these sites to the CAFE modelling procedures described earlier.

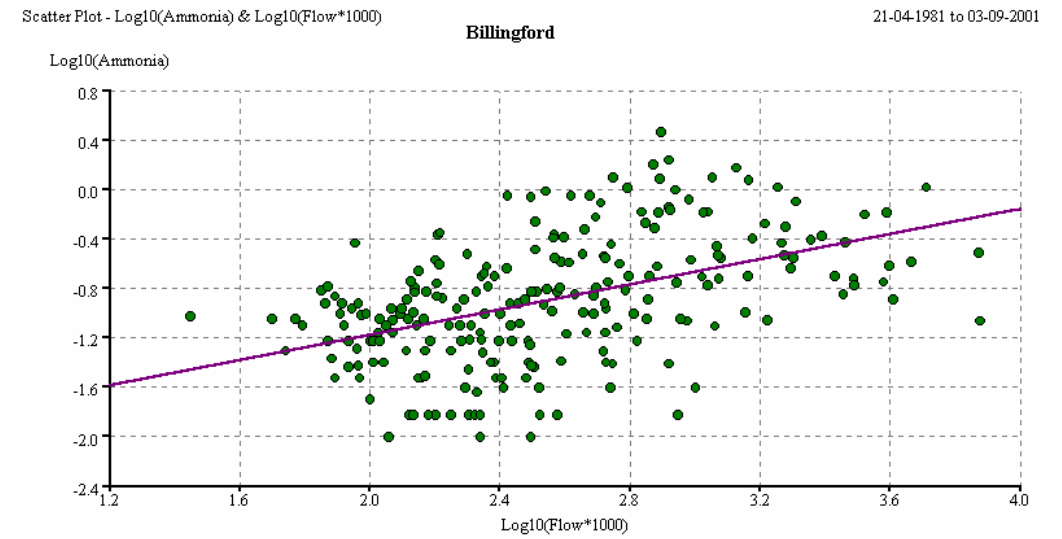
The model results for the logged data are summarised in Table 5.3. (Note that, as with the Thames analyses, the models for the logged data generally produced higher R values than those for the unlogged data.) Values not statistically significant ($P<0.05$) are greyed out. Scatter plots illustrating the three largest ammonia effects are shown in Figure 5.16. Three of the better models for DO% and BOD are shown in Figure 5.17.

Table 5.3 Summary of low-level results for Regions other than Thames

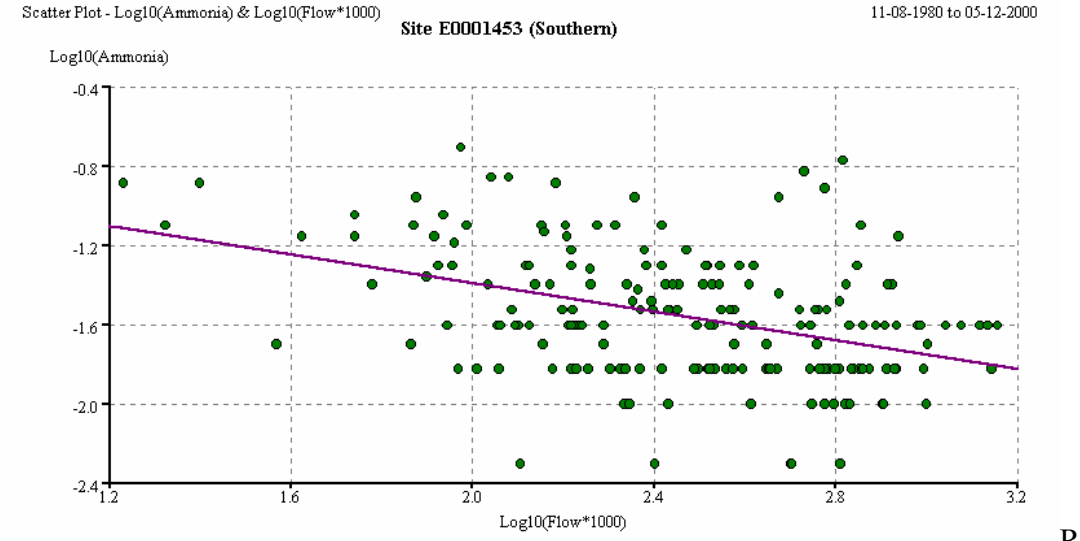
Region	GQA site	Correlation coefficients (logged data)		
		DO%	BOD	Ammonia
Anglian	Dernford	0.25	0.28	0.30
Anglian	Billingford Bridge	0.16	0.16	0.47
Midlands	Clifton on R. Avon	0.18	0.18	0.25
Midlands	Stanbridge Farm	0.27	0.31	0.40
Midlands	Water Lane	-0.02	-0.08	0.38
Southern	E0001453	0.41	-0.12	-0.41
Southern	E0001456	0.11	-0.10	-0.24
Southern	E0001462	0.24	0.05	-0.27

Generally, the R values seen here show a very similar spread to those obtained for the Thames sites. Only five of the 24 models have R values numerically greater than 0.35. The model with the highest R value (0.47) is that for ammonia at Billingford. This model accounts for 22% (= 0.47^2) of the variance in log(ammonia). Thus 78% of the variance is *un*accounted for by flow, and so the standard deviation of the scatter around the model is $\sqrt{0.78} = 0.88$ times the original standard deviation. As the discussion of Section 5.5 has indicated, a reduction in residual scatter as small as this means that the model is of little use for predictive purposes.

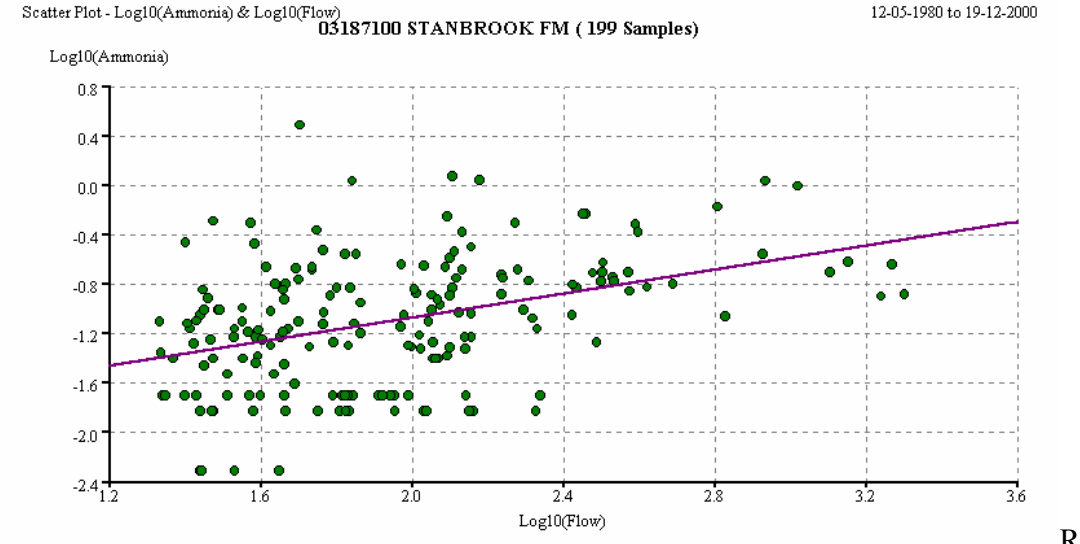
One interesting feature common to all sites is that the ammonia v. flow association is stronger than those for either DO% or BOD. There is also tentative evidence of a difference between Regions (at least in the way that the sites were selected), with increases in flow being associated with *increases* in ammonia at the Anglian and Midlands sites, but with *decreases* in ammonia at the Southern sites.



R=0.47

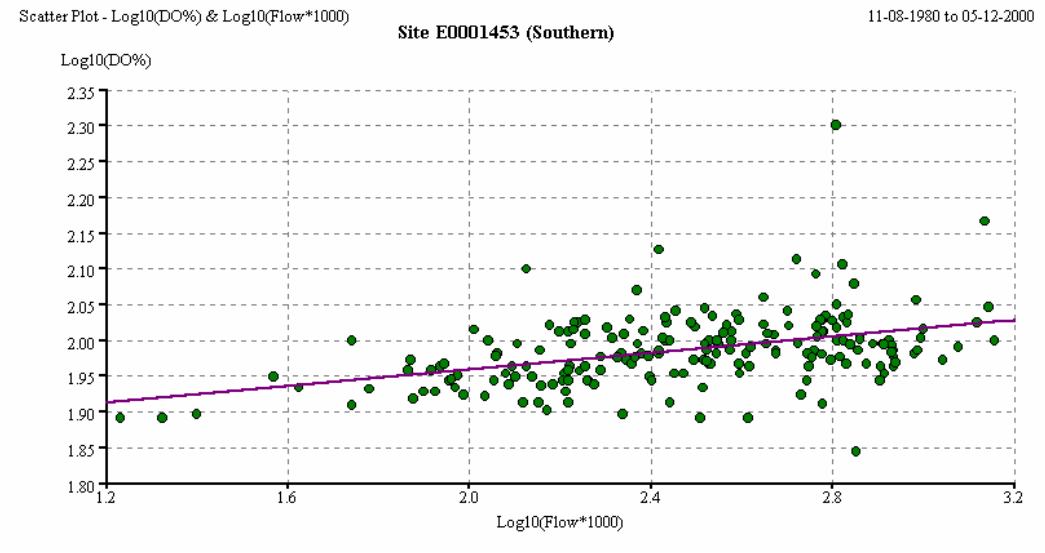


R = -0.41

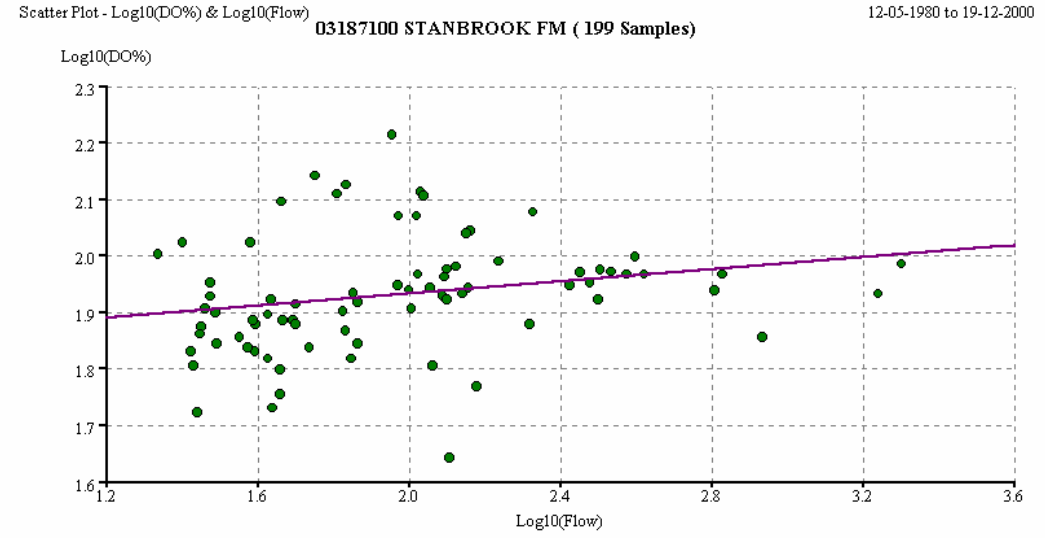


R=0.40

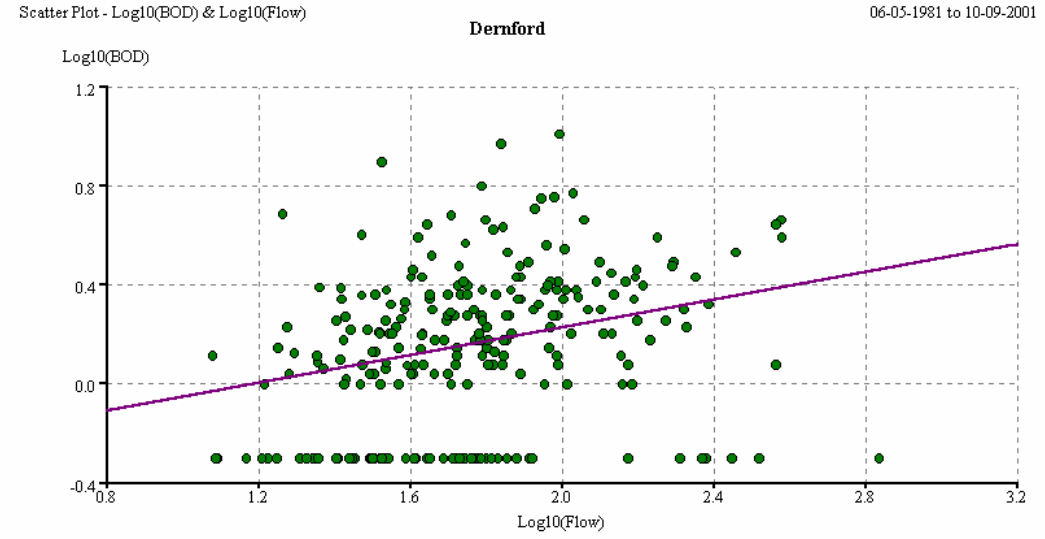
Figure 5.16 Examples of ammonia v. flow plots (logged data)



R=0.41



R=0.27



R=0.28

Figure 5.17 Examples of DO% v. flow and BOD v. flow plots (logged data)

6. HIGH-LEVEL RESULTS

6.1 Summary of GQA class data

We ran GCSE on all 565 GQA data sets. By way of illustration, a portion of the summary output is shown below.

Table 6.1 Illustration of the summary output from GCSE

```
PWRR0002,  3, R3Y from 80-82:, , , , ,F,E,E,D,D,D,D,E,E,E,E,E,C,C,D
PWRR0003,  1, R3Y from 80-82:,A,A,A,B, , , , ,A,A,A,A,A,A,A,A,A,A,A
PWRR0004,  2, R3Y from 80-82:, , , , , , , , ,B,C,B,B,B,B,B,B,B,B,A
PWRR0005,  1, R3Y from 80-82:,B,B,B,B,B,B,A,A,B,A,A,A,A,A,B,B,B,A,A
PWRR0006,  2, R3Y from 80-82:,E,E,E, , , , , ,D,D,D,D,D,C,C,C,D,D,D
PWRR0007,  2, R3Y from 80-82:,E,E,E, ,E,E,E,E,E,E,D,D,D,D,D,C,C,C,C
PWRR0008,  0, R3Y from 80-82:,A,A,A,A,A,A,A,A,A,A,A,A,A,A,A,A,A,A,A
PWRR0009,  1, R3Y from 80-82:, , , , , , ,A,A,B,A,A,A,A,A,A,A,A,A,A
PWRR0010,  1, R3Y from 80-82:, , , , , , , , ,B,B,B,B,A,A,A,A,A,A,A
PWRR0019,  2, R3Y from 80-82:,B,B,B, , ,A,A,A,B,C,C,A,A,A,A,A,A,A,A
PWRR0020,  0, R3Y from 80-82:, ,A,A,A,A, , , , ,A,A,A,A,A,A,A,A,A,A
PWRR0021,  1, R3Y from 80-82:,A,A,A,A,A,A,A,A,B,B,B,B,B,B,B,B,B,A
PWRR0023,  0, R3Y from 80-82:, ,A,A,A,A,A,A,A,A,A,A,A,A,A,A,A,A,A,A
PWRR0025,  1, R3Y from 80-82:, , , , , , , , ,B,B,B,A,A,A,A,A,A,A,A
PWRR0026,  1, R3Y from 80-82:,B,A,A, , ,A,A,A,B,B,B,A,A,A,A,B,B,A,A
PWRR0027,  0, R3Y from 80-82:, , , , , , , , ,A,A,A,A,A,A,A,A,A,A,A
PWRR0029,  1, R3Y from 80-82:, , , , , ,A,A,A,A,A,A,A,A,B,A,A,A,A,A,A
PWRR0032,  0, R3Y from 80-82:, , , , , , , , ,A,A,A,A,A,A,A,A,A,A,A
PWRR0035,  0, R3Y from 80-82:, , , , , , , , , , , , , ,A,A,A,A,A,A,A
PWRR0037,  2, R3Y from 80-82:, , , , , , , , , ,A,B,B,B,C,C,B,A,A,A
PWRR0039,  2, R3Y from 80-82:, , , , , , , , ,B,B,C,C,C,C,B,A,A,A,A
PUTR0196,  2, R3Y from 80-82:, , , , , , , , ,D,D,D,D,C,C,D,E,E,D,C
PUTR0009,  2, R3Y from 80-82:, , , ,C,D,D,D,D,D,D,C,C,D,D,E,D,D,C
PUTR0043,  1, R3Y from 80-82:,E,E,E,E, ,E,E,E,D,D,D,E,D,D,D,D,E,D,D
PUTR0057,  2, R3Y from 80-82:,E,E,E,E,E,E,E,D,D,D,D,D,D,C,D,E,E,D,C
PUTR0086,  2, R3Y from 80-82:, , , , , , ,C,C,C,C,C,C,B,A,B,C,C,B,B
PUTR0091,  1, R3Y from 80-82:, , , , , , , , , ,C,C,B,B,B,B,C,C,C,C,B
PUTR0093,  1, R3Y from 80-82:,C,C,C,C,C,C,B,B,C,C,C,B,B,B,C,C,C,B,B
PUTR0104,  3, R3Y from 80-82:, , , , , , , , , ,E,E,C,C,E,E,E,E,D,B
PUTR0108,  2, R3Y from 80-82:, , , , , , ,B,A,B,B,B,B,B,B,C,C,C,B
PUTR0013,  1, R3Y from 80-82:,B,A,A,A,B,B,B,A,A,A,A,A,A,A,B,B,A,A,A
PUTR0014,  1, R3Y from 80-82:,A,A,A,B, , ,A,A,A,A,A,A,A,A,A,A,A,A,A
PUTR0067,  1, R3Y from 80-82:, ,B,B,A,B,B,B,B,B,B,B,B,B,A,B,B,B,B,A
PUTR0069,  2, R3Y from 80-82:, , , , ,E,E,E,E,E,E,E,D,C,C,D,D,D,C,C
PUTR0070,  2, R3Y from 80-82:, , , , ,D,D,D,D,D,D,D,D,D,D,D,C,C,B
PUTR0072,  2, R3Y from 80-82:, , , , ,E,E,E,E,E,E,E,E,C,D,D,D,D,D,C
PUTR0127,  2, R3Y from 80-82:, , , , , , , , ,E,D,D,C,C,C,C,C,C,C,C
etc
etc
```

One of the supplementary statistics that GCSE calculates is the 'spread' of GQA classes seen at a site (converting A, B, C… into 1, 2, 3…). For example, a site that was always in class B would have a spread of 1; a site with classes ranging from B and E would have a spread of 4. The aim of doing this is to indicate the potential amount of variability in GQA class through

time - and hence the scope for identifying factors that influence GQA class (e.g. changes in the flow fingerprint). Clearly, a site that had a spread of only one over the past 10-15 years would hold out no hope for showing an effect of flow (or anything else) on water quality.

The results, summarised in Figure 6.1, show that there is an adequate degree of variability at the majority of sites. GQA class remained absolutely constant at only 27 of the 565 sites, whereas it changed by 3 classes or more at well over half of the sites.

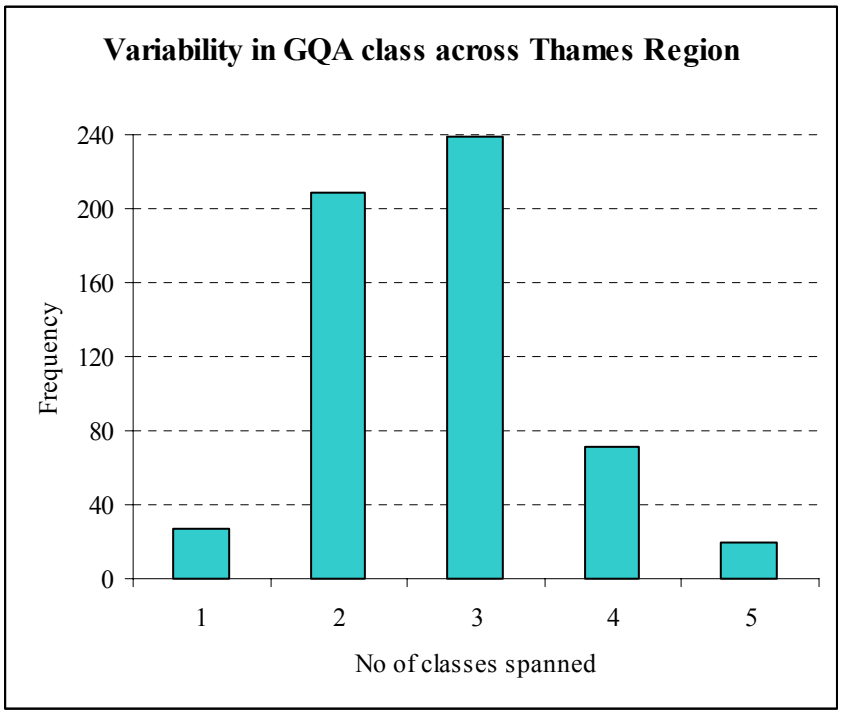


Figure 6.1 Variability in GQA class over time - index 1

One limitation of the simple index described above is that it makes no distinction between **A, A, A, B, A, A** and **A, A, A, B, B, B**, whereas in practice we would have a better chance of detecting a quality v. flow effect in the latter case (because there is a more even spread of data between the two levels of quality). To reflect distinctions of this sort, we developed a slightly more elaborate measure which worked as follows:

1. Look at GQA class for the six non-overlapping periods 82-84, 85-87, etc, 97-99. Suppose there are n values (ideally n = 6, but there may be missing data in the early years).

2. Determine the mode class (i.e. the most commonly occurring class), and count the number of values, m, equal to the mode.

3. Calculate the class variability index as the *percentage of class values that are different from the modal class* - that is, $100\times(1 - n/m)$.

So, for example, suppose a site had classes **B, (n/a), A, B, B, C**. Three of the five class values are equal to the modal class, B; thus 2 of the 5 class values are different from the mode, and so the class variability index is 40%.

Of the 565 sites, 63 had GQA data for fewer than four of the required three-year periods, and a further six had insufficient flow data. For the other 496 sites, Table 6.2 shows a breakdown according to their class variability index.

Table 6.2 Variability in GQA class over time - index 2

Class variability index	No of sites
0%	66
17 - 25%	145
33 - 50%	240
60 - 75%	45
Total	**496**
Too few GQA values	63
Insuff. flow data	6
Grand total	**565**

6.2 Trends in GQA class

The results of the GQA trend analysis are summarised in Table 6.3. Part (a) of the table shows the outcome of fitting a different mean for each 3-year period; part (b) then shows the results of fitting a linear trend. With both models the statistical significance of the results is extremely high ($P<0.001$) - both when viewing the full set of 496 sites, and when looking at the three separate groups of sites defined by having low, medium and high GQA variability. Note that the biggest across-the board changes would be thought most likely to occur within the sub-group of GQA sites that *individually* showed the biggest changes through time - and it is interesting to see that this is what does actually happen.

Overall, therefore, these findings provide formal statistical confirmation of the widespread improvement that has occurred in river quality in Thames Region over the past 20 years - though they do not of course provide any clue as to what factors might have driven this improvement.

Table 6.3 Summary of the time trend analysis of GQA results

(a) ANoVA results

Sites in analysis	F-value (5,n) df	P-value	Mean class (A=1, B=2, etc) over each 3-year period					
			82-84	85-87	88-90	91-93	94-96	97-99
All sites	62.2	<0.001	3.4	3.3	3.3	3.1	3.0	2.9
Sites with low GQA variability	10.7	<0.001	3.3	3.1	3.2	3.0	1.9	2.9
Sites with medium GQA variability	27.0	<0.001	3.3	3.2	3.3	3.1	3.0	2.8
Sites with high GQA variability	41.9	<0.001	4.4	4.3	4.3	3.4	1.9	1.7

(b) Linear regression results

Sites in analysis	F-value (1,n) df	P-value	Estimated slope per 3 years	Mean overall change of class in 18 yrs
All sites	189.4	<0.001	-0.12	-0.7
Sites with low GQA variability	22.5	<0.001	-0.07	-0.4
Sites with medium GQA variability	83.8	<0.001	-0.12	-0.7
Sites with high GQA variability	128.7	<0.001	-0.41	-2.4

6.3 Summary of 'Flow Fingerprint' measures

6.3.1 Detailed results

The summary output from FFION consists of an array of dimensions 25 × 4 × 116, namely 25 FFC statistics × 4 five-year periods × 116 flow sites. An example of the full output from FFION for a site is shown in Appendix C. The complete summary output file for all 116 sites has been provided as an Excel file.

6.3.2 Variation between sites

The first question addressed by the ANoVA related to the strength of site-to-site variation. The conclusions are summarised in Table 6.4. For each row (i.e. FFC statistic), the table shows the mean over all 116 sites, and the F-value for the Site effect. The table is sorted in decreasing order of the F-value column, so that the FFCs showing the biggest between-site differences are at the top of the table. All 25 effects are statistically significant at the P<0.001 level.

Table 6.4 Summary of the between-site ANoVA results for the FFCs

No	Description	Mean	F-value
2	Log of mean flow	-0.147	560.2
25	Log of summer mean flow	0.374	272.3
8	Lag-1 autocorr. coefficient	0.793	94.1
9	Lag-15 autocorr. coefficient	0.459	70.8
3	Relative st.deviation	1.131	62.5
10	Lag-30 autocorr. coefficient	0.353	59.3
15	Lag-1 autocorr. coeff. for ratio	-0.0060	22.7
5	5%ile:50%ile ratio	0.377	22.2
23	95%ile no of days in downward run	9.900	20.4
12	Mean F(i)/F(i-1) ratio	1.131	17.6
14	75%ile of ratio	1.058	14.9
19	95%ile no of days in upward run	4.875	13.7
18	75%ile no of days in upward run	2.274	13.4
22	75%ile no of days in downward run	4.317	13.3
4	Mean:50%ile ratio	1.786	12.5
6	95%ile:50%ile ratio	5.990	12.4
16	No of upward runs	340	11.4
20	No of downward runs	340	11.4
21	50%ile no of days in downward run	2.326	6.89
24	Rel.st.dev. of annual mean flow	0.276	6.20
17	50%ile no of days in upward run	1.424	5.34
13	Relative st.deviation of ratio	0.733	5.00
11	No of valid consec.day ratios	1708	3.14
1	No of valid daily flows	1712	3.07
7	95%ile:5%ile ratio	40.92	3.01

The statistics with overwhelmingly the biggest differences between sites (relative to their variation from one five-year period to another) are **log(mean flow)**, followed by **log(mean summer flow)**. This is just what we would expect, as these are the two FFC measures that are not scale-free: they are both directly related to the size of the river.

However, all other FFCs are defined as various forms of ratios or counts, and so are independent of the scale of the river. Consequently the ANoVA results for these are intrinsically more interesting. For example, the next four statistics are the lag-1, lag-15 and lag-30 autocorrelation coefficients, together with the relative standard deviation. This tells us that both the degree of persistence and the relative variability of flow are statistics that can vary substantially from one river to another.

There is then a substantial drop from these first six to the next cluster of entries in the list. However, even for the FFC showing the weakest evidence of variation between sites - the 95%ile:5%ile ratio - we reiterate the point that the effect is still statistically significant at the $P<0.001$ level.

6.4 Trends in flow

The second and more important question addressed by the ANoVA concerned the variations shown by the various FFC statistics through time. The outcome was surprising. We expected there to be few, if any, statistics which showed a statistically significant trend *across all sites*. In fact there were 11 (although two of these relate solely to sample numbers). We show the details in Table 6.5 - with the FFC statistics now ranked in decreasing order of the Time effect F-value. We also show in the final four columns of the table the mean values of each FFC statistic over the four five-year periods.

Table 6.5 FFC statistics showing statistically significant time trends across all sites

No	Description	F-value	1980-84	1985-89	1990-94	1995-99
24	Rel.st.dev.(annual mean flow)	102.6	0.187	0.217	0.328	0.361
25	Log of summer mean flow	96.4	-0.301	-0.333	-0.407	-0.456
3	Relative st.deviation (daily flow)	74.9	0.985	1.081	1.233	1.206
9	Lag-15 autocorr. coefficient	64.5	0.452	0.389	0.471	0.518
2	Log of mean flow	40.3	-0.106	-0.1354	-0.1565	-0.1834
8	Lag-1 autocorr. coefficient	36.8	0.768	0.783	0.801	0.819
10	Lag-30 autocorr. coefficient	30.0	0.338	0.307	0.379	0.386
6	95%ile:50%ile ratio	24.3	4.813	5.123	6.742	7.068
4	Mean:50%ile ratio	20.1	1.595	1.646	1.913	1.952
1	No of valid daily flows	7.03	1618	1660	1752	1743
11	No of valid consec.day ratios	7.01	1614	1658	1749	1739

The table paints a dramatic picture of Thames-wide trends in flow characteristics over the past 20 years. The main messages are as follows:

FFC no(s)	Comment
24, 3	Whether we look at year-to-year or day-to-day variability, flow has been much more variable in the 1990s than it was in the 1980s.
25	Mean summer flow has fallen markedly over each of the last three 5-year periods - by 7%, 16% and 11% respectively.
2	Mean annual flow has fallen steadily over each of the last three 5-year periods - by 6%, 5% and 6%.
8, 9, 10	Increases in several different measures of autocorrelation indicate that flow has been more persistent in the 1990s than it was in the 1980s.
6	Another measure of increased variability - the 95%ile:median ratio - has increased steadily from 4.8 to 7.1.
7	The increased skewness in flow is also borne out by an increase in the mean:median ratio.
1, 11	The final entries show merely that flow gauging was consistently more reliable in the 1990s than it was in the 1980s.

6.5 Characterisation of sites by their flow variability

Before we embark on a description of the results of the high-level modelling of GQA versus flow, there is one further aspect of flow to be addressed. This concerns the extent to which flow varies more at some sites than others.

We have already seen that, irrespective of what Flow Fingerprint Candidate we look at, there are statistically significant differences between sites. In particular, therefore, the *variability* of flow at a site (as measured by, say, FFC3 or FFC24) itself varies from site to site. From the standpoint of the project, the most useful sites are those with the greatest flow variability as they give the greatest scope for any effect of flow on quality to manifest itself - and also maximise the opportunity for it to be detected by statistical analysis. Conversely, sites at which flow varies very little are much less helpful. They can still be potentially useful, but only in the secondary sense of perhaps demonstrating that stable-flow sites tend to be those with little or no variation in GQA class.

This section therefore provides a characterisation of flow variability across the 116 flow gauging stations[1] - and so is analogous to the discussion in Section 6.1 of variability in GQA class.

Figure 1 presents three histograms. The top one shows the quantity **CoV of daily mean flow**. Over the 116 flow sites the values vary from 0.1 to 2.7 with a median value of 1.1.

The next histogram summarises the quantity **CoV of annual mean flow**. These values are of course substantially smaller than those in the top histogram, as all within-year variation is smoothed out by the process of calculating annual averages. Thus the CoV values range from 0.07 to 1.6 with a median of 0.32.

The final histogram shows the quantity **CoV of non-overlapping 3-year mean flow**. If annual mean flow CoVs were statistically independent, we would expect this histogram to be narrower than the preceding one by a factor of $\sqrt{3}$, viz 1.73. In fact, the smoothing effect of averaging over three years seems to be somewhat stronger than this, as the median CoV (0.12) is 2.6 times smaller than the preceding median value (0.36). This suggests a tendency for low-flow and high-flow years to be more evenly mixed than would occur with a purely random process.

[1] It would perhaps have been preferable to show the breakdown across the 565 GQA sites rather than across the 116 flow gauging stations. However, it was computationally much simpler to use the flow figures unweighted by their associated numbers of GQA sites, and in any case any distortion thereby introduced will not materially affect the conclusions drawn from them.

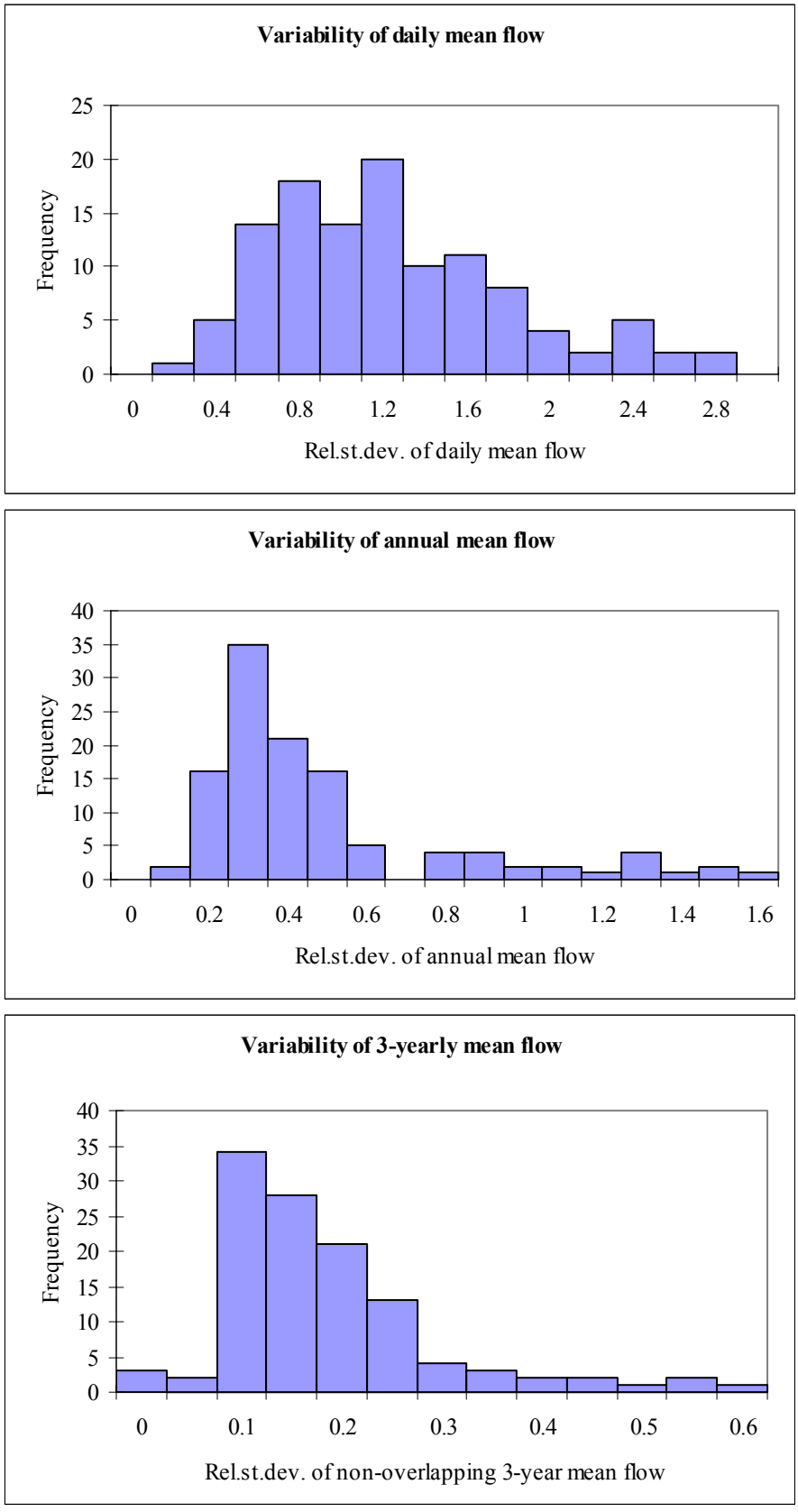


Figure 6.2 Flow variability for different averaging periods

6.6 GQA v. flow relationships

6.6.1 Classification of sites according to their potential to reveal a quality v. flow relationship

Following some preliminary modelling work, it became evident how important it was to put the result for any site into the context of how much or little variability (in flow and/or quality) had been seen at that site. It was decided, accordingly, to produce a classification table showing how the 565 sites subdivided into:

- sites with constant (or almost so) GQA class - and relatively *constant* flow;
- sites with constant (or almost so) GQA class - and relatively *variable* flow;
- sites with usefully variable GQA class - and relatively *constant* flow;
- sites with usefully variable GQA class - and relatively *variable* flow; and
- sites with insufficient quality data.

For judging the variability in quality we used the GQA class variability index described in Section 6.1. For flow, the most relevant measure of variability is that shown in the final histogram of Figure 6.2, namely the CoV of (non-overlapping) three-yearly mean flow. About a third of sites have a CoV greater than 0.15, and so this is a convenient criterion for identifying 'higher flow variability' sites.

On this basis we constructed Figure 6.3, which shows a plot of 3-yearly flow CoV against the GQA class variability index. If there *is* a correlation between sites with higher flow variability and sites with higher GQA variability, it is not very evident from the figure. The findings are slightly more promising when the data is presented in a two-way table, as in Table 6.6. The proportion of higher flow variability sites is below average (24%) for sites with constant GQA class, and above average (40%) for the sites with the greatest variation in GQA class. However, a statistical significance test shows that the difference between these proportions is only marginally significant. ('Fisher's exact test' gives $P = 0.06$.)

We can summarise this as follows: the extent to which GQA class varies at a site seems to be largely unrelated to the amount of variation in mean flow at that site. This finding did not augur well for the modelling exercise that we now describe.

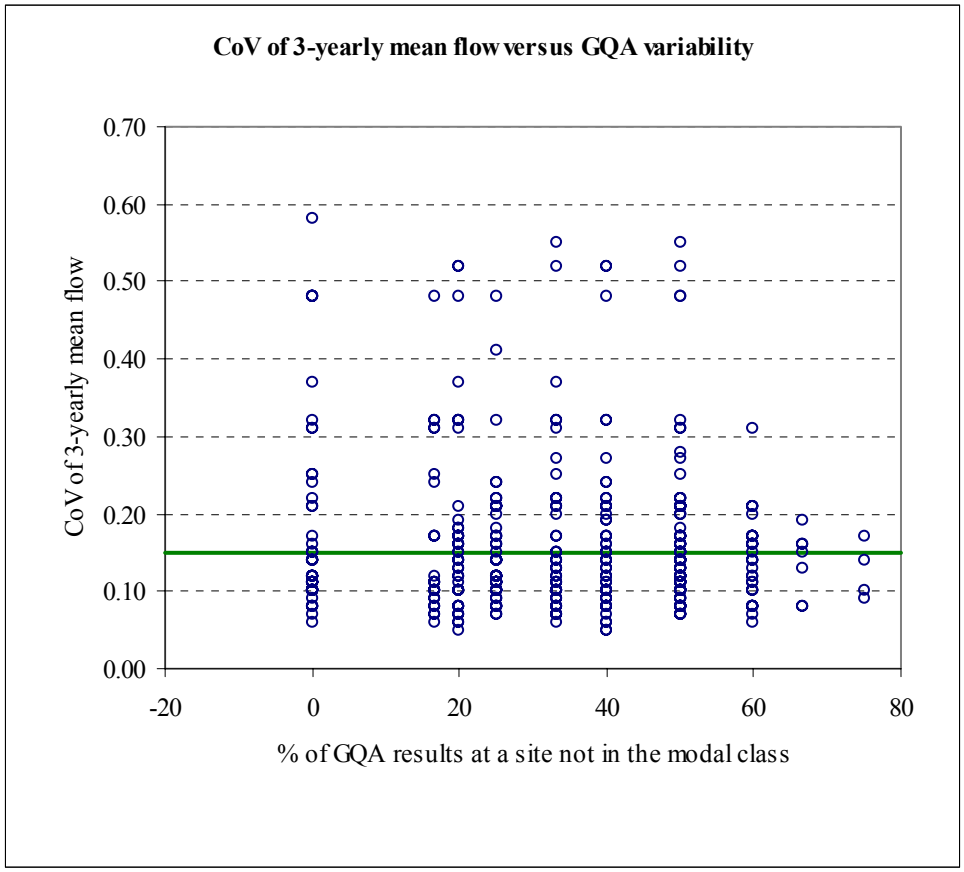


Figure 6.3 Flow variability plotted against GQA variability

Table 6.6 Classification of sites by flow variability and GQA variability

GQA Class variability index	Total frequency	No of sites with flow CoV > 0.15	% of sites with flow CoV > 0.15
0% (= least variable)	66	16	24
17 - 25%	145	50	35
33 - 50%	240	76	32
60 - 75%	45	18	40
Total	**496**	**160**	**32**
Too few GQA values	63		
Insufficient flow data	6		
Grand total	**565**		

6.6.2 Results for 3-year GQA class

The results of the 3-year GQA class v. FFC modelling are summarised in Table 6.7. For each row - i.e. Flow Fingerprint Candidate - the rightmost column shows the total number of sites for which the model was statistically significant (at the $P<0.05$ level). The four columns to the left then show how those significant sites were distributed across the low, medium and

high GQA variability categories. (Note that no models can ever be found for sites in the 'zero' column, as these represent the 66 GQA sites whose class remained constant over the 18-year period.)

Take for example the FFC2 row, describing the results for log (daily mean flow). A total of 33 sites showed a statistically significant relationship. Of these, 15 were sites with low GQA variability, 17 with medium GQA variability, and one with high GQA variability. As the total numbers of sites in the low, medium and high categories are 145, 240 and 45, the *proportions* of sites revealing a relationship are roughly 10%, 7% and 2%. Thus there is no tendency for sites with more variable flows to have a greater proportion of significant quality v. flow models - which reinforces the finding noted in the previous section.

We can make an overall assessment of the strength of evidence provided by Table 6.7 as follows. There are 430 sites for which there was non-zero variability in GQA class. The significance testing for each FFC was conducted at the $P<0.05$ level; and so out of those 430 sites we would have expected to get around 22 or 23 'false alarms', with a 19-in-20 statistical spread of about 16-30. There are only eight FFCs for which there is a statistically significant trend at more than 30 sites. Moreover, four of these are simply functions of the numbers of valid data points through time, and so can be discounted. (Flow monitoring has tended to become more reliable over the years, and river quality has tended to improve, but there is no possible causal link between the two.). Of the remaining FFCs, the one with the greatest number of sites showing statistically significant trends is FFC15; and for this, the number of significant sites - 34 - is only slightly beyond the expected window of background false-alarm rates. It must therefore be concluded that no worthwhile evidence has been found for a general high-level association between flow and GQA class.

One useful point does emerge. As log mean flow (FFC2) is close to being the best-performing FFC, with 33 sites showing a statistically significant correlation, it is reasonable to assume that we need give no further consideration to any of the more complex FFCs.

Table 6.7 Summary of the high-level modelling results for 3-year GQA class

FFC no	Description of FFC		No of sites, by GQA variability				
			Zero	Low	Medium	High	Total
			66	**145**	**240**	**45**	**496**
			No of models found (P<0.05)				
FFC1	Flow	N	0	22	9	5	36
FFC2		**Log10Mean**	**0**	**15**	**17**	**1**	**33**
FFC3		CoV	0	12	17	2	31
FFC4		Mean/P50	0	7	18	3	28
FFC5		P05/P50	0	7	18	4	29
FFC6		P95/P50	0	12	15	1	28
FFC7		P95/P05	0	11	11	2	24
FFC8		ACC1	0	4	7	2	13
FFC9		ACC15	0	3	9	2	14
FFC10		ACC30	0	10	1	1	12
FFC24		CoV(Ann.av)	0	7	15	9	31
FFC25		L_{10}SummerAv	0	8	14	3	25
FFC11	F(i)/F(i-1)	N	0	22	8	4	34
FFC12		Mean	0	4	9	1	14
FFC13		CoV	0	3	7	2	12
FFC14		P75	0	4	18	3	25
FFC15		ACC1	0	14	15	5	34
FFC16	UpRuns	N	0	7	20	7	34
FFC17		P50	0	9	7	1	17
FFC18		P75	0	8	1	0	9
FFC19		P95	0	13	5	3	21
FFC20	DownRuns	N	0	7	19	7	33
FFC21		P50	0	14	7	0	21
FFC22		P75	0	16	13	0	29
FFC23		P95	0	6	16	3	25

6.6.3 Results for 1-year GQA class

In view of the above findings from the full modelling exercise on 3-year GQA class, we restricted the modelling of variations in 1-year GQA class to the single explanatory variable FFC2, namely log (annual mean flow).

The results were somewhat stronger than those from the 3-year GQA modelling, with the number of statistically significant relationships increasing from 33 to 80. As Figure 6.4 shows, the R value for all but two of these is negative. (And one of the two positive cases is suspect, being based on only four years' data.) Thus, an *increase in mean flow* is consistently associated with a numerical decrease in GQA class - i.e. an *improvement in GQA class*.

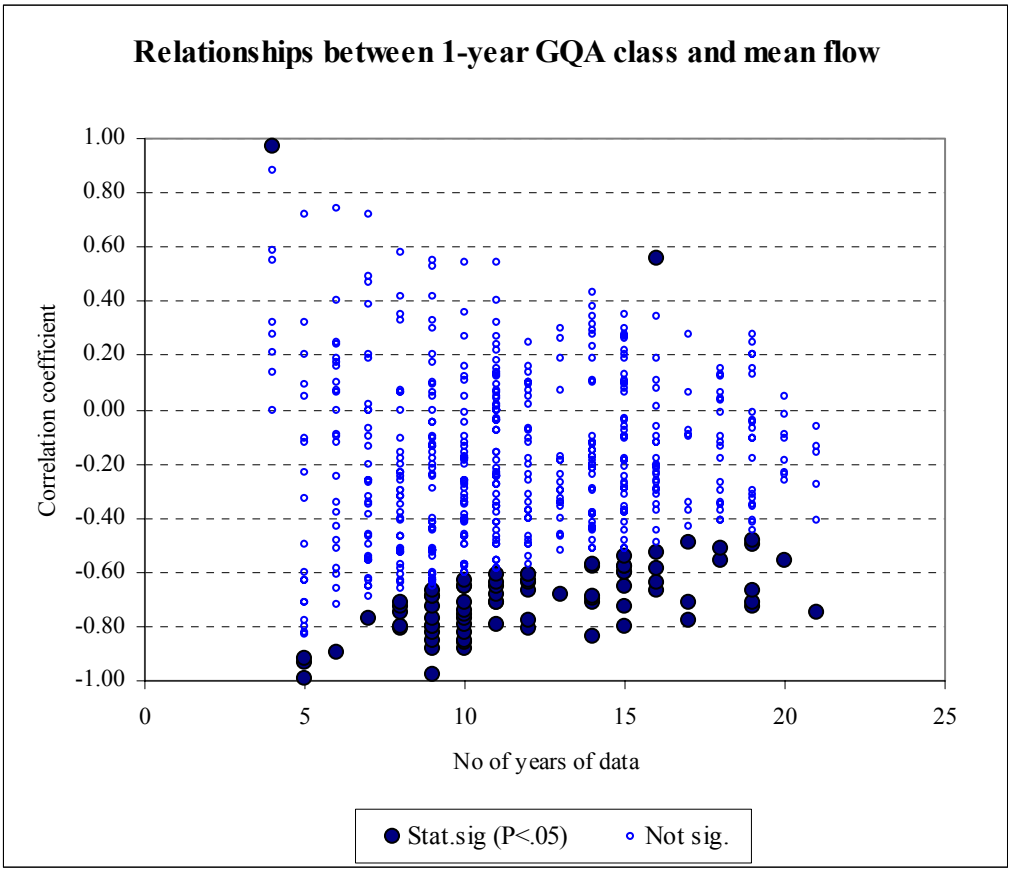


Figure 6.4 Summary of results for 1-year GQA class

Although the number of sites showing statistically significant effects is only about 19%, it could be argued that a *consistent* flow-induced change in quality at even one site in five would be quite enough to cause a large regional swing in GQA performance. For this to be so, however, it would be necessary for the GQA v. flow effects at those key sites to be not only consistent but also *substantial*. To test whether this was the case, we looked at the 78 sites for which a statistically significant negative relationship had been found, and used those models to predict the effect on 1-year GQA class of a 20% increase in mean flow. The results are shown in Figure 6.5. For the great majority of sites the predicted improvement is between 0.2 and 0.6 of a class, and improvements more marked than this are predicted for just six sites.

We conclude, therefore, that the 1-year GQA v. mean flow models identified here are too weak to account for more than a very small proportion of the observed 1-year GQA class changes. (It should also be noted that, even if the predicted effect had been strong, there would have remained the complication that there is no obvious way of extrapolating such effects from a 1-year to a 3-year assessment basis.)

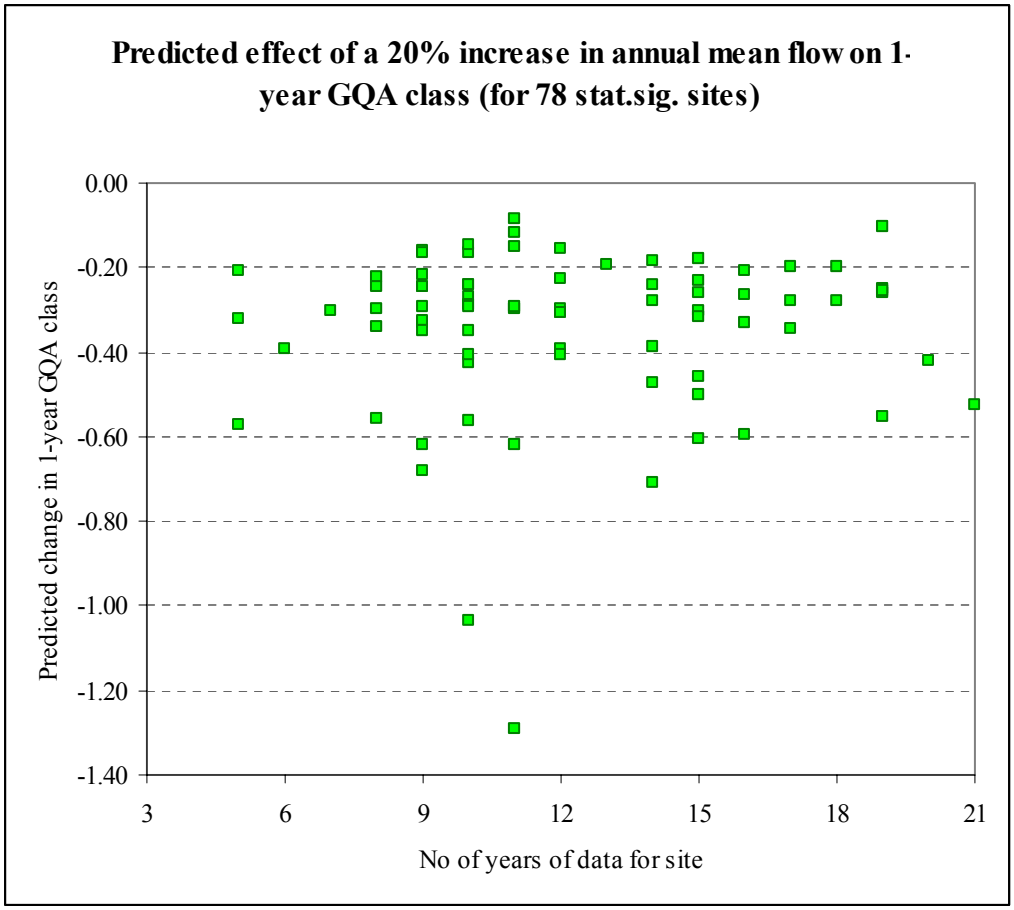


Figure 6.5 Predicted effect of a 20% increase in mean flow on 1-year GQA class

7. REGIONAL-LEVEL RESULTS

7.1 Introduction

The two preceding chapters have discussed the results obtained from:

(i) *low-level* analyses - where the modelling was done at the individual sample, determinand and site level with no data aggregation; and

(ii) *high-level* analyses - where the modelling was done using data aggregated both across *time* (by using 3-year summary statistics) and across *determinands* (by using GQA class).

In this final results chapter, we look briefly at some evidence of a quality v. flow effect at the most aggregated level, namely obtained from:

(iii) *'Regional-level'* analysis - where data is aggregated across time, determinands *and* sites.

Two interesting pieces of work were made available by members of the Steering Group. One supplied by Simon Bingham ('Paper A') used summary data drawn from all Regions to provide 'broad-brush' evidence for a quality v. flow effect. The other, an unpublished internal discussion paper written by Juliane Struve ('Paper B'), focused on summary measures of quality and flow for just one area of Thames Region.

7.2 Paper A

7.2.1 Data used

For each of the eight Regions, rolling 3-year data was available from 1992/94 to 1998/2000 on total river length in each class; and the statistic used to track Region-wide change was:

Net % improvement in class (by length) compared to 1988/90.

For each Region, flow data was obtained from a representative gauging station, and the statistic used to track Region-wide change was:

Change in 3-year mean flow as % of mean flow for 1988/90 .

7.2.2 Results and discussion

Figure 7.1 reproduces from the paper a plot of % GQA change against % flow change. (The only change that we have made is to use different symbols for each Region.) The plot certainly seems to show a fairly strong association. However, because it is based on rolling 3-year data some of the association will have been induced by autocorrelation in the data, and so we cannot use conventional significance testing methods (see the discussion in Appendix B).

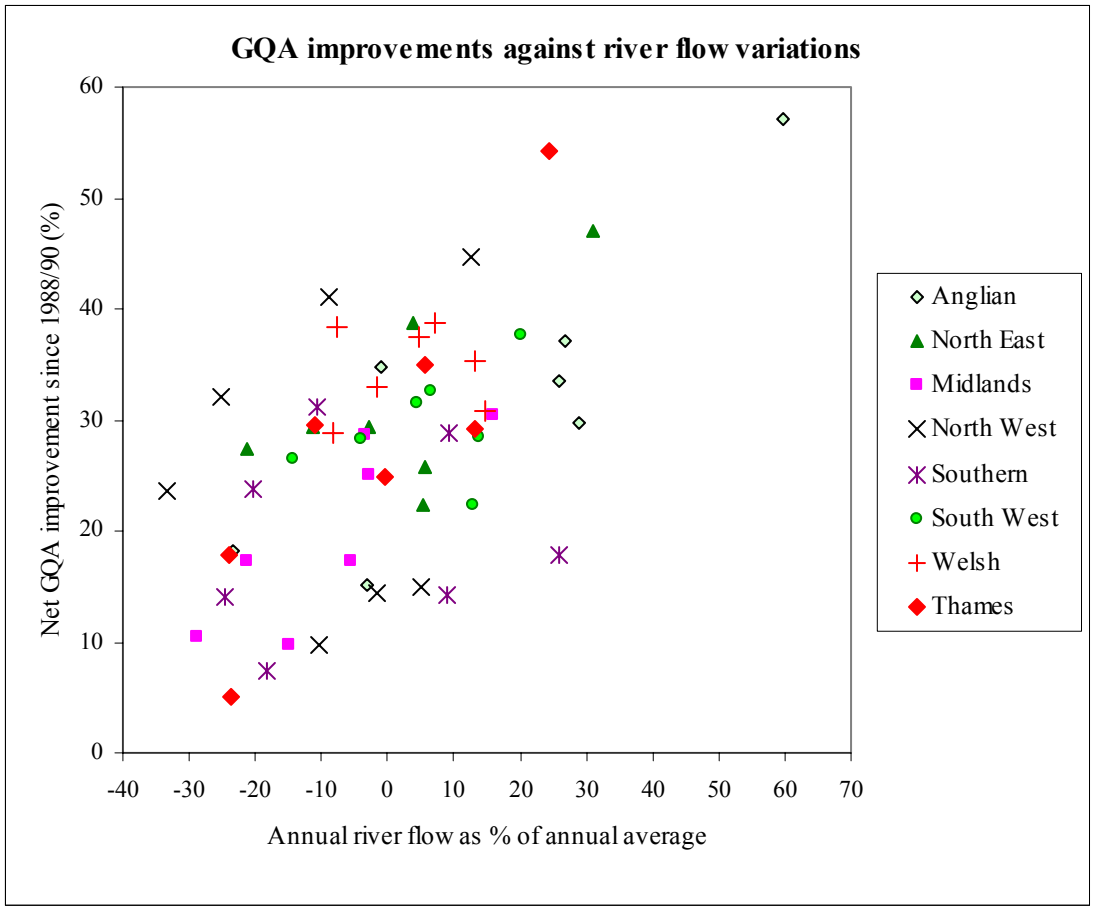


Figure 7.1 Results presented in Paper A

To explore the effect of autocorrelation, it is useful to focus on a single Region at a time. We start with Thames. Figure 7.2 (a) shows the Thames subset of Figure 1, with the seven points joined by a time line. This gives a good insight into the 'random walk' that is induced by the autocorrelation. Part (b) of the figure then shows the three *non-overlapping* points. These have a dramatically high R of 0.9999 (to 4dp), which exceeds even the 99% critical value (see

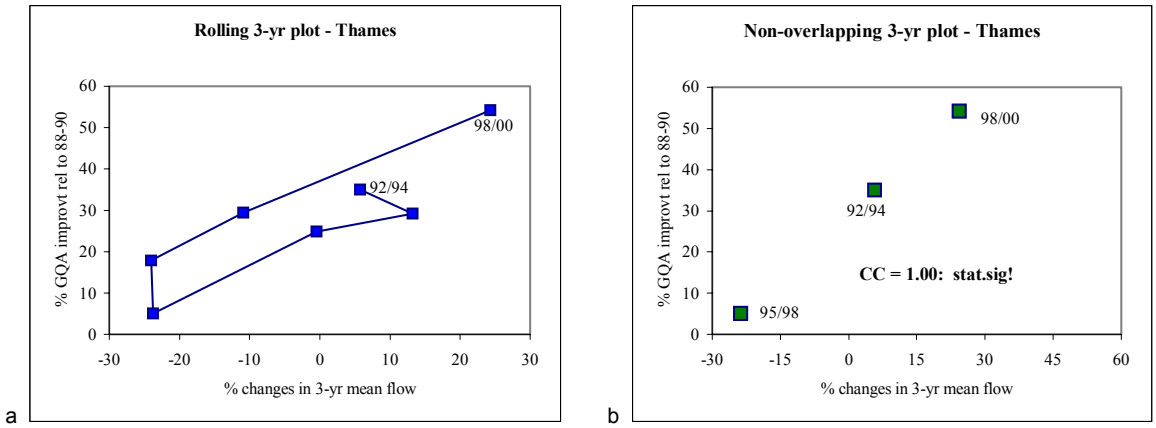


Figure 7.2 Thames subset of Figure 7.1

Appendix B). There is no doubt, therefore, that the association for Thames is statistically significant.

Figure 7.3 shows the corresponding pair of plots for the Anglian data. This time the R value is 0.97 - which, though apparently very high, is not in fact statistically significant. This may seem counter-intuitive - but remember that with 2 points the R value will always be a perfect 1.00, and we only have one more data point here…

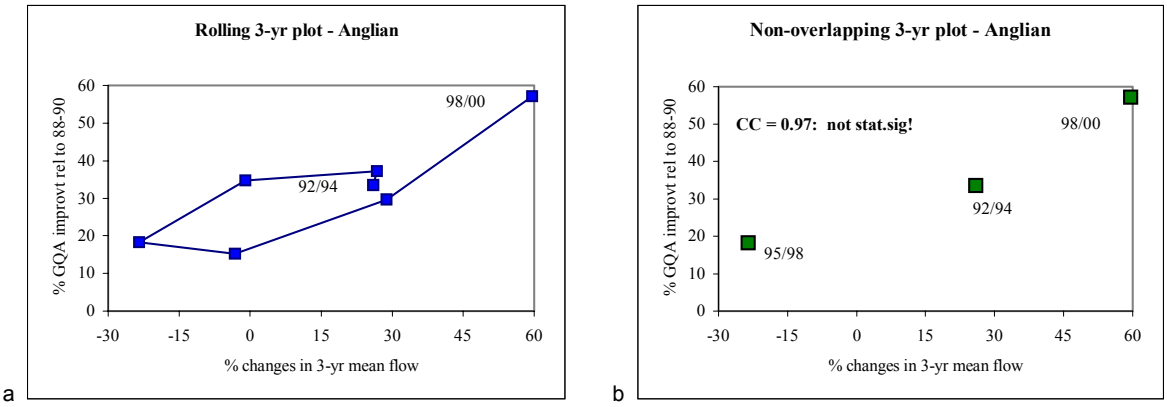


Figure 7.3 Anglian subset of Figure 7.1

The six pairs of plots for the other six Regions are shown in Figure 7.4 and Figure 7.5. (Note that the right-hand plots all have similar x-axis and y-axis scales to aid comparability between Regions.) We see from these that there is only one more very high (albeit not statistically significant) R value - the 0.98 value for North East. The other five Regions have R values ranging from 0.74 to -0.29, and although the plots show some upward movement from left to right (except for Welsh), none of them looks very convincing.

Of course, part of the problem for some Regions may be that the chosen gauging station is not particularly representative of the Region as a whole. Nevertheless, we can only comment on the data that was actually used in the paper. The upshot of the above analysis is that, on the basis of just three non-overlapping points, only for Thames is there strong *and statistically significant* evidence of a GQA v. flow association. There is strong but *non-significant* evidence for two other Regions - Anglian and North East; and there is only tentative evidence of an effect for four of the other five Regions. This does not of course mean that the effect does not exist, but simply that it cannot be detected from the very limited amount of data available.

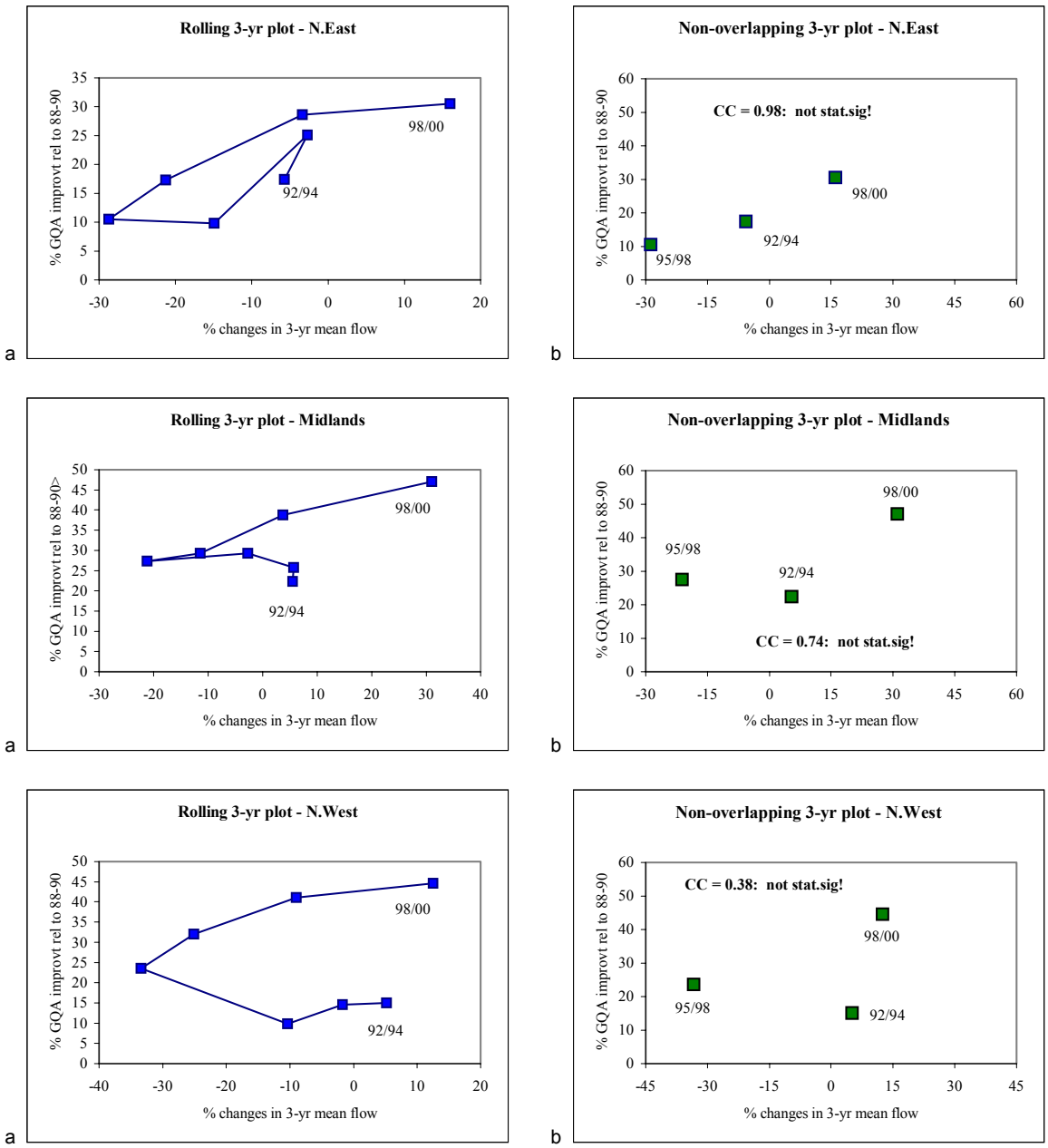


Figure 7.4 Corresponding plots for North East, Midlands and North West

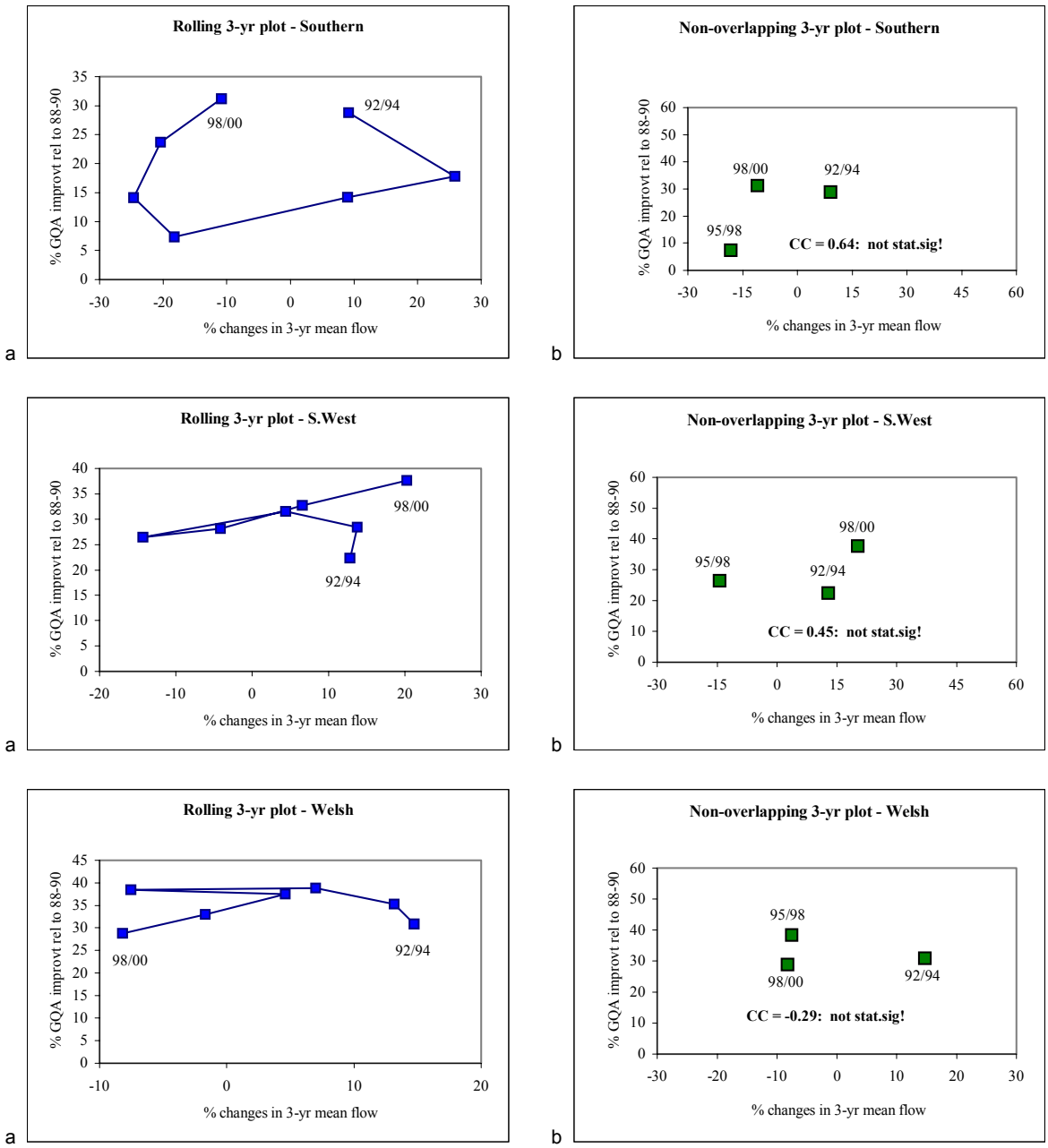


Figure 7.5 Corresponding plots for Southern, South West and Welsh

One obvious way of improving the power of the above analysis (other than waiting for another independent 3-year point to become available) would be to repeat it for smaller groups of sites than whole Regions. This would make it easier to get more representative flow data to go with the GQA data. In the meantime, however, one useful thing that we can do with the existing data is to pool the eight three-point graphs to produce a slimmed-down version of the plot shown in Figure 7.1. This is shown in Figure 7.6. The R value for this data is 0.75, and as this is based on 24 points it is highly statistically significant ($P < 0.001$). So the data *does* after all support the hypothesis of an overall association between increasing flow and improving GQA class - but only when all the Regional-level plots are themselves combined at an even higher level of aggregation.

This finding may seem at first sight somewhat counter-intuitive. However, it is simply a example of the general principle that the greater the number of samples, the greater is the chance of detecting a given underlying relationship. Figure 7.6 shows that there is a general tendency, broadly common to all Regions, for GQA improvements to be positively correlated with increases in mean flow But it also shows that there is a considerable degree of scatter in that relationship, and so it is not surprising that (with a couple of exceptions) it was not possible to detect this for individual Regions on the basis of just three data points.

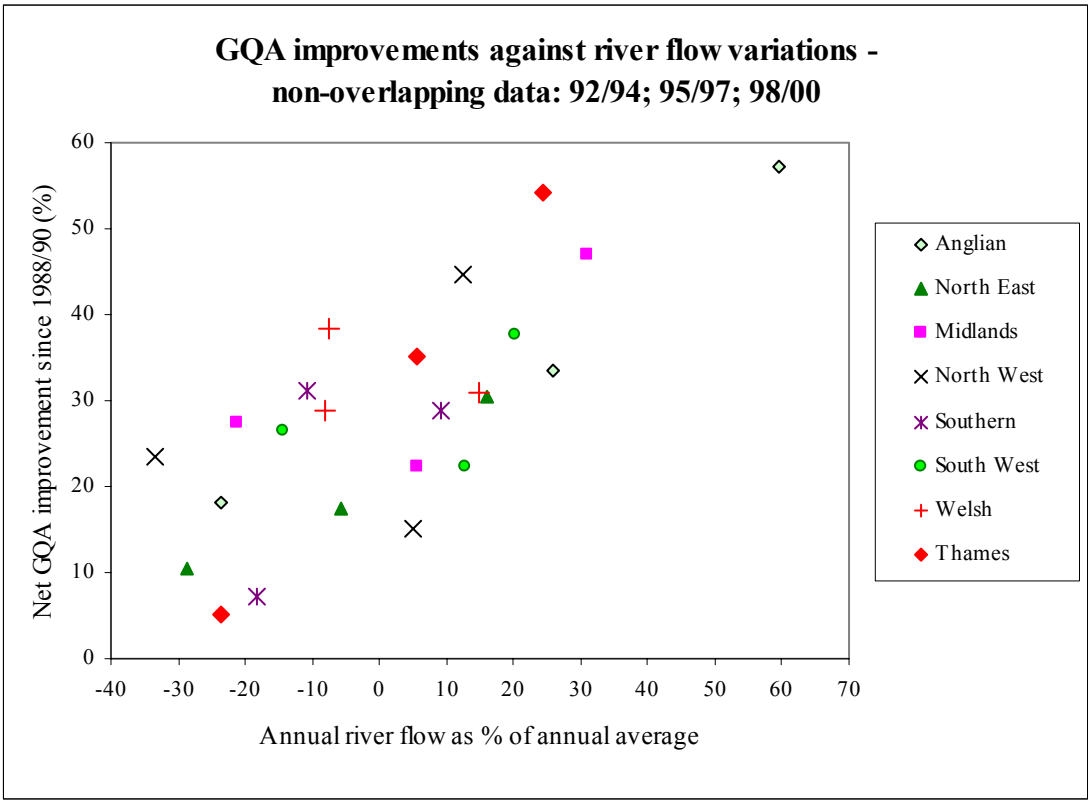


Figure 7.6 Equivalent of Figure 7.1 but showing just non-overlapping data

7.3 Paper B

7.3.1 Objective of paper

This internal discussion paper looked at the influence of river flow on river quality in the North-East area of Thames Region. It focused on the period 1988 - 1998, with particular emphasis on the impact of the drought from winter 1995/96 to winter 1996/98.

7.3.2 The GQA data

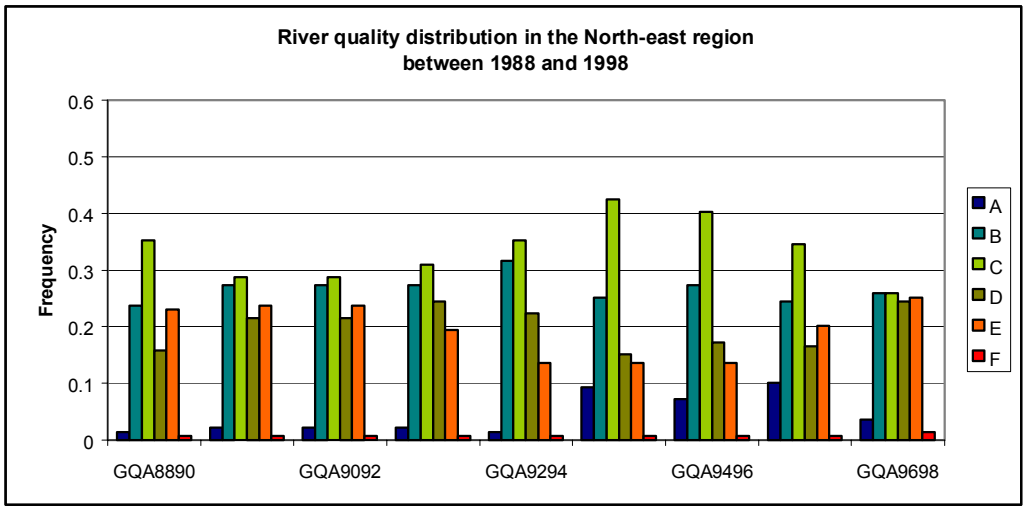
GQA results were obtained for all rivers in the area for each of the nine rolling 3-year periods from 1988-90 to 1996-98. The relative frequencies of classes A to F in each period are plotted in the top panel of Figure 7.7. The bottom panel shows the same information but aggregated into just two groups: A-C, and D-F. (Because of a few arithmetic errors, the two columns do not always add up to exactly 1.0.) It is clear from this that the proportion in classes A-C was lowest in the most recent period, namely 1996-98. To highlight this deterioration from earlier years, the middle panel shows what happens when the GQA results are expressed relative to the 1996-98 distribution. The effect is particularly marked for the proportions of rivers falling in classes A and C.

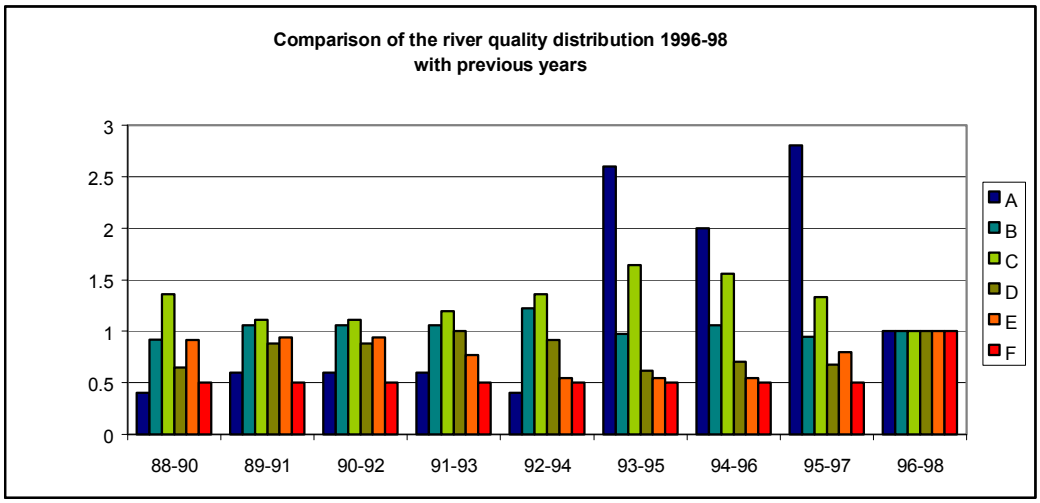
7.3.3 The flow data

For each river site, flow records were obtained for the nearest gauging station, and day-by-day rolling 3-year averages calculated. These are plotted in the top panel of Figure 7.8. In order to remove the effect of scale, the middle panel of Figure 7.8 re-plots the flow data as proportional deviations from each river's mean flow. That is, $Q_t' = (Q_t - Q_{ave})/Q_{ave}$. (Note that the y-axis label of the plot has lost its suffixes.)

7.3.4 Results and discussion

The summary information on changes in GQA class and 3-yearly mean flow is brought together in the final panel of Figure 7.8. This shows a positive correlation between the 'A-C to D-F' ratio and the relative deviation of 3-yearly mean flow from grand mean flow over the 11-year period. In other words, periods in which the proportion of sites in GQA classes A-C was higher than usual tend to be associated with periods in which mean flow was higher than usual. The paper concludes that 'variations in flow are partly responsible for shifts in the dominance of A-C reaches over C-D reaches'.





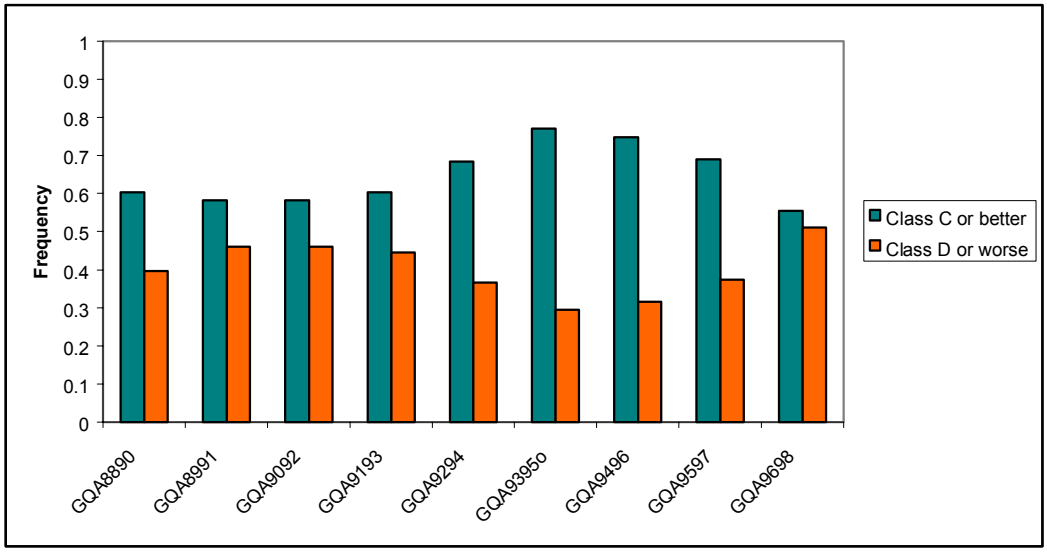
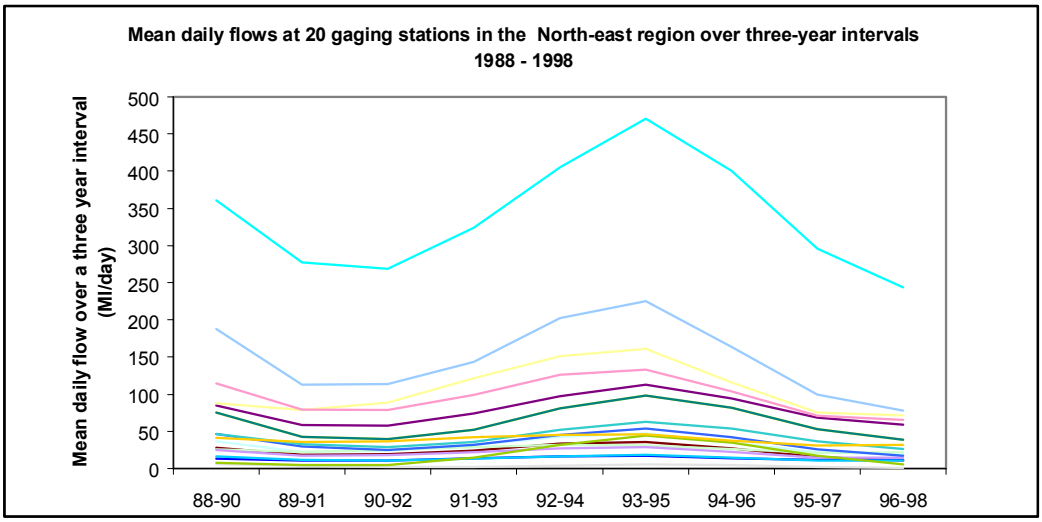
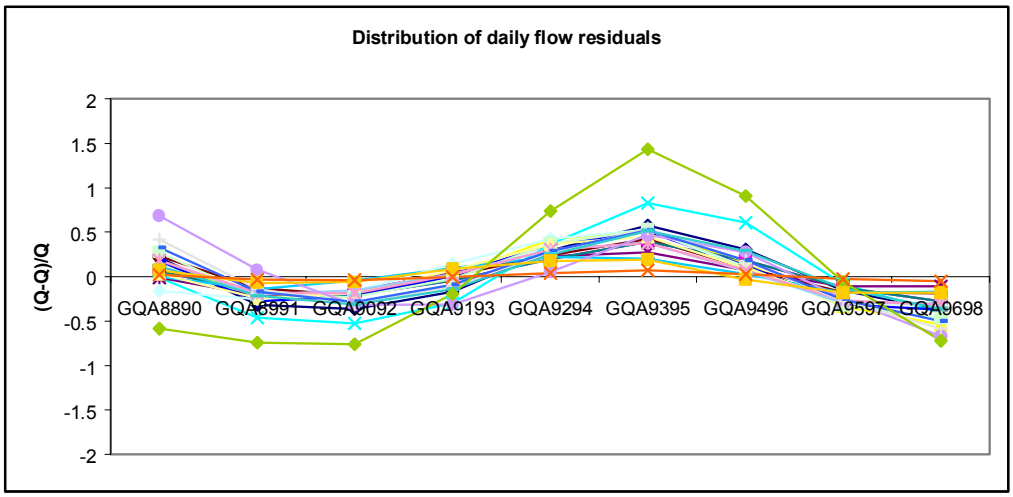


Figure 7.7 Figures 1, 2 and 3 reproduced from Paper B





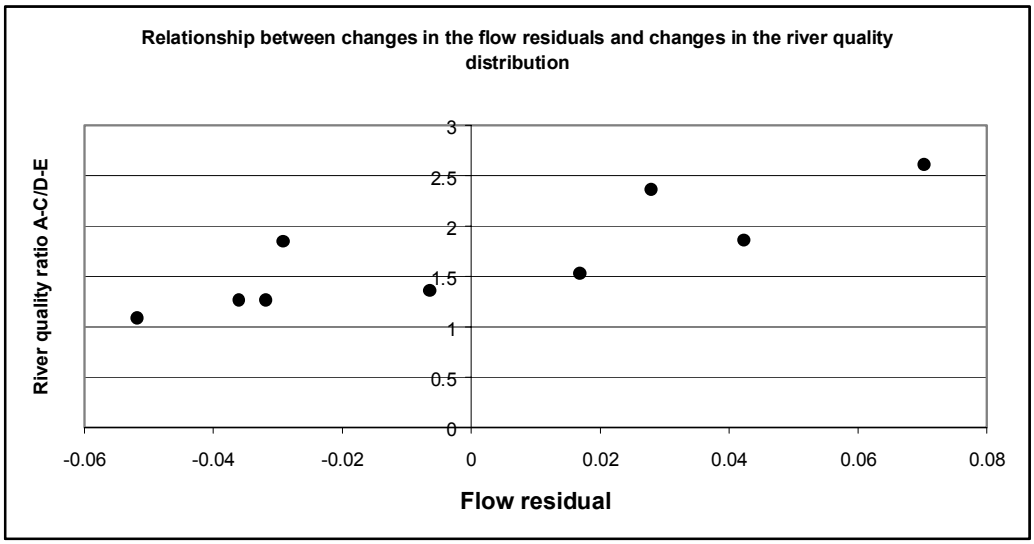


Figure 7.8 Figures 4, 5 and 6 reproduced from Paper B

The correlation coefficient for the association shown in the final panel of Figure 7.8 is 0.84. Had this been based on 9 *independent* pairs of data points, this would have been statistically highly significant (P<0.01). However, the plot shows rolling 3-year data, and so the same comment applies as that in the previous section for Paper A: the plot contains less information than it appears to because of the autocorrelation that exists between successive years.

The simplest way to correct for this is to calculate the correlation coefficient for just the *non-overlapping* data. This is illustrated in Figure 7.9. This shows the same nine data points as in Figure 6 of the paper, but with two additional features: to highlight the autocorrelation we have joined the points up by a time line; and we have shown how the data can be split up into three separate non-overlapping groups. For the first two groups, starting in 1988-90 (the squares) and 1989-91 (the triangles), the R values are 0.83 and 0.60. As noted earlier, the value would need to be at least 0.988 to achieve even mild significance (P<0.10), Thus these two groups provide no evidence at all of an effect.

For the group starting in 1990-92 (the circles), however, the R value is 0.9999, which is highly significant (P<0.01). This conclusion hinges largely on the very good GQA performance (and high flows) seen in 1993-95 coupled with the poor GQA results (and low flows) of 1996-98. It is also in good agreement with the conclusion reached from the results presented in Paper A for Thames region as a whole.

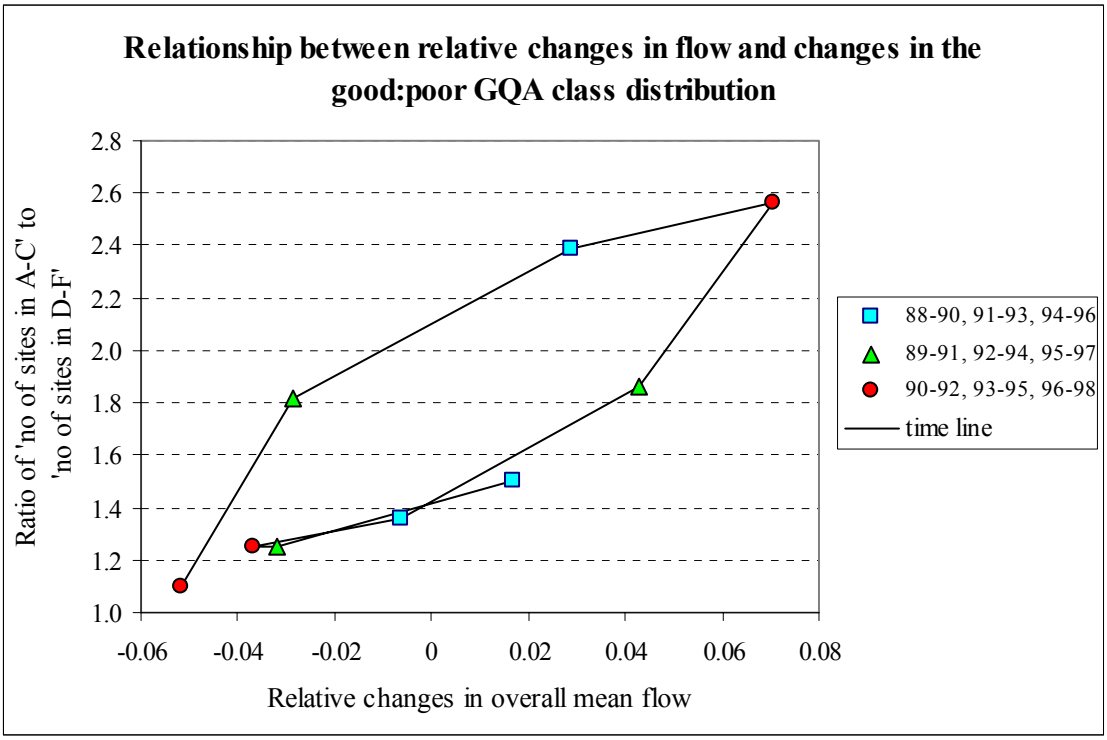


Figure 7.9 Figure 6 from Paper B revisited

8. FURTHER INVESTIGATIONS

8.1 Introduction

The evidence of the Regional-level papers discussed in Chapter 7 points to the association between GQA class and mean flow being at its strongest in the mid-to-late 1990s. This is also supported by the high-level results discussed in Chapter 6. The general improvement seen in GQA class over this period is confirmed by Figure 8.1, which shows that between 1995-97 and 1998-2000, only about 40 of the 565 sites showed a decline in quality (usually by just a single class), whilst nearly 300 sites experienced an improvement in class.

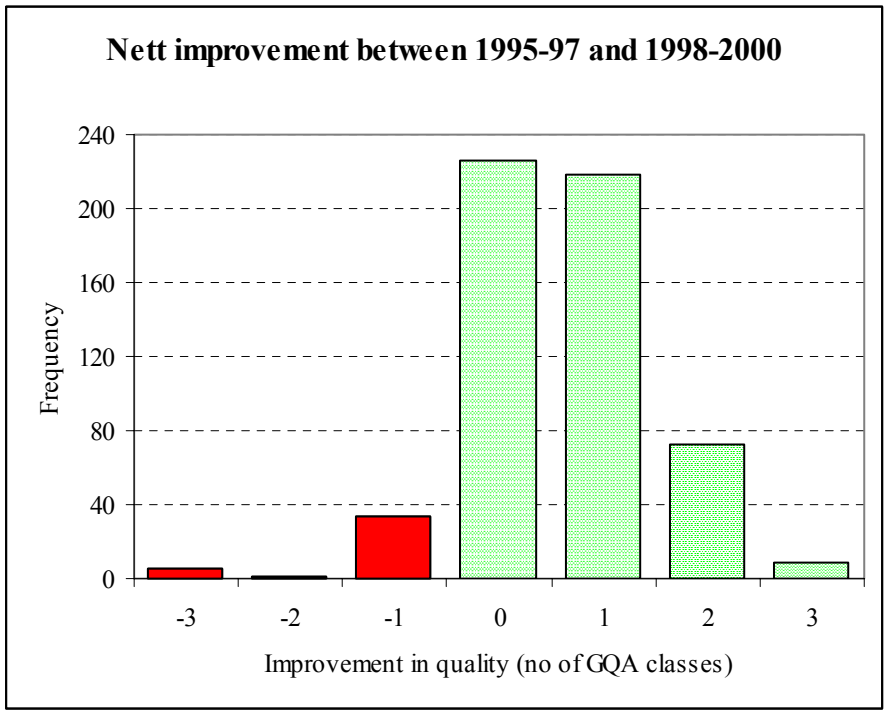


Figure 8.1 Nett improvement in GQA class for Thames Region between 1995-97 and 1998-2000

8.2 Judging the effect of restricting the date window

To explore the possibility that there had been a widespread change in the structure of quality v. flow models in the mid 1990s, we re-ran program CAFE restricting the analyses to the date window 1995-2000, and then compared the resulting correlation coefficients (Rs) with the corresponding R values obtained from analysis of the full data set (1980-2000). The results are plotted in Figure 8.2 for each of the three GQA determinands.

Look first at the top panel, showing the plot for DO%. Points that are tilted more steeply than the diagonal line - that is, points that are above the line in the positive (NE) quarter or below the line in the negative (SW) quarter - represent sites for which restricting the date span brings an improvement in the R value. Whilst there is a preponderance of elevated points in the

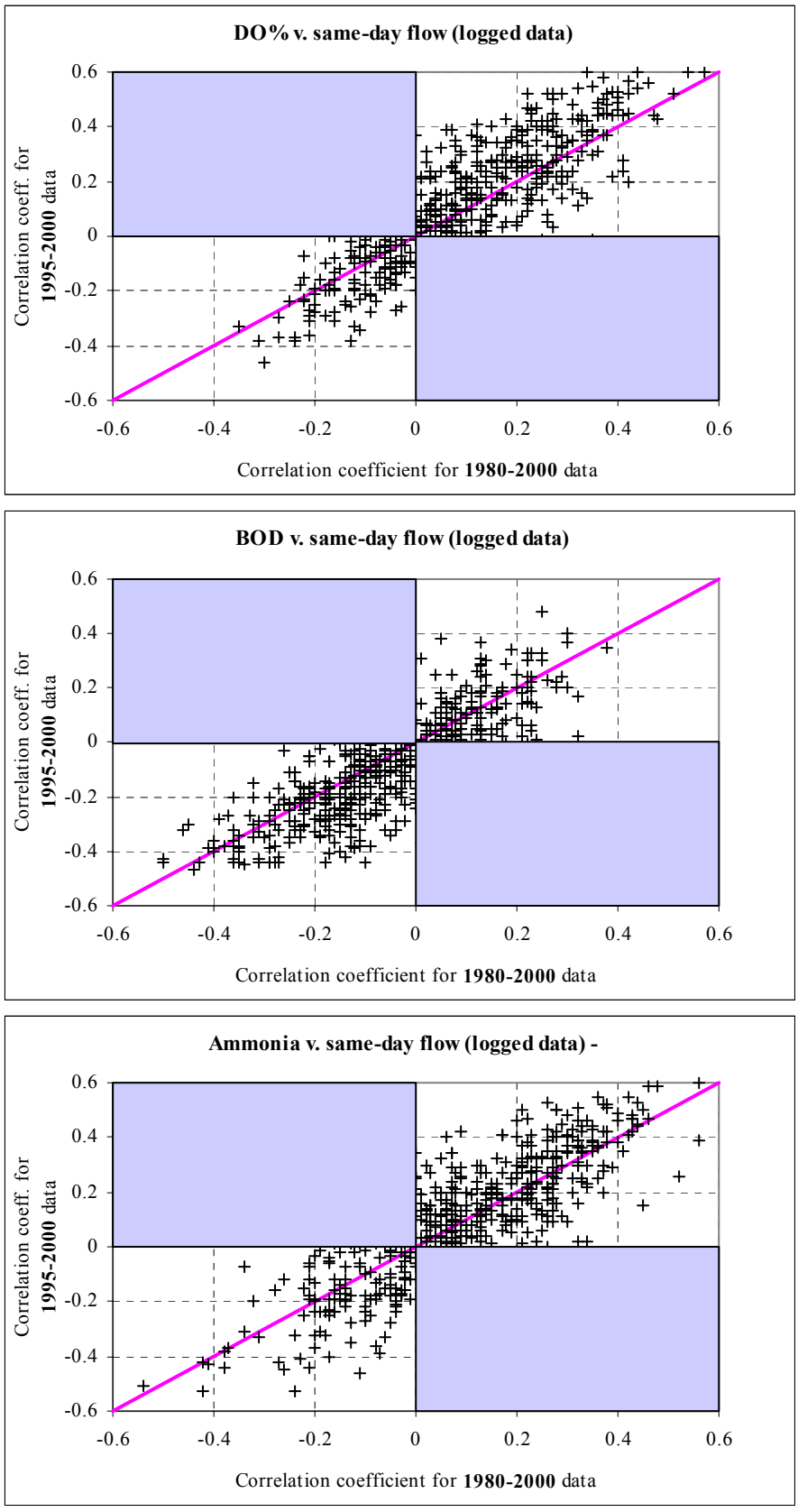


Figure 8.2 Effect of restricting the quality v. flow modelling data from the full 21 years to the most recent 6 years

positive quarter, the improvement in R is rarely more than 0.2. The same is true for the more extreme points in the negative quarter. Overall, therefore, there is no evidence of a dramatic improvement in the goodness-of fit of the models through use of the narrower date window.

The middle panel similarly shows the results for BOD. Here there is a more marked effect in the negative quarter than for DO%. However, the maximum negative correlation is seldom more extreme than -0.4, showing that even over a much shorter time period the BOD v. flow relationship is essentially very weak.

Finally, the bottom panel shows that, for ammonia, there is only a slight improvement in the typical goodness of fit in going from the full 21-year data set to the last six years.

Overall, therefore, there is little evidence that the relatively wide timespan adopted by the project (i.e. 1980-2000) has prevented stronger but more recent relationships from being identified.

8.3 Further investigation of sites showing large GQA improvements

8.3.1 Introduction

Previous sections have demonstrated that there was a widespread improvement in GQA class between one 3-year period in the mid-90s and the next, and that this was accompanied by a substantial increase in mean flow. Given this marked *high-level* association, it may seem anomalous that the *low-level* analyses found very few sites at which there was a strong relationship between flow and any of the three GQA determinands.

Before we describe the final element of the investigation, therefore, this is an appropriate point at which to present a simple example showing why a strong high-level association is no guarantee that there is a relationship at the level of individual data points. In the example, we suppose that the annual spread of BOD and ammonia values remain exactly the same over a six-year period, and that GQA class is driven solely by DO%. We also suppose for simplicity that, in any one year, flow is constant, but that the means show a steady increase through the six years.

The DO% values imagined to result from a monthly sampling programme over the six years are plotted against flow in Figure 8.3. The red blobs show the data for the first three years; and the red cross-hairs show 10%ile DO (43%) and mean flow (11 cumecs) over this period. The green squares show the data for the second three years. The corresponding green cross-hairs show that there have been big increases since the earlier three-year period both in 10%ile DO (71%) and mean flow (20 cumecs). However, the point of the example is that when we focus on the *individual* data points, there is absolutely no sign of a consistent association.

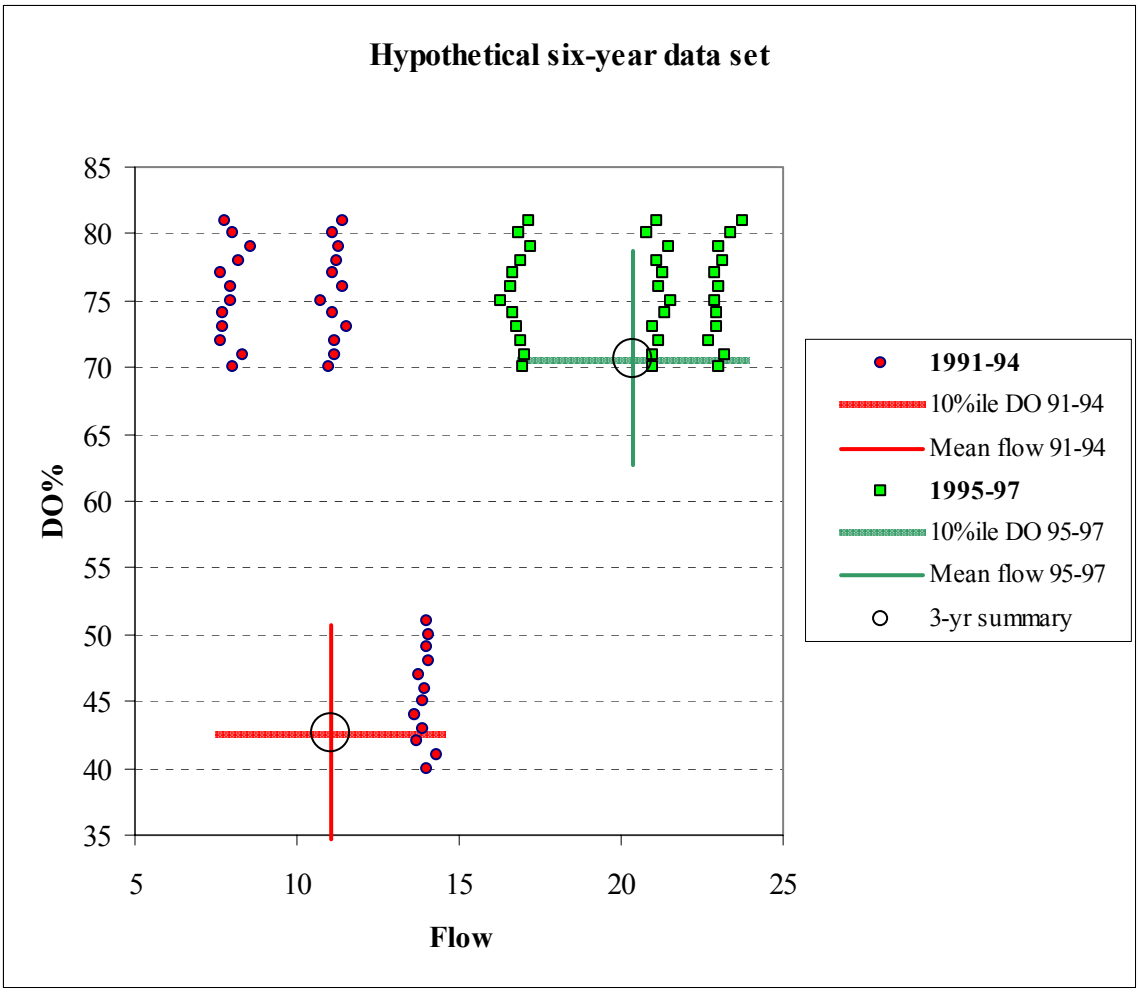


Figure 8.3 Hypothetical data set illustrating a strong high-level association between quality and flow despite no low-level association

Although that is just a hypothetical example, it clearly demonstrates the general point that the presence of a high-level association is no guarantee of a low-level association. Nevertheless, the presumption of the project has been that a low-level flow effect does exist. In a final attempt, therefore, to throw more light on why the low-level modelling produced so few useful quality v. flow relationships, we decided to look in detail at the data for the nine sites showing the greatest improvement (see Figure 8.1). It could be argued that we have distorted the investigation by picking such extreme sites. That may be so; but on the other hand, the low-level models have generally been so weak that we felt it important to focus on those sites which had the best possible chance of showing a worthwhile effect.

The details of the nine sites are shown in Table 8.1. In all but one case, the improvement is from class E to class B; and the final column shows that no one determinand is markedly more critical than another.

Table 8.1 Details of the nine sites improving by three GQA classes

Case	GQA site	Flow site	GQA class **1995-1997**				GQA class **1998-2000**				**Critical detd(s)**		
			BOD	Amm	DO%	All	BOD	Amm	DO%	All			
1	PUTR0104	0130THAM	A	A	**E**	**E**	A	A	**B**	**B**			DO%
2	PUTR0212	0790COLE	**E**	C	C	**E**	**B**	A	B	**B**	BOD		
3	PCHR0022	1420CHER	B	A	**E**	**E**	B	A	**B**	**B**			DO%
4	PTAR0115	1900THAM	**E**	**C**	B	**E**	**B**	**A**	B	**B**	BOD	Amm	
5	PWER0024	3020WEYN	B	B	**E**	**E**	A	B	**B**	**B**			DO%
6	PWER0089	3020WEYN	**E**	**C**	**E**	**E**	**A**	**A**	**B**	**B**	BOD	Amm	DO%
7	PRGR0119	5541BEAM	C	A	**E**	**E**	B	A	**B**	**B**			DO%
8	PWAR0060	4180WAND	B	**E**	E	**E**	B	**B**	B	**B**		Amm	
9	PRGR0011	5427CRIP	**F**	D	B	**F**	**C**	C	B	**C**	BOD		

8.3.2 Site PUTR0104

A cusum analysis identifies a marked step increase in DO% in Nov 1997. Unfortunately there were a lot of gaps in the flow record - and no flow values at all for 1999 and 2000 - and so we are unable to tell whether or not there was a corresponding increase in mean flow.

8.3.3 Site PUTR0212

Between the first 3-year period and the second, BOD 90%ile falls from 8.1 to 3.8 mg/l, and there is also a corresponding increase in mean flow. This is reflected by the statistically significant association between log(BOD) and log(flow), as shown in Figure 8.4. However, the relationship is weak, with an R value of only -0.33. It is also presents a somewhat mixed signal, with the worst BOD values all occurring in the *middle* flow range.

When we use the model to predict the decrease in mean BOD to be expected from the observed increase in mean flow, we find that this is only half that actually observed. Thus, whilst there does appear to be an association between flow and the class-critical determinand at this site, this is not strong enough to account for the whole of the improvement.

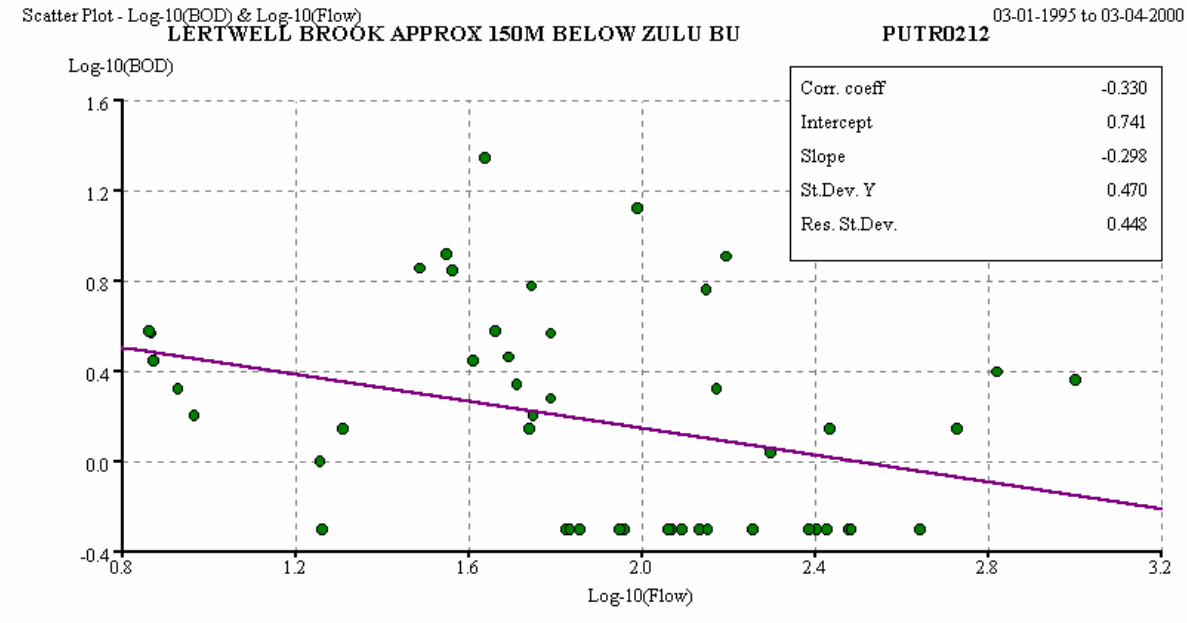


Figure 8.4 BOD v. flow relationship for Site PUTR0212

8.3.4 Site PCHR0022

Figure 8.5 illustrates the statistically significant relationship found between DO% and log(flow). As was the case with the model discussed in the previous section, the relationship is weak (R = 0.37). However, it is interesting to note the two clumps of low DO% values - one at low flows, and the other at medium flows. In contrast, DO rarely dips below 80% sat if log(flow) values are above 1.5 (corresponding to flows of $10^{1.5}$ = 32 cumecs). This suggests the possibility of a threshold-type mechanism, whereby the risk of obtaining low DOs increases sharply once flows have dropped below some critical value.

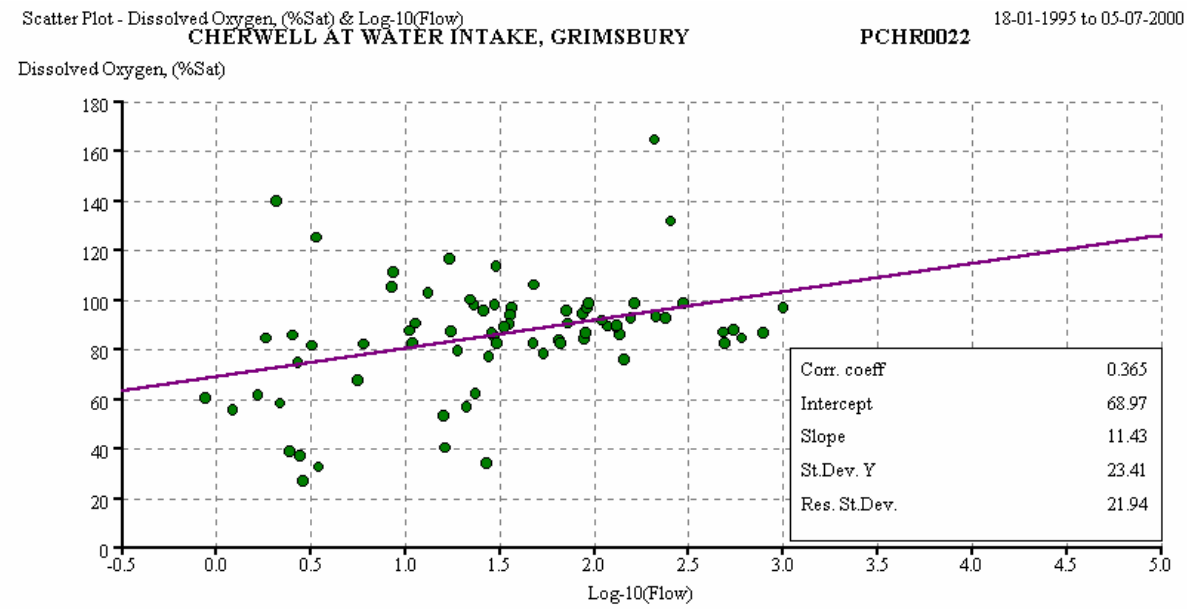


Figure 8.5 DO% v. flow relationship for Site PCHR0022

This is illustrated more clearly by the two time series plots shown in Figure 8.6. The low DOs in the first three years all occur at times of relatively low flow; conversely there are no low DOs in the last three years, and nor are there any prolonged periods of low flow.

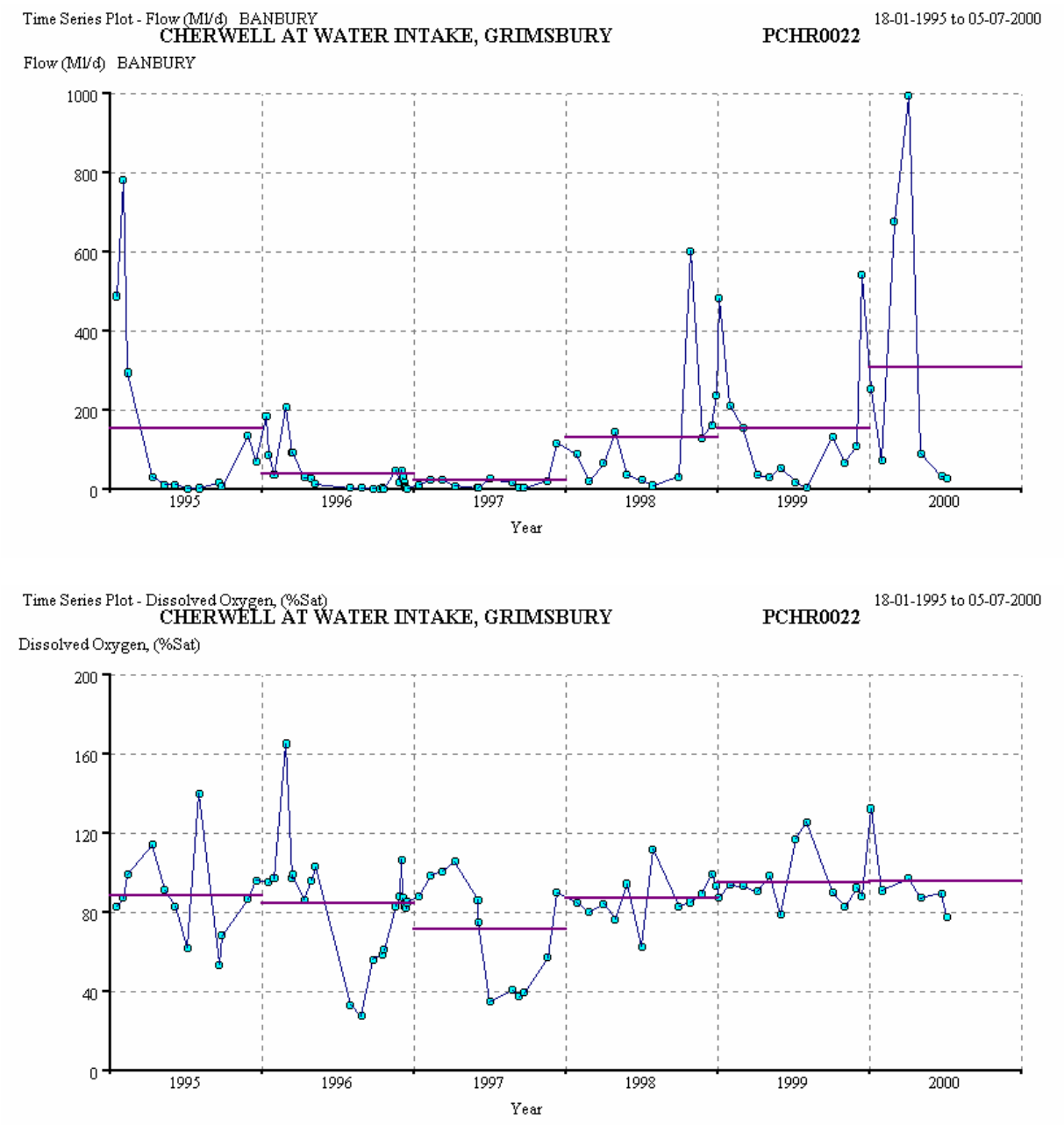


Figure 8.6 Time series of DO% and flow for Site PCHR0022

8.3.5 Site PTAR0115

The poor performance of this site during 1995-97 can be attributed to two outliers, as illustrated in Figure 8.7. For ammonia (the first plot), the improvement from class C to A was due to the outlier of 2.49 mg/l seen in 1997. The principal driver, however, was BOD (see the second plot), with the huge outlier of 33.1 mg/l in 1995 driving the site into class E for 1995-97. We found no sign of a BOD v. flow association - indeed, mean flow actually decreased in the second 3-year period.

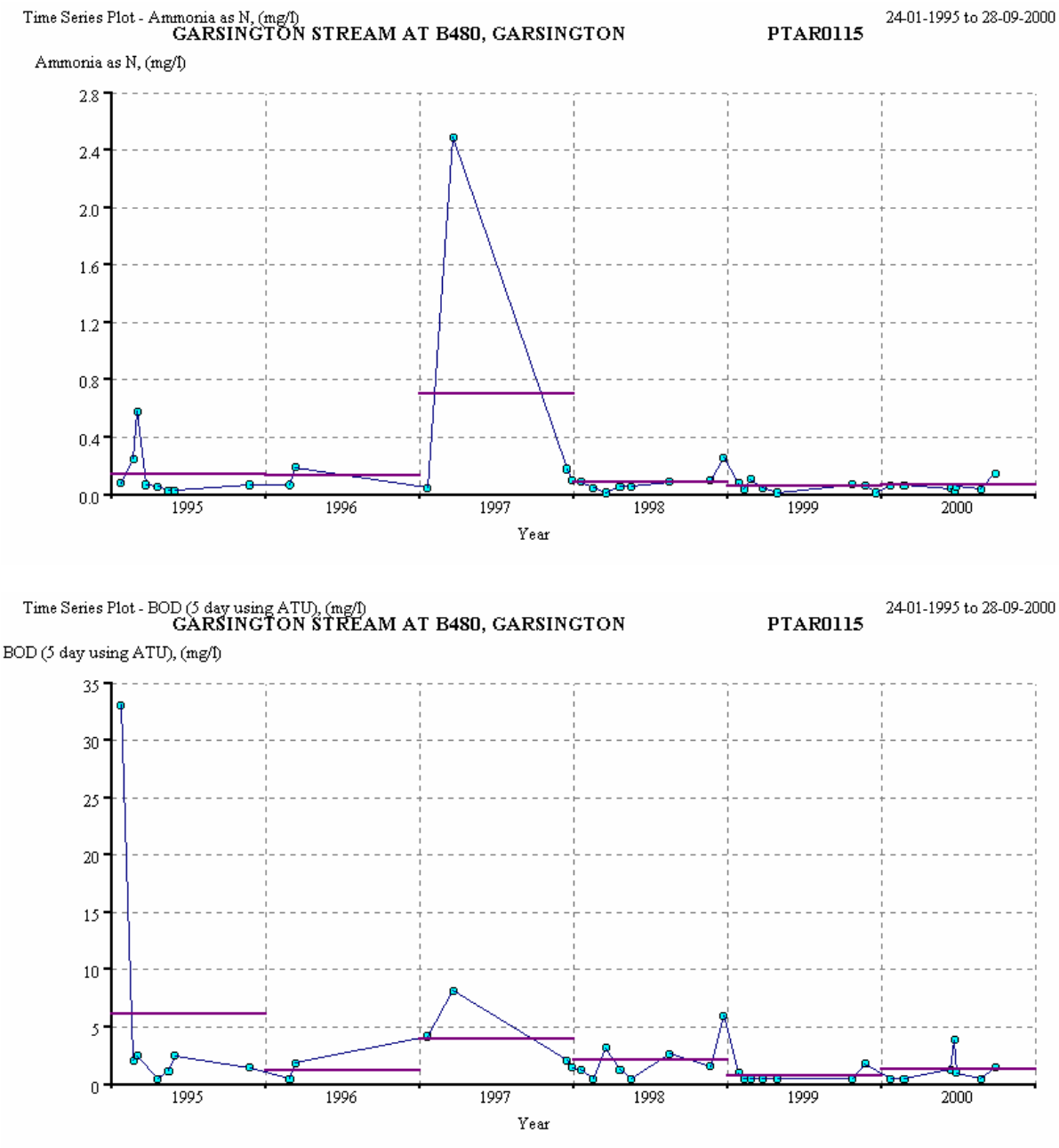


Figure 8.7 Ammonia and BOD time series for Site PTAR0115

8.3.6 Site PWER0024

For this site there is a relatively strong relationship between DO% and log(flow), as shown in Figure 8.8. In addition, there is again an indication of the 'threshold' effect that we noted earlier for Site PCHR0022, whereby low DOs only occur when flows are below a particular critical value. In this instance, the threshold seems to be when flow is about $10^{1.8}$, or 63 cumecs.

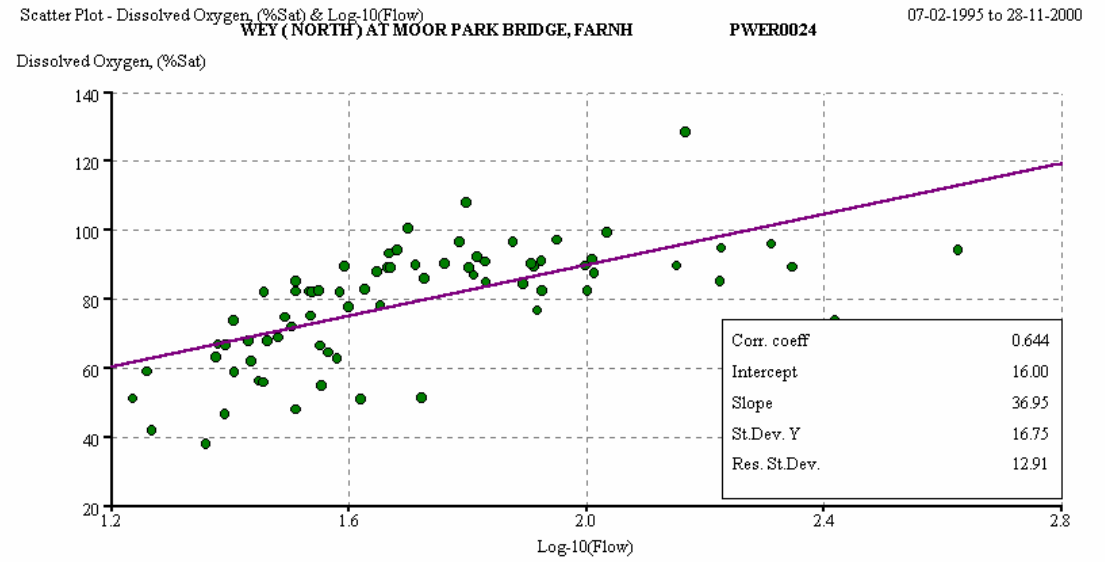


Figure 8.8 DO% v. flow relationship for Site PWER0024

We can see more clearly how this operates by looking at the two time series shown in Figure 8.9. The two years with the lowest DOs seem to be 1996 and 1997, and these are also the years with the most prolonged periods of low flow. At the other extreme, the customary summer trough in DO% is completely absent in 2000, and this is the year of highest flows.

For this site, therefore, the evidence of a threshold DO% v. flow effect is rather convincing.

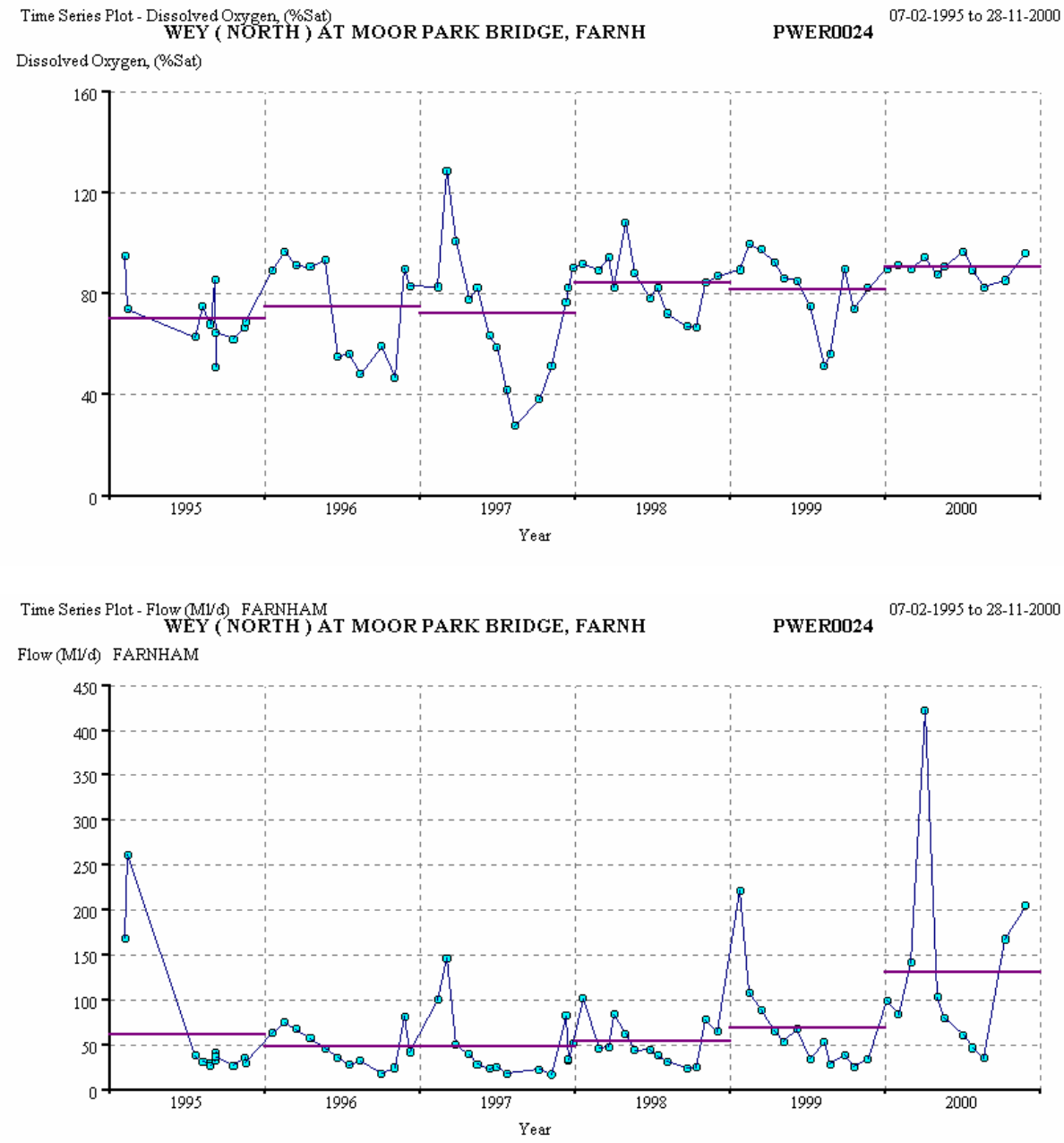


Figure 8.9 Time series of DO% and flow for Site PWER0024

8.3.7 Site PWER0089

The main driver for this site is BOD - and the poor quality in 1995-97 was primarily caused by a BOD outlier of 44 mg/l in 1996. The only hint of a BOD v. flow relationship is that BOD tends to be better at very high flows. However, this is of no relevance to GQA class. For flow to have a bearing on the behaviour of 90%ile BOD, we would need to see an association between certain conditions of flow and *poor* BOD quality.

For ammonia, quality was very bad in 1995, but greatly improved thereafter. There is no evidence here of a flow association.

Finally we looked at DO% - for which there had been an increase in class from E to B over the two periods. The relationship with log(flow) is shown in Figure 8.10. As before, the strength of the linear relationship is only modest (R=0.40) - but once again there is evidence of a threshold effect.

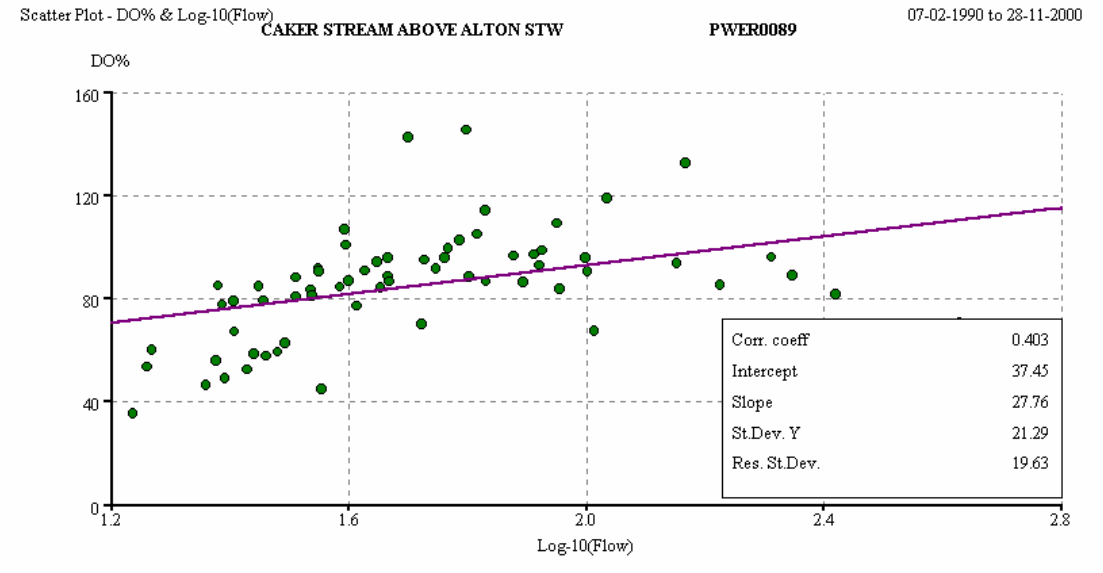


Figure 8.10 DO% v. flow relationship for Site PWER0089

8.3.8 Site PRGR0119

DO% is the critical determinand at this site, with DO% class improving from E in 1995-97 to B in 1998-2000. In this instance, however, there is no discernible association with flow. The R value between DO% and log(flow) is only 0.05, and the plot (not shown) shows little sign of the threshold effect seen at several other sites. This is reinforced by Figure 8.11, which shows that mean flows are generally very similar over the six years, whilst there is a marked drop in mean DO% over 1996-97.

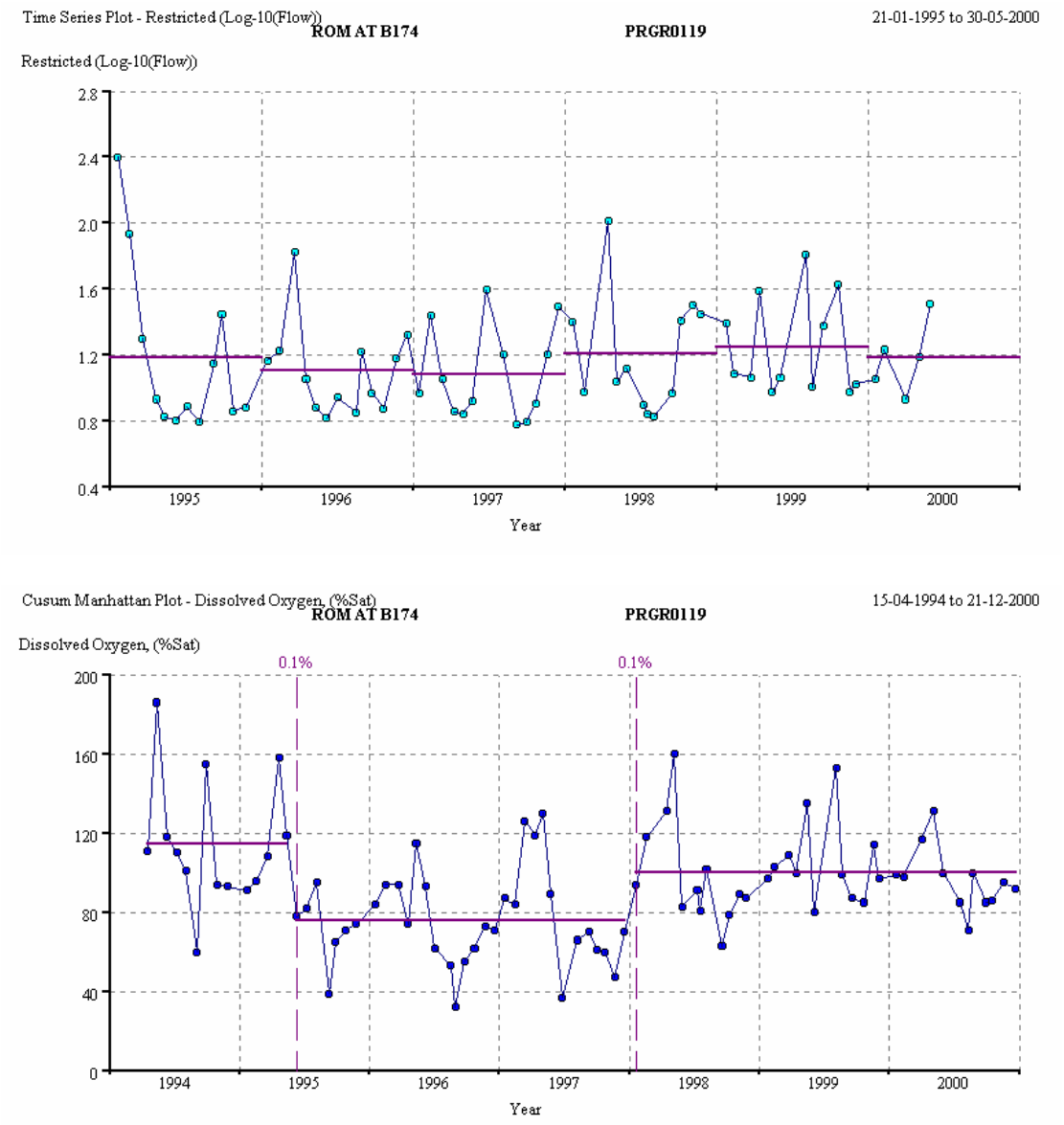
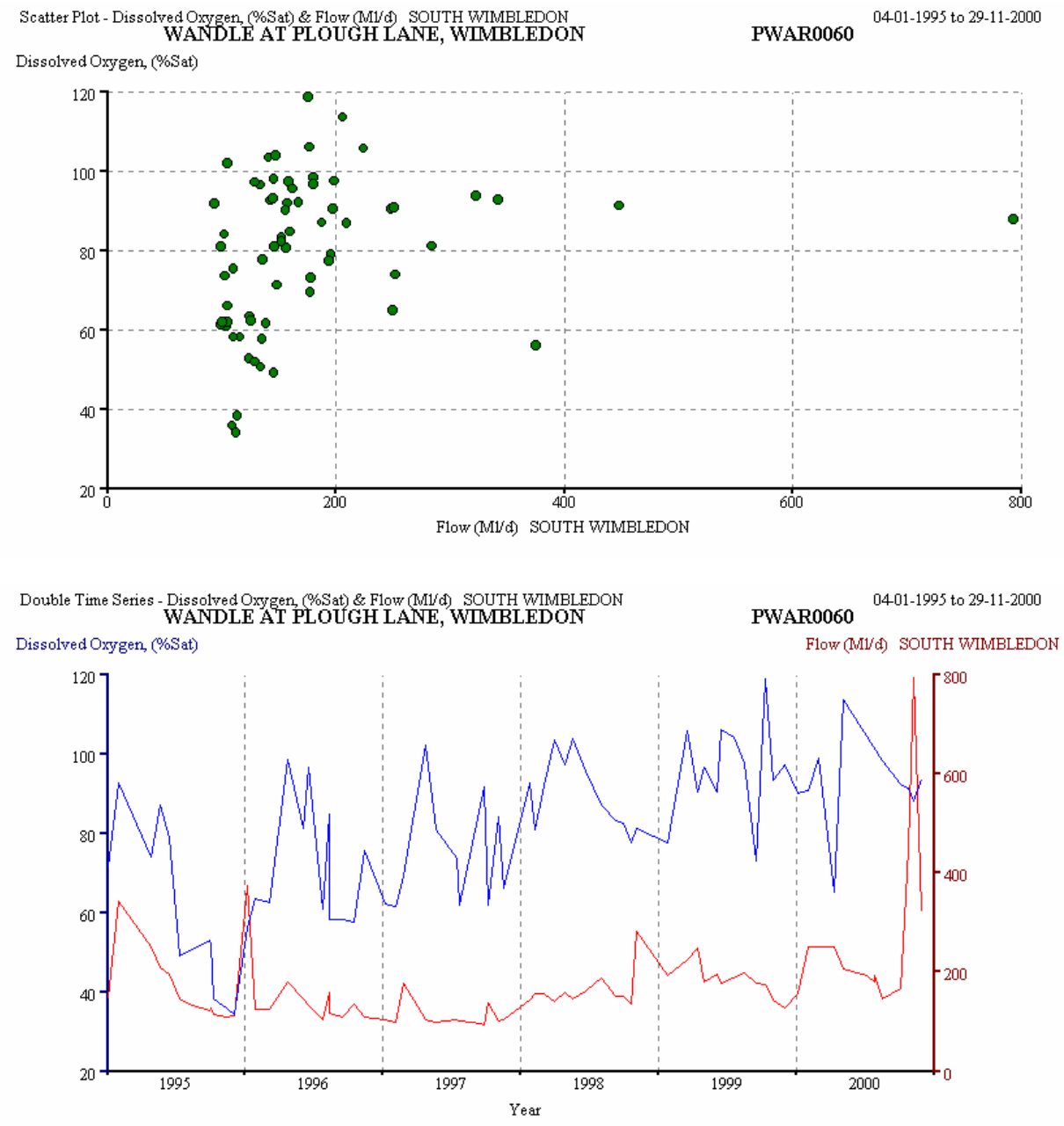


Figure 8.11 Time series of DO% and flow for Site PRGR0119

8.3.9 Site PWAR0060

Both ammonia and DO% improve from class E to class B at this site. We look first at DO%. The relationship between DO% and flow is shown in Figure 8.12. The first plot shows that, although overall the correlation is very weak, there is again evidence of a threshold effect. However, we see from the double time series in the second plot that the very low DO% values occurred only in 1995. There were periods of 1996 and 1997 when flows were just as low as they were in 1995, and yet DO% showed no tendency to dip below 60% sat. This suggests that, if low flows do have a causal effect on DO%, perhaps this manifests itself only when the low flows occur in conjunction with some additional influence (e.g. a weather-related factor).



Note: Flow is the lower of the two time series

Figure 8.12 Plots showing the DO% v. flow association at Site PWAR0060

The corresponding plots for ammonia are presented in Figure 8.13. The top plot shows that there is no tendency for the worst ammonia concentrations to occur at very low flows. Overall, the association between log(ammonia) and log(flow) is very weak - and, for the purpose of showing that quality improves with increasing flow, it is actually in the wrong direction!

The second plot illustrates the dramatic improvement that occurred in ammonia concentrations between the first and second 3-year periods - an improvement quite disproportionate to the slight increase seen in mean flow between the two periods.

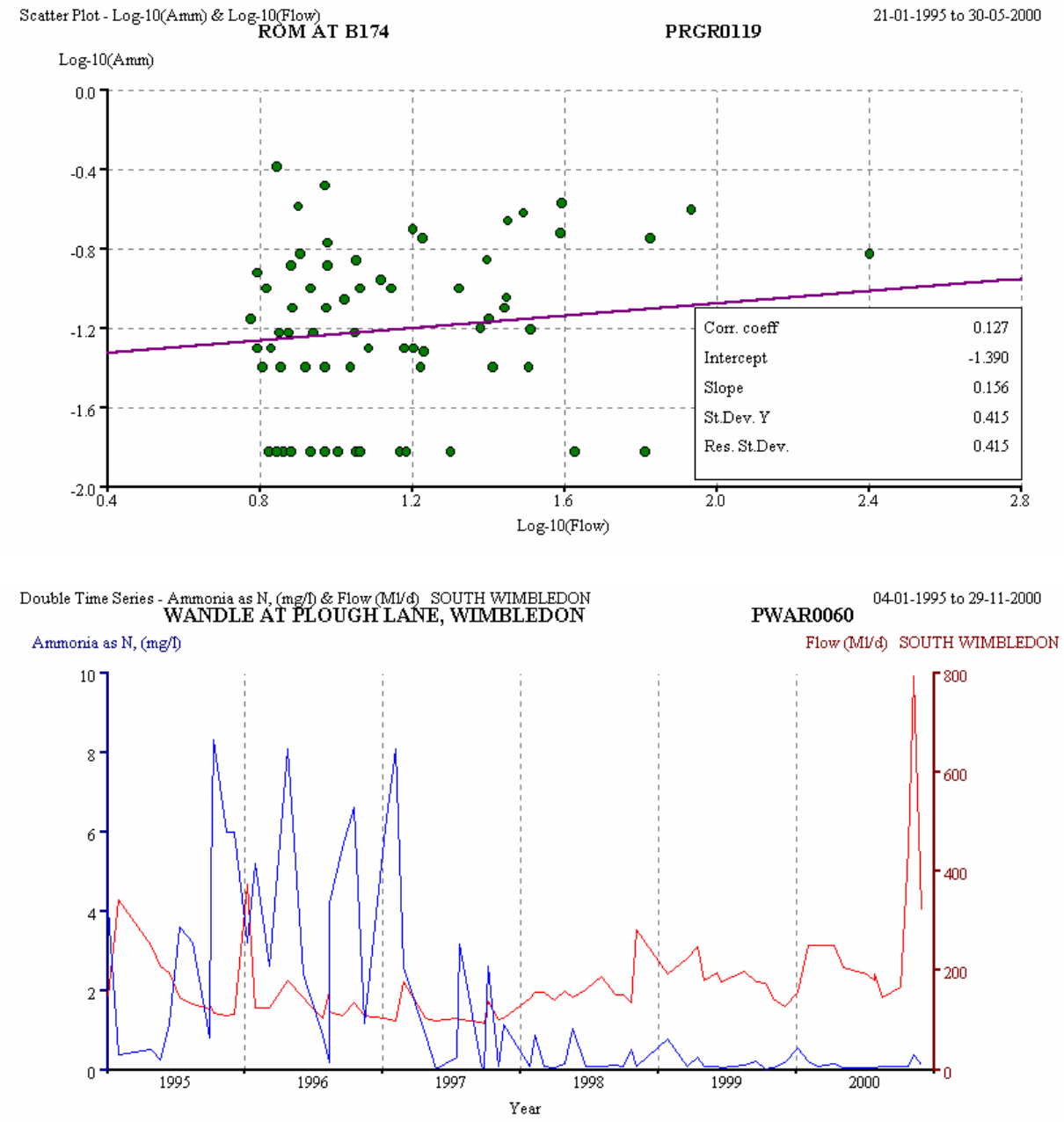


Figure 8.13 Plots showing the Ammonia v. flow association at Site PWAR0060

8.3.10 Site PRGR0011

The very poor quality seen in 1995-97 was entirely due to outliers of 273 and 47 mg/l early in 1997. These did not occur at unusually low flows. Moreover, no relationship at all was seen between BOD and flow for the remaining data.

8.3.11 Summary

Of the nine sites examined in detail, we can summarise the findings as follows (with the numbers of sites shown in brackets):

- missing flow records (1)
- a BOD improvement partially explained by flow (1)
- a low flow 'DO% threshold' effect (3)
- no DO% v. flow effect (1)
- no consistent DO% v. flow effect; and no ammonia v. flow effect (1)
- improvement due to presence of one or more outliers in first 3-year period (2)

Overall, therefore, clear evidence of a quality v. flow association emerged for only three of the nine sites. This is a rather low proportion - especially when we remember that these are the nine sites that showed the most dramatic improvement between 1995-97 and 1998-2000.

Nevertheless, for each site that did show an association with flow, this took the form of a 'DO% threshold' effect. This suggests that it might be worth searching systematically through the full data set for more widespread evidence of 'discontinuous' effects of this sort.

The existence of discontinuous effects in the case of DO% in particular would not be surprising, given that dissolved oxygen levels in rivers are subject to a range of phenomena – some potentially opposing in their effects – resulting from low flow. (For example, low flow could result in enhanced algal growth leading to enhanced photosynthetic activity and elevated oxygen levels in daytime – but also to enhanced respiration demand and lower oxygen at night. Again, low flows could also lead to increased demand for available dissolved oxygen, resulting from higher BOD concentrations consequent upon reduced dilution of effluents and run-off.)

9. SUMMARY OF FINDINGS

9.1 Introduction

Given the inevitably discursive nature of the detailed discussions of results in Chapters 5-8, we summarise here the principal findings for the convenience of the reader, and as a backdrop to the conclusions and recommendations which follow in Chapter 10. The order of the summary follows that of Chapters 5-8.

9.2 Low-Level Results

9.2.1 Quality v. flow relationships for DO%

- For the majority of sites - 422 out of the 565, or 75% - there was no statistically significant relationship between DO% and same-day flow. In only 4 cases out of 565 did the correlation coefficient (R) value exceed 0.4 - a value at which the standard deviation of the scatter around the model is barely 8% smaller than the original overall standard deviation, so the model would have no useful predictive value.
- Logging both variables increased the number of sites with statistically significant R values to nearly half. However, the highest R value was only 0.6, so that the uncertainty in DO% predictions would be at best 20% narrower than that from simply using the overall mean as the prediction.
- Whilst effects were weak, the majority (81%) of the significant R values were positive - that is, increasing flow tended to be associated with increasing DO%.
- Many antecedent flow measures were investigated but, even with the optimal one for each site, the increase in R was trivial or zero in the great majority of cases.

9.2.2 Quality v. flow relationships for BOD

- For about three-quarters of the sites there was no statistically significant relationship between BOD and flow. With a log-log model, this improved slightly: the proportion of non-significant sites fell from 72% to 63%.
- For sites with a statistically significant R value, the form of relationship was less clear-cut than with DO%. Increases in flow tended to be associated with decreases in BOD at about 2 sites out of 3, with the contrary at the remainder. Thus, unequivocal improvement in GQA class with increasing mean flow should not be expected where BOD is the class-critical determinand.
- As with DO%, the improvement gained from using antecedent flow measures was, in almost all cases, extremely slight or non-existent.

9.2.3 Quality v. flow relationships for ammonia - models using same-day flow

- The picture was very similar to DO% and BOD, with no statistically significant relationship between ammonia and flow at three-quarters of the sites.
- With logged data, the proportion of sites with statistically significant R values increased from a quarter to nearly half of the sites, but R values were again below 0.6. At 80% of those sites the relationship was positive (increased flow associated with increased ammonia), so an increase in flow would tend to have a potentially harmful effect on GQA class where ammonia is the class-critical determinand.
- Again, improvements gained from antecedent flow measures were negligible.

9.2.4 Adequacy and temporal stability of models

- Despite their poor predictive power, the log-log models provided an adequate representation of the relationship over all flows, in the great majority of cases.
- A step-change analysis of the model residuals found that there were commonly several different mean levels of the three determinands over the period, showing that there were time trends in quality which were *unrelated to flow*.

9.2.5 Predictive capability of models

- An analysis was conducted to confirm the generally weak predictive capability of the identified low-level models. It was shown that, for all three determinands, the predicted effect of a 20% increase in mean flow was commonly no greater than a tenth of a class-width, and almost never more than a quarter of a class-width. The models were therefore unable to account for more than a very small proportion of observed class-changes.

9.2.6 Quality v. AMP relationships

- For the 50 ‘AMP’ sites, time trends in BOD over an 11-year period were identified by cusum analysis. There was only suggestive evidence of a greater frequency of improvements - or a smaller frequency of deteriorations - during the ‘AMP scheme completion year’ than in other years. For ammonia and DO%, evidence for a positive effect of AMP activity was even weaker. Overall, therefore, the AMP schemes had no detectable effect over these 50 sites.
- Analysis of results for the 50 ‘Control’ sites was in striking agreement with that for the 50 AMP sites, suggesting that even the modest improvements seen for the AMP sites during the AMP scheme completion year could plausibly be attributed to a general improvement in river quality in the late 1990s. However, various reasons were noted for why the analysis failed to detect an effect of AMP improvements - such as imperfectly known completion dates, improvements that were gradual rather than step-change, and investments to protect against potential deterioration rather than to ameliorate existing problems.
- Both groups of sites showed similar relative proportions of DO% and BOD improvements to deteriorations for the 10 years outside the AMP/Control window. For ammonia,

however, the proportions were very different: at the Control sites there were over four times as many improvements as deteriorations, whereas at the AMP sites there were fewer than double. This has no bearing on the assessment of AMP improvements described above, but the increased propensity for ammonia deteriorations at AMP sites provides an insight into the reasons for such sites having being selected for AMP schemes in the first place.

9.2.7 Results for Regions other than Thames

- River quality and flow data was analysed for eight sites (2 Anglian, 3 Southern and 3 Midlands). Ammonia showed a consistently stronger association with flow than did DO% or BOD, with the relationship being negative for the Southern Region sites, and positive for sites in the other Regions. However, the highest R value for any of the three determinands was only 0.47, and generally the R values followed a similar scatter to that seen in the low-level analyses of the Thames data.

9.3 High-Level Results

9.3.1 Summary of GQA class data

- Analyses of the spread and variability of GQA classes seen at Thames Region sites showed that GQA class changed quite markedly at many sites over the period. Thus, for example, the class remained absolutely constant at only 27 of the 565 sites, and ranged over 3 or more classes at well over half of the sites.

- GQA trend analysis gave statistical confirmation of the widespread improvement that occurred in river quality in Thames Region over the past 20 years – but did not of course provide any clue as to what might have driven that improvement.

9.3.2 Summary of 'Flow Fingerprint' measures

- Some 25 measures were calculated and subjected to an ANoVA. All showed significant inter-site differences (at the $P<0.001$ level) but those with overwhelmingly the biggest differences (relative to their variation from one five-year period to another) were **log(mean flow)**, followed by **log(mean summer flow)**. This was expected, as both measures are directly related to river size.
- All the other measures were independent of the scale of the river, and intrinsically more interesting. The next four statistics (in terms of reflection of inter-site variability) were the lag-1, lag-15 and lag-30 autocorrelation coefficients, together with the relative standard deviation - showing that both the degree of persistence and the relative variability of flow can vary substantially from river to river.

9.3.3 Trends in flow

- The more important question addressed by the ANoVA concerned the variations of the various 'Flow Fingerprint' measures through time. The outcome was surprising, in that 11 of the measures showed a statistically significant trend across all sites.
- The analysis revealed a dramatic picture of Thames-wide trends in flow characteristics over the past 20 years, the main messages being as follows:
  - Both in terms of year-to-year and day-to-day variability, flow was much more variable in the 1990s than in the 1980s.
  - Mean summer flow fell markedly over each of the last three 5-year periods - by 7%, 16% and 11% respectively.
  - Mean annual flow has fallen steadily over each of the last three 5-year periods - by 6%, 5% and 6%.
  - Increases in several measures of autocorrelation show flow being more persistent in the 1990s than in the 1980s.
  - The 95%ile:median ratio increased steadily from 4.8 to 7.1, and increased skewness in flow was also borne out by an increase in the mean:median ratio.
  - Flow gauging was consistently more reliable in the 1990s than in the 1980s.

9.3.4 Characterisation of sites by their flow variability

- As noted above, all the 'Flow Fingerprint' measures showed statistically significant differences between sites; in particular, therefore, the variability of flow at a site itself varies from site to site. This was potentially helpful to the investigation, inasmuch as greater variability gives greater scope for any effect of flow on quality to manifest itself. (Sites with little flow variation could still be useful, but only through demonstrating that stable-flow sites tend to be those with limited variation in GQA class.)
- Flow variability was therefore characterised across the 116 gauging stations (analogously to assessment of GQA class variability), with the following findings:
  - The **Coefficient of Variation (CoV) of daily mean flow** over the 116 flow sites varied from 0.1 to 2.7, with a median value of 1.1.
  - The **CoV of annual mean flow** values were of course substantially smaller, as all within-year variation is smoothed out by calculating annual averages; they ranged from 0.07 to 1.6, with a median of 0.32.
  - The **CoV of non-overlapping 3-year mean flow** gave a median value of 0.12 - 2.6 times smaller than the median value of the **CoV of annual mean flow** (0.32). If annual mean flow CoVs were independent, the factor would be 1.73, so the smoothing effect of averaging over three years was somewhat stronger than expected, suggesting a tendency for low-flow and high-flow years to be more evenly mixed than would occur with a purely random process.

9.3.5 GQA v. flow relationships

Classification of sites by their potential to reveal a quality v. flow relationship

- A GQA class variability index was used described the quality variation, and the CoV of (non-overlapping) three-yearly mean flow used to described the flow variation. About one third of sites had a CoV > 0.15, the criterion used to identify 'higher flow variability' sites.
- A plot of 3-yearly flow CoV against the class variability index revealed no clear correlation. A two-way table showed the proportion of higher flow variability sites to be below average (24%) for sites with constant GQA class, and above average (40%) for the sites with the greatest variation in GQA class. However, a test showed the difference to be only marginally significant (P = 0.06). Thus the extent to which GQA class varies at a site seems to be largely unrelated to the amount of variation in mean flow at that site.

Results for 3-year GQA class

- Modelling of the 3-year GQA class v. 'Flow Fingerprint' measures revealed no tendency for sites with more variable flows to have a greater proportion of significant quality v. flow models, reinforcing previous findings. Overall, the high-level modelling produced a very low proportion of statistically significant result. The useful point did emerge, however, that log mean flow was almost the best-performing Flow fingerprint measure, indicating that the more complex measures could be discarded.

Results for 1-year GQA class

- In view of the above, modelling of variations in 1-year GQA class was restricted to the single explanatory variable log (annual mean flow). The number of significant relationships rose from 33 (3-year GQA modelling) to 80 - or 20% of all sites for which GQA variation was seen. The R value for all but two of these was negative, indicating that *increase* in mean flow is associated with an *improvement* in GQA class.

- However, the predictive capability of the models was generally very weak. For the great majority of sites the predicted improvement was between 0.2 and 0.6 of a class, whilst improvements more marked than this were predicted for just 6 sites. It was concluded that the 1-year GQA v. mean flow models identified here were too weak to account for more than a very small proportion of the observed 1-year GQA class changes.

9.4 Regional-Level Results

9.4.1 Sources of data

- In this set of analyses, data was aggregated across time, determinands and sites. The work explored two Agency discussion papers: 'Paper A' supplied by Simon Bingham (which used summary data from all Regions to provide 'broad-brush' evidence for a quality v. flow effect), and 'Paper B', an internal discussion document by Juliane Struve (which focused on summary measures of quality and flow for one area of Thames Region).

9.4.2 Paper A

- This used rolling 3-year data from 1992/94 to 1998/2000 on total river length in each class, for each of the eight Regions. The statistic used to track Region-wide GQA change was 'Net % improvement in class (by length) compared to 1988/90'. For each Region, flow data was obtained from a representative gauging station, and the statistic used to track Region-wide change was 'Change in 3-year mean flow as % of mean flow for 1988/90'.
- A plot in Paper A certainly seemed to show a fairly strong association between % GQA change and % flow change. However, because it was based on rolling 3-year data some of the association would have been induced by autocorrelation.
- To explore this, results were examined by Region, with the regression analysis using just the three *non-overlapping* points. The Thames R value was 0.9999, indicating a statistically significant association. For the other Regions, only the R values for North East (0.98) and Anglian (0.97) were high - and even these were not statistically significant, being based on just 3 points. Other R values ranged from 0.74 to -0.29.
- The eight 3-point Regional graphs were pooled, giving an R value of 0.75 (based on 24 points). This was highly significant ($P < 0.001$), and so the data did after all support the hypothesis of an association between increased flow and improved GQA class - but only when all Regional plots were aggregated in order to increase the statistical power of the analysis.

9.4.3 Paper B

- This looked at flow and quality in the North-East Area of Thames Region, focusing on 1988 - 1998, and in particular on the drought from winter 1995/96 to winter 1996/98. GQA results were obtained for all rivers in the Area for each of the nine rolling 3-year periods from 1988-90 to 1996-98, and for each site, flow records were obtained for the nearest gauging station and day-by-day rolling 3-year averages were calculated.
- Plotting the relative frequencies of the GQA classes in each period showed that the proportion in classes A-C was lowest in 1996-98. The effect was particularly marked for the proportions of rivers in classes A and C.
- Flow was plotted as day-by-day rolling 3-year flow averages. To remove the effect of scale, the flows were re-plotted as proportional deviations from each river's mean flow.
- Examination of changes in GQA class and 3-yearly mean flow showed a positive correlation between the 'A-C to D-F' ratio and the relative deviation of 3-yearly mean flow from grand mean flow over the 11-year period. Thus, periods in which the proportion of sites in classes A-C was higher than usual tended to be associated with periods in which mean flow was higher than usual. It was concluded that 'variations in flow are partly responsible for shifts in the dominance of A-C reaches over C-D reaches'.
- The R value for the association was 0.84, which - had it been based on 9 *independent* points - would have been highly significant ($P<0.01$). However, the plot shows rolling 3-year data, and so - as in Paper A - it contains less information than it appears to because of the autocorrelation between successive years.

- To correct for this, the R value was calculated for the data split up into three separate non-overlapping groups. For the first two groups, starting in 1988-90 and 1989-91, the resulting R values were 0.83 and 0.60 - substantially below the value of 0.988 necessary to achieve even mild significance ($P<0.10$).
- For the group starting in 1990-92, however, the R value was 0.9999, which is highly significant ($P<0.01$). This conclusion hinges largely on the very good GQA performance (and high flows) in 1993-95 coupled with the poor GQA results (and low flows) of 1996-98. It is also in good agreement with the conclusion reached from the results presented in Paper A for Thames Region as a whole.

9.5 Further Summarising Analysis

- The Regional-level analyses suggested that the association between GQA class and mean flow was strongest in the mid-to-late 1990s. This view was supported by the high-level analysis results: from 1995-97 and 1998-2000, only about 40 of the 565 sites showed a decline in GQA, whilst nearly 300 experienced an improvement.
- The possibility of widespread change in the structure of quality v. flow models in the mid 1990s was investigated by restricting analyses to the period 1995-2000, and comparing the R values with those obtained from the full data set (1980-2000). This showed no clear evidence that the relatively wide timespan of 1980-2000 had prevented stronger but more recent relationships being identified.
- A simple hypothetical example was used to show that a strong high-level association is no guarantee of a relationship at the level of individual data points. Nevertheless, in an further attempt to determine why the low-level modelling did not produce useful relationships, data for the nine sites showing the greatest GQA improvement was examined in detail (for all but one, the improvement in class was from E to B).
- For the nine sites examined the findings were as follows:
  - Missing flow records (1 site).
  - A BOD improvement partially explained by flow (1 site).
  - A low flow ‘DO% threshold’ effect (3 sites).
  - No DO% v. flow effect (1 site).
  - No consistent DO% v. flow effect, and no ammonia v. flow effect (1 site).
  - Improvement due to one or more outliers in first 3-year period (2 sites).
- Clear evidence of a quality v. flow association emerged for only 3 of the 9 sites - a rather low proportion, given that these are the nine sites with the most dramatic GQA improvement between 1995-97 and 1998-2000. However, for each of these 3 sites the association with flow appeared to take the form of a ‘DO% threshold’ effect, which suggests a suitable starting point for any further work in this area.

10. CONCLUSIONS AND RECOMMENDATIONS

10.1 Conclusions

- Significant correlations between individual GQA monitoring results (DO%, BOD and ammonia) and flow are found at only about one-half of Thames Region sites, and are weak, providing no convincing evidence of an effect of flow on GQA results. The majority of the significant correlations are positive for DO% and ammonia, but less clear-cut for BOD (negative at about 2 in 3 sites, and positive at the remainder).
- The predictive capability of these models is too weak to account for more than a very small proportion of the observed changes in GQA class.
- GQA monitoring data gives only suggestive evidence of a greater number of BOD improvements - or smaller number of deteriorations - during the AMP scheme completion years, with even weaker evidence for a positive effect of AMP activity on ammonia and DO%. However, there may be reasons why the analysis would not be expected to detect an effect of AMP improvements.
- GQA class changed quite markedly at many Thames Region sites over the past 20 years, and a widespread improvement has occurred in river quality in the Region over the period.
- Across the Thames Region, flow persistence and relative variability can vary substantially from river to river. Flow was much more variable in the 1990s than in the 1980s, mean summer and annual flows fell over each of the last three 5-year periods, and flow was more persistent in the 1990s than in the 1980s.
- The extent to which GQA class varies at Thames Region sites appears to be largely unrelated to the amount of variation in mean flow at that site, and sites with more variable flow do not have a greater proportion of significant quality v. flow models.
- Only 1 in 5 sites show significant associations between 1-year GQA class and log mean flow, although virtually all of those that are significant indicate that increase in mean flow is associated with an improvement in GQA class. However, the predictive capability of these models (as with the low-level models) is very weak.
- Previous Agency examinations of (a) aggregated data for each of the eight Regions, and (b) aggregated data for the North-East Area of Thames Region appeared to show quite strong associations between GQA change and flow, but were compromised by the presence of strong autocorrelation arising from the use of 3-year rolling GQA results. When analysed with correction for autocorrelation:
  - Paper (a) shows a significant GQA/flow association for only Thames Region individually, but also a highly significant association when data is pooled across all Regions.
  - Paper (b) shows no significant GQA/flow associations for two sub-groups of the data, starting in 1988-90 and 1989-91, but a highly significant correlation for the group starting in 1990-92 (primarily because of very good GQA performance and high flows in 1993-95, coupled with the poor GQA results and low flows in 1996-98).
- Although any association between GQA class and flow appears to have been strongest in the mid-to-late 1990s, there was no clear evidence that using data from 1980 to 2000

prevented identification of stronger, but more recent, relationships within the data from the 565 Thames Region sites.

- The absence of any clear evidence of strong relationships at the level of individual data points is not inconsistent with a strong high-level GQA/flow association. However, it obviously severely hampers attempts to explain, and justify any projected use of, such an association.
- No consistent pattern was discerned in the data for the nine Thames Region sites showing the greatest GQA improvement which were examined in detail in a final attempt to obtain more useful information from the individual site data. However, in three cases there was some evidence of a relationship between DO% and flow, in the form of a threshold-type mechanism, whereby the risk of obtaining low DOs increases sharply once flows have dropped below some critical value.
- In summary, it is concluded that an association between GQA and flow does exist at some sites, but that it is not readily discernible at the individual determinand and site levels, for several likely reasons:
  - The limitations of monthly GQA data, particularly in relation to relatively infrequent events of limited duration (e.g. the postulated 'DO% threshold' effect).
  - The complexities of behaviour of the GQA component determinands. This is especially likely for dissolved oxygen, which may be subject to a range of phenomena - some potentially opposing in their effects - resulting from low flow.
- In particular, the project has found no convincing evidence to support previous suggestions by the Agency that a deterioration in national GQA results could mainly be attributed to low flows.

10.2 Recommendations

1) Broad evidence of some form of GQA v. flow relationship has been obtained by the analysis of aggregated (i.e. Regional-level) data, but not by the analysis of either GQA data (high level) or individual determinand data (low-level) at site level. One obvious way of developing the existing analysis of aggregated data would be to repeat it for smaller groups of sites than whole Regions. Such 'Areal-level' analysis would allow more representative flow data to be used with the aggregated GQA data, and would also give the opportunity for differences to be seen in the strength of association within Regions as well as between Regions.
2) Such further analysis of aggregated data at finer geographical levels would not, however, provide an explanation of the association, and one way to tackle that need would be to conduct a detailed search for patterns of behaviour such as the low flow v. low DO% 'threshold' effect postulated here. (Whilst this should focus primarily on DO, it could also be combined with exploratory searches for potentially relevant patterns in BOD and ammonia data.) Such a search would need to be automated, using suitable software and statistical tools, given the numbers of sites involved.
3) Finally, given the limitations of 'once per month' GQA data in relation to events that are likely to be relatively infrequent or transient, consideration could be given to examination of river quality records involving more frequent sampling (where these exist), especially in relation to DO. Thus, for example, records of continuous DO monitoring could be

examined, if they are available for sites which show substantial GQA and flow variations. Such investigation could, conceivably, be combined with modelling of DO behaviour using available data on flow, BOD and algal populations.

Given the results to date, and the needs of the Agency, item 2) would seem the most potentially cost effective and fruitful, and is therefore recommended.

11. REFERENCES

Ellis, J.C. (1989) Handbook on the Design and Interpretation of Monitoring Programmes. WRc Report NS29.

Ellis, J.C., van Dijk, P.A.H., and Kinley, R.D. (1993) Codes of Practice for Data Handling. NRA Report No. R&D 241.

APPENDIX A TESTING A QUALITY V. FLOW MODEL FOR ADEQUACY AND STABILITY

A.1 MODEL RESIDUALS

In testing how well-behaved a model is, the approach is based on calculating the 'residuals' from the model - that is, the deviations between actual and fitted concentrations - and then testing these for randomness in various ways.

Suppose we label the residuals $Resid_T$ - the 'T' denoting the fact that, by virtue of the input quality data being ordered through time, the residuals are themselves in time order. If we take a copy of these and sort them into increasing flow order, we produce a new set of values, $Resid_F$.

We then carry out a cusum analysis on the $Resid_F$ series. If the model fits the data satisfactorily at all levels of flow, we would expect the mean of the residuals to remain close to zero. If, however, the cusum identifies one or more step changes in mean, this indicates that the model is tending to under-predict quality at some flow bands and or over-predict it at others.

We also carry out a cusum analysis on the $Resid_T$ series. Here the interpretation runs as follows. The $Resid_T$ values can be seen as showing the variations in quality *after removing the effects of flow*. Thus, if there are no trends in quality other than those induced by trends in flow, we would expect the mean of the residuals to remain close to zero over all years. It follows, therefore, that if the cusum does identify step changes in the mean residual, this is evidence of time trends in quality *unrelated to changes in flow*.

We illustrate the methodology below using the example of orthophosphate ('OrthoP') in the River Thames at Eysey.

A.2 MODEL ADEQUACY

Figure A.1 shows an example of a model that clearly provides a poor fit (even though, for a *valid* linear model, a correlation coefficient of -0.43 would be statistically highly significant). The cusum analysis of the flow-sorted residuals produces the result shown in Figure A.2. It is evident that, for the great majority of the flow range, the residuals are persistently negative (i.e. below the solid zero line); whilst to compensate for these there are clumps of pronouncedly positive residuals at either end of the plot (i.e. at very low and very high flows). These findings - that the model badly under-estimates quality at either end of the plot and over-estimates quality at middling flows - are of course obvious at a glance from Figure A.1. The point of the cusum analysis, however, is that it produces its conclusions automatically *without the need for visual inspection*; and this is important when there are three determinands at over 500 sites to be analysed.

In this particular example, the problem is readily solved by fitting the model in the 'log' world - that is, regressing log(OrthoP) against log(Flow) rather than OrthoP against Flow - as shown in Figure A.3 This was confirmed by a cusum analysis of $Resid_F$, which found no step changes in mean.

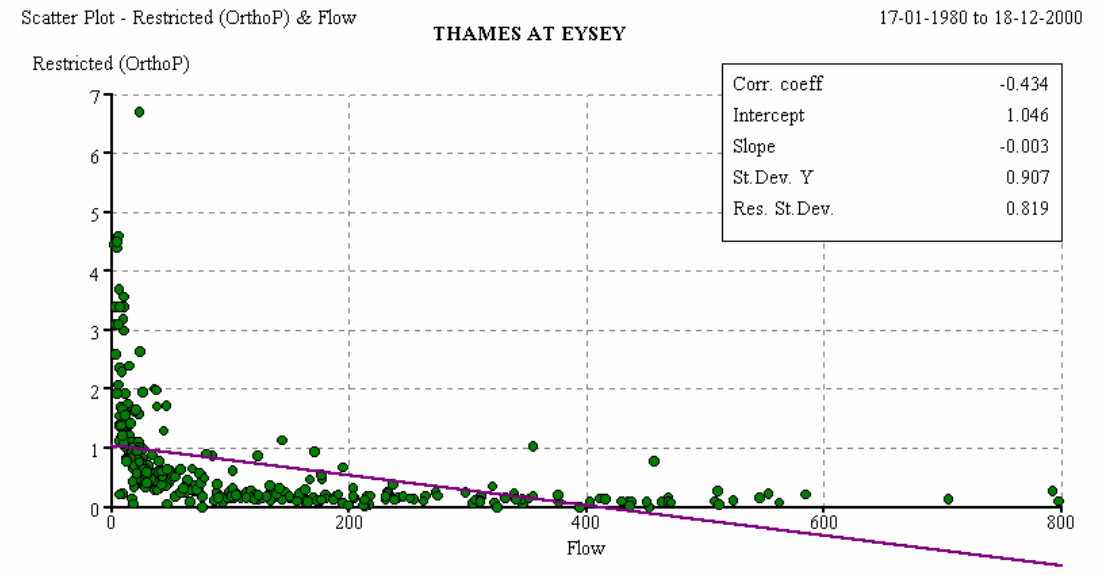


Figure A.1 Example of an inadequate model

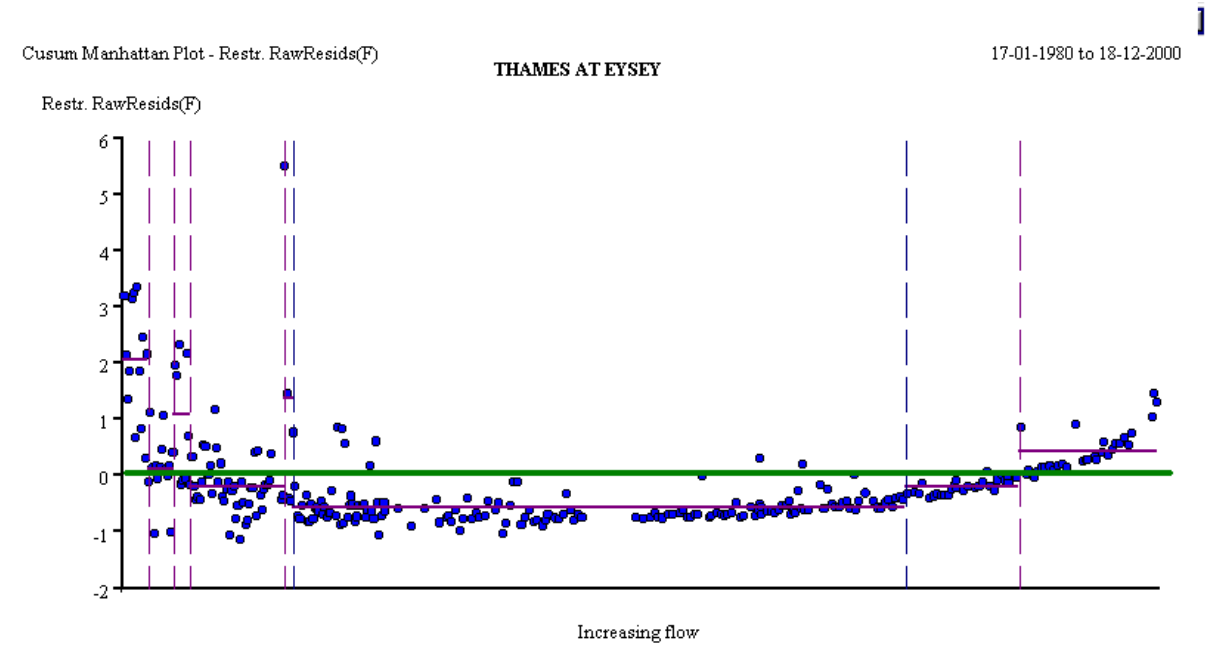


Figure A.2 Result of cusum analysis of flow-sorted residuals from Figure A.1 model

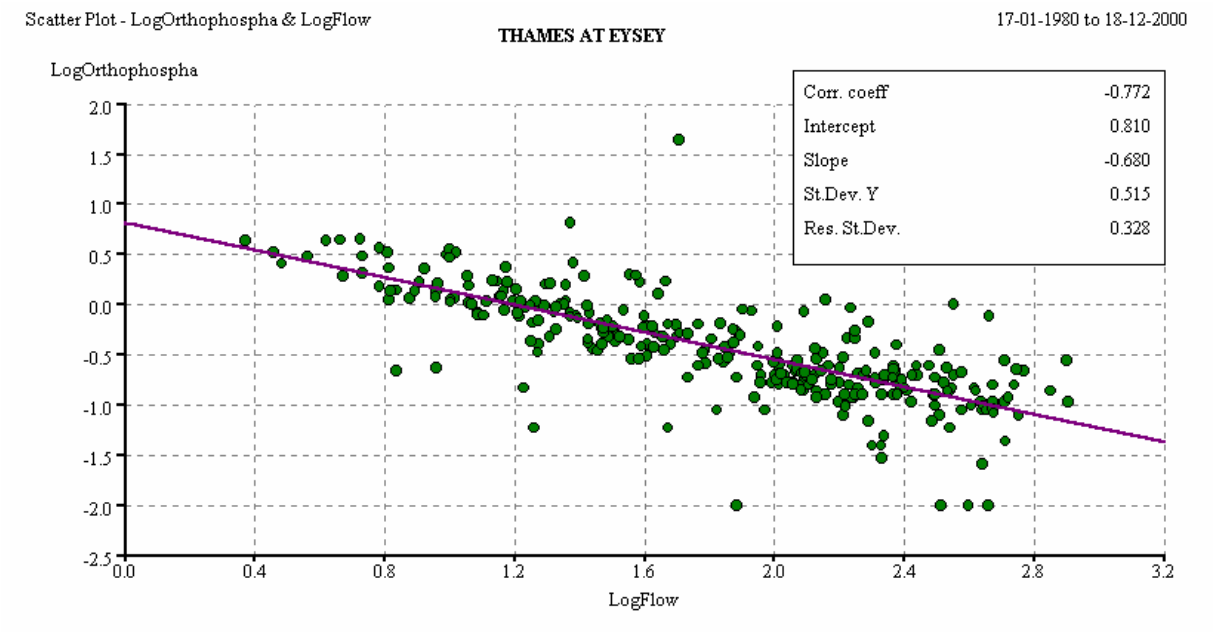


Figure A.3 Example of an adequate model (using the same data as in Figure A.1)

There does nevertheless remain the possibility that if we were to break the time series down into shorter periods and fit a separate model to each, we would find marked changes in the slope of the quality v. flow model. To test this we have added a further refinement to the analysis. After fitting the overall model and calculating the residuals, we split these into four equal series of length N/4 (each typically spanning about 5 years), and *then* sort each of these separately in flow order. If, over any of these shorter periods, the best local model should happen to be steeper or shallower than the overall model, this would make that particular subset of the residuals from the *overall* model autocorrelated. We therefore calculate the lag-1 autocorrelation coefficients (AC1) for the four N/4 sets of residuals; and Consequently, an *absence* of statistically significant AC1 values would indicate that the slope of the flow model was broadly the same over the whole time period.

A.3 MODEL STABILITY THROUGH TIME

Just as the flow-sorted residuals are used to test the adequacy of the model across the whole *flow* range, so can the time-based residuals are used to test whether or not the model has drifted through *time*. An example is given in Figure A.4. This shows the outcome of the cusum analysis of the ResidT values from the model in Figure A.3. We see that the residuals were well behaved for the great majority of the 19 years from 1981 onwards, except for a 5-year period in the mid-90s. In other words, OrthoP was persistently higher than usual during this period *after allowing for the effect of flow*.

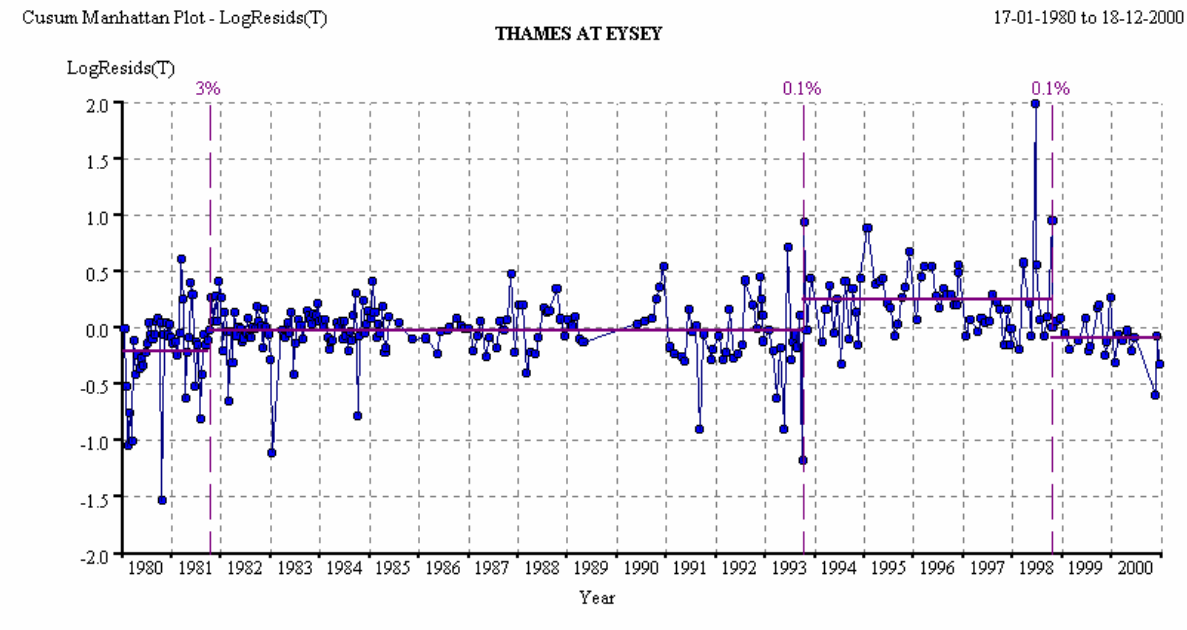


Figure A.4 Result of cusum analysis of time-based residuals from Figure B.3 model

APPENDIX B QUANTIFYING THE EFFECT OF AUTOCORRELATION

To demonstrate the complicating effect of autocorrelation on regression analysis, we wrote a short simulation program to work through the following steps:

(a) Generate two Normally distributed independent random time series, Q and F, of length n = 12. (These represent the situation in which annual quality Q and annual flow F have zero underlying correlation. Whatever the *observed* correlation may be, this will be due purely to random sampling error.)

(b) Calculate 3-year rolling means for the Q and F series. These series will have length n - 2 = 10.

(c) Pick out the *non-overlapping* 3-year means for the Q and F series. These series will consist of the 1st, 4th, 7th and 10th values, and so have length 4.

(d) Calculate and save the correlation coefficients (Rs) for:

- the 12 pairs of annual means in a);
- the 10 pairs of rolling means in b); and
- the 4 pairs of *non-overlapping* means in c);

(e) Repeat steps a to d 5000 times.

(f) Sort each of the 3 sets of Rs into increasing order (first stripping away all the minus signs), and hence read off various high percentiles from the 90%ile to the 99%ile.

(g) Compare the resulting empirical 'critical values' with those tabulated in statistical tables for the corresponding 2-tailed probability points.

The results are shown in Table B.1. For both the annual means and the non-overlapping 3-year means there is very close agreement between the theoretical critical values and those determined empirically by simulation. (This is to be expected, and merely serves to demonstrate that the simulation is working correctly.) The interesting part of the table is the pair of columns for the rolling 3-year means, where the empirically determined critical points are substantially higher than those that would apply for 10 *independent* pairs of Q and F values. For example, the 90% critical value - the R that is exceeded only one time in ten - would be 0.55 for independent data, but is 0.71 for rolling 3-year data. In other words, the R value for a set of 10 pairs of rolling 3-year Q and F values would need to be as high as 0.71 before we could even start to think that it might be something more than a chance effect. For us to be 95% confident that it was evidence of a real effect, the R value would need to be at least 0.78; and so on.

Table B.1 Critical values of the correlation coefficient for various types of data
(see text for details)

2-tailed significance probability	Annual means n = 12		Rolling 3-yr means n = 10		Non-overlapping 3-yr means n = 4		n = 3
	Theory	Simul'n	~~Theory~~	Simul'n	Theory	Simul'n	Theory
0.10	0.497	0.502	0.549	**0.711**	0.900	0.903	0.988
0.05	0.576	0.570	0.632	**0.779**	0.950	0.949	0.997
0.02	0.658	0.657	0.715	**0.848**	0.980	0.981	0.999
0.01	0.708	0.706	0.765	**0.881**	0.990	0.990	0.9999

The simplest way of avoiding the autocorrelation problem is to pick out just the *non-overlapping* data. But we then encounter another problem: we end up with a lot less data. In particular, our sample size in the simulation falls from 10 to 4. Unfortunately this leads to a very large increase in the critical values of R. Specifically we need to observe an R value of at least 0.90 for it to be statistically significant even at P<0.10. And for Rs based on only 3 pairs, the situation is even more severe: even for a mild level of significance (P<0.10) R needs to be at least 0.988!

APPENDIX C EXAMPLE OF THE OUTPUT FROM THE FLOW ANALYSIS PROGRAM FFION

```
\CvF\Floze\0130THAM.SCF
 ----------------------------------------
 "0130 THAM" "EWEN"                        "RIVER THAMES"
 Date range: 1980 - 1984

 Total no of records:   1827
 No of zeroes       :    409
 No of -ves         :    413
 No of OK records   :   1005

 Code counts:    0    1    2    3    4    5    6    7    8    9   10
             1414    0    0    0    0    0    0    0  413    0    0
 No of Up runs:   83
 No of Down runs:   86
 Flow statistics...
 Mean, st.dev & CoV:      .5051     .5608   1.110

  lag     N      ACC
   1      995     .997
   5      963     .944
  10      925     .849
  15      893     .736
  20      863     .625
  25      830     .531
  30      795     .429
 Flow(i)/Flow(i-1) statistics...
 Mean, st.dev & CoV:     1.0131     .4011    .396

  lag     N      ACC
   1      985     .127
   5      954     .096
  10      915     .017
  15      883     .015
  20      853     .000
  25      820     .002
  30      785     .010
 ----------------------------------------
 "0130 THAM" "EWEN"                        "RIVER THAMES"
 Date range: 1985 - 1989

 Total no of records:   1826
 No of zeroes       :    214
 No of -ves         :    265
 No of OK records   :   1347

 Code counts:    0    1    2    3    4    5    6    7    8    9   10
             1561    0    0    0    0    0    0    0  265    0    0
 No of Up runs:  120
 No of Down runs:  120
 Flow statistics...
 Mean, st.dev & CoV:      .5102     .5748   1.127

  lag     N      ACC
   1     1338     .981
   5     1307     .851
```

```
 10     1267     .688
 15     1231     .625
 20     1196     .549
 25     1161     .452
 30     1130     .395
Flow(i)/Flow(i-1) statistics...
Mean, st.dev & CoV:     1.0224     .6047    .591

 lag     N      ACC
  1     1329     .135
  5     1298     .020
 10     1258     .007
 15     1221     .007
 20     1186     .005
 25     1151     .005
 30     1119    -.002
---------------------------------------
"0130 THAM" "EWEN"                        "RIVER THAMES"
Date range: 1990 - 1994

Total no of records:   1826
No of zeroes       :    365
No of -ves         :    867
No of OK records   :    594

Code counts:    0    1    2    3    4    5    6    7    8    9   10
              959    0    0    0    0    0    0    0  867    0    0
No of Up runs:   49
No of Down runs:   50
Flow statistics...
Mean, st.dev & CoV:      .3040     .4056   1.334

 lag     N      ACC
  1      574     .947
  5      498     .799
 10      417     .592
 15      349     .534
 20      289     .531
 25      238     .521
 30      195     .565
Flow(i)/Flow(i-1) statistics...
Mean, st.dev & CoV:     1.0325     .6487    .628

 lag     N      ACC
  1      554     .075
  5      480     .055
 10      403     .010
 15      334     .008
 20      274     .012
 25      225     .018
 30      183    -.008
---------------------------------------
"0130 THAM" "EWEN"                        "RIVER THAMES"
Date range: 1995 - 1999

Total no of records:   1816
No of zeroes       :    450
No of -ves         :    870
No of OK records   :    496
```

```
 Code counts:    0    1    2    3    4    5    6    7    8    9   10
              939    0    0    0    0    0    0    0  877    0    0
 No of Up runs:   21
 No of Down runs:   27
 Flow statistics...
 Mean, st.dev & CoV:      .2380     .3553   1.493

  lag     N      ACC
   1      476     .939
   5      413     .693
  10      351     .396
  15      302     .316
  20      260     .282
  25      230     .241
  30      214     .251
 Flow(i)/Flow(i-1) statistics...
 Mean, st.dev & CoV:     1.0165     .4310    .424

  lag     N      ACC
   1      458     .152
   5      398     .051
  10      335     .024
  15      286     .012
  20      246     .030
  25      221     .009
  30      204     .007

 ====================
 Grand summary for the  4  5-yr periods...
    .27    .51    .56    .43    .67    .29    .76    .46    .51    .45
.30
    .34    .14    .19    .51    .28    .31    .12    .09    .52    .00
.00
    .00    .00    .00
   .303   .348   .491   .657   .000   .000   .000   .000   .000   .000
 134 262 183 214 212 183 279 269 355 261  69  75 139 136 175
  78 123 187  48  60   0   0   0   0   0
 ACC values for Flow...
   1   .997   .981   .947   .939
   2   .989   .954   .897   .883
   3   .976   .922   .862   .825
   4   .961   .887   .830   .759
   5   .944   .851   .799   .693
   6   .925   .815   .763   .628
   7   .906   .779   .721   .574
   8   .888   .745   .670   .512
   9   .869   .715   .611   .449
  10   .849   .688   .592   .396
  11   .829   .664   .572   .374
  12   .806   .653   .553   .356
  13   .783   .643   .536   .336
  14   .759   .636   .535   .327
  15   .736   .625   .534   .316
  16   .712   .613   .533   .305
  17   .689   .598   .532   .305
  18   .667   .583   .534   .298
  19   .646   .567   .534   .285
  20   .625   .549   .531   .282
  21   .605   .529   .527   .275
  22   .587   .508   .528   .271
```

```
 23   .568   .487   .531   .263
 24   .550   .469   .530   .253
 25   .531   .452   .521   .241
 26   .512   .437   .537   .244
 27   .492   .424   .555   .251
 28   .471   .413   .567   .258
 29   .450   .404   .569   .246
 30   .429   .395   .565   .251
Percentile values for Flow...
 1.   .003   .003   .003   .002
 5.   .010   .017   .012   .009
10.   .019   .028   .019   .016
25.   .052   .076   .044   .034
50.   .325   .293   .124   .072
75.   .700   .755   .380   .281
90.  1.410  1.330  1.020   .669
95.  1.710  1.750  1.220   .891
99.  2.250  2.540  1.940  1.910
ACC values for Flow(i)/Flow(i-1)...
  1   .127   .135   .075   .152
  2   .072   .073   .072   .075
  3   .064   .045   .065   .075
  4   .050   .045   .046   .063
  5   .096   .020   .055   .051
Percentile values for Flow(i)/Flow(i-1)...
 1.   .571   .832   .500   .600
 5.   .859   .900   .857   .855
10.   .916   .917   .895   .895
25.   .945   .942   .939   .937
50.   .970   .971   .966   .967
75.  1.004  1.016  1.000  1.000
90.  1.082  1.115  1.092  1.113
95.  1.191  1.190  1.188  1.259
99.  1.954  1.750  2.188  1.900
Percentile values for Up flow run lengths...
 1.    1.0    1.0    1.0   -1.0
 5.    1.0    1.0    1.0    1.0
10.    1.0    1.0    1.0    1.0
25.    1.0    1.0    1.0    2.0
50.    1.0    1.0    2.0    2.0
75.    4.0    4.0    4.0    6.0
90.    9.0   10.0    8.0   17.0
95.   12.0   14.0   15.0   23.0
99.   23.0   28.0   -1.0   -1.0
Percentile values for Down flow run lengths...
 1.    1.0    1.0    1.0   -1.0
 5.    1.0    1.0    1.0    1.0
10.    1.0    1.0    1.0    1.0
25.    1.0    1.0    2.0    4.0
50.    3.0    4.0    5.0    8.0
75.   12.0   11.0    9.0   18.0
90.   22.0   22.0   17.0   43.0
95.   33.0   32.0   29.0   54.0
99.   61.0   44.0   63.0   -1.0
========================================
```