DEFRA

DEMONSTRATION TEST CATCHMENTS: AN EXPERIMENTAL DESIGN AND MONITORING STRATEGY

INTERIM REPORT – OCTOBER 2009

WRc Ref: UC8104.01

DEMONSTRATION TEST CATCHMENTS: AN EXPERIMENTAL DESIGN AND MONITORING STRATEGY

Report No.: UC8104.01

Date: 5 October 2009

Authors: Andrew Davey

Contract Manager: Ian Codling

Contract No.: 15328-0

RESTRICTION: This report has the following limited distribution:

Any enquiries relating to this report should be referred to the authors at the following address:

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

Telephone: + 44 (0) 1793 865000
Fax: + 44 (0) 1793 865001
Website: www.wrcplc.co.uk

This document has been produced by WRc plc.

CONTENTS

SUMMARY		1
1.	INTRODUCTION	2
2.	HOW SHOULD MITIGATION MEASURES BE TRIALLED?	3
2.1	Scale of manipulation	3
2.2	Choice of measures	3
2.3	Individual vs combined measures	4
3.	EXPERIMENTAL DESIGN	6
3.1	Introduction	6
3.2	Spatial and temporal controls	6
3.3	Replication and representativeness	8
3.4	Idealised design	9
3.5	Design variants	10
4.	WATER QUALITY MONITORING STRATEGY	12
4.1	Introduction	12
4.2	Water quality data requirements for assessing effectiveness	12
4.3	Sampling strategy options	16
4.4	Sampling frequencies	19
4.5	Flow monitoring	19
REFERENCES		21

LIST OF TABLES

Table 2.1	A 2x2 factorial design to test the independent and combined effects of two measures	4
Table 3.1	Experimental design variations	11
Table 4.1	Pros and cons of alternative sampling strategies	17

LIST OF FIGURES

Figure 3.1	Comparison of TC and BACI approaches to assessing effect of mitigation measures (black hatching) on water quality at monitoring points (blue triangles)	7
Figure 3.2	Illustration of how a BACI experimental design can reveal the effectiveness of a measure	7
Figure 3.3	Idealised experimental design to test effectiveness of three measures in one catchment	10
Figure 4.1	Instantaneous loads at control and impact sites before and after mitigation measures	16
Figure 4.2	Three alternative sampling strategies	17

SUMMARY

This interim report presents options for an experimental design and associated monitoring strategy to underpin the Demonstration Test Catchments project. Its goal is to support the planning phase of the project by guiding and stimulating discussion and highlighting the key decisions that need to be taken. This document will continue to be revised and updated to take on board the views of stakeholders and should not be interpreted as being a prescriptive or final set of recommendations.

A full executive summary will be produced when the report is finalised.

1. INTRODUCTION

The overall objective of the Demonstration test Catchments (DTC) project is to provide a research platform to develop an integrated assessment of the effectiveness of potential mitigation measures on diffuse pollution from agriculture. Four catchments have been selected to host research commissioned by Defra, the Environment Agency and other funders. Surface water quality, groundwater quality and ecology will be monitored at multiple spatial scales and a fine temporal resolution in order to distinguish diffuse and point sources of pollution, characterise patterns of environmental quality, and to assess the of pollution mitigation measures at a catchment scale.

Following a meeting of the DTC project team at Nobel House on 24 August, a list of ten questions was drafted to guide the development of an appropriate experimental design and associated monitoring strategy:

1. How many sub-catchments do we need to monitor per catchment?
2. How many monitoring points are required per sub-catchment?
3. How do we deal with lack of baseline data / providing sufficient controls?
4. How can we disaggregate the effects of individual measures whilst looking at the potential mitigation from combinations of measures?
5. What monitoring is required to link water quality to ecology?
6. How frequently should ecological measurements be taken,
7. What are the requirements for flow gauging? Every site?
8. How do we deal with groundwater quality?
9. Load vs concentration?
10. What lag time can we expect between installing measures and detecting an improvement in water quality?

This report provides guidance and recommendations to enable the Environment Agency and other project partners to develop a statistically robust and cost-effective programme of monitoring in each of the four catchments. Section 2 starts by considering some general principles for trialling mitigation measures, including the most appropriate scale of manipulation, the choice of measures, and options for testing measures individually and in combination. Section 3 focuses on the overall experimental design by developing an idealised design that provides adequate control of spatial and temporal variations and appropriate replication. Variations on this basic design are described and their implications discussed. Section 4 addresses the monitoring strategy to be implemented at each monitoring site, in particular: what type of data is required to characterise environmental status and to assess the effect of specific mitigation measures; the pros and cons of alternative sampling strategies; and how frequently measurements should be taken.

2. HOW SHOULD MITIGATION MEASURES BE TRIALLED?

2.1 Scale of manipulation

Assessing the effectiveness of measures to reduce diffuse water pollution from agriculture (DWPA) is challenging and the DTC study should be designed to give the best possible chance of demonstrating an effect, at least at a local scale. The effect of mitigation measures will inevitably attenuate downstream due to a combination of dilution and in-stream processes, and it follows that detecting the effect of individual measures will become more difficult with increasing distance between the mitigation measure and the monitoring point. It is therefore recommended that:

- Measures should be trialled as far up the catchment as possible, ideally on first or second order streams; this will help to eliminate potentially confounding factors, ensure that DWPA is the main source of load, and mean that mitigation measures will have a larger proportional effect on water quality. The catchment characterisation exercise will be important in identifying hot spots of diffuse agricultural pollution where there is the highest potential of achieving a water quality improvement.

- Measures should be trialled intensively over a small, focused area; implementing changes that affect a tiny proportion of the area draining to the monitoring point will make detecting an effect much harder.

- Monitoring points should be located immediately downstream of locations where mitigation measures are being trialled, in order that the full effect is quantified before any attenuation.

2.2 Choice of measures

Ideally, one would draw up a list of measures to be trialled and implement the changes in a controlled experimental fashion so that the independent and consistent effect of each measure can be quantified. In reality, however, the choice of measures will inevitably be constrained to some extent by the nature of the DWPA issues within each catchment and sub-catchment. This means that it may not be appropriate or even possible to repeat the same manipulation in different sub-catchments, which may limit the generality of any conclusions that can be drawn about its effectiveness.

Furthermore, some measures can easily be trialled in multiple locations – for example, fields can have presence/absence of cover crops and be cultivated across/downslope – whereas other measures can only be trialled in specific locations – for example those that relate to manure and slurry storage, dirty water systems etc. For the latter group, it will again be more difficult to provide a robust assessment of their effectiveness.

Recommendation: DTC should concentrate only on those measures that are relevant to pressures operating in all sub-catchments and that can be trialled in multiple locations.

If the same measure is trialled in different catchments then it could be possible to test whether that measure has a consistent effect in different catchments. However, it may be more practicable to treat each catchment as a separate experiment.

2.3 Individual vs combined measures

One of the main objectives of DTC is to provide evidence on the effectiveness of mitigation measures individually and in combination.

Mitigation measures applied in locations with separate sources and transport pathways will be expected to have a simple, additive effect on water quality. For example, if measure 1 in location A achieves a reduction in load of X, and measure 2 in location B achieves a reduction in load of Y, then the total reduction in load to the river will be X+Y. If, however, measures are applied in the same location (e.g. loosen compacted soil layers and reduce stocking density), or measures interact hierologically (e.g. a gateway is re-sited and the adjacent river fenced off, both to limit runoff), then there may be some redundancy, in which case the combined improvement in water quality may be less than the sum of that achieved by each measure individually. It is this latter scenario that is least well understood.

A classical approach to assessing the combined effect of two measures is to construct a factorial experimental design in which the presence /absence of measure 1 is crossed with the presence/absence of measure 2, as shown in Table 2.1.

Table 2.1 A 2x2 factorial design to test the independent and combined effects of two measures

	No change	**Measure 2**
No change	Treatment 1 (control)	Treatment 2
Measure 1	Treatment 3	Treatment 4

This 2x2 design generates four unique treatments, which permits the individual and combined effects of the two measures to be analysed readily by a factorial analysis-of-variance (ANOVA) statistical model.

The advantage of testing measures in this way is that it is possible to assess the independent effect of each measure. The limitation of this approach is that the number of treatments quickly snowballs as the number of different measures increases – 3 measures require a total of 2x2x2 = 8 treatments, 4 measures require 16 treatments and so on. A second disadvantage is that individual measures in isolation may have a relatively modest impact on water quality, which will be harder to detect.

An alternative approach is to trial measures in sets rather than individually (e.g. measures 1 and 2 vs measures 3 and 4). This will reduce the number of different treatments required and increase the chances of detecting a water quality response, but it fails to quantify the independent effect of each individual measure.

Recommendation: Although it would be preferable to trial measures independently, a safer and more practical approach would be to trial sets of measures. Individual research projects are probably best placed to investigate the individual effect of each measure at a local scale.

3. EXPERIMENTAL DESIGN

3.1 Introduction

Assessing the effectiveness of a particular measure of programme of measures faces the problem of determining if a desired change in water quality has occurred as a result of the measure(s), against a background of innate water quality variability. Any assessment has to address the need to ensure that:

1. possible causes of changes to the relevant water quality variable(s) between the 'before' and 'after' periods are adequately controlled – i.e. so that the extent of their influence can be isolated from the effects of the programme itself;

2. the experiment is properly replicated so that inferences can be drawn about water quality at other sites or other times;

3. the experiment focuses on a clearly defined population and is representative of conditions at broader spatial and temporal scales.

3.2 Spatial and temporal controls

A simple approach to evaluating effectiveness would be to compare water quality at a site before and after implementation of a mitigation measure. This is termed a Temporal Comparison (TC) strategy and is illustrated in Figure 3.1a. This is a weak design because it fails to control a multitude of factors that could potentially influence water quality at the downstream monitoring point. Differences in farming practices, rainfall, flow etc before and after the measures were implemented could act to exaggerate or mask changes due to the mitigation measures. Although it is possible to statistically factor out the effect of variations in flow, such an approach is fraught with difficulties and relies on being able to establish meaningful and consistent relationships between concentration and flow. Even if the effect of flow could be accounted for in the analysis, it would be impossible to anticipate and measure all of the other potentially confounding factors. For these reasons, a simple TC assessment of water quality before and after the measures were implemented is not recommended because it could lead to an incorrect assessment of effectiveness.

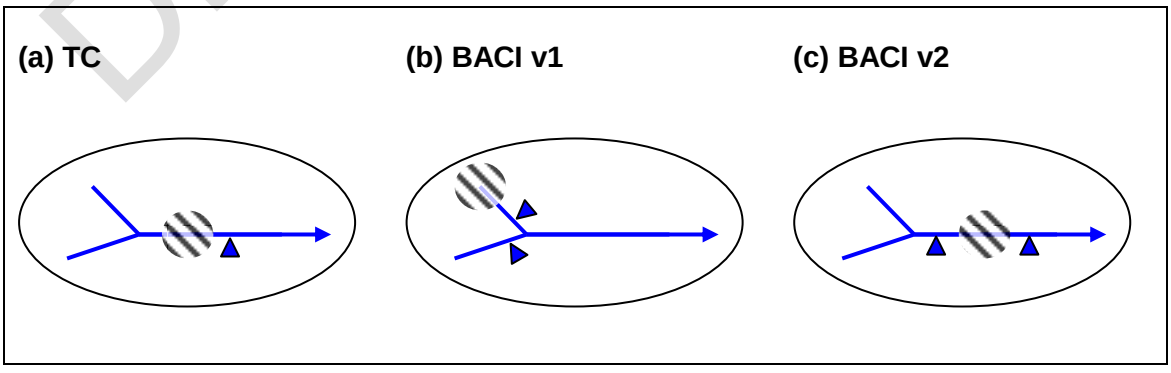

Figure 3.1 Comparison of TC and BACI approaches to assessing effect of mitigation measures (black hatching) on water quality at monitoring points (blue triangles)

A better approach is a Before-After, Control-Impact (BACI) comparison, which involves comparison of a manipulated with a non-manipulated stream before and after implementation of a mitigation measure (Figure 3.1b). Like a TC design, measurements taken pre-mitigation provide a baseline against which to compare post-mitigation conditions in the focal stream; the key difference is that an adjacent control stream provides an additional spatial reference that can be used to factor out confounding effects of changes in land use, rainfall flow etc.

The value of the BACI approach is illustrated in Figure 3.2 below. The impact stream containing the mitigation measure yields the same load both before and after the manipulation, which suggests on first sight that the measure has not been effective. However, the adjacent control stream shows an increase in load over the same time period; as the impact stream did not show a corresponding increase, this provides evidence that the measure has been successful in reducing the load in the impact stream. Thus, although the BACI approach requires two monitoring points rather than one (essentially doubling monitoring costs compared with a TC approach), this extra investment is more than justified by its superior ability to distinguish an effect of a specific measure.

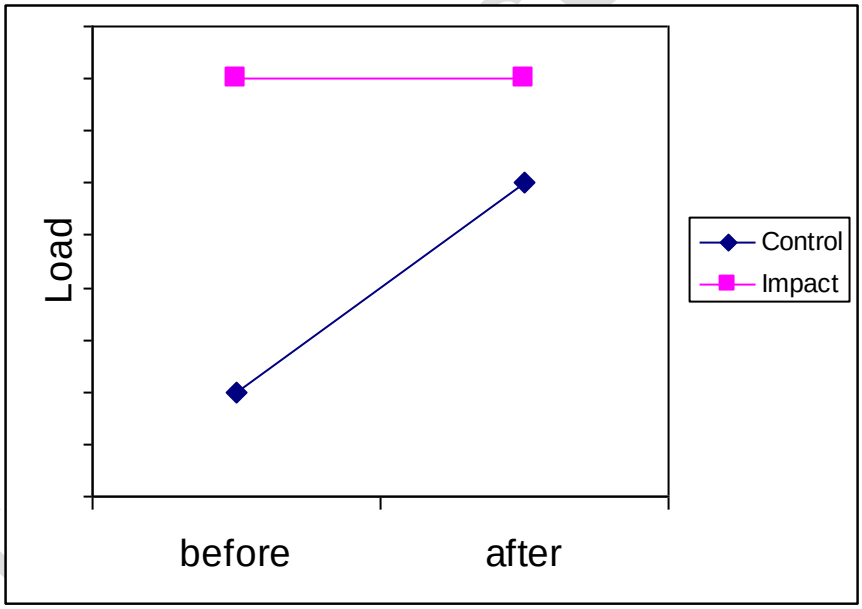

Figure 3.2 Illustration of how a BACI experimental design can reveal the effectiveness of a measure

The BACI approach relies on the assumption that any background changes in the focal stream are mirrored in the control stream and that the two streams experience the same weather conditions and respond to rainfall events in a similar fashion. In practice, it can be very difficult to find appropriate control streams that meet these assumptions. As a general

rule of thumb, the closer together the control and manipulated areas, the more similar they are likely to be, and the more powerful will be the resulting test of the measure. This logic argues for separate control streams in each sub-catchment, rather than separate control and manipulated sub-catchments.

A variant on the BACI approach is to use a monitoring point upstream of the mitigation area as the control (Figure 3.1c). This acts to isolate the part of the stream that is subject to the mitigation measures and means that a separate control stream is no longer necessary (although the assumption is now that the upper part of the stream and the part between the two monitoring points are fundamentally similar). The main disadvantage compared with the conventional BACI approach is that the loading coming from the manipulated area is not estimated directly, but rather as the difference in water quality between the upstream and downstream monitoring points; because two sets of measurements are involved, each with its own error, this will invariably increase the uncertainty in the result.

Recommendation: The DTC project should adopt a BACI design, with measures trialled at the scale of 1 km2 sub-catchments. Adjacent, unmanipulated 1 km2 sub-catchments in each 10 km2 sub-catchment should also be monitored to control for factors that may potentially confound the effect of the mitigation measure(s).

3.3 Replication and representativeness

If the goal of the study is simply to assess whether a measure has been effective at a particular location, then the experiment need not be replicated elsewhere. But if we want to know whether the same measure might be effective elsewhere, then the experiment must be replicated in two more independent locations so that we can get some idea of how consistent the measure is in its effectiveness from place to place.

As an example, suppose that a buffer strip is created along one tributary stream and that the loading to the stream decreases by 30% as a result. Without repeating the experiment elsewhere we have no way of knowing how reliable or representative that result is: is 30% typical for that type of stream, is it on the high side, or on the low side? Now suppose that a buffer strip is created on a second stream in the same sub-catchment and that the loading there decreases by 20%. Now we can say that buffer strips reduce the loading by 25% on average *in that sub-catchment*, and put some confidence intervals on this result. Note, however, that we still cannot say whether buffer strips are equally effective in other sub-catchments because we have not tested them there; to do this we would have to repeat the experiment in another sub-catchment. Similarly, to be able to extrapolate the results to other catchments, we would have to repeat the experiment in at least two catchments selected at random from those in the broader population that we are attempting to draw some inferences about.

The key thing is that the experimental units (be they buffer strips, fields, farms or whatever) must be *representative* of conditions within the sub-catchment (and possibly beyond too) so that inferences can be drawn about processes operating and broader spatial and temporal scales. This is essential if the results are to be used to inform policy decisions that affect not just the three DTC but other catchments across the country.

Recommendation: Each measure or set of measures should ideally be trialled in at least two different sub-catchments in each catchment, and in at least two catchments, in order to gauge the consistency of effect from place to place.

If the catchment is heterogeneous and has different areas with contrasting characteristics, then it would be useful to select replicate sub-catchments in each area so that inferences can be made about the effectiveness of a measure under different conditions.

The same rules apply when considering temporal replication. One year of pre-measure monitoring provides a baseline against which to evaluate any future changes, but relies on that year being a normal, representative year. Similarly, one year of post-measure monitoring can tell us about the effectiveness of a measure in that year, but to generalise to other years we need to monitor for at least two years so that we can gauge how consistent the effect of the measure is from year to year.

Recommendation: Monitoring should *ideally* be undertaken for at least two years before and two years after the implementation of mitigation measures in order to gauge the consistency of the effect from year to year.

3.4 Idealised design

In planning an experimental approach, it is helpful to start by developing an idealised experimental design that provides the most robust assessment of the effectiveness of a set of measures. Variations on the design can then be explored and consideration given the trade-offs that are often necessary to reconcile scientific rigor with practicality and resource availability.

As our starting point, Figure 3.3 shows a hypothetical, idealised experimental design to test the effectiveness of three measures (denoted by shaded circles) in one catchment. There are four 10km2 sub-catchments (dashed ovals) – two located in one geological area and two in another geological area. Within each sub-catchment there are four 1km2 mini-catchments. One mini-catchment is an unmanipulated control (white circle) and one measure is trialled in each of the other three (shaded circles), thus making it a BACI design. The four treatments are randomly allocated to mini-catchments within each sub-catchment. The same three measures are trialled in each sub-catchment, thereby allowing their effectiveness to be evaluated at a sub-catchment, geological area and catchment scale.

Water quality would be monitored immediately downstream of each mini-catchment (16 blue triangles). To trace any water quality changes downstream then it would also be necessary to monitor at the downstream end of each sub-catchment (four green triangles) and at the downstream end of the catchment (one red triangle). Monitoring would be undertaken for two years before and at least two years after the implementation of mitigation measures.

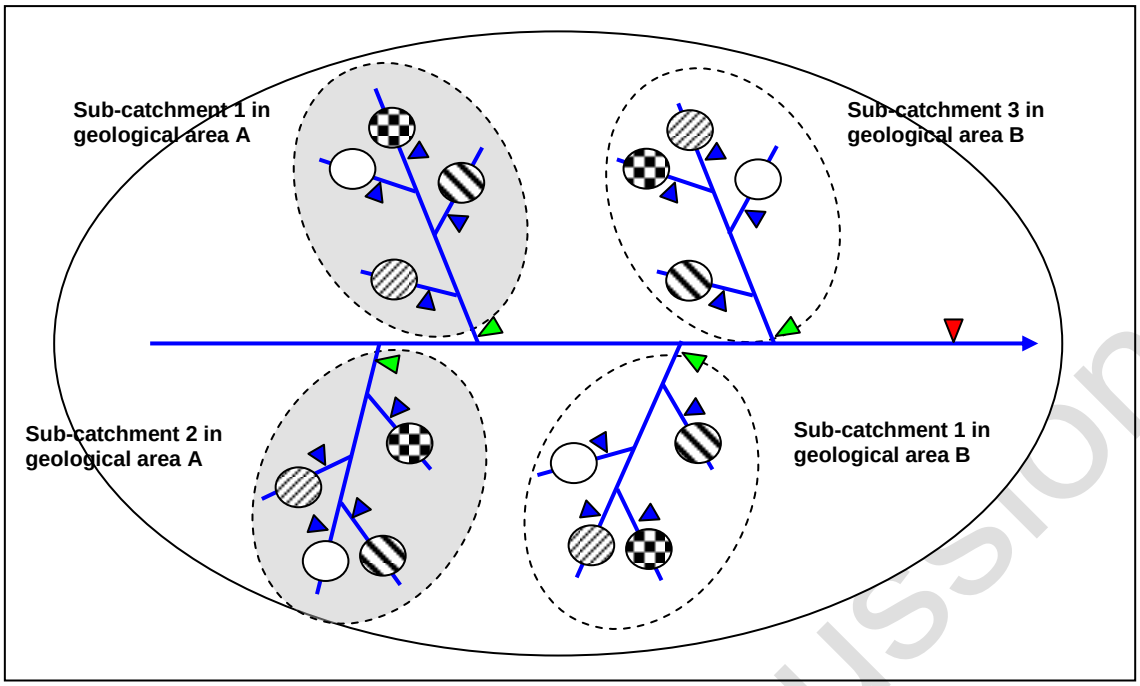

Figure 3.3 Idealised experimental design to test effectiveness of three measures in one catchment

This design fulfils all the main design requirements discussed in Section 3 above, namely:

- a control mini-catchment in each sub-catchment provides a spatial reference against which to compare the manipulated sub-catchments;
- the treatments are replicated in two sub-catchments to explore the generality of any effect from place to place;
- the experiment is replicated in two different geological areas within the catchment to test whether the effectiveness varies with geology;
- pre-mitigation monitoring characterises the hydrochemistry of the catchment and provides a baseline against which to assess any changes in water quality over time;
- monitoring is conducted over a number of years to measure both between- and within-year variation in hydrochemistry and to test the consistency of any effect from season to season and year to year.

3.5 Design variants

But of course Figure 3.3 is an *idealised* design, and there are numerous reasons why it might not be possible to implement such a design in practice. Where this is the case, the design will need to be altered in some way, which may necessitate compromising on one or more of the design features and making trade-offs between replication and generality. Some likely

scenarios are listed in Table 3.1 together with suggestions for how the design could be altered, and the implications of those changes.

Table 3.1 Experimental design variations

Scenario	Alteration to design	Implication
Insufficient resources to monitor for two years before and after implementation of mitigation measures.	Monitor for only one year before and after implementation of mitigation measures.	Cannot test whether effect of measures is truly season-specific or judge how consistent effect of measures is from year to year.
Each sub-catchment is in a different geological area.	No change, expect that the four sub-catchments are no longer grouped into two sets of two.	No longer possible to distinguish cause of variation in effectiveness among sub-catchments.
Sub-catchments have contrasting diffuse pollution issues OR number of measures to be trialled exceeds number of mini-catchments available in each sub-catchment.	Different sets of measures are trialled in each sub-catchment (e.g. measures A, B and C in sub-catchment 1; measures A, C and D in sub-catchment 2; and so on).	More difficult to compare effectiveness of alternative measures because they are not trialled in each and every sub-catchment.
	Completely different measures are trialled in each sub-catchment (e.g. measures A, B and C in sub-catchment 1; measures D, E and F in sub-catchment 2; and so on).	Each sub-catchment becomes an isolated experiment; no longer possible to judge generality of results from place to place, and increased risk that differences between mini-catchments will mask or exaggerate effect of measures. Not possible to compare effectiveness of alternative measures because they are trialled in different sub-catchments.
Different numbers of min-catchments in each sub-catchment.	Some measures cannot be trialled in every sub-catchment.	More difficult to compare effectiveness of alternative measures because they are not trialled in each and every sub-catchment.

4. WATER QUALITY MONITORING STRATEGY

4.1 Introduction

Whereas the experimental design addresses the spatial arrangement and replication of experimental treatments and the overall spatial and temporal extent of monitoring, the monitoring strategy addresses the question of how to measure water quality at each individual monitoring site.

This section of the report considers:

- what type of data is required to characterise and quantify water qaulity and to assess the effect of specific mitigation measures;
- the pros and cons of alternative sampling strategies;
- how frequently measurements should be taken.

It does not:

- advise which water quality determinands should be monitored;
- make recommendations on the most appropriate methods or technology to undertake water quality and flow monitoring.

The focus of this section is on water *quality* monitoring, with particular reference to the statistical difficulties of assessing the effectiveness of a specific management intervention, but the challenges of measuring flow and linking these readings to water quality are also discussed. Biological monitoring will be dealt with separately because of the different issues posed by this type of monitoring.

4.2 Water quality data requirements for assessing effectiveness

The goal of monitoring should be to deliver the right type and amount of information to allow the effectiveness of specific mitigation measures to be evaluated in a statistically robust way. We therefore need to consider three key questions:

- Should water quality be characterised in terms of a concentration or load?
- Should water quality be quantified instantaneously, seasonally or annually?
- How should the resulting data be analysed?

4.2.1 Concentration or load?

The choice of concentration or load (the product of concentration and flow) depends upon how the water quality monitoring data is to be used and interpreted.

Concentration may be more appropriate when the goal is to assess the effect of mitigation measures on water quality with reference to specific standards. Most water quality standards are expressed as average concentrations or as percentiles. For example, UKTAG standards for Soluble Reactive Phosphorus (SRP) in rivers are set as (time-weighted) annual average concentrations as this provides a good measure of the availability of nutrients and hence the severity of eutrophication (UKTAG, 2008). Similarly, the UKTAG standards for ammonia are set as a 90th percentile concentration; because ammonia is toxic to fish and macroinvertebrates, a percentile approach is more appropriate because it seeks to limit the proportion of time that organisms are exposed to harmful concentrations.

Load may be more appropriate when seeking to gauge directly the effectiveness of mitigation measures in reducing the transport of pollutants from land to the water. The Diffuse Pollution Inventory (Cuttle et al. 2006) gives estimates of the effectiveness of 44 measures designed to reduce losses of nitrate, phosphate and faecal indicator organisms (FIOs) from agricultural land. Reductions are expressed either as kg/ha/year or as a % change in load. Changes in loadings can be inputted into catchment models to predict downstream changes in concentration.

4.2.2 Instantaneous, seasonal or annual?

The simplest way to describe water quality is as an *instantaneous* concentration or load – i.e the concentration or load observed at a particular moment in time when a water sample was taken. By treating each sample as a separate observation, one has a large number of 'snapshots' of water quality that can provide insight into the hydrochemical behaviour of the river and provide a reasonably large dataset for statistical analysis

These instantaneous measurements can also be used to generate *summary statistics* describing water quality over a longer period of time – say quarterly or yearly. Concentration is commonly quantified as:

1. a percentile concentration – for example, a 10^{th} or 90^{th} percentile;
2. a time-weighted average concentration – the average concentration over a defined period of time (e.g. a year);
3. a flow-weighted average concentration – the average concentration over a defined volume of discharge (e.g. annual total discharge).

Similarly, load is commonly summarised as:

1. an annual average instantaneous load (i.e. g/sec);
2. an annual total load (i.e. kg/year)

The main advantage of summary statistics is that they condense a lot of data into a single number that can be compared to a water quality standard or used to calibrate a catchment model. The main disadvantage is that estimates of summary statistics are not always very accurate or precise. Depending on the frequency of sampling, level of natural variability in water quality, and timing of sampling, summary statistics can systematically over- or under-estimate water quality, or be estimated with very wide confidence intervals (Johnes 2007; WRc 2008).

4.2.3 Statistical methodology

Detecting a potentially subtle improvement in water quality and attributing it unambiguously to specific mitigation measures is not an easy task. Careful experimental design can help to eliminate some potentially confounding factors, but there will still be a need to factor out the uncontrollable natural variation in water quality in order to isolate the effect of the measures. Extensive analysis of water quality data collected as part of the England Catchment Sensitive Farming Delivery Initiative (ECSFDI) monitoring and evaluation programme has revealed which statistical approaches are best suited to analysing this type of data to assess the effectiveness of diffuse pollution mitigation measures (WRc 2009a,b).

The ECSFDI monitoring and evaluation programme has been undertaken since 2007 in nine selected catchments to find out whether the Initiative has been successful in improving water quality. Weekly spot samples of water quality are taken at 8-10 sites in each catchment. These spot samples are supplemented at some sites by hourly samples collected by autosamplers during high flow events. Flow are measured every 15 minutes at sites with gauging stations and mean daily flows are estimated by extrapolation for ungauged sites.

WRc (2009b) explored three different approaches to assessing the effectiveness of CSF activities:

A. comparison of annual summary statistics at a single site;

B. analysis of flow-concentration relationships at a single site;

C. analysis of instantaneous loads or concentrations at paired sites.

Approach A tests for differences in annual summary statistics (e.g. total annual load) from year to year. A BACI approach may be used, with a neighbouring control site used to factor out year to year variation in weather and flow, which might otherwise mask more subtle influences of mitigation measures. The problem with this approach is that high temporal variability in water quality, particularly in rivers with a low Base Flow Index (BFI), makes it difficult to generate precise summary statistics of water quality. As a consequence, comparison of annual summary statistics (e.g. total annual load) from year to year provides a weak test for water quality improvements.

Approach B attempts to use paired measurements of instantaneous concentration and flow to characterise the hydrochemistry of an individual site and to construct a statistical model to test whether concentration is lower *at a given flow* post-mitigation than pre-mitigation. This approach relies on there being a strong and consistent relationship between concentration and flow but unfortunately this is often no the case; seasonal variations in loading (reflecting seasonal farming practices), complex hysteresis effects (concentrations are typically higher on the rising limb of a storm hydrograph than on the falling limb) and idiosyncratic storm effects (individual storm events can display markedly different flow-concentration relationships due to differences in the intensity and duration of rainfall and variable antecedent conditions) all combine to make it very difficult to factor out the effect of variation in flow.

Finally, Approach C compares *instantaneous* loads (or concentrations) at paired control and impact sites before and after the implementation of mitigation measures (Figure 4.1). Here, the control site acts as a reference, recording changes in instantaneous load over time in response to fluctuations in flow; assuming that the same changes are mirrored at the impact site, this neatly factors out the effect of short-term weather-driven variations in load that can

otherwise mask or exacerbate the effect of mitigation measures. If all load measurements are log-transformed, then the instantaneous load at the 'impact' site can be modelled as:

$$\log_{10}(I_t) = \log_{10}(C_t) + BeforeAfter_t$$

where:

I_t = the instantaneous load on day *t* at the 'impact' site downstream of the mitigation area;

C_t = the instantaneous load on day *t* at the 'control' site downstream of the non-mitigation area;

$BeforeAfter_t$ = whether day *t* is before of after the implementation of the mitigation measures.

A seasonal factor could also be included in the model if loads varied systematically with season.

Using this model, a negative coefficient for the $BeforeAfter_t$ term indicates that the instantaneous load at the impact site was lower post-mitigation after controlling for changes in load at the control site. Because the loads are on a log scale, the $BeforeAfter_t$ term actually estimates the % change in instantaneous load at the impact site. Although this simple model assumes that the % change is the same at both high and low loads, it would be a relatively simple matter to include an interaction to test this assumption.

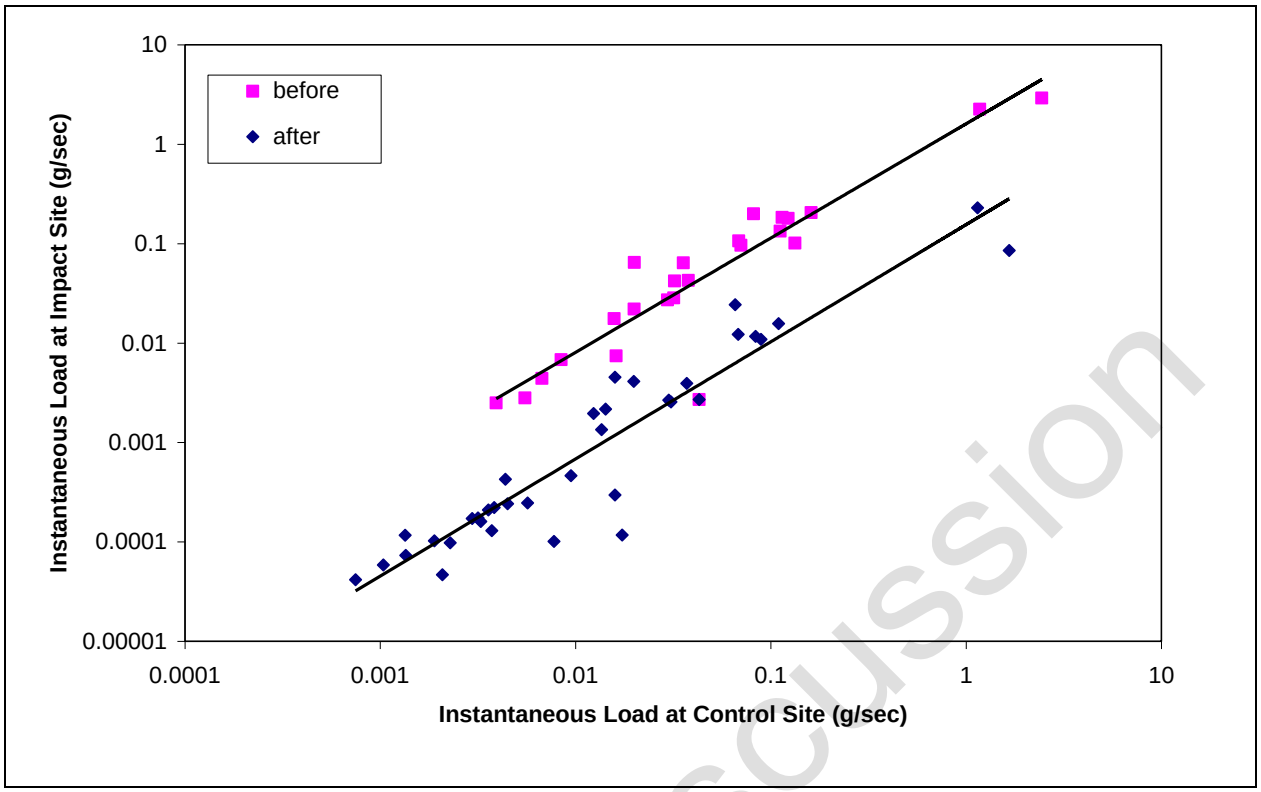

Figure 4.1 Instantaneous loads at control and impact sites before and after mitigation measures

4.3 Sampling strategy options

There are three main types of sampling strategy that could be used to measure water quality at a monitoring site: regular, stratified and flow-proportional. These are illustrated in Figure 4.2.

The *regular* strategy is the simplest: samples are taken at regular time intervals, irrespective of flow, such that each sample represents an equal period of time. This strategy is suitable if the goal is to estimate the time-weighted average concentration, but it is likely to produce unreliable estimates of the annual load. This is because high flows, when the majority of load transport takes place, occur for a relatively short period of time and very few will happen to coincide with a regular sample (unless the sampling frequency is very high, say at least daily). In this situation, the annual load will tend to be under-estimated (Walling & Webb 1985; Johnes 2007). Worse still, with sample sizes of only 30 or 40, there is a real risk that high flows will be entirely unrepresented, with the consequence that the sample variability will badly under-estimate the true variability in water quality, and so lead to an unrealistically optimistic statement of precision on the load estimate (Walling & Webb 1985; Ellis 1989).

One way to try to overcome these limitations is to adopt a *stratified* sampling strategy. This approach focuses sampling effort on periods of high flow when the majority of load transport takes place. Once the flow reaches a defined threshold, say Q_{10} or Q_5, the sampling frequency is stepped up until such time as the flow drops back below the threshold. An average

concentration or load can then be calculated separately for the low and high flow periods and the two results weighted by the duration of low and high flows to give an overall result for entire period of monitoring. This strategy provides concentration and load yields with less bias and better precision than the regular sampling strategy.

Whatever method is used to take samples under high flow conditions, it is important that it operates consistently above the flow threshold. Some instruments have a limited capacity to store samples, with the result that the start of every high flow event is sampled but the tail-end of some events is missed. This greatly complicates the statistical analysis of the data and introduces additional uncertainty into the result. A trade-off must therefore be made between the frequency of sampling and the flow threshold, and the decision tailored to the hydrology of the site and the instrument being used.

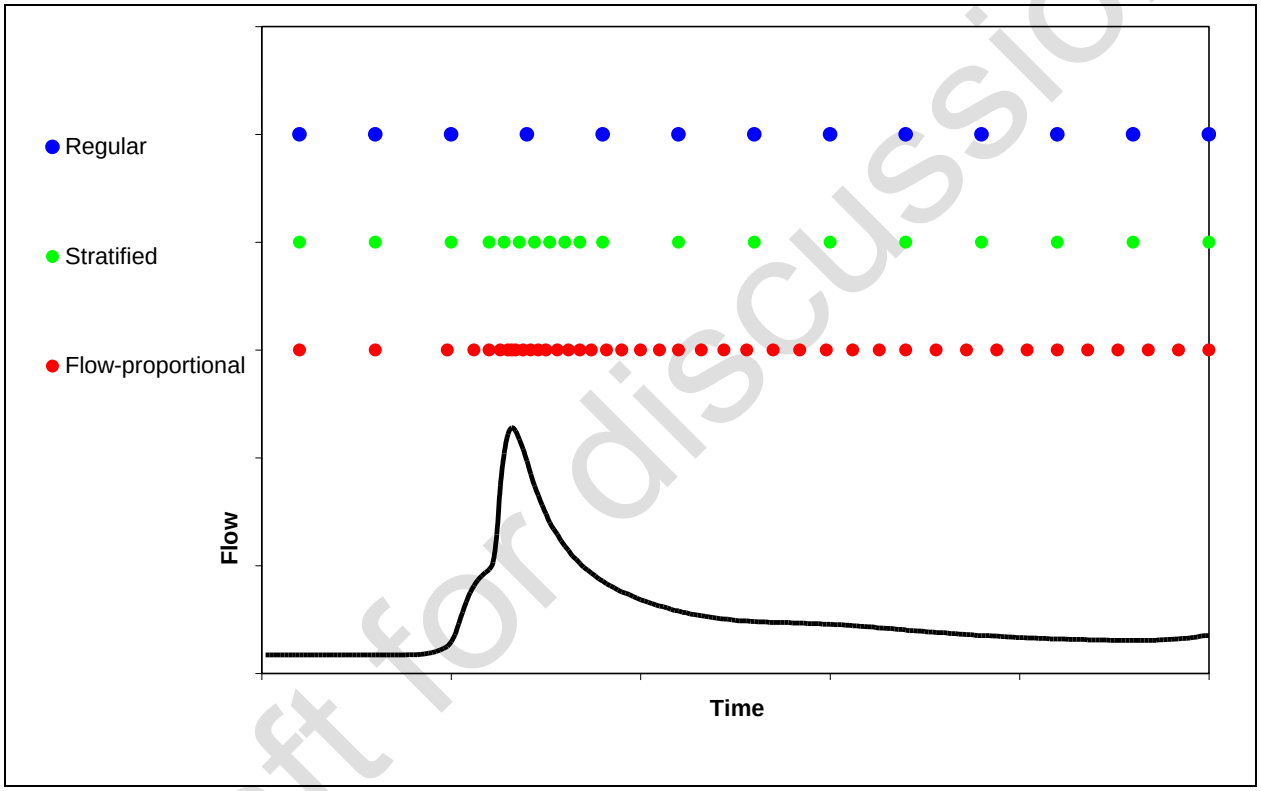

Figure 4.2 Three alternative sampling strategies

The third option is the *flow-proportional* sampling strategy. In this case, the frequency of sampling is proportional to flow, and each sample represents an equal volume of discharge. This strategy is the best approach for estimating flow-weighted average concentrations and average or total loads: the mean of the concentration readings gives a flow-weighted average concentration, which can multiplied by the total discharge to give a total load. The total load can then be divided by time to give an average instantaneous load. Using this method, loads are estimated with negligible bias but the level of precision will still depend upon the level of sampling effort.

The pros and cons of these three sampling strategies are summarised in Table 4.1.

Table 4.1 Pros and cons of alternative sampling strategies

Strategy	Pros	Cons
Regular	Simple; does not require flow measurements or expensive equipment. Straightforward way to estimate time-weighted average annual concentration. Easy to synchronise measurements at paired sites.	Risk of biased and imprecise estimates of flow-weighted average concentration and annual load, particularly when sampling frequency is low.
Stratified	Yields less biased and more precise estimates of flow-weighted average concentration and annual load.	Requires flow measurements to trigger sampling. Careful planning needed to ensure that entire period of all high flow events is captured.
Flow-proportional	Yields unbiased and more precise estimates of flow-weighted average concentration and annual load.	Requires flow measurements to trigger sampling. Careful planning needed to ensure that frequency of sampling is directly proportional to flow, even during flood events. May be difficult to synchronise measurements at paired sites.

The choice of strategy will depend upon the goals of the study. If the main goal is to estimate annual or seasonal loads of pollutants then flow-proportional sampling would be the most efficient option. Flow-proportional sampling would also provide data that could be used to test the effectiveness of mitigation measures using any of the three approaches described in Section 4.2.3. If, however, paired control and impact sites differ in the timing, duration and peakiness of high flow events, then one would have to choose between (i) taking samples that are perfectly synchronised across the two sites but which are not perfectly flow-proportional at each site, or (ii) taking samples that are perfectly flow-proportional at each site but which are not perfectly synchronised across sites. Obviously, this dilemma would not arise if sites could be selected in adjacent mini-catchments with near-identical hydrological characteristics.

By contrast, if the main goal is to assess the effect of mitigation measures on water quality, then a simpler stratified approach may be a more attractive option because it is easier to collect synchronised samples of water quality at paired control and impact sites. The important thing is to take plenty of measurements during high flow periods as this is when concentrations and loads tend to be highest and, correspondingly, when the effects of any mitigation measures are likely to be most apparent. If annual summary statistics of water quality are not important at all, then there is not even a need to monitor throughout the year; one could conduct intensive bursts of monitoring – say one month in each season – and treat

these instantaneous measurements as being representative of conditions in each three month period.

4.4 Sampling frequencies

In general, increasing the frequency of sampling will improve the accuracy and precision of estimates of average concentration and annual load (useful if trying to estimate summary statistics describing water quality) and will provide more information with which to assess the effectiveness of mitigation measures. The question is: how much sampling is enough?

If using a regular sampling strategy, the frequency of sampling required to achieve a given level of accuracy and precision will depend upon the summary statistic to be estimated and how much water quality changes over time. Rivers with a low BFI (i.e. a 'flashy' flow regime) will require a higher frequency of sampling to achieve the same level of accuracy and precision. Although increasing the sampling frequency will improve precision, it will be more effective to invest that additional effort in a stratified time-proportional or flow-proportional sampling strategy.

With a stratified sampling strategy, the sampling frequency during high flow periods will be influenced by the characteristics of the sampling instrument. For example, if autosampler that has a capacity of eight samples and the longest high flow event is expected to be 24 hours, then the sampling frequency should be every 3 hours so that all high flow events are monitored in full.

Similarly, a flow-proportional sampling strategy should ensure that the instrument is set up to sample at maximum frequency at maximum flow. Otherwise if the flow sometimes exceeds the capacity of the instrument to take samples then high flow events will be slightly under-represented.

4.5 Flow monitoring

Some form of flow monitoring is essential to complement measurements of water quality. Flow is required to calculate loads of pollutants, to estimate flow-weighted average concentrations, and to trigger the operation of automatic samplers (used as part of a stratified time-proportional or flow-proportional strategy).

Ideally flow would be measured directly and continuously (i.e. 15 minute intervals) at every water quality monitoring site. If this is impractical or too costly then there are a number of options.

1. Use predictive models to hindcast flows at ungauged sites from neighbouring gauged sites. Depending on the model, it may be difficult to predict 15-minute flow accurately, and prediction of mean daily flows may be a more realistic aim. However, this would obviously compromise the estimation of water quality summary statistics, and reduce the sensitivity of any analysis testing for an effect of mitigation measures.

2. Monitor intermittently. If estimating annual statistics is not important, then conducting a few short, intensive bursts of monitoring may generate enough data to be able to conduct an assessment of effectiveness (see Section 4.3).

3. Monitor some sites and not others. If a strong correlation could be established between flows at a pair of sites, then flow monitoring could be dropped at one site and the flows there predicted from those at the other site using an empirical relationship. Alternatively, if it could be assumed that a pair of neighbouring rivers had similar hydrological characteristics (i.e. even through they might differ in size, the pattern and timing of flow was similar) then one could use the flow at the gauged site as a surrogate of flow at the ungauged site. Although it would then not be possible to estimate actual loadings, it would still be possible to examine relative changes over time in water quality.

REFERENCES

Cuttle S.P., Haygarth P.M., Chadwick D.R., Newell-Price P., Harris D., Shepherd M.A., Chambers B.J. and Humphrey R. 2006. An inventory of measures to control diffuse water pollution from agriculture: a user manual.

Ellis, J.C. 1989. Handbook on the design and interpretation of monitoring programmes. WRc Report NS29.

Johns P.J., 2007. Uncertainties in annual riverine phosphorus load estimation: Impact of load estimation methodologies, sampling frequency, baseflow index and catchment population density. *Journal of Hydrology*, 332, 241-258.

UKTAG. 2008. UK Environmental Standards and Conditions (Phase 1) – Final report, April 2008.

Walling D.E. and Webb B.W. 1985. Estimating the discharge of contaminants to coastal waters by rivers: some cautionary comments. *Marine Pollution Bulletin*, 16, 488-492.

WRc 2008. Load determination for the England Catchment Sensitive Farming Delivery Initiative (ECSFDI) surface water quality monitoring programme. WRc Report EA7650 to the Environment Agency. EA Science project SC060046.

WRc 2009a. Load determination for the England Catchment Sensitive Farming Delivery Initiative (ECSFDI) surface water quality monitoring programme: 2008. WRc Report EA8038 to the Environment Agency.

WRc 2009b. Assessment of CSF effectiveness 2007-2008. WRc Report UC8087 to the Environment Agency. [Currently in draft]