crew

Scotland's centre of expertise for waters

Monitoring guidance to assess the effectiveness of the Rural Diffuse Pollution Plan

crew
Scotland's centre of expertise for waters

Published by CREW: Scotland's Centre of Expertise for Waters. CREW connects research and policy, delivering objective and robust research and expert opinion to support the development and implementation of water policy in Scotland. CREW is a partnership between the James Hutton Institute and all Scottish Higher Education Institutes supported by MASTS. The Centre is funded by the Scottish Government.

This document was produced by:
Ioanna Akoumianaki and Jannette MacDonald
The James Hutton Institute, Cragiebuckler, Aberdeen, AB15 8QH, Scotland
and Jackie Potts (Biomathematics and Statistics Scotland, BioSS).

Please reference this report as follows: Akoumianaki, I, Potts, J & MacDonald J 2016, *Monitoring guidance to assess the effectiveness of the Rural Diffuse Pollution Plan. CD2014/14*, Available online at: crew.ac.uk/publications.

While every effort is made to ensure that the information given here is accurate, no legal responsibility is accepted for any errors, omissions or misleading statements. All statements, views and opinions expressed in this paper are attributable to the author(s) who contribute to the activities of CREW and do not necessarily represent those of the host institutions or funders.

Cover photograph courtesy of: Clare Neely, James Hutton Institute

The James Hutton Institute

Scottish Government
gov.scot

Abbreviations

BACI	Before-After/Control-Impact
CSF	Catchment Sensitive Farming
DP GBR	Diffuse Pollution General Binding Rules
DPMCs	Diffuse Pollution Monitored Catchments
DTC	Demonstration Test Catchments
FIOs	Faecal Indicator Organisms
MDC	Minimum detectable change
RBMP	River Basin Management Plans
SRDP	Scotland Rural Development Programme
TDML	Total Daily Maximum Load
WFD	Water Framework Directive

Acknowledgements

We thank SEPA staff for constructive discussions in the preparation of this report, and specifically: Brian McCreadie (lead contact for the project); Fiona Napier (contact for chemistry monitoring and reporting); Alison Bell (contact for ecological monitoring and reporting); Alasdair Hardie (contact for flow monitoring and reporting), and; Susan Campbell (contact for FIO monitoring and reporting). We also wish to thank Dr Leah Jackson Blake for kindly sharing with CREW the daily SRP daily data set from the Tarland catchment. We acknowledge constructive comments from Dr Nikki Dodd (CREW).

Contents

Executive summary

The question

What is the best monitoring for assessing the effectiveness of the Rural Diffuse Pollution Plan?

Key findings

- Trial data and a review of the literature showed the need for a statistically robust monitoring design, longer monitoring duration and higher sampling frequency to enable change in pollutants and ecology to be quantified at a waterbody scale in SEPA's priority catchments.
- Monitoring in river and bathing water catchments should be based on a Before-After/Control-Impact (BACI) design.
- Simultaneous concentration and flow measurements should be taken to enable a reliable flow adjustment of concentrations and load estimation of key pollutants.
- Pollutant, diatom and macroinvertebrate monitoring should ideally be undertaken for more than four years before and more than four years after the introduction of diffuse pollution measures to enable changes from year to year to be detected.
- In-stream pollutants should ideally be monitored on a weekly basis with spot sampling or the automated time composite sample method to account for background variation.
- In Bathing Waters, routine spot and event-based automated sampling should be combined to separate the effects of measures and rainfall on Faecal Indicator Organisms (FIOs).
- Diatoms and macroinvertebrates should be monitored on a biannual basis.

The table below outlines the major components of the monitoring strategy developed here.

Background

The Rural Diffuse Pollution Plan was launched in 2011 in Scotland as part of the national response towards achieving WFD objectives. The Diffuse Pollution plan promotes the uptake of measures to help reduce diffuse pollution from rural sources. Predictive modelling has indicated significant declines in pollutants following the implementation of measures. Monitoring must be able to detect the predicted improvements and capture expected improvements in stream ecology. If changes occur and go unnoticed, this will have important implications for evaluating the effectiveness of the Diffuse Pollution Plan and communicating its environmental outcomes, or lack of, with stakeholders.

Research undertaken

We used trial data collected for WFD and Bathing Water classification from four priority catchments: Lemno Burn, Eye Water, Cessnock Water, and River Ayr. We also used event-based data from the Cessnock Water, one of the two Diffuse Pollution Monitored Catchments (DPMCs) in Scotland, where automated sampling is carried out. Daily phosphorus data from the Tarland catchment were provided by the James Hutton Institute. The questions of metric, duration, sampling frequency and suitability were assessed using a Before-After statistical design. Procedures were demonstrated in R code and were based on five statistical tests: (i) trend; (ii) step-change and flow adjustment; (iii) minimum detectable change; (iv) sample size; and (v) autocorrelation.

Trials showed:

- Significant post-implementation trends could not be detected.
- Significant step-change between before and after the introduction of measures could not be detected due to small sample size (i.e. number of samples).
- Long-term monitoring data are required to enable a sufficient number of samples to be collected; therefore, it is unrealistic to expect model predicted reductions to be detected within the first five years after launching the measures because
- At the Cessnock Water, event-based sediment concentrations were up to ten times greater than the maximum concentration observed with spot sampling. In general, the current automated event-based pollutant monitoring at the DPMC is insufficient to detect change reliably because of irregular sampling frequency.

Table i Major components of the enhanced monitoring developed here

Component	Enhanced monitoring to enable a statistically robust detection of change
Design	BACI design in river and bathing water catchments. Trend design in groundwater/lochs.
Flow	All river waterbodies and bathing waters for flow-adjustment and load estimation.
Metric	*Pollutants:* Concentrations for pollutants in groundwater and surface waters.
	Biota: Ecological Quality Ratios (EQRs); biomass; species composition / richness.
Duration	Long-term, i.e. more than four years pre- and post-implementation.
Frequency	*In-stream pollutants:* Weekly spot sampling or weekly retrieval of time-composite samples.
	FIOs: Bathing season spot sampling with event-based samples retrieved daily.
	Diatoms / Macroinvertebrates: Biannual sampling.
	Groundwater: Quarterly (highly permeable aquifers) or annually (less permeable aquifers).
	Nutrients-sediments (lochs and transitional waters): Seasonal spot sampling.

- There is a great mismatch in the sampling frequency of pollutant and ecological data, which precludes the establishment of cause-effect relationships for the interpretation of diatom response.

Recommendations

- Flow data from existing flow gauging stations should be assessed to determine their suitability for use in reliable flow-adjustment of concentrations and load estimation of key pollutants.
- Ancillary research projects should be carried out to help to understand how diatoms and macroinvertebrates respond to different pollution sources and measures, and at what spatial scales. For example, monthly diatom data should be collected at two sites or waterbodies representative of sources of phosphorus with different bio-availability (e.g. arable land versus septic tanks) to help to understand how the measures and source influence diatom response.
- The enhanced monitoring should be applied in all priority catchments to inform the weight-of-evidence method developed to evaluate the effectiveness of the Diffuse Pollution Plan (Akoumianaki *et al*. 2016). But if this is not feasible, the enhanced monitoring with long term-duration, weekly frequency of key pollutants and simultaneous flow measurements should be targeted at waterbodies representative of land management (i.e. implementation of improvement measures) and land use.

Reference

Akoumianaki, I, Potts, J, Baggio, A, Gimona, A, Spezia, L, Sample, J, Vinten, A, & MacDonald J 2016, *Developing a Method to Monitor the Rural Diffuse Pollution Plan: Providing a Framework for Interpreting Catchment Data, CRW2014/13*. Available from: crew.ac.uk/publications.

1 Introduction

Water quality is generally good across Scotland. Yet, SEPA estimates that around 30% of water bodies are expected to be at less than the good status required by the Water Framework Directive (WFD) at the end of 2015 due to the adverse effects of rural diffuse pollution (SEPA 2015). More than 200 of these waterbodies are rivers. The areas affected also include protected areas for bathing waters, shellfish waters, drinking water areas, and designated areas for wildlife conservation. The most widespread diffuse pollution pressures which remain for water quality are losses of nutrients, pesticides and faecal indicator organisms (FIOs) in runoff from a variety of rural land uses. Intensive arable and livestock farming are the dominant sources, but inputs of pollutants from forestry, septic tanks and low-intensity hill farming and sheep grazing can also contribute.

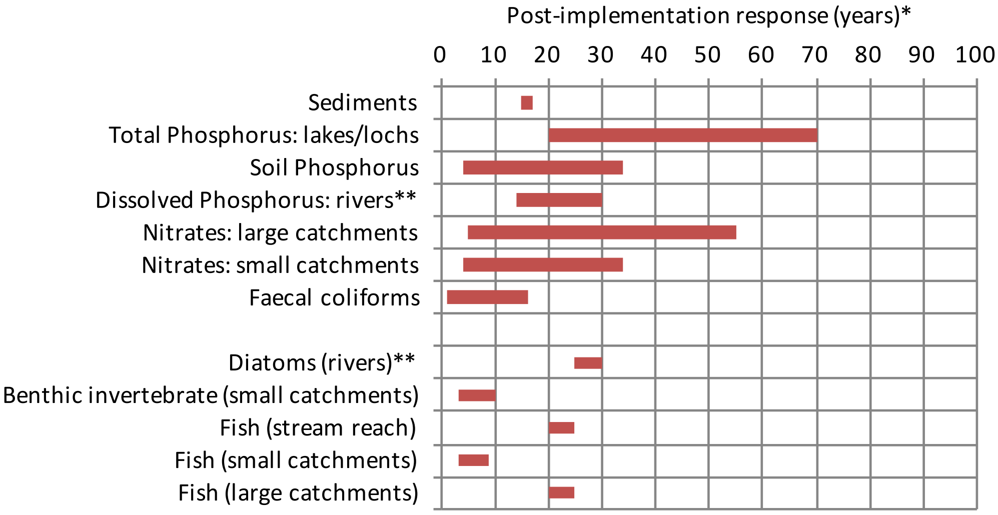
The Rural Diffuse Pollution Plan (DPMAG, 2011) was launched in 2011[1] to ensure effective reduction of these pressures and delivery of the good status. It includes a "national awareness raising campaign", and a "priority catchment approach" in catchments needing more focused land management intervention. This approach involves evidence gathering; predictive modelling of the effectiveness of measures; and one-to-one farm visits to deliver advice on good land management practices and to document the uptake of regulatory measures, such as the Diffuse Pollution General Binding Rules (DP GBRs)[2].

Specific actions are planned, monitored and reported in these catchments by SEPA, the main focus being on linking the improvements expected from the implementation of measures with water quality monitoring data. In this respect, water quality monitoring plays a strategic role in SEPA's evaluation of the effectiveness of the measures.

Fourteen priority catchments were taken forward in the first River Basin Management Planning (RBMP) cycle. Modelling has predicted significant declines in pollutants from the first years of the implementation of the Diffuse Pollution Plan (ADAS 2008). Assessments showed notable increases in DP GBR uptake after 2011, and WFD status improvements. However, a straightforward procedure for detecting the reductions in pollutants predicted by the model was not possible. WFD status and status improvements are unsuitable to provide information on change in pollutants and ecology between before and after launching the measures.

Within priority catchments, each year's classification and any subsequent updates are based on a three-year data window for pollutants (i.e. 36 samples for nutrients and sediment) and biology (i.e. six samples for benthic diatoms and four to six for invertebrates) and a four-year data window for faecal indicator organisms (FIOs) in bathing waters (i.e. about 80 samples). This sample size is sufficient to capture reliably any breaches of compliance with the specified standards and to prevent background variability from confounding status classification, as shown by Kelly *et al*. (2009a) for diatoms; Clarke (2013) for benthic invertebrates; and Skeffington *et al*. (2015) for physio-chemical parameters. However, detecting the predicted reductions in pollutants requires addressing this background variability.

Studies clearly show there is a lag time between implementation of diffuse pollution control measures and the expected improvements. The lag time varies with pollutant, biological indicator, catchment size, and waterbody type (Figure 1).



*Response refers to reaching regulatory targets except for sediments, where response refers to achieving export of sediment from in-stream storage sites.

**The range comes from a very limited number of success stories from Jarvie et al. 2013 and the Nonpoint Source Success Stories web site (http://water.epa.gov/polwaste/nps/success319/#progress).

Figure 1 Time required to detect water quality and ecological improvements in response to diffuse pollution control measures, based on the outcomes of mitigation projects implemented in Europe, New Zealand, Canada and the USA. This evidence was compiled using the reviews by Meals *et al*., 2010; Hamilton 2012; Gabel *et al*., 2012; Jarvie *et al*., 2013; Bechmann *et al*., 2008; Gitau *et al*., 2010; Palmer *et al*., 2014; Clements *et al*., 2010; Yates *et al*., 2007; and the Nonpoint Source Success Stories web site (http://water.epa.gov/polwaste/nps/success319/#progress).

[1] Members include SEPA, the Scottish Government, National Farmers Union of Scotland, Scottish Land and Estates, the Tennant Farmers Association, the Scottish Crofting Foundation, Forestry Commission Scotland, SNH, Scottish Environment LINK and Scottish Water.

[2] In addition to regulatory measures, it includes supplementary measures such as support via the Scotland Rural Development Programme (SRDP), where the regulatory baseline has been complied with; the new 2014-2020 SRDP agri-environment-climate scheme is targeted to ensure delivery of Scotland's environmental and biodiversity objectives. https://www.ruralpayments.org/publicsite/futures/topics/all-schemes/agri-environment-climate-scheme/

Lag times may result from:

- The time required for a measure to become functional, e.g. vegetative riparian buffers progressively mature, resulting in gradual effects expressed over time (Newbold *et al*. 2008).
- The time needed for a measure to become effective at the catchment scale, e.g. measures reducing faecal bacteria require longer to become effective at a catchment scale than at a field-scale (Kay *et al*. 2012).
- Several natural catchment processes, rainfall and runoff delaying or counteracting the effects of measures (Hamilton, 2012). For example, pollutants that have accumulated in the soils and stream banks or stream-bed sediments in the past (i.e. legacy pollutants) continue to enter watercourses post-implementation through biogeochemical cycling and mobilisation during short-lived storm events.
- Land use (Gitau *et al*., 2010). Certain crop choices may have adverse effects on water quality, e.g. potatoes and winter cereal increase erosion risk, and may potentially lead to losses of nutrients and sediments in runoff.
- Non-linear ecological recovery trajectories, which may preclude the detection of ecological response until long after pollutants have been reduced (Withers *et al*., 2014).
- Monitoring design (Meals *et al*., 2010). Sample size determines the magnitude of change that can be detected with adequate statistical power. A large sample size, as a result of long-term monitoring, high **sampling frequency**, or both, enables the detection of a relatively small statistically significant change. A small sample size, such as that collected with operational WFD monitoring (EU, 2000, Annex V, Sect. 1.3), allows for the detection of a relatively large statistically significant change (Clarke 2013; Skeffington *et al*. 2015). Being able to detect a change larger than that actually occurring or predicted by modelling of the effectiveness of measures, translates into delays in documenting improvements. Thus, the sample size for WFD classification has the potential to introduce an additional "statistical" lag time.

Accounting for lag times is of major importance for SEPA because lag times control the time-scales required for water quality improvements to take place. A sound monitoring design should provide the basis for distinguishing between lag times (i.e. **noise**) and the true response to the measures (i.e. **signal**). It should also ensure sufficient sample size to minimise the potential for a statistical lag time and account for the actual time needed for a particular measure to be effective. Planning for a suitable sample size is a crucial challenge facing SEPA's water quality monitoring strategy. If small improvements occur unnoticed, this will have important implications for evaluating the Diffuse Pollution Plan, justifying its cost and communicating its environmental outcomes, or lack of, with stakeholders.

SEPA asked CREW, Scotland's Centre of Expertise for Waters, to provide expert opinion on the suitability of currently available monitoring data to reliably detect change and identify true response to diffuse pollution control measures. This report provides recommendations on the following issues:

- What statistical analyses are needed to identify change reliably?
- What is more suitable to show change reliably:
  - The concentration or load metric?
  - The spot or automated composite sampling technique?
  - The flow-proportional (i.e. sampling at fixed levels of flow or for fixed water volume) sample method or time-composite (sampling at fixed time intervals) sample method?

This report aligns with a parallel CREW report that developed a weight-of-evidence method to underpin the evaluation of the Rural Diffuse Pollution Plan (Akoumianaki *et al*., 2016). More specifically, the weight-of-evidence method assesses direction of travel using three criteria of effectiveness towards: achieving sufficient uptake of measures; compliance with Water Framework Directive (WFD) standards; and modelled reductions in pollutants. Diffuse pollution risks are also assessed by monitoring indicators of catchment change, i.e. fertiliser inputs, erosion risk crops, livestock, rainfall and other diffuse pollution pressures. Uncertainties in water quality data are addressed by estimating the sample size (number of samples) required to detect the improvements expected assuming 100% uptake of measures. Overall, the weight-of evidence method was found to be essential for understanding the interplay among the major catchment factors influencing water quality and where further action is needed. However, the weight-of-evidence method requires enhanced monitoring for pollutants and flow measurements to enhance the certainty of evaluations.

The monitoring recommendations developed here build on data from four priority catchments and demonstrate the shortcomings of currently available monitoring data, and the feasibility of the recommended statistical approaches, in the form of detailed tutorials. The recommendations are also based on published evidence and expert judgement to balance feasibility, practicality and scientific rigour. Lastly, the monitoring recommendations are discussed in the context of implications for SEPA's monitoring strategy for the evaluation of the Diffuse Pollution Plan.

2 Trial water quality and ecological data

Analysis was carried out on data on diffuse source pollutants and biological communities impacted by diffuse pollution, which were collected from 2007 to 2014 for WFD classification. The post-implementation period included data from 2011 to 2014. However, data from a pre-implementation period – collected at the same sampling frequency as post-implementation[3] – were also required. Therefore, data from the years 2007 to 2010 (sampled for WFD status classification before the start of the first RBMP) were selected to establish a baseline.

This project analysed priority catchment data from three waterbodies: the Lemno Burn, Eye Water (ID: 5011) and the Cessnock Water; and two bathing water catchments: the Eye Water and the River Ayr.

Background information about these catchments is described in detail by Akoumianaki *et al*., 2016. In brief, Cessnock Water is a Diffuse Pollution Monitored Catchment (DPMC) where a higher frequency sampling is carried out in parallel to monitoring for WFD classification. Arable land dominates land cover in the Lemno Burn and Eye Water with arable and mixed farming comprising the main diffuse pollution pressure. Other important pressures include sewage disposal at the Lemno Burn and livestock (i.e. poultry) at the Eye Water. Improved grassland dominates Cessnock Water and River Ayr, with livestock comprising the main diffuse pollution pressure.

In addition, DP GBR uptake post-implementation was greater than 50% at the Lemno Burn and Eye Water but lower than 50% at the Cessnock Water and the River Ayr (Brian McCreadie, pers.com). A DP GBR uptake greater than 50% has been considered sufficient to benefit water quality, on the grounds that half the amount of predicted reductions is sufficient to result in compliance with WFD standards in waterbodies at moderate status because of failing nutrients or FIOs. However, it is recognised that this is simplifying a rather complex relationship between measures and water quality improvements.

Trend and step-change analysis, analysis of minimum detectable change and sample size, and analysis for autocorrelation (Appendix 1) were performed using the following priority catchment data:

- Monthly spot samples of **dissolved phosphorus, ammonium and suspended sediment** from the **Lemno Burn and Eye Water waterbodies**.
- Biannual spot samples for Diatoms for Assessing River Ecological Status (DARES) and the Proportion of Sediment-sensitive Invertebrates (PSI) from **Lemno Burn and Eye Wate**r.
- Monthly spot samples of suspended sediment and ammonium from the **Cessnock Water**.
- Event-based samples of suspended sediment and ammonium from the **Cessnock Water**.
- Continuous flow measurements from the **Cessnock Water and the River Ayr**.
- Bathing season (May to September) spot samples of FIOs, i.e. Faecal Coliforms (FC) and Faecal Streptococci (FS) from **Eyemouth (Eye Water)**.
- Bathing season spot samples FIOs from the **River Ayr**.

A data set of daily spot samples of soluble reactive phosphorus from the **Tarland catchment**, which is out with the priority catchment approach, was also analysed to demonstrate the test for autocorrelation.

The trial catchments were selected by SEPA on the grounds of availability of the baseline data (pre-2011) required to develop the weight-of-evidence method for the evaluation of the Diffuse Pollution Plan (Akoumianaki *et al*., 2015). The current report does not determine whether the degree of DP GBR uptake, land use and other pressures affect the detection of change. This report uses the trial data to assess whether sample size is sufficient to enable change to be detected and to demonstrate what statistical analyses can be done with currently available data to assess response to measures.

[3] The pre- and post-implementation periods can have slightly different sample sizes but should have the same sampling frequency (Spooner *et al.*, 2011).

3 Estimation of change in WFD data

3.1 Developing a monitoring programme

WFD prescribes three types of monitoring to enable assessments to be made about status classification, compliance with the specified standards and the causes of change, or lack of change (Box 1).

BOX 1 Types of monitoring prescribed in the Water Framework Directive (WFD)

WFD distinguishes between three types of monitoring:

- Surveillance monitoring aims to assess long-term changes resulting from widespread anthropogenic activity.
- Operational monitoring is carried out to establish the status of those water bodies identified as being at risk of failing to meet their environmental objectives and to assess any changes in the status of such water bodies resulting from the programmes of measures.
- Investigative monitoring is carried out where the reason of any exceedance for ecological and chemical status is unknown; where surveillance monitoring indicates that the objectives for a water body are not likely to be achieved (and determine the causes); or to ascertain the magnitude and impacts of "accidental" pollution.

Source: EU, 2000, Annex V, Sect. 1.3; Borja *et al*. 2008.

WFD does not specify the spatial (i.e. number of sites, locations in the waterbody) and temporal (i.e. frequency and duration) scales of monitoring. **However, the most frequently asked questions when developing a water quality monitoring study are "How many samples and for how long?" Unfortunately, the correct response is: "It depends." (USDA-NRCS 2003).**

Indeed, there is no formula for determining the number and location of sites, the frequency and duration of any particular monitoring programme. The guidance for the design of a monitoring programme clearly prescribes that sampling frequency and duration must be determined once the objectives of monitoring have been determined and once an analysis of any background information on the water quality problem being addressed has been carried out. Vague or inaccurate statements of objectives lead to programme designs that provide too little or too much data, thereby failing to meet objectives or costing too much.

The objectives may vary from identifying the effectiveness of a single diffuse pollution measure (e.g. effectiveness of buffer strips) to assessing the effectiveness of a diffuse pollution mitigation programme such as the Diffuse Pollution Plan to monitoring to identifying compliance with specified standards as in WFD operational monitoring. The scale, frequency and duration of a monitoring programme varies with the objectives, and depends on site-specific natural variability, with frequency (sampling interval) being inversely proportional to the natural variability of the system (USDA-NRCS 2003). For example, monitoring to assess the effectiveness of a particular buffer strip installed at a particular reach of a stream requires sampling at the plot or field scale, with samples taken upstream and downstream of the buffer strip. Table 1 provides a summary of the general guidelines for the spatial scale, frequency and duration of statistical analysis that should be taken into account when developing any monitoring programme, which were followed here.

3.2 Developing a monitoring design for robust statistical analysis

The monitoring designs required to meet different objectives for the same waterbody may differ considerably.

The **first step** in developing the monitoring design for robust water quality assessments is to decide whether the primary evaluation tool is parameter estimation or hypothesis testing (US EPA 1997). As an example, parameter estimation can

Table 1 General characteristics of monitoring depending on objectives of a monitoring programme
Source: US EPA (1997 and literature cited therein); USDA-NRCS 2003

Objective	Scale of sampling	Frequency	Duration
Assess effectiveness of Diffuse Pollution mitigation programmes	Waterbody; River basin	High to medium	Long to very long
Assess effectiveness of individual mitigation measures	Farm plot; Field	High	Usually medium*
Establish baseline conditions	River basin; Waterbody	Low	Short to medium
Validation of model predictions	Farm plot; Field; Waterbody	High	Usually medium to long
Assess fate and transport of pollutants or relationship between pollutants and biota	Farm plot; Field; Waterbody; River Basin	High	Short
Long-term exposure (e.g. surveillance WFD monitoring)	Waterbody; River Basin	Low	Long
Compliance (e.g. operational WFD monitoring)	Waterbody; River Basin	Variable*	Dependent on regulations
Investigation / Research (e.g. investigative WFD monitoring)	Farm plot; Field; Waterbody; River Basin	Medium to high	Greater than project duration

* The frequency of compliance monitoring should be approximately equal to the frequency at which a pollutant or a biological indicator exceeds a specified standard, e.g. if exceedances occur once a month then monthly monitoring would be suitable (USDA-NRCS 2003).

Where,
Low frequency = Quarterly to annual for pollutants / Bi-annual to annual for biota
Medium frequency = Monthly for pollutants) / Quarterly (seasonal) for biota
High frequency = Daily to fortnightly for pollutants / Weekly to monthly for biota

Short duration = up to 2 years
Medium duration = 2 to 5 years
Long duration = more than 5 years
Very long duration = more than 10 years

be applied in assessments to determine pollutant loads from various sources, or waterbody status classification. Hypothesis testing is used in the evaluation of effectiveness of a single measure or of a diffuse pollution mitigation programme. Balanced (a.k.a. symmetrical) designs, i.e. two or more sets of data with the same number of observations in each set, are suitable and desirable for hypothesis testing, whereas parameter estimation can be carried out using unbalanced (a.k.a. asymmetrical) data, i.e. different number of measures from each waterbody (Gaugush, 1986 cited in US EPA 1997). Hypothesis testing is the appropriate approach to assess the effectiveness of the Diffuse Pollution Plan and typically requires more intensive databases than those needed for waterbody status classification.

The **second step** in developing the monitoring design is to decide whether sampling stations will be selected on the basis of a probabilistic or a targeted design (US EPA 1997). Probabilistic designs refer to random selection of stations (sampling sites/stream-reaches and/or sampling events to provide an unbiased assessment of the waterbody. In targeted designs, stations (sampling sites/stream-stretches) are selected on the basis of known effects (e.g. implementation of measures or not) or knowledge of upcoming events in the waterbody (e.g. waterbody-wide installation of various types of DP GBRs). Box 2 summarises the types of catchments in the priority catchment context.

BOX 2 Types of catchments for a robust monitoring design to identify step change in Scotland

"Impact" catchments (or waterbodies) are those catchments within the priority catchments where DP GBRs are sufficiently implemented. For example, sufficient implementation may be assumed when the degree of uptake exceeds 50% of farms, as suggested by Akoumianaki *et al*. (2016) in order to support the evaluation of the Diffuse Pollution Plan.

"Control" catchments (or waterbodies) are those catchments within the priority catchments where the DP GBRs are not in place yet or sufficiently implemented. For example, insufficient implementation may be assumed when the degree of uptake is below 50% of farms, as suggested by Akoumianaki *et al*. (2015) in order to support the evaluation of the Diffuse Pollution Plan.

"Reference" catchments (or waterbodies) are those catchments where there are no diffuse pollution pressures. "Reference" waterbodies represent baseline conditions and can be used when "control" data before and after introduction of measures are not available.

"Impact", "control" and "reference" catchments should be characterised by similar background conditions to enable changes that have occurred because of the measures to be revealed. Background conditions may be: land use as in LCM-07 (see footnote 5); rainfall regime; presence of particular pressure; or implementation of a particular measure or group of measures.

See also: Smith 2002; US EPA 1997: Chapter 2; Underwood 1994.

The most common and effective way to evaluate whether or not the measures have reduced pollutant concentrations and improved ecology in a waterbody, and to estimate magnitude of the effect, is by means of a Before-After/Control-Impact (BACI) design (Davey 2010; Green 1979; Smith 2002; Stewart-Oaten *et al*. 1986; Underwood 1991, 1994; US EPA 1997). Commonly applied variations of targeted BACI designs[4] in river and stream monitoring sites are presented below.

The "Single-Waterbody/Before-After" design is the simplest design. It requires data from a single location many times (depending on sampling frequency per year) before and after installing the measures (Green 1979). The design is confounded: the difference between before and after may not be related to the measures but due to response of pollutants and ecology to other factors and seasonality (Smith 2002). Effectiveness assessments based on this design cannot be transferred to other waterbodies (US EPA 1997).

The "Multiple-Waterbodies/Before-After" design is an improvement to the "Single-Waterbody/Before-After" design (Green 1979). This design can treat waterbodies as replicates, thus accounting for the true variability in the response of "impact" waterbodies to the measures (Smith 2002; US EPA 1997). This is essential to account for the variability from year-to-year and among waterbodies (Smith 2002; US EPA 1997: Chapter 2). The design may entail studying a number of "impact" waterbodies which have similar measures in place to examine whether response(s) to particular measures can be generalised. However, the Multiple-Waterbodies/Before-After" design may be challenging because there is no "control" waterbody to calibrate the response: the difference between before and after may not be caused by the measures, if the measures are not targeted properly, but by a factor (e.g. rainfall) that causes a common temporal response in the "impact" waterbodies. A solution to this problem could be to compare certain "impact" waterbodies that have a particular measure implemented with other "impact" waterbodies that do not have this measure implemented, to examine the effect of specific measures in the catchment context. A prerequisite for the latter approach is that the "impact" waterbodies have similar LCM-07-based land use[5].

The "Before-After/Upstream-Downstream" paired design uses data collected many times (depending on the sampling frequency) before and after installing the measures (US EPA 1997). In this design, data are collected from a location upstream and downstream from the area where the diffuse pollution measures are implemented within one or many "impact" waterbodies. Adding "control" waterbodies with upstream and downstream data is essential for a robust interpretation of change. This design is useful when it is necessary to locate monitoring sites above known point sources or areas within a waterbody where the measures are not implemented to remove their effects as confounding influences (US EPA 1997). Box 3 presents the potential of the "Before-After/Upstream-Downstream" design in bathing water catchments in Scotland.

[4] The term BACI design refers to a number of variations of the basic Before-After/Control-Impact design, which involves comparisons between one "impact" site and one "control" site one time before and one time after the impact. Here we present the major variations separately to prevent a misreading of the term "BACI design".

[5] The LCM 2007 (Land Use map) can be derived from a dataset consisting 23 target classes produced by the Centre for Ecology and Hydrology (http://www.ceh.ac.uk); it was used in the development of indicators to assess the effectiveness of the Diffuse Pollution Plan with the weight of evidence method described in Akoumianaki et al. (2015).

BOX 3 Potential for applying the "Before-After/Upstream-Downstream" design in priority catchments

The "Before-After/Upstream-Downstream" design uses data from a location upstream and downstream of the area where the diffuse pollution measures are implemented, within one or many "impact" waterbodies.

This design can be applied in bathing water catchments to help distinguish the effect of measures and other processes taking place outwith the operational area. In this case, downstream waterbodies will be those within the operational area, and will account for the effect of measures ("impact" sites); upstream waterbodies will be those outwith the operational area in the same bathing water catchment and will help isolate the effect of measures.

This design can be applied in bathing water catchments to distinguish between the effect of measures and losses from septic tanks. In this case, data from waterbodies with high septic tank effects from a particular bathing water catchment could be compared with data from waterbodies without (or with lower) septic tank effects in the same bathing water catchment.

Source: US EPA (1997).

The typical Before-After/Control-Impact (BACI) paired design compares one "impact" waterbody and a "control" waterbody, with data collected several times (depending on sampling frequency) both before and after installing the measures (Green 1979; Stewart-Oaten *et al.* 1986). The design identifies the effect of two different sources of variation: (i) change in pollutants and ecology caused by differences between before and after the measures; and (ii) change caused because of differences between "control" and "impact" waterbodies with similar land use (see footnote 5). Then, it estimates change due to the interaction of these two sources of variation. BACI assumes that other factors influencing water quality remain unchanged or, more realistically, change in a similar way without accounting for waterbody-specific variability. However, the BACI design applied at one "control" and one "impact" waterbody is confounded: any difference from before to after the measures may have been caused by differences between the selected "control" and "impact" waterbodies and not due to the measures (Underwood 1994).

The "Multiple-Control BACI" design (Underwood 1994) is a better variation of the typical BACI paired design. It compares impact and multiple "control" waterbodies. In this case, "Control" catchments may be selected on the basis of being representative of environmental conditions (i.e. similar land use, as in footnote 5) in the vicinity of an "impact" waterbody to represent background noise in the data (Underwood 1994). The major constraint of this design is that, if the implementation of measures is extensive, it would be difficult to identify multiple "control" waterbodies.

The "Multiple-Waterbodies BACI" design is described as a substantially effective factorial design in "effectiveness of measures" assessments (Smith 2002; US EPA 1997). This design accounts for all sources of variation: among "impact" catchments; among "control" catchments; and between before and after. As a result, this design has the potential for a robust characterisation of the effect of measures. As in the "Multiple-Waterbodies/Before-After" variation of BACI, the "Multiple-Waterbodies BACI" design may entail monitoring "impact" waterbodies with similar measures in place to examine whether response to particular measures can be generalised. A possible constraint, however, is that if the uptake of measures across a waterbody or a river basin is extensive, it would be difficult to identify a proper "control". However, certain "impact" waterbodies that have a particular measure implemented can be compared with other "impact" waterbodies that do not have this measure implemented, and therefore can be regarded as "control" for this particular measure or group of measures. It must be recognised that the "Multiple-Waterbodies BACI" design may involve complicated statistical analyses (Smith 2002; Underwood 1991) and requires careful planning to be effective.

The "Two- Waterbodies post-implementation" design, involves data collected many times (depending on sampling frequency) but only post-implementation and from two waterbodies at a single location in each one of them (US EPA 1997). This is a simple design but it does not account adequately for the effect of the measures as there is no "control". Improvements in pollutants and ecology may have been caused by other factors and not due to the measures. If there are no baseline data for "impact" waterbodies, it would be more informative to compare an "impact" waterbody and a similar (representative of environmental conditions) "reference" waterbody. However, the design would still be confounded, as any difference might have been caused by random changes in one of the two types of waterbodies used.

The "Multiple-Waterbodies post-implementation" design requires data from a single location at each of as many as possible, and definitely many more than two, waterbodies. This is also a very effective design in general (US EPA 1997). For a robust meaningful comparison, this design can include "impact" and "reference" waterbodies with samples collected during the same period of time to calibrate for the effects of rainfall and/or land use (as in LCM 07, see footnote 5) on the land management practices implemented (i.e. measures). For example, "impact" and "reference" waterbodies with similar hydrological regime can be used to isolate the recovery process from the effects of catchment hydrology. Alternatively, "impact" waterbodies that have particular measures in place at a particular type of land use (e.g. grassland) can be compared with "reference" waterbodies dominated by grassland.

The trend design may be more suitable for monitoring to assess the effects measures in groundwater, lochs and estuaries. The trend design requires data to be collected at a single location in a waterbody. It is the most suitable approach to detect change when the available data are collected on a long-term basis (as in Table 1 this report) with few gaps and medium to low frequency and by consistent sampling techniques (Hirsch 1988). It is also suitable when gradual, slow-rate water quality changes are expected (US EPA 1997), as in groundwater waterbodies, and a single station is available as in the cases of lochs and transitional waters. Baseline data are not necessary but if available they would be useful to distinguish the effect of measures from confounding processes.

The designs, and the associated data needs, described in this section are summarised in Table 2.

Table 2 Monitoring designs applied to assess effectiveness of diffuse pollution measures
Sources: Green 1979; Smith 2002; Stewart-Oaten *et al*. 1986; Underwood 1991 1994; US EPA 1997

Design	Data needs		Is the design confounded*?	Waterbody (as in WFD)
	Types of catchments (as in Box 2)	Baseline data		
"Single-Waterbody/Before-After"	"Impact"	Yes	Yes	River
"Multiple-Waterbodies/Before-After"	"Impact"	Yes	Depends	River
"Before-After/Upstream-Downstream"	"Impact"/"Control"	Yes	No	River
Typical BACI	"Impact"/"Control"	Yes	Yes	River
"Multiple-Control BACI"	"Impact"/ "Control"	Yes	No	River
"Multiple-Waterbodies BACI"	"Impact" /"Control"	Yes	No	River
"Two-Waterbodies post-implementation"	"Impact"/"Reference"	No	Yes	River
"Multiple-Waterbodies post-implementation"	"Impact"/"Reference"	No	No	River
Trend design	"Impact"	Not essential	Yes	Groundwater Lochs Transitional waters

* A design is confounded when it fails to distinguish between the effects of measures and other environmental and catchment processes, such as land use and rainfall.

3.3 Monitoring for the evaluation of the Diffuse Pollution Plan

Monitoring for the evaluation of the Diffuse Pollution Plan must answer four broad questions:

1. **Have pollutant concentrations declined since the Diffuse Pollution Plan was launched? Or has ecology improved since pollutant concentrations started declining?**
   Answering these questions requires performing trend analysis, which is suitable when data have been collected according to the trend design (see section 3.2). Testing for gradual and continuing changes, either increases or reductions in values (i.e. monotonic trends) can be performed on data collected since the Diffuse Pollution Plan was launched. Trend analysis was performed in the data from the trial catchments and the results are described in Section 3.3.1.

2. **Have pollutants concentrations changed significantly between before and after the implementation of the Diffuse Pollution Plan? Or has ecology improved between before and after the implementation of the Diffuse Pollution Plan?**
   Answering these questions involves a comparison of data between two non-overlapping periods of land management, i.e. pre- and post-implementation the measures. The analysis is known as step-change analysis. A robust identification of step-change requires a suitable design such as the Before-After design applied at multiple waterbodies, the BACI design, or the Multiple-Waterbodies design applied on post-implementation data (see section 3.2).
   Step change analysis was performed on the data from the trial catchments and is described in Section 3.3.2.

3. **How much change must be measured in pollutant concentrations or biological quality element to be considered statistically significant?**
   The answer here requires calculation of the minimum detectable change (MDC) between before and after the introduction of measures and is illustrated in Section 3.3 using data from the trial catchments.

4. **Is the current monitoring programme and number of samples sufficient to detect the change in pollutants concentrations predicted by the model assuming 100% DP GBR uptake?**
   This question can be answered with the calculation of the number of samples (a.k.a. sample size analysis) required to detect a pre-specified change in the mean pollutant concentration between data collected in the periods before and after the introduction of measures. Sample size analysis on data from the trial catchments is demonstrated in section 3.4.

The following sections analyse the advantages and disadvantages of trend analysis and step change analysis for the evaluation of the Diffuse Pollution Plan and provide recommendations in view of their suitability, or not, to detect significant change in pollutant concentrations and biological quality elements.

3.3.1 Trend analysis

Trend analysis has many **advantages**:

- It requires only post-implementation data.
- It is suitable in large river catchments with widespread and extensive implementation of control measures but with only one monitoring station in a receiving waterbody (Meals *et al*., 2011); **in the Scottish context, it may be the only feasible option for evaluating change in lochs, transitional waters, bathing waters and groundwater drinking waters post-implementation, as long as there are long-term data**.
- It is useful when available data come from waterbodies where the measures have long lag times (e.g. riparian buffers); in this case, trend analysis enables the assessment of the gradual changes occurring in parallel to the implementation of measures.

On the other hand, the **major limitation** of trend analysis is the requirement for long-term data (as in Table 1), which is rarely available for each catchment or pollutant of interest. **Additional key considerations include**:

- Trend analysis is not suitable for understanding the cause or causes of a trend.
- A pollutant trend may be confounded by seasonal cycles, flow variation and land management, or may be artificially produced by intensive frequency sampling due to the tendency of closely sampled pollutant concentrations to be similar, a.k.a. autocorrelation (see Section 3.5).

The shortcomings of trend analysis are illustrated in the trial data from the priority catchments, i.e.

1. The measures were introduced only four to five years ago; therefore, trend analysis is generally not suitable in the currently available data sets.
2. Significant 2007 to 2014 trends could be detected only for phosphorus (i.e. reduction by 8%) at the Eye Water (Figure 2b) and sediment (i.e. increase by 12%) at the Cessnock Water (Figure 2c). No trends could be detected in ammonium concentrations at any of the trial catchments (Figure 2a, b, c). Finally, no significant trend could be detected for FIO data from the Eye Water (Figure 2d) and River Ayr (Figure 2e).
3. The post-implementation trends of pollutants were not significant, despite the relatively large changes detected in phosphorus and sediment, which indicates that a data record of four years is too short to enable a significant trend to be detected.
4. The available data record for ecology is insufficient. No significant trends could be detected in the biological data, because of both the relatively short length of the record and the sampling frequency, resulting in only sixteen data points overall.

Flow data to enable concentrations to be adjusted for flow variation were available only at the Cessnock Water for sediment data and the River Ayr for FIO data. The magnitude of increase in sediments at the Cessnock Water was reduced from **+12% without flow adjustment to +5% with flow adjustment**. This indicated that a considerable amount of sediment increase was due to an increase in flow. However, it remains unknown how and what other factors (e.g. land management practices) are influencing sediment variation. No significant FIO trend could be detected with flow adjustment. Trend analysis in R code is demonstrated in Appendix 2.

To sum up, the findings show that trend analysis requires that post-implementation monitoring of water quality, ecology and flow data is longer than four years. The currently sample size (i.e. four years of data) is insufficient for a robust statistical analysis and a meaningful interpretation of water quality and ecological response to measures with long lag times.

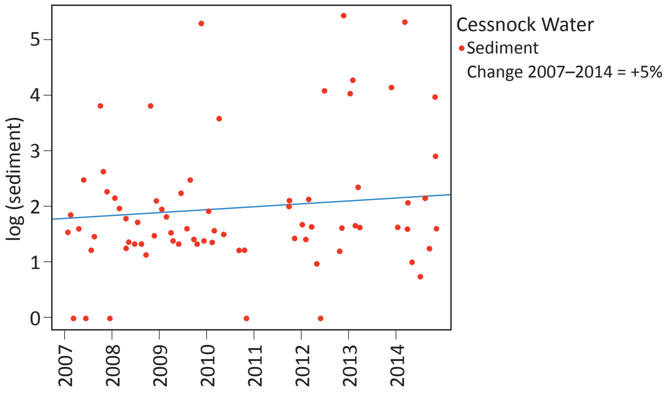
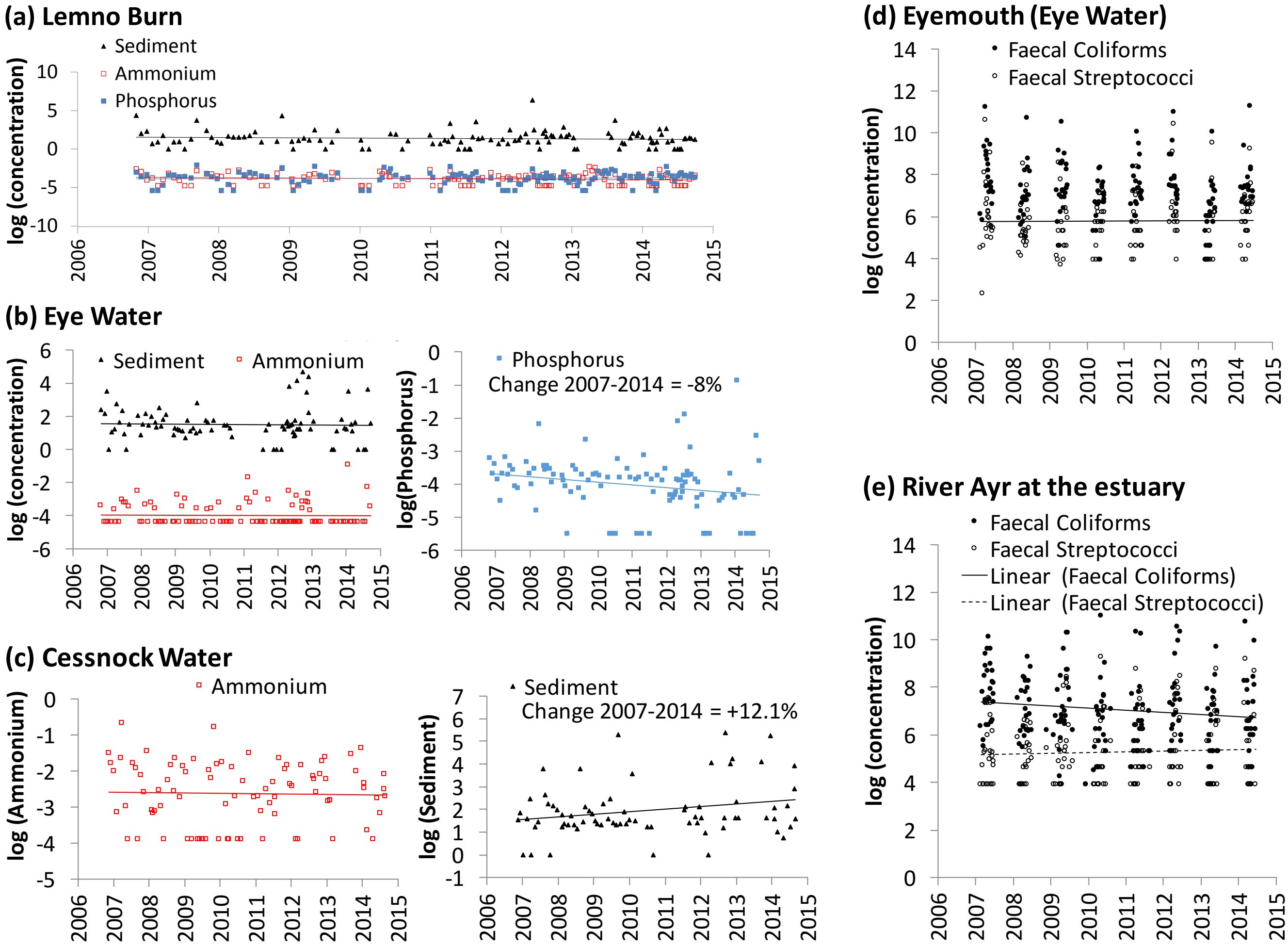


Figure 3 Estimated trend for seasonally and flow-adjusted sediment concentrations collected with spot sampling at the Cessnock Water.


(a) Lemno Burn

(b) Eye Water

(c) Cessnock Water

(d) Eyemouth (Eye Water)

(e) River Ayr at the estuary

Figure 2 Estimated trends without seasonal or flow adjustment based on log transformed data (Appendix 1). (a) Trends of dissolved phosphorus, ammonium and sediment concentrations at the Lemno Burn ; (b) Trends of ammonium and sediment, and dissolved phosphorus, which displayed a significant reduction by 8%; (c) Trends of ammonium and sediment, which displayed a significant increase by 12%; (d) Trends of faecal coliforms and streptococci at the Eyemouth; (e) Trends of faecal coliforms and streptococci at the River Ayr, where coliforms slightly declined but not significantly.

3.3.2 Step change analysis

The greatest advantage of step-change analysis is that it provides a **quantification of the effect of measures for a given period of time** (e.g. five years since launching the measures) in a particular waterbody.

There are also many **limitations in step-change analysis**.

Firstly, the interpretation of causes of step change, or lack of, depends on the design that has been selected to assess step-change (Underwood 1994). For example, step-change identification with the BACI design may be confounded by the lack of proper calibration of the assessment. In the BACI design "control" and "impact" waterbodies must be exposed to the same conditions (i.e. diffuse pollution pressures, geomorphology, hydrology, etc.) to enable true variability to be assessed. Identifying a proper "control" is difficult mainly because certain catchment processes remain unquantified, so the degree of similarity is uncertain. It is also possible that because the implementation of measures may be spread on a wide, river basin, scale, most or all waterbodies are treated as "impact" waterbodies. In the priority catchment context pressures and waterbody characteristics have already been assessed and DP GBR uptake is already tracked and reported. Thus, it would be easy to understand whether a proper "control" waterbody is available for a meaningful step-change analysis, or not.

Secondly, establishing cause-effect relationships between and the concentrations of pollutants when a step-change has been detected requires a carefully planned design. A reliable interpretation of step-change requires that the right measures have been installed at the right places across a waterbody so as to allow for the observed **changes in water quality to be linked reliably to the effects of measures** (e.g. reductions in FIO losses to be attributed to extensive installation of riparian fencing within a catchment). This can be addressed by selecting a design that includes multiple "impact" waterbodies (see section 3.2 for more detail). These designs have the potential to reduce the time needed to detect a significant step-change by increasing the power of analyses and removing background noise in the data caused by waterbody-specific response (Smith 2002; Underwood 1994).

Thirdly, long-term (as in Table 1) data are needed for a meaningful and statistically robust identification of step change. For example, long-term baseline monitoring is essential to enable the year-to-year differences between one or more "control" waterbodies and one "impact" waterbody to be understood in the BACI design. In general, monitoring duration must strike a balance between the time needed for a measure "signal" to be detected as a significant step-change, the availability of monitoring resources, and policy targets. Statistically insufficient short time-scales may be too long for policy deadlines and stakeholders, or too costly to implement.

Finally, a high-frequency water quality sampling (as in Table 1) is required for an unbiased identification of step-change. A high sampling frequency for pollutants (as in Table 1) has the potential to increase sample size and enable the detection of a small step-change (see section 3.3.6) as well as to enable the difference between temporary and long-lasting effects of the measures to be discerned). Identifying step-change with a balanced design would require samples to be collected at a high frequency in both "impact" and "control" waterbodies in a BACI design, and also in "reference" waterbodies, if baseline data are not available (as in the Multiple-Waterbodies post-implementation design) but this would certainly increase the cost of monitoring.

These limitations show that careful planning and targeting of resources (budget, equipment, staff) is essential to strike a balance between cost and the need to demonstrate that the measures are effective in practical time-scales. In addition, these limitations influenced the choice of step-change analyses performed on the trial data in the following ways:

1. It was not possible to apply the Before-After design in multiple waterbodies because the data came from waterbodies with different diffuse pollution pressures and measures, or to use a BACI design as data came from waterbodies with different environmental conditions. For example, it was not appropriate to compare FIOs between Eyemouth and River Ayr since these two catchments have different land use as well as different regional rainfall regimes (see section 2.0). According to Met Office data average annual rainfall in the period from 2007 to 2014 was 1132 mm in East Scotland (where Eyemouth is located) and was considerably higher in West Scotland (where the River Ayr is located), i.e. 1878 mm.

2. SEPA's data on specific DP GBR uptake rates were not available during the course of this project. Therefore it was uncertain whether the measures for reducing a specific pollutant were in place, or not. For example, it was not appropriate to consider the Eye Water as an "impact" catchment with respect to FIO mitigation because it was unknown whether uptake of DP GBR-livestock management was sufficient or whether the high SRDP spend for hedgerows targeted FIO sources (see also Akoumianaki *et al*., 2015).

3. The weight-of-evidence method applied in the water quality and land use data from the trial catchments demonstrated that land use and rainfall have the potential to counteract the measures (Akoumianaki *et al*. 2016). Therefore the results of BACI would be confounded by unaccounted sources of variation. For example, it was not appropriate to consider the Eye Water (waterbody 50101) as "impact" and the Cessnock Water as "control" waterbody with respect to phosphorus mitigation because of the different land use (i.e. Cessnock Water is dominated by improved grassland whereas Eye Water is dominated by arable land) and the East versus West difference in rainfall regime (see also Akoumianaki *et al*. 2015). Similarly, a Before-After design with multiple "impact" waterbodies, such as the Lemno Burn and the North Ugie Water, could not be used with respect to phosphorus because, despite similar land use, baseline phosphorus data from the North Ugie Water were not available.

To tackle these limitations, and in consultation with SEPA, step change analysis was carried out only between before and after launching the Diffuse Pollution Plan in 2011 and at each waterbody separately, despite advise by US EPA (1997: page 2-23) that this approach should be generally avoided in "effectiveness of measures" assessments. The method is demonstrated in Appendix 2 for pollutants with seasonal and flow-adjustment using sediment and flow data from the Cessnock Water and for ecological data with seasonal adjustment using diatom (DARES) and invertebrate (PSI) data from the Lemno Burn.

3.3.3 Minimum detectable change (MDC)

Step-change was not significant in any of the trial data with or without seasonal adjustment, with or without flow-adjustment, and with instantaneous or daily averaged flow measurements. To determine if this result was caused by ineffective implementation of the measures at the Eye Water and the Lemno Burn or was caused by a small sample size that is insufficient to enable the robust estimation of a significant change,it was necessary to calculate the **minimum detectable change (MDC)** with WFD monitoring.

The Minimum Detectable Change (MDC) is the minimum change in a pollutant or biological quality element, between before and after the measures are implemented, for a given sample size to be considered statistically significant and not an artefact of system variability. The **major advantage** of MDC is that it allows the evaluation of current or proposed water quality and ecological monitoring designs in terms of their effectiveness in detecting improvements post-implementation the measures. The same formula and the same datasets can be used to calculate the sample size required to detect a pre-specified change between before and after the measures are implemented in a given system (see also section 3.4 for Sample size analysis).

The **key considerations for MDC calculation with a Before-After design in priority catchments** are:

- **System variability**. As a rule of thumb, the higher the system variability, the larger the MDC that can be detected; but the point is to enable the calculation of smaller MDCs. In fact, MDC is proportional to the standard deviation pre-implementation of the measures (Spooner *et al*., 2011). Consequently, adjusting for sources of variation such as season and flow or land use before launching the measures will serve to reduce the MDC and increase the ability to detect a real change in water quality due to the measures.

- **Sample size (duration and frequency)**. MDC decreases with an increase in the number of samples. Increasing sample size by increasing the number of years of monitoring benefits interpretation of the MDC as it confirms that observed changes are not artefacts of unmeasured factors and increases. Increasing sample frequency reduces the MDC but only after adjusting for autocorrelation (see section 3.5).

- Feasibility of a monitoring programme/regime with the same sampling frequency **before and after the introduction of measures**. Calculation of MDC assumes that there is the same sampling frequency before and after the measures are implemented but this is not always possible. Fewer samples are expected to be available in the pre-implementation period due to the nature of operational monitoring (see also Box 1), which focuses on sites characterised as at risk. Also, fewer concentration data collected in combination with flow are expected to be available on a waterbody basis because measurement of flow is not required in WFD status classification.

- **Monitoring technique**. The way samples are collected affects sample variability and therefore the magnitude of MDC. For example, variability is usually higher with event-based sampling than with spot sampling, and higher with spot sampling than using the time-composited sample method (Stone *et al*., 2000). Data from time-composited samples also have a lower degree of autocorrelation than data from spot samples.

The effects of variability and sample size were illustrated in the trial data (Figure 4) and are also demonstrated in the tutorials shown in Appendix 2. To have an 80% probability of detecting a significant change with current sample size in the trial catchments, the change would have to be larger than the values listed in Table 3 below.

Table 3 Minimum detectable change (MDC) between before and after the implementation of the Diffuse Pollution Plan with current sampling

Parameter	Without flow-adjustment	With flow-adjustment
Dissolved phosphorus	29 to 37%	No flow data
Sediments	44 to 45%	38%
Ammonium	30 to 38%	37%
FIOs	42 to 49%	37 to 39%
Benthic invertebrates	15 to 17%	No flow data
Diatoms	26 to 44%	No flow data

These MDC values clearly showed that flow-adjustment with the current sample size can reduce MDC. Flow-adjustment was more effective for sediments and FIOs but not so important in reducing ammonium variation.

A smaller baseline than post-implementation sample size was available for phosphorus, sediments and ammonium at the Lemno Burn and Eye Water (Figure 4). However, at the Cessnock Water and River Ayr (Mainholm), where flow data were available, MDC calculation involved a smaller sample size post-implementation.

The most important finding of this analysis is that the magnitude of MDC in pollutants and ecology with current sample size on the basis of WFD monitoring is **unrealistic**. This is because the estimated significant changes are much larger than the reductions predicted by modelling of the effectiveness of the DP GBRs and the agri-environment schemes in Scotland implemented together (Gooday *et al*. 2014). The expected changes are about 30% for phosphorus and FIOs, and 5% for sediments. This analysis simply demonstrated that reducing MDC requires increasing the duration of sampling and/or sampling frequency.

3.3.4 Sample size analysis

The **usefulness of sample size analysis** lies in its ability to identify the time-scales of detection of expected improvements with the currently available WFD monitoring data. Using the modelled reductions as the assumed change, sample size analysis helped to estimate the sample size required to detect the expected improvements for water quality. Specifically, the model predicted reductions of 15 to 25% for phosphorus, about 17% for FIOs, and 2% for sediments (Gooday *et al*. 2014). For sediment, sample size could be estimated for above 5% change.

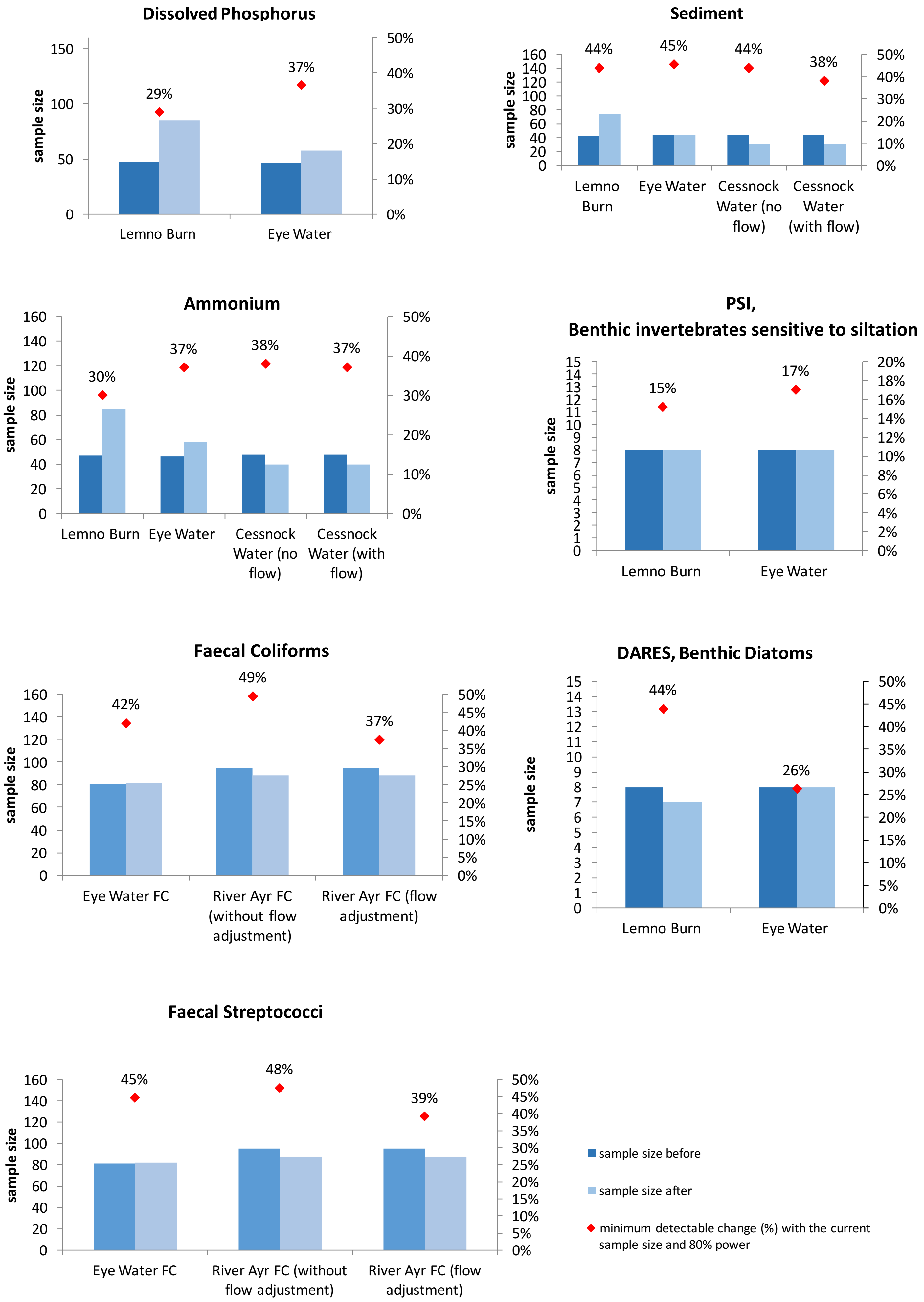


Figure 4 Sample size for WFD classification and minimum detectable change (%) with 80% statistical power, i.e. the magnitude of significant change that could have been detected reliably with current sample frequency and duration of monitoring.

The results of sample-size analysis using the trial data are illustrated in Figure 5. The analysis is demonstrated in Appendix 2. It is important to note that the available diatoms and invertebrate data were insufficient to enable estimation of the sample size for smaller changes.

The findings clearly showed that the magnitude of MDC is substantially reduced by simply increasing the years of baseline and post-implementation monitoring without changing sampling frequency (Figure 5). For a 20% reduction in phosphorus and FIOs and a 5% reduction in sediment to be detected, monitoring with current sampling frequency and assuming equal number of samples before and after the introduction of measures would require (Figure 5):

- Eleven to twelve years WFD monitoring post-2011 for FIOs at the River Ayr using flow-adjusted concentrations.
- 24 to 28 years WFD monitoring post-2011 for FIOs at the Eyemouth without flow-adjustment.
- Ten to sixteen years WFD monitoring post-2011 for phosphorus without flow-adjustment.
- More than 50 years WFD monitoring for sediments regardless of flow-adjustment.

The predicted reductions are intended to show what is feasible in terms of mitigation and not what change is needed to achieve compliance with WFD standards. The latter depends on the degree of impairment in each water body of interest, the levels of pollutant concentrations, and the constituent biological communities. Detecting the modelled reductions in each catchment with 80% statistical power is important in ensuring that the measures are effective towards delivering WFD objectives. Trial data from the first fourteen priority catchments showed that detecting an expected improvement post-2011 in FIOs and phosphorus (without flow-adjustment) requires longer time-scales than those required to meet the policy targets at the end of the first or even the second RBMP cycle, i.e. six and twelve years. Flow-adjustment reduced MDC for FIOs and sediment, but was insufficient to allow a step-change to be detected. **Consequently, the current monitoring introduces a considerable "statistical" lag time in the**

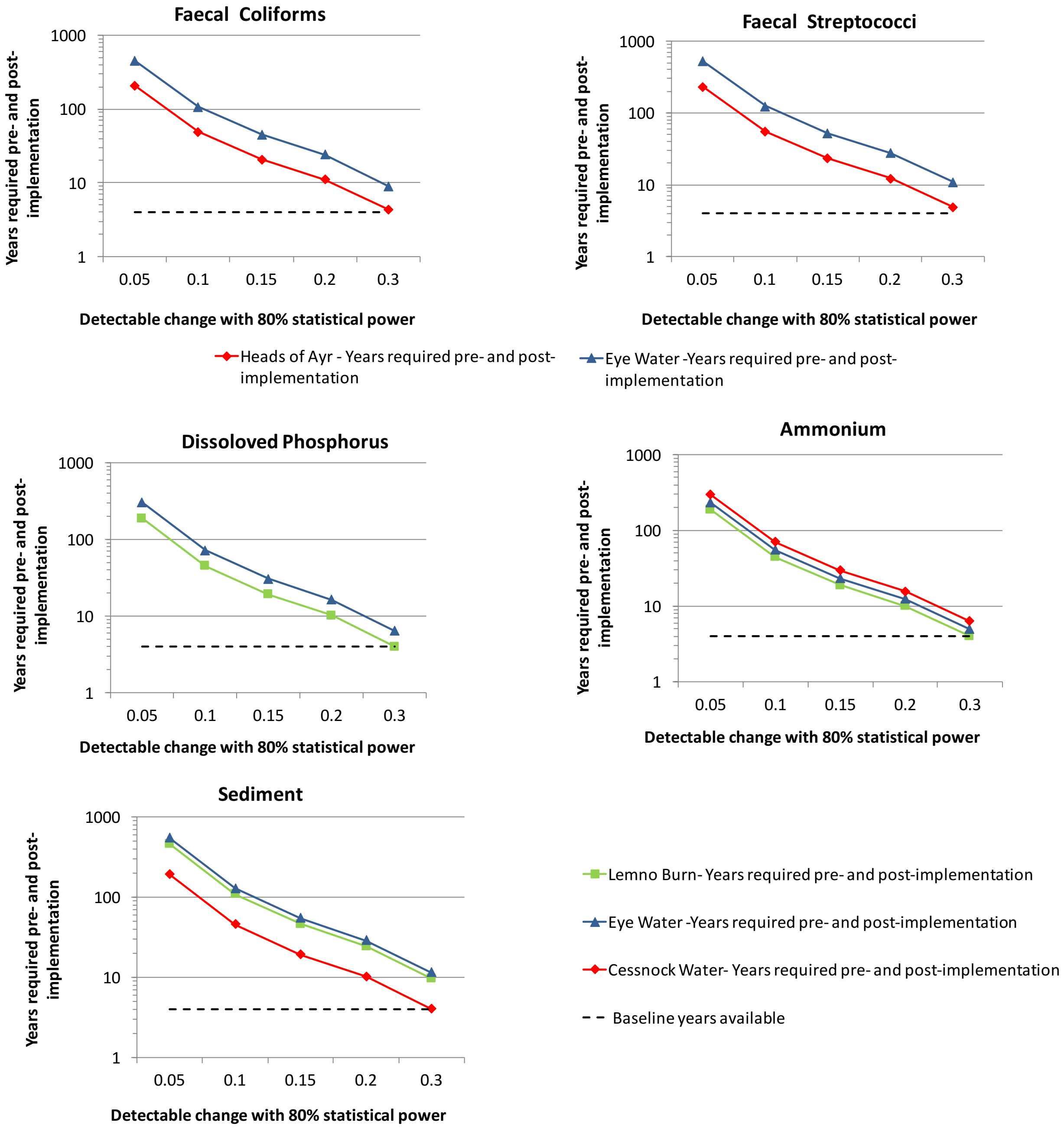


Figure 5 Detectable change with longer-term monitoring but current sampling frequency, presented as years of WFD monitoring.

evaluation of the effectiveness of measures, with the potential for a "pessimistic" bias.

It must be recognised that these considerations refer to the background variation in the trial catchments. There may be catchments where phosphorus and FIOs need shorter or longer term monitoring to show a significant reduction by 20% in response to DP GBR uptake because of different spatial effects of the measures and biogeochemical lag times. In addition, the relation of MDC and sample size with the DP GBRs implemented to reduce a specific pollutant remains unknown without considering multiple "impact" catchments, or "control" or "reference" catchments to account for background variation. The findings are useful in understanding the shortcomings of current monitoring but cannot be extrapolated.

To sum up, compliance with DP GBRs had little or no effect on MDC with current samples size or sample size with current sampling frequency. This is indicative of the mismatch between the purpose of WFD (i.e. status classification) and the need for assessing whether the Diffuse Pollution Plan has been effective in reducing the amount of pollutants. If the time-scales required for detecting a 20% change in phosphorus and FIOs, and a 5% reduction in sediments with current sampling frequency are impractical, we need to consider whether a higher sampling frequency is meaningful and feasible.

3.3.5 A tale of caution and autocorrelation

Sample size can also be increased by altering sampling frequency to enable the detection of smaller changes with a specified statistical power (i.e. 80% in this report). However, the benefit from increasing the sampling frequency may not be as great as that from increasing duration because of the effects of autocorrelation. **Autocorrelation** exists if an observation at a given time is correlated with observations taken at previous or subsequent times; as such, autocorrelation is a function of the lag time between observations. If there is significant autocorrelation, it needs to be accounted for in trend analysis and in the estimation of step change. Otherwise the standard error of the trend or step-change will tend to be underestimated; this may result in the response/change being regarded as significant when in fact it is not. The effect of autocorrelation can be visualised by plotting the partial autocorrelation function (ACF) to assess whether observations are independent in time. The test for autocorrelation using the Tarland dataset is demonstrated in Appendix 3.

The effect of autocorrelation was illustrated using daily soluble reactive phosphorus data from the Tarland catchment. Partial ACF plots show that the lag one autocorrelation coefficient is around 0.7 but values at subsequent lags are inside or only very slightly outside, the significance level indicated by the dashed lines (Figure 6a). This roughly translates to having only 64 **independent observations** to detect change instead of the 365 daily observations actually taken. Such discrepancies between sampling effort and efficacy show the importance of planning to account for the effect of autocorrelation and the risk of wasting monitoring resources.

On the other hand, autocorrelation was not significant in the monthly sediment data (about 12 observations per year) from the Cessnock Water. ACF plots clearly showed that with monthly sampling all the autocorrelation coefficients lie within the dashed lines (Figure 6b), but as sample size analysis showed, monthly sampling frequency is insufficient in detecting a significant step change or trend in short-term time-scales.

3.3.6 The sampling technique question: spot, flow-proportional or time-composited samples for pollutants?

The sampling techniques for the evaluation of diffuse pollution mitigation measures have been widely described and evaluated for their ability to provide robust estimates of change in water quality. There are two broad categories of sampling techniques:

- Spot-sampling, widely used by SEPA in priority catchments for status classification.
- Automated sampling, used in SEPA's Diffuse Pollution Monitoring Catchments (see section 2.0); it includes the flow-proportional and the time-composite sample methods

In the flow-proportional (FP) sample method, a flow signal, usually but not always high (hence "event-based" is a misnomer of the term "flow-proportional"), indicates when a fixed volume increment of water has passed the flow-meter of the auto-sampler. As this signal comes in pulses, variability of

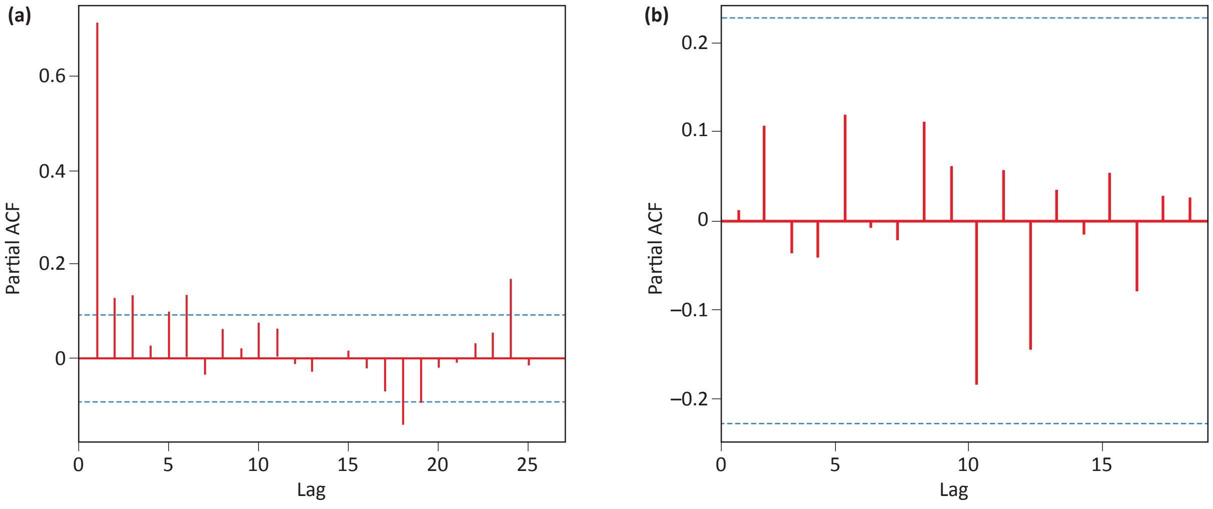


Figure 6 Partial autocorrelation function (ACF) plots. (a) Partial ACF for the log of daily measurements of soluble reactive phosphorus from the Tarland catchment in the period 2004-2005. Lag one is significant but subsequent lags are inside or only slightly outside the dashed lines, which represent the significance threshold. (b) Partial ACF for the log of monthly measurements of sediment concentrations from the Cessnock Water in the period 2007–2014.

flow over a period of interest leads to varying sample volumes being collected. If storm events are frequent, as in Scotland, then the period of collecting and retrieving samples must be short. *A priori* knowledge of the range of flows for a given period is critical in planning the frequency of retrieval or the number of aliquots required to build a composite sample, but this is rare. Alternatively, base-flow and storm flow signals can be used in combination with retrieval of composite samples on a weekly basis (e.g. for sediments as in Abtew and Powell, 2003; National Research Council US, 2000).

In the time-composite (TC) sample method, the auto-sampler is programmed to collect time-based composite samples comprising sub-samples (aliquots) taken at fixed-time intervals. The advantage over the FP sample method is that planning is easier because programming for an unknown range of flow signals is unnecessary. A widely accepted approach involves weekly compositing of 24 samples collected at 7-hour intervals, also known as the "24/7 solution" (Jordan and Cassidy, 2011; see also section 3.3.8 this report).

A review of the benefits of the FP and TC methods used for the evaluation of the effectiveness of measures in the US can be found in Abtew and Powell, 2003; National Research Council US, 2000; and Stone *et al*., 2000.

Both spot and automated sampling techniques provide the potential for collecting robust data. **Data from spot sampling provide the benchmark against which data from automated sampling will be assessed**. The key considerations for selecting sampling technique include:

- **Representativeness across a range of flows** (baseflows and stormflows) at a frequency that minimises the risk of autocorrelation in the data and reduces the magnitude of MDC (US EPA 1990; 1997). In this respect, **automated sampling techniques are essential** to capture representatively all sources of variation in pollutant concentrations and to achieve a sufficient sample size to detect change.
- **Site accessibility and landowner cooperation** in data collection efforts to enable a frequent, fit-for-purpose frequency to be applied (US EPA 1997).
- **Feasibility**. Automated sampling depends on many factors, the most important being set-up cost to cover a network of stations and manpower for retrieving the samples from all stations at the appropriate intervals.

It must be noted that a fixed-date spot sampling on monthly or fortnightly basis is by definition non-representative of the elevated pollutant loads during events and it is unable to capture the highest concentrations and lowest concentrations. If stormflow concentrations remain undocumented, it is impossible to distinguish between their immediate and long-lasting effects on stream biota. Additionally, certain measures such as riparian buffer strips have the potential to reduce losses of pollutants; depending on local circumstances and type of pollutant, some measures may reduce losses during storm runoff and some may reduce leaching after the event. Event-based sampling accounts only for high concentrations during storm-events and not for the concentrations affected by leaching and subsurface (delayed) flow outwith the wet spells signal. **Consequently, without documenting the amount of pollutants during both events and low flows it is questionable whether the effects of measures can be actually understood and evaluated**.

Additional considerations for the use of automated sampling techniques are summarised in Table 4.

The trial data illustrated the limitations of the spot and event-based sampling using sediment concentrations from fixed-date spot sampling and irregularly-collected, daily maximum

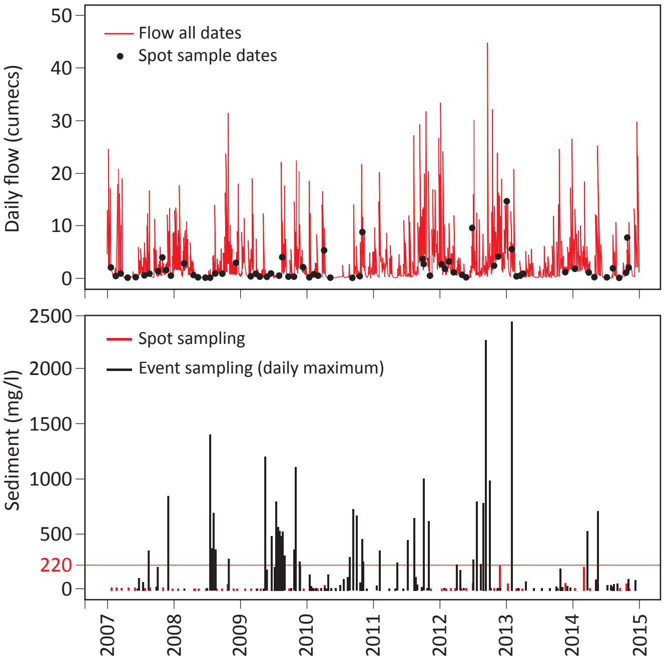


Figure 7 Effect of monthly spot and event based sampling on sediment concentrations at the Cessnock Water. Red line denotes the maximum concentration (average of three values) observed with spot sampling.

Table 4 Key considerations for the collection of in-stream pollutants with automated sampling
Source: Abtew & Powell 2003; National Research Council US 2000; and Stone et al. 2000; and US EPA 1997

Type of sampling	Key considerations	Suitable for
Flow-proportional (FP) sampling	Prior knowledge of the range of flows at a site Programming the auto-sampler for both low and storm flows Refrigeration of the samples before retrieval Retrieval of samples when needed	Nutrients / sediments / FIOs
Event-based (storm flows) sampling	Must be combined with weekly or fortnightly spot sampling No specific resource constraints for an unpredictable number of events Refrigeration of the samples before retrieval Daily retrieval of event samples Accessibility	FIOs
Time-composite (TP) sampling	Refrigeration of the samples before retrieval Compositing should be at a frequency to minimise risk of autocorrelation and to reduce MDC Auto-samplers should be located near or at flow gauging stations	Nutrients / sediments

concentrations from event-based sampling[6] from the Cessnock Water (Figure 7). Spot sampling captured a few elevated flow events but it practically missed several events in each year. The maximum spot-sampled concentrations (i.e. average: 220 mg/l) were observed only three times throughout the data record. However, event-based concentrations were two to ten times greater than this and occurred several times in each year.

An important problem in the event-based trial data alone and in combination with spot sample data was the different sample size and frequency pre- and post-2011. This irregularity resulted from spot sampling at monthly intervals being insufficient to cover the gaps of event-based sampling. Also the long gaps between event-based samples **indicated the need for better planning of the retrieval of event based samples to enable all events to be sampled properly.**

3.3.7 The flow question: concentrations, flow-adjusted concentrations or loads?

Concentrations vary with flow but each pollutant has a different relationship with flow. In this respect, monitoring should account for both concentrations and flows to enable the effects of flow on pollutant variation to be understood and uncertainties to be minimised and assessed. However, such sampling is resource intensive and rarely available for each pollutant and at each waterbody across Scotland. Therefore, it is worth considering why and where flow measurement is an essential component of the monitoring for the evaluation of measures.

Removing the variation in concentrations caused by flow requires flow-adjustment, a common statistical technique that prevents the identification of a change in pollutant concentrations when it is the result of correlation with flow. The benefits of flow-adjustment in improving the ability to detect a significant trend and reduce the magnitude of MDC are widely accepted and have already been clearly demonstrated in the trial data (Section 3.1 to 3.4). Removing the flow-related variation from the data makes step-change tests more powerful and prevents the identification of a step-change in concentration when it is the result of correlation with flow. In this respect, flow measurement is essential.

The major challenge for reliable flow-adjustment is collecting concentration and flow data (i.e. paired measurements) with a sufficient sampling frequency and at a waterbody scale. Ideally, flow measurements should be taken concurrently with samples for pollutant concentrations from the same waterbody to ensure that flow-adjustment removes the site-specific effect of flow on water quality. If temporally and spatially concurrent data are not available, as in priority catchments, then step-change estimation is subject to uncertainties. In such cases, it is questionable whether flow-adjustment adds meaningful information or bias; if errors remain unquantified, simple concentration data will be more reliable, especially when collected with appropriate sampling technique (e.g. automated sampling and time-compositing) and frequency.

Integrating concentrations and flows involves the use of a different metric, i.e. loads (concentration multiplied by flow). **Load estimation does not remove variation due to flow therefore is not suitable in detecting the signal of the measures. The availability of flow data determines whether load estimation provides reliable, additional information, or not** (Webb *et al*., 1997). Load estimation requires paired, preferably concurrently collected, measurements of concentrations and flows at regular and relatively frequent intervals, such as monthly or more frequent than monthly intervals. This is to enable effective representation of the relationship between flow and concentrations (Littlewood, 1992).

Good load estimates can usually be derived from continuous flow data and intermittent frequency (high or medium as in Table 1) data on pollutant concentrations. Although sampling frequency requirements depend on system variability, quarterly concentration observations are generally inadequate; monthly observations will probably not yield reliable load estimates; and even weekly observations may not be satisfactory, especially if very accurate load estimates are required (Meals *et al*. 2013). In Swiss streams, for example, more than 30-50 samples must be collected per year to detect a change in load estimates of SRP to increase statistical power (Moosmann *et al*. 2005). However, it has been shown that a high sampling frequency of paired concentration-flow measurements (e.g. fortnightly or weekly) which increases the chance of capturing flood events, alongside monthly pollutant-only monitoring, effectively reduces bias in load estimation (Bieroza *et al*. 2014; Cooper & Watts 2002; Johnes 2007; Skarbovik *et al*. 2012).

Alternatively, flow data from gauging stations in, more or less, adjacent water bodies with similar hydrological characteristics (a.k.a. analogue catchments) may be used for load estimates based on statistical modelling. The reliability of such approaches is questionable and relies on two main issues. **First**, load estimation should consider how closely the underlying assumptions of "similarity" represent the actual pollutant-flow relationship at a specific water body. For example, the concentrations of certain dissolved pollutants transported to watercourses through runoff and leaching are not necessarily proportionally correlated with flow. Therefore in addition to rainfall-runoff processes, soil permeability, land use and the specific control measures in place, must be also considered in catchment selection. **Second**, load estimation depends on the selected load estimation method, e.g. simple averaging, ratio or linear interpolation and use of rating curves (for a review of methods see: Littlewood 1992; Webb *et al*. 1997).

Interestingly, **the amount of flow-related uncertainties that can be tolerated depends on the water quality issue under consideration and not on the use of metric** (Hirsch *et al*., 1991):

- In mass balance approaches of **water quality regulation** it is essential to ensure that load assessments refer to concentrations and flows from a particular water body. The total maximum daily loads (TMDL) approach developed by US EPA[7] exemplifies this need for site-specific measurements. At each waterbody, TMDLs must clearly and reliably identify the links between the pollutants failing standards; the source apportionment of pollutants; and the pollutant load reductions needed to meet the applicable water quality standards.

- In **source apportionment studies that answer research questions** with potential applications in catchment management and

[6] The cause of irregular sampling is that the auto-sampler used for sampling pollutants was programmed to be triggered above a certain flow-threshold, which has an irregular distribution over time.

[7] http://water.epa.gov/lawsregs/lawsguidance/cwa/tmdl/overviewoftmdl.cfm

modelling, errors due to the use of concentrations and flow measurements from different catchments can be tolerated, as long as the "similarity" assumption is satisfied and load estimation is based on a suitable method. Examples include mass balance approaches for all source and sinks of pollutants in a given catchment to validate source apportionment models or models of export or removal of pollutants, e.g. sediments, from a catchment (e.g. Chen *et al*., 2013).

- In studies aiming to understand the **cumulative effects of diffuse pollution** in systems with long-residence times of pollutants, load estimation does not require high frequency sampling of flows, if inflow rates are relatively stable and the flow-concentration relationship is known. However, it does benefit from measurements of inflow rates into the system of concern to provide a robust basis for calculation of the pollutants imported into a system. Examples include the monitoring of effects of conservative[8] pollutants (e.g. sediments, toxic substances) on biological communities in low water turnover receiving waters, such as lochs and baseflow dominated areas for the conservation of wildlife, fishing and shellfish harvesting (Hirsch *et al*., 1991).

3.3.8 Pollutant sampling frequency

The 2007–2014 trial data available for this report were insufficient for a comparison of step-change and MDC estimates using data collected at different sampling frequencies. In any case, there is evidence that daily pollutant data would be autocorrelated (see section 3.6). A reasonable question arises: what is the sampling frequency that enabled the detection of a significant change where diffuse pollution mitigation projects were implemented elsewhere?

The methods and results from a variety of mitigation and status characterisation projects have been collated and the information for effective sampling frequency summarised by type of catchment or waterbody in Table 5a and by pollutant type in Table 5b. In general, sufficient sampling frequency to represent background variability and detect change is found to be inversely proportional to catchment size, and water residence time (e.g. Bertram and Balance 1996; National Research Council, 2000; US EPA 1997). In the UK, however, catchment hydrology (e.g. baseflow index) exerts a stronger influence than catchment size on water quality response and therefore on sampling frequency (Cassidy and Jordan 2011; Skeffington *et al*. 2015).

Table 5a Best sampling frequency options for assessing the effectiveness of diffuse pollution mitigation programmes in specific types of waterbodies

Type of waterbody	Best sampling frequency option
Flashy / small streams	Fortnightly or weekly sampling*
Baseflow-dominated streams	Fortnightly or weekly sampling
Groundwater	High productivity formations: Monthly to quarterly sampling Low productivity formations: Annual sampling
Standing waters (lochs)	Seasonal (quarterly or biannual) sampling

*Sampling technique (spot or automated sampling) is addressed in Table 4 and Table 5b.

References as in text: Bertram and Balance, 1996; US EPA 1997; Kronvag and Bruhn 1996; National Research Council 2000; Skeffington *et al*. 2015; USGS 2006.

As shown in Table 5a, high-frequency concentration sampling i.e. daily, weekly or fortnightly monitoring, is suitable for flashy and small streams (Bieroza *et al*. 2014; Kronvag and Bruhn 1996; US EPA 1997). On the other hand, low-frequency (i.e. monthly to quarterly for small highly permeable aquifers and annual for less permeable aquifers or large, high productivity formations) concentration data are suitable for groundwater monitoring (Bentram and Balance 1996; US EPA 1997; USGS 2006). Sampling in standing waters (e.g. lochs) for both pollutants and biota is usually seasonal, i.e. quarterly or biannual, but it may require increased monitoring to enable impacts to be adequately characterised (Bentram and Balance 1996).

Different sampling frequencies and strategies may be needed for each pollutant (Table 5b). High sampling frequencies (i.e. at daily, weekly or fortnightly intervals) have been found to enable representative sampling of nutrients, both dissolved and particulate, and sediments across a range of flow regimes in small rivers and streams (Bieroza *et al*. 2014). Brauer *et al*. (2012) found that suitable sampling frequency is site-specific and may vary from two days to monthly for total nitrogen (TN) and daily to fortnightly for total phosphorus (TP) and turbidity. Generally, dissolved phosphorus, turbidity and sediments have been found to require a higher frequency than TN, but weekly or fortnightly spot sampling is regarded as the best frequency for all nutrients, turbidity and sediments (Bieroza *et al*. 2014; Brauer *et al*. 2012; Pott *et al*. 2014; Thompson *et al*. 2014; Grove *et al*. 2015).

One of the best sampling frequency options for nutrients and sediments is the "24/7 solution" described by Jordan and Cassidy (2011). The "24/7" sampling frequency solution has

Table 5b Sampling frequency to detect change in water quality (concentrations or loads) between before and after the implementation of measures: best monitoring option per pollutant

Pollutant	Best options for sampling frequency to assess effectiveness
Suspended sediments	Weekly compositing of 7-hourly samples collected by auto-sampler Weekly or fortnightly spot sampling
Total Phosphorus	Weekly compositing of 7-hourly samples collected by auto-sampler Weekly spot sampling
Dissolved phosphorus, SRP, o-phosphate	Weekly compositing of 7-hourly samples collected by auto-sampler Weekly or fortnightly spot sampling
Total nitrogen	Fortnightly spot sampling or weekly compositing of 7-hourly samples collected by auto-sampler Weekly or fortnightly spot sampling
Nitrates in groundwater	Monthly to quarterly in highly permeable aquifers Annually in less permeable aquifers Use of predictive modelling in unmonitored sites
Turbidity	Fortnightly spot samples Event-based sampling, if a paired-catchment design is applied
Nitrates and pesticides in groundwater	Monthly to quarterly in highly permeable aquifers Annually in less permeable aquifers Use of predictive modelling in unmonitored sites
Faecal Indicator Organisms (FIOs)	In streams: Combined routine spot sampling (in bathing water season in bathing waters) and event-based (during events and for 72 hours after rainfall).

References as in text: Bertram and Balance 1996; Brauer *et al*. 2009 Vinten *et al*. 2011; Francy *et al*. 2000, 2006; Grove *et al*. 2015; Jordan and Cassidy 2011; National Research Council US, 2000; Neal *et al*., 2011; Pott *et al*., 2014; Thompson *et al*., 2014; USDA-NRCS 2003; USGS 2006..

[8] Not altered by the biological processes that occur in natural waters.

been shown to be the best practical way forward (Table 5b) when automated samplers can be installed in a number of river waterbodies. It requires 24 samples to be taken at 7-hour intervals and composited on a weekly basis, resulting in 52 composite samples a year (Jordan and Cassidy 2011; Neal *et al*. 2011). Interestingly, the "24/7" solution and weekly spot sampling, result in the same number of samples for nutrient analysis. However, automated time-composite sampling has a much greater potential to capture the background noise influencing pollutant variation than weekly spot sampling (Jordan and Cassidy 2011). Thus, compositing has the potential to provide the basis for more reliable estimates for a given high frequency and, with a proper BACI design, for a shorter (but not short as in Table 1) monitoring duration.

Table 5b also shows best sampling frequency options for pollutants impacting groundwater systems, such as nitrates and pesticides. In general, groundwater pollutants are sampled on a monthly to quarterly basis in high productivity formations and on an annual basis in low productivity formations and large aquifers (e.g. USDA-NRCS 2003). However, it must be acknowledged here that the collection of long-term and consistent data from a number of sites within a catchment has also been found to be essential to assess long time lags and long-term exposure. For example, a persistent pesticide can remain in groundwater long after its use is discontinued because of the low rate of groundwater flow and the resulting long residence times. In such cases, pesticides would require long-term monitoring (i.e. more than five years as in Table 1) in combination with high-frequency monitoring in certain seasons to provide early warnings and for updating and improving models (USGS 2006).

Sampling for FIOs in bathing waters is typically restricted to bathing season with fixed-date spot sampling. The best option for effectiveness monitoring of FIOs is to combine fixed-date (regulatory) sampling with event based sampling (Table 5b). For example, Francy *et al*. (2006) report sampling during the bathing water ("recreational") season which represented "a range of conditions: during dry, calm weather; after a light or heavy rainfall; and during increased wave heights." Francy *et al*. (2000; 2006) concluded that FIO sampling should also be carried out during events (event-based) and immediately after the event, for the next 72 hours, to enable FIO losses to be assessed and modelled adequately. FIO sampling in streams should generally include both fixed date sampling, with up to 18 to 20 samples a year, and event-based sampling (Francy *et al*. 2006). Combining event-based and routine spot sampling for FIOs in bathing waters greatly increases monitoring effort. However, it is recognised that without FIO event-sampling in bathing waters, it is impossible to understand whether measures such as fencing and riparian buffer strips have been effective in reducing FIO losses in agricultural runoff, or not (Francy *et al*. 2000; 2006). FIO sampling in groundwater could vary depending on recharge rates and flow paths (Francy *et al*. 2000), therefore, there is no general guidance.

Evidence on a suitable sampling frequency for pesticides in rivers depends on local circumstances, type of product, toxicity and seasonal use. Therefore, no best option guidance could be included for in-stream pesticides in Table 5b. In some countries (e.g. US), national-scale modelling is being used for predicting pesticide levels in unmonitored streams (USGS 2006). Such spatial extrapolation is fundamental to extending the evaluation of sources and factors that affect pesticide occurrence – such as pesticide use, climate and soil – in unmonitored areas. It must also be noted that pesticide sampling may take place at two or more sites within catchment and a high frequency sampling may be applied when intense nutrient or pesticide use coincides with periods of high runoff to the system (USGS 2006).

Interestingly, in many cases more than one sampling technique is combined to deliver a result with sufficient statistical power. In addition to accuracy and precision, combinations are dictated by practical issues such as cost-effectiveness and lack of knowledge on flow range in a site, e.g.

- In a pasture-dominated river basin in north-western Arkansas, US, event-based sampling was combined with fortnightly spot sampling (or weekly, in flashier waterbodies) of pollutant concentrations (e.g. Gitau *et al*., 2010). This approach was the best available option due to lack of evidence on flow range and flow measurements at each water body.
- In small lowland streams in Denmark, monthly sampling in summer (i.e. low-flow season) and fortnightly in winter (i.e. high-flow season) has been found to represent background variation well (Kronvag and Bruhn, 1996).
- In the Lunan Water, Scotland, turbidity measurements during 130 events were found to be adequate for assessing the effect of measures (Vinten *et al*. 2011). However, it must be recognised that this was enabled using a paired-catchment design.
- In England, as part of the Initiative for Catchment Sensitive Farming, The frequency of routine (WFD) spot sampling was increased (from monthly to weekly or twice-weekly) and automatic water quality samplers collected additional samples during high flow events (in more 'flashy' catchments). Without weekly spot and event-based sampling, pollutant loads would have been under-estimated by an average of 17 % and subject to additional uncertainty of at least +/- 40 % (CSF Team, 2011).

3.3.9 Ecological sampling

Ecological monitoring is essential to identify how well a waterbody supports aquatic life, what kind of life and, if not, to understand the type of pressures on biota. This is because aquatic organisms integrate the exposure to various or specific stresses over time. It must be emphasised, however, that the objective of a monitoring programme (e.g. compliance or effectiveness of mitigation monitoring) must be taken into account before assessing the suitability of the frequency and duration of ecological sampling.

More specifically, monitoring diatoms and invertebrates to assess compliance with WFD standards and to identify ecological status, using the most recent three-year window of data sampled biannually with one replicate at a specific reach of a waterbody, is sufficient to inform WFD status classification. In addition, the ecological quality ratios (EQRs)[9], developed as indicators of specific types of stress (e.g. eutrophication, sedimentation, dissolved oxygen) in the WFD classification procedure, represent the composition of species in a given community compared with a reference representing minimum

[9] The overall status class for a waterbody is based on the use of EQRs, estimated status classes and rules for combining metrics and status classes for one or more sampled (or surveyed) biological quality elements (BQEs), namely fish, macroinvertebrates, phytoplankton, phytobenthos and macrophytes.

impairment due to excess of a specific pollutant. Therefore, interpretations of status classification do not require paired ecological monitoring at "reference" and "impact" sites to identify WFD ecological status. This is remarkably useful given resource constraints (i.e. manpower, cost, taxonomic skill).

However, monitoring for assessing WFD ecological status is increasingly regarded as being insufficient to address ecological response to changes in land management and pollutants because it assumes that the ecological recovery is a linear and immediate response to the temporal patterns of pollutants (Jarvie *et al*. 2013). For example, Kelly *et al*. (2009a) showed that six samples (collected through biannual sampling for three years in one site per waterbody) are sufficient to prevent seasonality, spatial differences and background variability from confounding WFD status classification. Kelly *et al*. (2009b) also warned that "diatomists are good at describing diatom species ecology in terms of a few variables that are easy to measure but are, in the process, missing many nuances of the interactions between physical and chemical environments, of the importance of the speciation of nutrients and of the effects of short-term variability in the chemical environment".

In the same line, Clarke (2013), simulating a variety of invertebrate data from the UK, including some data provided by SEPA, stated that "If the sole aim was to estimate average waterbody-wide quality over the three year period, a statistically efficient strategy might be to take a sample from one site in each year, as this provides some temporal replication even though with such a scheme (and no external data for this or similar waterbodies) we cannot distinguish the relative importance of spatial and temporal variability in the observed metric values".

The Before-After analyses on trial data (see section 3.3 of this report) and comparisons between pollutant and biological data (Figure 8) provided further support to the argument that monitoring to assess WFD status is insufficient to assess response to change in land management and pollutants. Indeed, benthic diatom and invertebrate sample sizes were too small to enable the detection of a significant step change with adequate certainty. In addition, comparing phosphorus and sediment concentrations from monthly spot sampling with DARES and PSI sampled biannually showed that it is difficult to relate biological response to the effect of stressors because of the low sampling frequency for diatoms and invertebrates (Figure 8). Importantly, pre- and post-2011 phosphorus averages complied with the WFD standard, which is at 0.065mg/l at the Lemno Burn; DARES and PSI[10] also complied with WFD standards post-2011. The DARES-phosphorus graph clearly demonstrated that DARES values below the specified standard (deterioration) are not temporally related to phosphorus "spikes" (Figure 8). DARES might have been affected by unmeasured "spikes" in phosphorus or other unmeasured impacts on the overall phytobenthic community, resulting in the lack of any discernible cause-effect pattern. Likewise, a direct response of the PSI to sediments could not be discerned (Figure 8).

These findings also show that the assumption that spring and autumn biological samples capture different intensity of diffuse pollutant losses, and are therefore representative of the range of impacts over time, may be inappropriate. Firstly, rainfall, and hence losses in runoff, have no strong seasonal pattern[11]; and secondly, legacy pollutants have unknown temporal patterns of release. **Subsequently, the ecological monitoring data used for WFD status classification are also inappropriate to account for the effects of background variation on biota.**

Several challenges therefore arise for the development of a robust ecological monitoring programme to assess the effectiveness of the Diffuse Pollution Plan.

A major challenge we need to consider towards a reliable estimation of step change in stream ecology is that biological communities respond to changes in water quality in a different way than pollutants, and definitively more slowly than pollutants respond to the diffuse pollution measures. Ecological monitoring should capture short-term and chronic stresses as well as reductions in the levels of pollutants.

Short-term temporal ("ephemeral") impacts of diffuse pollution may include sudden increases (a.k.a. "spikes") of phosphorus because of septic tank failure or delayed pollutant runoff or leaching after wet spells (Jarvie *et al*. 2013). These types of

[10] For PSI there is a provisional indicative standard for the Priority Catchment process in Scotland.

[11] In Eastern Scotland where the Lemno Burn is, the range of winter rain from 2007 to 2014 is 273 to 587 mm and the summer range is 183 to 403 mm (MetOffice.gov.uk).

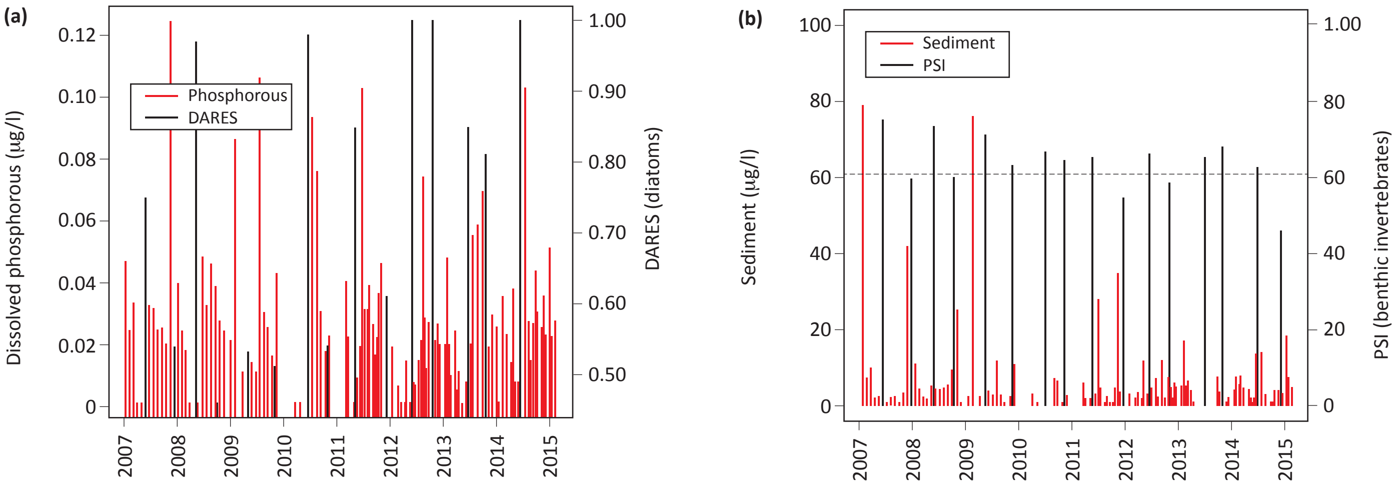


Figure 8 Water quality and ecological sampling at the Lemno Burn. Left: DARES and dissolved phosphorus; dashed line: threshold for moderate to good status in DARES standard. Right: PSI and sediment; dashed line: provisional threshold for moderate to good status showing only natural sedimentation levels.

losses cannot be adequately detected with only ecological data taken into account or if only event-based or only routine (monthly) spot water quality sampling carried out. The solution to this problem may involve sampling stressors at a higher frequency and at the same sites as biota, to enable a better understanding of ecological response (Box 4).

BOX 4 Effect of small-scale temporal phosphorus variation on Diatom Trophic Index

A recent study at the Eden Demonstration Test Catchments (DTC) showed that the Diatom Trophic Index integrates changes in nutrients and flow that have taken place (only up to 21 days) before diatom sampling.

Diatoms were sampled monthly. Nutrients and flow were monitored on a near-continuous basis at the same site as diatoms. It was shown that the calculated Trophic Diatom Index integrates the effects of:

- Stream flow variation over the preceding 18–21 days.
- Total phosphorus variation over the preceding 7–21 days.

In the light of these findings, the effect of phosphorus "spikes" on diatoms will go unnoticed if sampling fails to adequately capture phosphorus variability over a range of flows. The study also shows that it is essential to take phosphorus data into account when assessing change in diatoms.

Source: Snell *et al*., 2014.

Chronic impacts of low-level stresses usually refer to highly variable diffuse pollutant losses as a result of the interaction between rainfall, legacy pollutants and land use (Scheffer *et al*. 2001). A study of the chronic impacts of low-level stresses requires long term monitoring and largely depends on the specific effects of each stressor on the biological community as a whole, rather than an indicator species or group of species (e.g. diatoms) (US EPA 1997: Biological Monitoring of Aquatic communities; Clements *et al*. 2010).

Step-change in pollutants as a result of improvements in land management post-implementation diffuse pollution mitigation measures will eventually lead to ecological improvements. As mentioned in section 1.0, the lag time between change in pollutants and ecological improvement is variable and its causes are also variable. In addition, it is not well understood how EQRs respond to diffuse pollution stressors. EQRs do not always detect the main stressor driving diffuse pollution impacts. A plethora of recent studies worldwide provide data on why a mismatch between the indicators of a diffuse pollution impact and the actual degree of that impact may occur (Jarvie *et al*., 2013, Palmer *et al*., 2014; Yates *et al*., 2007; Scheffer *et al*. 2001; Yates & Bailey, 2010; Clements *et al*., 2010).

Three plausible explanations could be given for this mismatch:

1. **There may be non-linear trajectories in ecological recovery** after the reduction of the pollutant causing impairment. In some communities, an apparent recovery in abundance or species richness can occur despite loss of functional resilience. Such communities display increased sensitivity to background variability and "normal stress" and thus may never return to pre-pollution conditions, as predicted by the Humpty–Dumpty model of ecological recovery (Pimm 1991). A practical implication is that ecological recovery might not occur until pollutant levels are reduced far below those that triggered impairment in species composition. In addition, recovery endpoints may be very different from the original undisturbed state used for developing the EQR.

2. **EQRs are regarded as insufficient to address decoupling between a biological community and a pollutant.** EQR models, such as the DARES, are built around the assumption that it is possible to infer the level of impacts from the taxonomic composition and relative abundance of the taxa present. EQRs do not address functional aspects of impairment by diffuse pollution and, as noted by Clarke *et al*. (1996), they are of little use without some understanding of the sources and sizes of the sampling error and other uncertainties in their estimation. Decoupling in the diatom-phosphorus relationship has been found to be a cause of lack of ecological improvements despite intensive diffuse pollution mitigation efforts (Box 5).

BOX 5 Critical processes causing decoupling in the phosphorus-diatom relationship

The main processes that remain unaccounted in developing EQRs but can cause decoupling in the phosphorus-diatom relationship include:

- Variability in the bioavailability of phosphorus, especially in areas with stream-bed legacy pollutants, high sediment inputs and sewage disposal.
- Variable nutrient demand by diatom communities in the presence of other impacts.
- Nitrogen and phosphorus co-limitation confounding expected diatom response.
- Physical controls, such as light availability.

Source: Jarvie et al. 2013.

3. **The implementation of a diffuse pollution mitigation programme may bring about a sharp change on overall habitat quality and ecosystem, rather than a continuous, gradual change in indicator species or groups of species with increasing degree of implementation**. More specifically, it has been shown that species composition of fish, macroinvertebrate and diatom communities exhibit a sharp change after establishing extensive uptake of the measures (a.k.a. "threshold effect", Clements *et al*. 2010). Clements *et al*. (2010) showed that community composition was more sensitive than abundances of sensitive species to short-term variation in stressor levels and might represent a more effective endpoint for assessing mitigation success. Thresholds are determined by the relationship between ecological responses and time since "stressor removal", which refers to pollutant reductions below specified standards. **In this respect, a multiple-assemblage (e.g. total benthic community) approach is regarded as more useful than EQRs** to evaluate the effectiveness of measures at the catchment (waterbody) scale.

A number of strategies have been put forward to tackle the challenges relating to ecological monitoring that aims to assess the effectiveness of diffuse pollution mitigation programmes. Review of the literature on the approaches to ecological analyses has not been extensive because the focus of this report is to identify a suitable field monitoring strategy. It can be concluded that strategies to enable ecological change to be detected refer to solutions that increase the intensity of taxonomic and statistical analyses (Jarvie *et al*. 2013; Yates *et al*. 2007; Yates & Bailey 2010; Clements *et al*. 2010; Clarke 2013) and the sampling frequency of pollutants-stressors (Sneel *et al*. 2014), rather than the sampling frequency of ecological sampling.

The field practices and analyses used for assessing ecological change and response to measures at a catchment scale are summarised as follows.

A common practice in ecological monitoring for the assessment of the effectiveness of diffuse pollution programmes is to carry out monitoring on a long-term, biannual basis with simultaneous examination of biota and pollutants-stressors at the same site (waterbody reach) (US EPA 1997). High frequency monitoring of biota is not necessary to assess effectiveness, unless the relationship between biota and stressors is unknown. For example, if the diatom response to phosphorus from agricultural and sewage sources is unknown, high-frequency, targeted diatom sampling should be carried out in tandem with monitoring/sampling phosphorus from different sources, to understand the time-scales and magnitude of response. In any case, increased sampling frequency for pollutants-stressors is essential to assess whether there are cause-effect relationships between specific pollutants and biota (Sneel *et al*. 2014; See also Box 3). Long term monitoring of biota (US EPA 1997) is also required to allow chronic impacts and non-linear recovery trajectories to be discerned. Ideally, long-term (as in Table 1) biannual benthic diatom data (e.g. DARES) should be juxtaposed against weekly phosphorus data, and benthic invertebrate PSI data should be juxtaposed against weekly sediment data, as shown in Figure 8.

In addition, the emphasis of ecological monitoring should be on selecting an appropriate design and spatial and temporal scales to test the hypothesis that the measures are effective, rather than on increasing sampling frequency for biota. Commonly used designs enable comparisons among multiple catchments to increase the statistical power of ecological assessments and account for waterbody-specific variation. “Impact” catchments may be compared with “control” catchments, as in the typical BACI design; or “reference” catchments, if baseline data are not available; or other “impact” catchments, if it is difficult to use data from “control” and “reference” catchments (See also Box 2 and Box 3 for general guidance on potential designs).

In the US, for example, the monitoring strategy is based on the design developed by US EPA (1990) to underpin the Total Maximum Daily Load (TDML) approach. The design examines ecological data from a network of waterbody-stations with similar hydrological and diffuse pollution pressure characteristics from two seasons, spring and autumn, and uses three types of waterbodies: “control”, “impact”, and “reference” waterbodies to allow for flexibility because of lack of baseline (“before”) data or “control” data for all catchments. Gabel *et al*. (2012) used a variation of the BACI design, i.e. repeated measures analysis of variance (RP-ANOVA), with type of waterbody (“control”, “impact” or “reference)” as the fixed effect and “year” as the random effect, with “season” nested in “year” because there were no baseline data available.

In addition to collecting evidence to inform assessments based on single-indicators (e.g. EQRs), the focus of ecological monitoring is also on community structure and alternative metrics at appropriate spatial scales. For example, diatom communities can also be assessed using changes in the dominant taxa at a genus or species level (e.g. Kelly *et al*. 2009b and literature cited therein) and chlorophyll *a* measurements (e.g. Gabel *et al*. 2012). A growing body of evidence also shows that diatom growth is influenced by variability in phosphorus bioavailability, grazing, temperature, light availability, change in nitrogen to phosphorus ratio, flow regime, riparian tree-cover and toxic substances (Sponseller *et al*. 2001; Jarvie *et al*. 2013 and literature cited therein). The Environment Agency in England reports light availability and flow in tandem with diatom estimates to help understand ecological response to the effects of Catchment Sensitive Farming (CSF Team 2014). Jähnig *et al*. (2011) have also suggested the use of indicators of public perceptions of ‘meaningful’ ecological metrics, such as the presence, or not, of ‘murky waters’ or riparian landscape aesthetics, to assess the effects of measures from a civic perspective and a local, sub-waterbody, scale.

Likewise, macroinvertebrates can be assessed using diversity indices (e.g. Shannon) but also biomass, and community species composition (e.g. Yates *et al*. 2007; Clements *et al*. 2010; Yates and Bailey 2010; Gabel *et al*. 2012). Results, however, should be interpreted with caution. For example, Sponseller *et al*. (2001) reviewed evidence from a number of studies which showed that in-stream pollutants are influenced by waterbody-wide land use but macroinvertebrate density and species richness respond only to local upstream land use (i.e. 200 m distance from monitoring station).

Finally, multivariate analyses such as principal components analysis or canonical correspondence analysis are usually performed to assess effectiveness when abundance data from all species of a community (e.g. total phytobenthos instead of only diatoms, or total benthos) are available. These analyses use species abundances over the years from a variety of taxa; the relationship of the total biological community and a pollutant can then be explored by examining the relationship between a pollutant and the principal component or the canonical variable explaining the greatest amount of variation in species composition. This approach has been found to deliver a better understanding of the effects of both diffuse pollution pressures and the measures at a catchment scale (e.g. Yates *et al*. 2007; Clements *et al* 2010; Yates and Bailey 2010).

3.3.10 Proxy variables to assess the effectiveness of the Diffuse Pollution Plan

Relationships between different pollutants but also between pollutants and the EQRs, i.e. DARES and PSI were explored. No significant relationships could be identified. This indicated that the use of a proxy variable to reduce sampling effort before thorough examination of the data is inappropriate. This has already been stressed by SEPA (Greig *et al* 2004) for data from the Cessnock Water. However, a relationship between turbidity and storm event Total P has been established for a number of storm events in the Baldardo catchment (Lunan Water) by Vinten *et al*. (2009). US EPA (1997), however, suggest that the measurement of a proxy variable, e.g. total suspended solids or turbidity, may help increase the precision of total phosphorus estimates; therefore, the “proxy” variable could be used to improve understanding rather than substitute the measurement of the actual pollutant. Sound evidence, demonstrating a significant and meaningful relationship over time between the pollutants of interest, must be established before initiating proxy monitoring for the evaluation of improvements in priority catchments, to prevent misleading conclusions to be drawn. At the catchment scale, it is uncertain whether positive relationships between the pollutant of interest and a proxy variable can be detected reflects similar response to the measures or other underlying and unmeasured processes.

3.3.11 Should the spatial network of monitoring sites be revised?

According to the guidance summarised in Table 1, assessing the effectiveness of a single measure at a time requires monitoring at the plot or field scale. Assessing the effectiveness of a combination of mitigation measures implemented across a catchment and a waterbody can be carried out at the waterbody scale. SEPA's network of monitoring sites is compatible with this guidance. Therefore, it can be considered that monitoring at the bottom of a priority catchment enables the integrated effect of measures implemented concurrently upstream of each monitoring site in each waterbody to be assessed.

But how do we know that pollutants and biota respond to pressures and land management improvements at the waterbody scale and not at the plot or field scale? Do measures implemented far upstream from the monitoring sites contribute to changes in water quality in the same way as the measures implemented immediately upstream of the monitoring sites? Do pollutants and biota have similar responses to the scale of land management improvements? As Sponseller *et al*. (2011) showed, macroinvertebrates and phytobenthos respond to small spatial scales in land use patterns (i.e. within a stream reach), but pollutants respond to large spatial scales in land use (i.e. catchment scale). Unfortunately, such data cannot be extrapolated and indeed it is very difficult, if not impossible to choose the appropriate spatial scales of sampling to detect an effect without rigorous sampling at different spatial scale (Underwood 1992). In humid areas, such as Scotland, waterbody catchments should generally be smaller than 5 square miles to obtain a uniform area with the measures in place and similar climatic conditions (USDA-NRCS 2003).

Several voices with both a research and regulatory background have suggested the introduction of replicates to increase confidence of WFD class. Clarke (2013) for example showed that increasing spatial replication (e.g. sampling in two or more stretches within the same waterbody to avoid pseudo replication) has the potential to improve substantially confidence in the WFD status classification outcome). In this respect, research into improving the monitoring design for WFD classification is also contributing to the pressure for a larger sample size to enable the short-term WFD targets for improvements in ecological status to be delivered.

Overall, the objective of monitoring to assess effectiveness of measures is not to increase confidence in WFD status classification but to enable a reliable detection of step-change due to the measures implemented. Within-waterbody spatial replication is insufficient to increase cost-effectively sample size in a BACI design because it does not address variation among waterbodies with similar measures in place, or land use or rainfall regime. In addition, replication does not account for the spatial scales of the effects of pressures and the measures. Assessing effectiveness at the waterbody scale requires knowledge of whether the response to DP GBRs is waterbody-specific or whether it can be generalised to inform the development of typologies of catchment response to the measures. In this respect, accounting for within-waterbody variability is unnecessary.

4 Implications for SEPA's monitoring strategy: Recommendations

This report analysed water quality and ecological trial data from priority catchments collected by SEPA and research data from the James Hutton Institute. The analyses assessed the suitability of SEPA's data to detect improvements in water quality following the installation of diffuse pollution measures. The findings clearly showed that currently available monitoring data in Scotland provide only part of the evidence required to assess the effectiveness of measures.

The reasons for this shortcoming are:

1. Inappropriate monitoring design. Currently monitoring is designed to assess WFD status classification at a waterbody scale, but assessing the effectiveness of measures reliably requires comparisons between waterbodies with different land management and pressures to distinguish between the effects of measures and background variability (as in the BACI design).
2. Small sampling size. Currently the monitoring duration and frequency pre- and post-implementation the measures are sufficient to support WFD status classification, but assessing the effectiveness of measures requires accounting for lag times in the effectiveness of measures and change in pollutants across a range of flow regimes.
3. Lack of concurrent flow-pollutant measurements, which are required for flow-adjustment of pollutant concentrations.
4. Failure to assess pollutant-biota relationships. Current monitoring is designed to collect and treat data with the aim to assess compliance of each pollutant and EQR with WFD standards, but understanding ecological recovery in response to water quality improvements requires statistically robust estimates of pollutant-biota interactions.

Overall key finding

Trial data and a review of the literature showed the need for a statistically robust monitoring design, longer monitoring duration and higher sampling frequency to enable change in pollutants and ecology to be quantified at a waterbody scale in SEPA's priority catchments.

Monitoring methods used for the evaluation of agricultural diffuse pollution mitigation projects elsewhere were also reviewed. Within this report, the importance of identifying the objectives of a monitoring programme and making decisions about the design and sampling duration and frequency accordingly was stressed. Studies were reviewed to identify the optimal sampling frequency and duration required to ensure detection of change with adequate statistical power in catchments where diffuse pollution control measures are implemented. Comparisons were also made of the potential of the currently available pollutant data collected for WFD classification to match model predicted reductions in pollutants. This is an unusual but very constructive approach to assessing the suitability of the WFD monitoring for the evaluation of the effectiveness of the Diffuse Pollution Plan.

The major problem in the peer reviewed literature was the paucity of information: studies identifying best monitoring approach towards detecting a significant pollutant trend or quantifying flow variability rarely mentioned or explored whether the measures were effective. This is probably because, usually, the production of clear positive outcomes is a prerequisite for an academic publication whereas diffuse pollution mitigation programmes show a slow progress or no change for considerably long-term periods (see Table 1). Technical summaries or reports, however, were used to derive the essential information.

In addition, the majority of the ecological studies pertaining to the effectiveness of measures analyse the theories explaining ecological lag times, rather than issues of monitoring frequency, to assess effectiveness. Many ecological studies also assess the inadequacy of ecological indicators, and what sampling is needed to enhance confidence to WFD classification, but questions about WFD confidence are not related to the question "what is the best monitoring strategy to assess the effectiveness of measures?"

Taking into account the results from the analysis of trial data and the review of methods, a number of recommendations have been developed to underpin the evaluation of the effectiveness of waterbody management with diffuse pollution measures. The recommendations and their implications are presented in the following sections.

4.1 Implications for monitoring and statistical design

SEPA's monitoring strategy should be based on any of the BACI design variations described in section 3.2 (see also Table 2) to monitor pollutants, macroinvertebrates and diatoms. Data documented pre-implementation will provide the appropriate baseline against post-implementation data. Use of "control", "reference" or multiple "impact" catchments, depending on what is feasible, will help to factor out the effects of differences in land use and other catchment processes not related to the measures.

Finding 4.1 BACI Design

SEPA's monitoring strategy and analysis of step-change in river waterbodies and bathing water catchments should be based on a BACI design that reduces the risk of confounding, such as:

- "Multiple-Waterbodies/Before-After"
- "Before-After/Upstream-Downstream"
- "Multiple-Control BACI"
- "Multiple-Waterbodies BACI"
- "Multiple-Waterbodies post-implementation"

Data should be collected from multiple "control", "reference" and "impact" catchments to allow for flexibility in the application of a fit-for-purpose BACI design.

Waterbody selection can be based on the evidence that has already been gathered by SEPA on a national level to inform the priority catchment approach. In addition, the degree of DP GBR uptake is already documented at a waterbody scale. The weight of evidence method developed to underpin catchment monitoring for the evaluation of the Diffuse Pollution Plan will also be useful in providing the data needed to identify which catchments share similar land use (as in LCM-07, see footnote 5) and rainfall characteristics (Akoumianaki *et al*. 2016). The majority of "impact" and "control" catchments are monitored (operational monitoring) and data are also collected from "reference" catchments (surveillance monitoring) to inform WFD classification. Therefore, there is no need for extra monitoring sites within a waterbody or at river catchment scale for either pollutant or ecological monitoring. Consequently, we recommend a new typology for waterbodies, as "impact", "control" and "reference" to help select waterbodies for the statistical analyses of both currently available monitoring data and monitoring data to be collected in the next RBMP cycles.

In the event that step-change is estimated according to the "Single-Waterbody/Before-After" design, a meta-analysis of the resulting step changes would be useful to assess what percentage of waterbodies show a reduction, and what magnitude of reduction, and whether there is any regional variation. Results, however, should be treated with caution because the Single-Waterbody/Before-After" design is confounded (see section 3.2), and its application increases the risk of misleading the evaluation of the Diffuse Pollution Plan.

Trend analysis can be performed on data from groundwater, lochs and transitional waters once long-term post-implementation data become available.

4.2 Implications for flow measurements

The analyses and the literature review showed that flow measurement should be incorporated in to water quality monitoring to allow for flow-adjustment (i.e. to remove the effects of flow variation on pollutant concentrations), and for reliable load estimation, flow and pollutants should be sampled concurrently i.e. same day and waterbody. Ideally, flow should come from continuous flow measurements. In this respect, it is essential that water quality and ecological monitoring stations are located near (same waterbody) or at a SEPA flow gauging station, when possible, due to the paramount importance of obtaining accurate flow records for estimating pollutant concentrations and loads.

Finding 4.2 Flow measurements

Simultaneous concentration and flow measurements should be taken in each sampling time to enable flow adjustment of concentrations and a reliable load estimation of key pollutants.

Flow is not measured in each waterbody but SEPA has already established 392 gauging stations[12] for the measurement of water levels throughout Scotland. Flow from these sites can be used only when there is sound evidence that its use is not introducing unquantifiable errors in flow-adjustment and load estimation.

Recommendation: Flow measurements

Flow data from existing flow gauging stations should be assessed for their suitability to be used in reliable flow-adjustment of concentrations and load estimation of key pollutants.

Annual load estimation will be useful in linking modelled source apportionment for each pollutant with robust estimates of **in-stream annual pollutant loads**. The unbiased load estimation is essential in the validation of modelled reductions in pollutant loads in priority catchments (Bowes *et al*. 2006). In this respect, requirements for reliable flow-measurements are as demanding as for the identification of step-change. **However, we do not recommend the use of load estimation for step-change analyses to show the effect of measures because we need to factor out the effect of flow to enable detection of change**.

4.3 Implications for sample size: monitoring duration and frequency

Analyses on trial data and a review of the literature showed that provisions should be made for a larger sample size to understand whether the measures are effective in bringing about improvements, or not. A larger sample size for pollutants will reduce the magnitude of change that can be detected and will enable the reductions predicted by the model to be detected. **For pollutants, spot sampling is a straight-forward sampling technique but, if feasible, the automated time-composite sample method is the best option in the Scottish context, because of the wide and variable range of flows**. Spot sampling could be carried out on a weekly or fortnightly basis but to align spot sampling frequency with automated time-composite sampling, which should be on a weekly basis, weekly sampling for both spot and time-composite sampling is recommended. Event-based sampling is suitable when daily retrieval is feasible.

Increasing sample size in pollutants requires:

- More than four years of baseline and post-implementation monitoring for all pollutants in groundwater and surface waterbodies, and for diatoms and invertebrates.
- High sampling frequency for key in-stream pollutants with long-term monitoring to account for the key considerations mentioned in section 3.3.6 and 3.3.8 for the suitability of spot and automated sampling.
  - Spot sampling, where suitable, should be carried out on a weekly basis (52 samples per year) for nutrients (particulate or dissolved phosphorus and ammonium) and sediment.
  - Automated time-composite sampling, where suitable, should involve weekly retrieval of samples collected every 7 hours every day.
  - Routine bathing season spot sampling (about 20 samples per year) for FIOs should be combined with event-based sampling and immediate retrieval (including sampling for the next three days following rainfall), i.e. samples can be collected every 7 hours on event days and then be composited on a daily basis.
- Long-term monitoring for diatoms and macroinvertebrates on a biannual basis, not only post-implementation of the measures, but also post-reduction of pollutants.

[12] http://www.sepa.org.uk/environment/water/water-levels

Increasing information to improve interpretations requires:

- The use of additional multi-species ecological metrics.
- Simultaneous examination of ecological and pollutant data to enable cause and effect relationships to be explored.
- Ancillary projects to investigate specific questions, depending on waterbody and modelling needs, e.g. what is the effect of different sources and measures on the phosphorus bio-availability and the associated diatom response?

Additional considerations include:

- Pollutants should be collected at the same sample frequency before and after the installation of measures and at the waterbodies compared, depending on the BACI design.
- Diatoms and macroinvertebrates should be collected at the same sample frequency before and after the installation of measures and at the waterbodies compared, depending on the BACI design.
- **Monthly spot sampling for pollutants should continue after the introduction of weekly automated composite sampling for at least a year.** Thus, interpretations of "old" spot data can be analysed for evaluations of comparative reliability. Once a sufficient concurrent record of both types of monitoring (spot and composite exists), a wide range of advanced statistical techniques[13] can be used for retrospective modelling and analyses of historical data.
- **Pollutant concentrations from weekly or fortnightly spot samples, weekly composited samples and event-based samples should be tested for autocorrelation** before use for assessments of step-change for the evaluation of the effectiveness of measures.
- Diatom (i.e. DARES) change, or lack of, has to be further investigated

Finding 4.3 Monitoring duration

Pollutant, diatom and macroinvertebrate monitoring should ideally be undertaken for more than four years before and more than four years after the introduction of diffuse pollution measures to enable changes from year to year to be detected. The duration of biological monitoring should also account for the temporal scales of the processes acting to promote or delay ecological recovery following reductions in pollutants.

Finding 4.3 Monitoring frequency

In-stream pollutants should ideally be monitored on a weekly basis with spot sampling or the automated time composite sample method to account for the effects of background variation. In Bathing Waters, routine spot and event-based automated sampling should be combined to separate the effects of measures and rainfall on Faecal Indicator Organisms (FIOs). Groundwater pollutants should be collected on a quarterly or annual basis. Pollutants in lochs and transitional waters should be monitored on a seasonal basis. Macroinvertebrates and diatoms should be monitored on a biannual basis.

Finding 4.3 Sampling technique

Automated sampling should be carried out with the time-composite method. Samples should be collected every 7 hours and composite samples should be retrieved on a weekly basis for nutrient sand sediment. In bathing waters, fixed-date spot sampling should be combined with automated time-composite sampling during events with daily retrieval for FIOs. If the installation of automated samplers is not feasible in all catchments, then the remainder of waterbodies should be sampled (i) with the weekly spot sampling method for nutrients and sediment; and (ii) with the combination of fixed-date spot sampling and spot sampling during events in the bathing season for FIOs.

Finding 4.3 Analysis of ecological data

Changes in Ecological Quality Ratios (e.g. DARES) should be analysed in tandem with changes in pollutants and other physical factors (e.g. phosphorus bioavailability, grazing, temperature, light availability, change in nitrogen to phosphorus ratio, flow, riparian tree-cover, toxic substances) and in combination with additional ecological metrics, such as species composition, diversity and biomass to enable the effects of the measures to be understood and detected on a habitat level.

Recommendation: Ancillary research projects

Ancillary research projects should be carried out to help to understand how diatoms and macroinvertebrates respond to different pollution sources and measures and at what spatial scales. For example, monthly diatom data should be compared between two sites representative of sources of phosphorus with different bio-availability (e.g. arable land versus septic tanks) within a waterbody or a river catchment to help understand how the measures or the type of source influence diatom response.

4.5 Targeting the enhanced monitoring design

These findings will enhance the intensity of:

- **Planning the siting of auto-samplers and retrieval of composite samples in the context of a fit-for-purpose BACI design.** This is because identifying step-change (i.e. using the BACI design) requires that the same sampling frequency is applied in the waterbodies that are compared. So, if auto-samplers are installed in particular "impact waterbodies" then it is essential to install auto-samplers in the selected "control" or "reference" waterbodies, depending of which BACI design is fit-for-purpose.
- **Field work for in-stream pollutants**. This is because of the weekly (spot or composite) sampling on a long-term basis. Event-based FIO sampling will also require the staff to be on stand-by to retrieve FIOs on a daily basis during and in the immediate aftermath of events.
- **Laboratory work**. This is because a greater number of pollutant samples will have to be analysed.
- **Taxonomic analyses**. For example, instead of only diatoms the whole phytobenthic community species composition will have to be identified.

If resource constraints prevent the implementation of enhanced monitoring for in-stream pollutants in all "impact" and "control" or "upstream" waterbodies and in "reference" waterbodies, depending on the BACI design applied (see section 3.2), **we recommend the targeting the enhanced**

[13] Outwith the scope of this report.

monitoring to selected, representative waterbodies within and out with priority catchments. Targeting will enable:

(i) The effects of measures to be understood.
(ii) The efficacy of this monitoring paradigm to be demonstrated to stakeholders.
(iii) A suitable, fit-for-purpose BACI design to be tested and optimised.

The targeted waterbodies should typically be representative of different land management types, (as in LCM-07; see also footnote 5 this report). The weight-of-evidence method developed to evaluate the effectiveness of the Diffuse Pollution Plan (Akoumianaki *et al*. 2015) can also help with the selection of waterbodies of interest on the basis of degree of DP GBR uptake and the risks related to land use (fertiliser use, erosion risk) and rainfall.

Recommendation: Targeting enhanced monitoring

The enhanced monitoring developed here should be applied and tested in all catchments. If this is not possible, enhanced monitoring should be targeted at selected waterbodies representative of land management, through implementation of appropriate measures, and land use within and out with the priority catchments.

A shared understanding with the Environment Agency on the practical implications of tackling diffuse pollution would benefit decision making on the monitoring strategy to assess the Diffuse Pollution Plan. Enhanced monitoring with state-of-the-art infrastructure ("outdoor laboratory") and use of the BACI design are already implemented by EA and DEFRA as part of the Demonstration Test Catchment (DTC) project (CSF Team 2014). Therefore, establishing links with the DTC project may be useful.

5 Conclusions

Trial data and a review of the literature showed the need for a statistically robust sampling design, longer monitoring duration and higher sampling frequency to quantify change in pollutants and ecology at a waterbody scale in SEPA's priority catchments. Within this report, the minimum detectable change with current frequency and the sample size needed to detect the improvements predicted by modelling to demonstrate the problems were calculated. The following were specified: the metric; the sampling frequency and technique; the parameters essential to assess effectiveness; the design; and ways to link the findings with research and policy needs (Table 6). Finally, a demonstration of the statistical tests in R-code to enable the analyses to be applied in other catchments was provided (Appendix 2 and 3).

The key findings can be summarised as follows:

- Trial data and a review of the literature showed the need for a statistically robust monitoring design, longer monitoring duration and higher sampling frequency to enable change in pollutants and ecology to be quantified at a waterbody scale in SEPA's priority catchments.
- Monitoring in river and bathing water catchments should be based on a Before-After/Control-Impact (BACI) design.
- Simultaneous measurements of concentration and flow should be to enable a reliable flow adjustment of concentrations and load estimation of key pollutants.
- Pollutant, diatom and macroinvertebrate monitoring should ideally be undertaken for more than four years before and more than four years after the introduction of diffuse pollution measures to enable changes from year to year to be detected.
- In-stream pollutants should ideally be monitored on a weekly basis with spot sampling or the automated time composite sample method to account for background variation.
- In Bathing Waters, routine spot and event-based automated sampling should be combined to separate the effects of measures and rainfall on Faecal Indicator Organisms (FIOs).
- Diatoms and macroinvertebrates should be monitored on a biannual basis.

The following **recommendations** are also provided for existing data, ancillary research projects and targeting of the enhanced monitoring design and frequency in the case of resource constraints:

- Flow data from existing flow gauging stations should be assessed for their suitability to be used in reliable flow-adjustment of concentrations and load estimation of key pollutants.
- Ancillary research projects should be carried out to help understand how diatoms and macroinvertebrates respond to different pollution sources and measures and at what spatial scales. For example, monthly diatom data should be collected at two sites or waterbodies representative of sources of phosphorus with different bio-availability (e.g. arable land versus septic tanks) to help understand how the measures and source influence diatom response.
- The enhanced monitoring should be applied in all priority catchments to inform the weight-of-evidence method developed to evaluate the effectiveness of the Diffuse Pollution Plan (Akoumianaki *et al*. 2015). If this is not feasible, the enhanced monitoring with long term-duration, weekly frequency of key pollutants and simultaneous flow measurements should be targeted at waterbodies representative of a range of land management improvement measures, and land use types.

The enhanced monitoring developed here is in line with the monitoring practice applied internationally in programmes assessing the effectiveness of measures. The enhanced monitoring means increased monitoring effort but this is essential for detecting reliably whether water quality has improved or not. If changes in pollutants can be detected reliably, the weight-of-evidence method will help to interpret this change in the context of WFD standards, modelled predictions and of changes in the catchment between before and after launching the measures, i.e. DP GBR uptake, fertiliser use, livestock density, erosion risk and rainfall. Conversely, if changes in pollutants have not been observed, the weight of evidence method will help to interpret the lack of change in the context of WFD standards, modelled predictions and risks in the catchment because of gaps in the implementation of measures or deterioration of catchment pressures.

Table 6 Major components of the monitoring strategy to enable a statistically robust detection of water quality and ecological change in priority catchments

Component	Enhanced monitoring to enable a statistically robust detection of change
Design	Before-After/Control-Impact design in rivers and bathing water catchment.
Flow	Essential in all river waterbodies and bathing waters for flow-adjustment and load estimation.
Pollutant metric	Concentrations for pollutants in groundwater and surface waters.
Ecological metric	Ecological Quality Ratios (EQRs); biomass, and; species composition and richness.
Duration	Long-term, i.e. more than four years pre- and post-implementation.
Frequency	*Nutrients – sediment (in-stream)*: Weekly spot sampling or weekly time-composite sampling. *FIOs*: Bathing season spot sampling with event-based daily composite sampling. *Diatoms / Macroinvertebrates*: Biannual sampling. *Pollutants in groundwater*: quarterly (highly permeable aquifers) or annually (less permeable aquifers). *Nutrients-sediments (lochs and transitional waters)*: seasonal spot sampling

References

Abtew, W & Powell, B 2003, *Cost-effective Water Quality Sampling Scheme for Variable Flow Canals at Remote Sites*. South Florida Water Management District.

ADAS 2008, "Initial evaluation of effectiveness of measures to mitigate diffuse rural pollution". *Report produced for Scottish Government under project ADA/011/07*. Available from www.gov.scot/resource/doc/256611/0076192.pdf [February 2015].

Akoumianaki, I, Potts, J, Baggio, A, Gimona, A, Spezia, L, Sample, J, Vinten, A, & MacDonald J. 2015, *Developing a Method to Monitor the Rural Diffuse Pollution Plan: Providing a Framework for Interpreting Catchment Data, CRW2014/13*. Available from: crew.ac.uk/publications.

Bechmann, M, Deelstra, J, Stalnacke, P, Eggestad, HO, Øygarden, L & Pengerud, A 2008, "Monitoring catchment scale agricultural pollution in Norway: Policy instruments, implementation of mitigation methods and trends in nutrient and sediment losses". *Environmental Science Policy* 11:102–114. doi:10.1016/j.envsci.2007.10.005.

Bertram, J, & Balance, R 1996, "A practical guide to the design and implementation of fresh water quality studies and monitoring programmes". *Published on behalf of United Nations Environmental Programme (UNEP) and World Health Organization (WHO), E & FN Spon publishers*, 172-177. Published on behalf of United Nations Environment Programme (UNEP) and the World Health Organization (WHO). Available from: www.nrcs.usda.gov/Internet/FSE_DOCUMENTS/stelprdb1044775.pdf. [October 2014].

Bieroza, MZ, Heathwaite, AL, Mullinger, NJ & Keenan, PO 2014, "Understanding nutrient biogeochemistry in agricultural catchments: the challenge of appropriate monitoring frequencies". *Environmental Science: Processes & Impacts*, 16(7), pp.1676-1691.

Borja, Á, Tueros, I, Belzunce, MJ, Galparsoro, I, Garmendia, JM, Revilla, M, Solaun, O & Valencia, V 2008, "Investigative monitoring within the European Water Framework Directive: a coastal blast furnace slag disposal, as an example". *Journal of Environmental Monitoring*, 10(4), pp.453-462.

Bowes J. *et al*. 2006, "Diffuse Pollution Screening Tool: Stage 3. WFD77". *Sniffer report*. Available from: www.envirobase.info/PDF/SNIFFER_WFD77.pdf [November 2014].

Brauer, N, O'Geen, AT, & Dahlgren, RA 2009, "Temporal variability in water quality of agricultural tailwaters: Implications for water quality monitoring". *Agricultural water management*, 96(6), 1001-1009.

Cassidy, R & Jordan, P 2011, "Limitations of instantaneous water quality sampling in surface-water catchments: comparison with near continuous phosphorus time-series data". *Journal of Hydrology*, 405, 182–193.

Clarke, RT 2013, "Estimating confidence of European WFD ecological status class and WISER Bioassessment Uncertainty Guidance Software (WISERBUGS)". *Hydrobiologia*, 704(1), pp.39-56.

Clarke, RT, Furse, MT, Wright, JF & Moss, D 1996, "Derivation of a biological quality index for river sites: comparison of the observed with the expected fauna". *Journal of Applied Statistics*, 23(2-3), pp.311-332.

Chen, D, Dahlgren, RA & Lu, J 2013, "A modified load apportionment model for identifying point and diffuse source nutrient inputs to rivers from stream monitoring data". *Journal of Hydrology*, 501, pp.25-34.

Clements, WH, Vieira, NK, & Sonderegger, DL 2010, "Use of ecological thresholds to assess recovery in lotic ecosystems". *Journal of the North American Benthological Society*, 29(3), 1017-1023.

Cooper, DM, & Watts, CD 2002, "A comparison of river load estimation techniques: application to dissolved organic carbon". Environmetrics, 13(7), 733-750.

CSF Team 2011, *England Catchment Sensitive Farming Delivery Initiative (ECSFDI) Phase 1 & 2 Evaluation Report*. Available from: publications.naturalengland.org.uk/publication/5329340644458496. [April 2015].

CSF Team 2014, *England Catchment Sensitive Farming Evaluation Report – Phases 1 to 3 (2006–2014)*. Available from: publications.naturalengland.org.uk/file/5083194468597760 [April 2015].

Davey, A 2010, *Demonstration Test Catchments: an experimental design and monitoring strategy*. WRc Ref: DEFRA 8104.03. http://randd.defra.gov.uk/document.aspx?document=wq0207_8763_frp.doc. [October 2015].

DPMAG 2011, *Rural diffuse pollution plan for Scotland*. Available from: http://www.sepa.org.uk/water/river_basin_planning/diffuse_pollution_mag.aspx Scotland.

EU 2000, "Directive 2000/60/EC of the European Parliament and of the Council of 23 October 2000 establishing a framework for Community action in the field of water policy". *Official Journal of the European Communities, L327, 1–70*.

Francy, DS, Bertke, EE, Finnegan, DP, Kephart, CM, Sheets, RA, Rhoades, J &Stumpe, L 2006, *Use of spatial sampling and microbial source-tracking tools for understanding fecal contamination at two Lake Erie beaches: U.S. Geological Survey Scientific Investigations Report 2006–5298*, 29 p.

Francy, DS, Myers, DN & Helsel DN 2000, *Microbiological Monitoring for the U.S. Geological Survey National Water-Quality Assessment Program. Water-Resources Investigation Report 00-4018* http://oh.water.usgs.gov/reports/wrir/wrir.00-4018.pdf

Gabel, KW, Wehr, JD, & Truhn, KM 2012, "Assessment of the effectiveness of best management practices for streams draining agricultural landscapes using diatoms and macroinvertebrates". *Hydrobiologia*, 680(1), 247-264.

Gaugush, RF 1986, Statistical methods for reservoir water quality investigations. *Available from the National Technical Information Service, Springfield VA. 22161. Instruction Report E-86-2, June 1986. Final Report. 216 p, 23 fig, 44 tab, 23 ref, append.*

Gitau, MW, Chaubey, I, Gbur, E, Pennington, JH, & Gorham, B 2010, "Impacts of land-use change and best management practice implementation in a Conservation Effects Assessment Project watershed: Northwest Arkansas". *Journal of soil and water conservation*, *65*(6), 353-368.

Gooday, R, Anthony, S, Calrow, L, Harris, D & Skirvin, D, 2014, "*Predicting and understanding the effectiveness of measures to mitigate rural diffuse pollution*". SNIFFER Project DP1 Draft Final Report, November 2014 ADAS UK Ltd.

Green, RH 1979, *Sampling design and statistical methods for environmental biologists*. John Wiley & Sons.

Greig, S, Dawson, P. & Craig, B. 2004, "Trends in diffuse pollution: data report to assist with the design and implementation of effective diffuse pollution monitoring programmes". *Diffuse Pollution Report Initiative No. 21*. DPI No. 21/SG/Sept 04.

Grove, MK, Bilotta, GS, Woockman, RR, & Schwartz, JS 2015, "Suspended sediment regimes in contrasting reference-condition freshwater ecosystems: Implications for water quality guidelines and management". *Science of The Total Environment*, 502, 481-492.

Hamilton, SK 2012, "Biogeochemical time lags may delay responses of streams to ecological restoration". *Freshwater Biology, 57*(s1), 43-57).

Hirsch, RM, 1988, "Statistical methods and sampling design for estimating step trends in surface-water quality". *JAWRA Journal of the American Water Resources Association*, 24(3), pp.493-503.

Hirsch, RM, Alexander, RB & Smith, RA 1991, "Selection of methods for the detection and estimation of trends in water quality". *Water resources research, 27*(5), pp.803-813.

Jähnig, SC, Lorenz, AW, Hering, D, Antons, C, Sundermann, A, Jedicke, E, & Haase, P 2011, "River restoration success: a question of perception". *Ecological Applications, 21*(6), 2007–2015. Available from: < http://www.esajournals.org/doi/abs/10.1890/10-0618.1> [September 2015].

Jarvie, HP., Sharpley, AN, Withers, PJ, Scott, JT, Haggard, BE, & Neal, C 2013, "Phosphorus mitigation to control river eutrophication: Murky waters, inconvenient truths, and "postnormal" science". *Journal of Environmental Quality, 42*(2), 295-304.

Johnes, PJ 2007, "Uncertainties in annual riverine phosphorus load estimation: impact of load estimation methodology, sampling frequency, baseflow index and catchment population density". *Journal of Hydrology*, 332(1), 241-258.

Jordan, P & Cassidy, R 2011, "Technical Note: Assessing a 24/7 solution for monitoring water quality loads in small river catchments". *Hydrology and Earth System Sciences, 15*(10), pp.3093-3100.

Kay, D, Crowther, J, Davies, C, Edwards, T, Fewtrell, L, Francis, C, Kay, C, McDonald, A, Stapleton, Watkins, J & Wyer, M 2012, "Impacts of agriculture on water-borne pathogens". *Environmental Impacts of Modern Agriculture, 34*, p.83.

Kelly, M, Bennion, H, Burgess, A, Ellis, J, Juggins, S, Guthrie, R, Jamieson, J, Adriaenssens, V & Yallop, M, 2009a, "Uncertainty in ecological status assessments of lakes and rivers using diatoms". *Hydrobiologia*, 633(1), pp.5-15.

Kelly, M, King, L & Ní Chatháin, B, 2009b, "The conceptual basis of ecological—status assessments using diatoms". In *Biology and environment: proceedings of the Royal Irish Academy* (pp. 175-189). Royal Irish Academy.

Kronvag, B, & Bruhn, AJ 1996, "Choice of sampling strategy and estimation method for calculating nitrogen and phosphorus transport in small lowland streams". *Hydrological processes*, 10, 1483-1501.

Littlewood, IG 1992, *Estimating contaminant loads in rivers: a review*. Institute of Hydrology.

Meals, DW, Dressing, SA, & Davenport, TE 2010, Lag time in water quality response to best management practices: A review. *Journal of Environmental Quality*, 39(1), 85-96.

Meals, D. W., Spooner, J, Dressing, SA & Harcum JB 2011, "Statistical analysis for monotonic trends, Tech Notes 6, November 2011". *Developed for U.S. Environmental Protection Agency by Tetra Tech, Inc., Fairfax, VA*, 23 p. Available from: www.bae.ncsu.edu/programs/extension/wqg/319monitoring/tech_notes.htm. [April 2015].

Meals, DW, Richards PR & Dressing SA 2013, "Pollutant load estimation for water quality monitoring projects. Tech Notes 8, April 2013". Developed for U.S. Environmental Protection Agency by Tetra Tech, Inc., Fairfax, VA, 21 p. Available from: www.bae.ncsu.edu/programs/extension/wqg/319monitoring/tech_notes.htm.

Moosmann, L, Müller, B, Gächter, R, Wüest, A, Butscher, E & Herzog, P 2005, "Trend-oriented sampling strategy and estimation of soluble reactive phosphorus loads in streams". *Water resources research*, 41(1). W01020, doi:10.1029/2004WR003539.

National Research Council US 2000, "Watershed management for potable water supply: assessing the New York City strategy". *Committee to Review the New York City Watershed Management Strategy*. National Academies Press.

Neal, C, Reynolds, B, Norris, D, Kirchner, JW, Neal, M, Rowland, P, & Wright, D 2011, "Three decades of water quality measurements from the Upper Severn experimental catchments at Plynlimon, Wales: an openly accessible data resource for research, modelling, environmental management and education". *Hydrological Processes*, 25(24), 3818-3830.

Newbold, JD, Herbert S, Sweeney, BW & Kiry P 2008, "Water quality functions of a 15-year-old riparian forest buffer system. p. 1–7. *In* Riparian ecosystems and buffers: Working at the water's edge". *AWRA Summer Specialty Conf. Am. Water Resources Assoc.*, Virginia Beach, VA.

Palmer, MA, Hondula, KL, & Koch, BJ 2014, "Ecological restoration of streams and rivers: shifting strategies and shifting goals". *Annual Review of Ecology, Evolution, and Systematics*, 45, 247-269.

Pimm, SL 1991, *The balance of nature?: ecological issues in the conservation of species and communities*. University of Chicago Press.

Pott, CA, Jadoski, SO, Schmalz, B, Hörmann, G, & Fohrer, N 2014, "Temporal variability of nitrogen and phosphorus concentrations in a German catchment: water sampling implication". *Revista Brasileira de Engenharia Agrícola e Ambiental*, 18(8), 811-818.

Scheffer, M, Carpenter, S, Foley, JA, Folke, C, & Walker, B 2001, "Catastrophic shifts in ecosystems". *Nature*, 413(6856), 591-596.

SEPA 2015, *A public consultation to inform the development of the second river basin management plan for the Scotland river basin district*. Available from http://wwwsepaorguk/environment/water/river-basin-management-planning/publications/ [October 2015].

Skarbøvik, E, Stålnacke, P, Bogen, J & Bønsnes, TE, 2012, "Impact of sampling frequency on mean concentrations and estimated loads of suspended sediment in a Norwegian river: implications for water management". *Science of the Total Environment*, 433, pp.462-471.

Skeffington, RA, Halliday, SJ, Wade, AJ, Bowes, MJ & Loewenthal, M 2015, "Using high-frequency water quality data to assess sampling strategies for the EU Water Framework Directive". *Hydrology and Earth System Sciences, 19*(5), pp.2491-2504.

Smith, EP, 2002, "BACI design". *Encyclopedia of environmetrics*. Available from: http://www.stat.sfu.ca/~cschwarz/Stat-650/Notes/Handouts.readings/smith-2002-EES-baci.pdf. [April 2015].

Snell, MA, Barker, PA, Surridge, BWJ., Large, ARG., Jonczyk, J, Benskin, CMH, Reaney, S, Perks, MT, Owen, GJ, Cleasby, W & Deasy, C 2014, "High frequency variability of environmental drivers determining benthic community dynamics in headwater streams". *Environmental Science: Processes & Impacts*, 16(7), pp. 1629-1636.

Sponseller, RA, Benfield, EF & Valett, HM 2001, "Relationships between land use, spatial scale and stream macroinvertebrate communities". *Freshwater Biology, 46*(10), pp.1409-1424.

Spooner, JS, Dressing, A & Meals DW 2011, "Minimum detectable change analysis. Tech Notes 7, December 2011". *Developed for U.S. Environmental Protection Agency by Tetra Tech, Inc., Fairfax, VA*, 21 p. Available from: www.bae.ncsu.edu/programs/extension/wqg/319monitoring/tech_notes.htm [October 2014].

Stewart-Oaten, A, Murdoch, WW & Parker, KR 1986, "Environmental impact assessment:" Pseudoreplication" in time?". *Ecology, 67*(4), pp.929-940.

Stone, KC, Hunt, PG, Novak, JM, Johnson, MH, & Watts, DW 2000, "Flow-proportional, time-composited, and grab sample estimation of nitrogen export from an eastern Coastal Plain watershed". *Transactions of the ASAE, 43*(2), 281-290.

Thompson, J, Cassidy, R, Doody, DG, & Flynn, R 2014, "Assessing suspended sediment dynamics in relation to ecological thresholds and sampling strategies in two Irish headwater catchments". *Science of the Total Environment*, 468, 345-357.

Underwood, AJ 1991, "Beyond BACI: experimental designs for detecting human environmental impacts on temporal variations in natural populations". *Marine and Freshwater Research, 42*(5), pp.569-587.

Underwood, AJ 1992, "Beyond BACI: the detection of environmental impacts on populations in the real, but variable, world". *Journal of experimental marine biology and ecology, 161*(2), pp.145-178.

Underwood, AJ 1994, "On beyond BACI: sampling designs that might reliably detect environmental disturbances". *Ecological applications, 4*(1), pp.3-15. [April 2015].

USDA-NRCS 2003, *National Water Quality Handbook*. www.nrcs.usda.gov/Internet/FSE_DOCUMENTS/stelprdb1044775.pdf [October 2014].

US EPA 1990, "Environmental Monitoring and Assessment Program: Ecological Indicators". *US EPA, Office of Research and Development, Washington*, DC EPA /600/3-90/060. September.

US EPA 1997, *Monitoring Guidance for Determining the Effectiveness of Nonpoint Source Controls*. http://www2.epa.gov/polluted-runoff-nonpoint-source-pollution/monitoring-guidance-determining-effectiveness-nonpoint [April 2015].

USGS (US Geological Survey) 2006, *Pesticides in the Nation's Streams and Ground Water, 1992–2001—A Summary (Fact Sheet 2006–3028)* pubs.usgs.gov/fs/2006/3028/ [October 2015].

Vinten, A, Futter, M, Dunn, S, Stutter, M, Blackstock, K, Friburg, N, ... & Simpson, E 2009, *Monitored Priority Catchment Project Lunan Water*. Available from http://www.programme3.ac.uk/water/Lunanthirdyearreport3.pdf [November 2014].

Vinten, A, Stutter, M, & Potts J. 2011, "Assessing diffuse pollution and land management impacts on water quality in the Lunan Water, using event-based monitoring". *Ecosystems and biodiversity, Food, health and wellbeing* http://www.knowledgescotland.org/briefings.php?id=212 [October 2014].

Webb, BW, Phillips, JM, Walling, DE, Littlewood, IG, Watts, CD & Leeks, GJL 1997, "Load estimation methodologies for British rivers and their relevance to the LOIS RACS (R) programme". *Science of the Total Environment*, 194, pp.379-389.

Withers, PJ, Neal, C, Jarvie, HP, & Doody, DG 2014, "Agriculture and eutrophication: Where do we go from here?" *Sustainability, 6*(9), 5853-5875. Available from http://www.mdpi.com/2071-1050/6/9/5853/htm [October 2014].

Yates, AG, & Bailey, RC 2010, "Covarying patterns of macroinvertebrate and fish assemblages along natural and human activity gradients: implications for bioassessment". *Hydrobiologia, 637*(1), 87-100.

Yates, AG, Bailey RC &. Schwindt JA 2007, "Effectiveness of best management practices in improving stream ecosystem quality". *Hydrobiologia* 583: 331–344.

Appendices

Appendix 1
Outline of the statistical analyses performed.

1a. Step change and trend: The change between the ‘before’ and ‘after’ periods was assessed by fitting a regression model with a dummy variable for the two periods. If the p-value for the t-test on the slope of this dummy variable is <0.05 then this indicates a significant change. Trends were similarly assessed by fitting a regression model with the date (effectively the number of days from some starting point) as an explanatory variable. For the chemical data, adjustment for seasonal effects was carried out by adding a harmonic cycle to the regression model. This was done by calculating sin(2π ×day of year/365) and cos(2π ×day of year/365) and including both terms in the regression model. For the ecological data, the adjustment for seasonal effects was made by simply adding a dummy variable in the regression for spring/autumn. Where flow at the time of sampling was available, this was also included as a covariate in the regression model (flow adjustment). The chemical data are highly skewed and as an assumption of regression is that the data are normally distributed, the data were transformed by taking natural logarithms prior to the regression analyses. Without this transformation, the results would be strongly influenced by a small number of outliers. If a change of –d is found in the log transformed data, this corresponds to a percentage decrease of 100×(1–e^{-d})%.

1b. Sample size analysis: The sample size that would be required to have an 80% probability of being able to detect a given percentage change (80% power), assuming equal sample sizes ‘before’ and ‘after’ the introduction of measures, was found using standard statistical software for calculating the power of a t-test. It was assumed that the variance in both periods was equal to the residual variance from the model with seasonal adjustment or seasonal and flow adjustment. Flow adjustment will generally decrease the residual variance and therefore mean that fewer samples are required to detect a given change than without flow adjustment. It should be noted that if there are fewer samples in the ‘before’ period than in the ‘after’ period (which is likely to be the case since ‘before’ samples cannot be collected retrospectively) then even larger sample sizes would be required than those calculated. The magnitude of change that could be detected with 80% power, given the ‘before’ and ‘after’ sample sizes that are currently available, was also calculated. If this had been calculated for 50% power rather than 80% power, it would have given the MDC. These calculations were also based on the assumption that there is no temporal autocorrelation between one sample and the next. The more frequent the sampling, the greater the autocorrelation. So, for example, a greater number of weekly samples than monthly samples would be needed to detect a given change.

1c. Autocorrelation tests: A set of daily soluble reactive phosphorus data, belonging to the James Hutton Institute, collected from the Tarland catchment in 2004–2005, were used for performing this test. A plot of the partial autocorrelation function indicates at which lags the autocorrelation is significant. Coefficients outside the dashed lines on the plot are significant. In R, it is possible to fit a regression model with autocorrelated errors by using the gls function in the package nlme. An autoregressive lag 1 (AR1) model is commonly used. An AR1 model is appropriate when the autocorrelation coefficient at lag one is significant but those at other lags are not. The R code for fitting this is shown in Appendix 3. For an AR1 model the effective sample size is approximately $\frac{1-p}{1+p}$ times the actual sample size. So, for example, if the autocorrelation coefficient for daily data is 0.7, having 365 daily observations is roughly equivalent to having 64 independent observations.

Appendix 2
R code and output for fitting regression models with and without seasonal and flow adjustment, involving a step change or trend

```
library(season)
library(Kendall)
library(pwr)
library(nlme)

#Read in data
data<-read.csv(“Cessnock.csv”)

#Log transform variable
y<-log(data$SuspSolids_mg_L)

data<-data[complete.cases(y),]
y<-y[complete.cases(y)]

#Log transform flow
logflow<-log(data$flow)

#Set up dummy variable for before/after
date<-as.Date(data$Date_Taken,"%d-%b-%y")
beforeafter<-date>as.Date(“31-DEC-2010”,”%d-%b-%Y”)

#Set up harmonic terms for seasonal effects
cos<-cos(2*pi*yrfraction(date))
sin<-sin(2*pi*yrfraction(date))

#Models without seasonal or flow adjustment
#Step-change
mod1<-lm(y~beforeafter)
summary(mod1)
plot(date,y)
pred<-mod1$coef[1]+mod1$coef[2]*beforeafter
lines(date,pred)
percentchange<--100*(1-exp(mod1$coef[2]))
percentchange
rms1<-anova(mod1)[[“Mean Sq”]][2]
#Trend
mod2<-lm(y~date)
summary(mod2)
plot(date,y)
abline(mod2)
annualpercentchange<--100*(1-exp(mod2$coef[2]*365))
annualpercentchange
MannKendall(y)

#Models with seasonal adjustment
#Step-change
mod3<-lm(y~beforeafter+sin+cos)
summary(mod3)
plot(date,y)
pred<-mod3$coef[1]+mod3$coef[2]*beforeafter
lines(date,pred)
percentchange<--100*(1-exp(mod3$coef[2]))
percentchange
rms3<-anova(mod3)[[“Mean Sq”]][4]
#Trend
mod4<-lm(y~date+sin+cos)
summary(mod4)
plot(date,y)
abline(mod4)
annualpercentchange<--100*(1-exp(mod4$coef[2]*365))
annualpercentchange
#Models with seasonal and flow adjustment
#Step-change
mod5<-lm(y~beforeafter+sin+cos+logflow)
summary(mod5)
plot(date,y)
pred<-mod5$coef[1]+mod5$coef[2]*beforeafter
lines(date,pred)
percentchange<--100*(1-exp(mod5$coef[2]))
percentchange
rms5<-anova(mod5)[[“Mean Sq”]][5]
#Trend
mod6<-lm(y~date+sin+cos+logflow)
summary(mod6)
plot(date,y)
abline(mod6)
annualpercentchange<--100*(1-exp(mod6$coef[2]*365))
annualpercentchange

#Sample size

#change detectable with actual sample size
nbefore<-sum((1-beforeafter)*(1-is.na(y)),na.rm=TRUE)
nafter<-sum(beforeafter*(1-is.na(y)),na.rm=TRUE)
pwrresult<-pwr.t2n.test(n1 = nbefore, n2= nafter,sig.level =
0.05,power=0.8)
effsize<-pwrresult$d
#without flow adjustment
pcdetectable3<-100*(1-exp(-(effsize*sqrt(rms3))))
pcdetectable3
#with flow adjustment
pcdetectable5<-100*(1-exp(-(effsize*sqrt(rms5))))
pcdetectable5

#sample size required to detect chosen change with 80%
power
#specify change required e.g. 0.1 for 10% change, 0.3 for 30%
change
#without flow adjustment
reqchange<-0.3
c<-log(1-reqchange)/sqrt(rms3)
sampsizeresult<-pwr.t.test(d=c, sig.level=0.05,
power=0.8,type=c(“two.sample”))
sampsizeresult$n

#with flow adjustment
reqchange<-0.3
c<-log(1-reqchange)/sqrt(rms5)
sampsizeresult<-pwr.t.test(d=c, sig.level=0.05,
power=0.8,type=c(“two.sample”))
sampsizeresult$n
```

Output

```
> library(season)
> library(Kendall)
> library(pwr)
> library(nlme)
>
> #Read in data
> data<-read.csv(“Cessnock.csv”)
>
> #Log transform variable
> y<-log(data$SuspSolids_mg_L)
>
> data<-data[complete.cases(y),]
> y<-y[complete.cases(y)]
>
> #Log transform flow
> logflow<-log(data$flow)
>
> #Set up dummy variable for before/after
> date<-as.Date(data$Date_Taken,”%d-%b-%y”)
> beforeafter<-date>as.Date(“31-DEC-2010”,”%d-%b-%Y”)
>
> #Set up harmonic terms for seasonal effects
> cos<-cos(2*pi*yrfraction(date))
> sin<-sin(2*pi*yrfraction(date))
>
> #Models without seasonal or flow adjustment
> #Step-change
> mod1<-lm(y~beforeafter)
> summary(mod1)

Call:
lm(formula = y ~ beforeafter)

Residuals:
  Min   1Q Median   3Q   Max
-2.2613 -0.6085 -0.2461 0.3327 3.5594

Coefficients:
        Estimate Std. Error t value Pr(>ltl)
(Intercept)    1.7389   0.1742  9.981 3.2e-15 ***
beforeafterTRUE  0.5224   0.2736  1.909  0.0602 .
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.156 on 72 degrees of freedom
Multiple R-squared: 0.04819,  Adjusted R-squared: 0.03497
F-statistic: 3.645 on 1 and 72 DF, p-value: 0.06022

> plot(date,y)
> pred<-mod1$coef[1]+mod1$coef[2]*beforeafter
> lines(date,pred)
> percentchange<--100*(1-exp(mod1$coef[2]))
> percentchange
beforeafterTRUE
    68.6072
> rms1<-anova(mod1)[[“Mean Sq”]][2]
> #Trend
> mod2<-lm(y~date)
> summary(mod2)

Call:
lm(formula = y ~ date)

Residuals:
  Min   1Q Median   3Q   Max
-2.1510 -0.5952 -0.2827 0.3394 3.4356

Coefficients:
       Estimate Std. Error t value Pr(>ltl)
(Intercept) -2.6960596 2.2539058 -1.196  0.2356
date     0.0003130 0.0001515  2.065  0.0425 *
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.151 on 72 degrees of freedom
Multiple R-squared: 0.05593,  Adjusted R-squared: 0.04282
F-statistic: 4.265 on 1 and 72 DF, p-value: 0.0425

> plot(date,y)
> abline(mod2)
> annualpercentchange<--100*(1-exp(mod2$coef[2]*365))
> annualpercentchange
  date
12.10073
> MannKendall(y)
tau = 0.0989, 2-sided pvalue =0.21603
>
> #Models with seasonal adjustment
> #Step-change
> mod3<-lm(y~beforeafter+sin+cos)
> summary(mod3)

Call:
lm(formula = y ~ beforeafter + sin + cos)

Residuals:
  Min   1Q Median   3Q   Max
-2.1597 -0.7168 -0.1113 0.2021 3.1475

Coefficients:
        Estimate Std. Error t value Pr(>ltl)
(Intercept)    1.7255   0.1706 10.112 2.51e-15 ***
beforeafterTRUE  0.4514   0.2703  1.670  0.0993 .
sin       -0.1930   0.1754 -1.101  0.2748
cos        0.3997   0.2042  1.958  0.0543 .
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.13 on 70 degrees of freedom
Multiple R-squared: 0.1149,  Adjusted R-squared: 0.07692
F-statistic: 3.028 on 3 and 70 DF, p-value: 0.0351

> plot(date,y)
> pred<-mod3$coef[1]+mod3$coef[2]*beforeafter
> lines(date,pred)
> percentchange<--100*(1-exp(mod3$coef[2]))
> percentchange
beforeafterTRUE
    57.05685
> rms3<-anova(mod3)[[“Mean Sq”]][4]
> #Trend
> mod4<-lm(y~date+sin+cos)
> summary(mod4)

Call:
lm(formula = y ~ date + sin + cos)

Residuals:
  Min   1Q Median   3Q   Max
-2.2917 -0.7197 -0.1003 0.2521 3.0534

Coefficients:
       Estimate Std. Error t value Pr(>ltl)
(Intercept) -2.0404124 2.2272863 -0.916  0.3628
date     0.0002659 0.0001499  1.774  0.0805 .
sin     -0.1706742 0.1752186 -0.974  0.3334
cos     0.4020756 0.2033057  1.978  0.0519 .
```

```
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.127 on 70 degrees of freedom
Multiple R-squared: 0.1192,  Adjusted R-squared: 0.08141
F-statistic: 3.157 on 3 and 70 DF, p-value: 0.03004

> plot(date,y)
> abline(mod4)
Warning message:
In abline(mod4) : only using the first two of 4 regression
coefficients
> annualpercentchange<--100*(1-exp(mod4$coef[2]*365))
> annualpercentchange
  date
10.19194
>
> #Models with seasonal and flow adjustment
> #Step-change
> mod5<-lm(y~beforeafter+sin+cos+logflow)
> summary(mod5)

Call:
lm(formula = y ~ beforeafter + sin + cos + logflow)

Residuals:
   Min    1Q  Median    3Q   Max
-3.04424 -0.49428 0.01801 0.49001 2.25401

Coefficients:
         Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.93749  0.13404 14.454 < 2e-16 ***
beforeafterTRUE 0.10332  0.21207  0.487  0.628
sin       0.12084  0.14016  0.862  0.392
cos       -0.20792  0.17800 -1.168  0.247
logflow     0.63488  0.08688  7.308 3.95e-10 ***
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8582 on 68 degrees of freedom
 (1 observation deleted due to missingness)
Multiple R-squared: 0.5043,  Adjusted R-squared: 0.4751
F-statistic: 17.29 on 4 and 68 DF, p-value: 7.896e-10

> plot(date,y)
> pred<-mod5$coef[1]+mod5$coef[2]*beforeafter
> lines(date,pred)
> percentchange<--100*(1-exp(mod5$coef[2]))
> percentchange
beforeafterTRUE
    10.88477
> rms5<-anova(mod5)[[“Mean Sq”]][5]
> #Trend
> mod6<-lm(y~date+sin+cos+logflow)
> summary(mod6)

Call:
lm(formula = y ~ date + sin + cos + logflow)

Residuals:
   Min    1Q  Median    3Q   Max
-3.07384 -0.47149 0.02537 0.49794 2.14853

Coefficients:
        Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.0474392 1.7101004 -0.028  0.978
date      0.0001365 0.0001150  1.187  0.239
sin      0.1277760 0.1384007  0.923  0.359
cos      -0.2156385 0.1766205 -1.221  0.226
logflow   0.6291767 0.0849312  7.408 2.6e-10 ***
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8509 on 68 degrees of freedom
 (1 observation deleted due to missingness)
Multiple R-squared: 0.5126,  Adjusted R-squared: 0.484
F-statistic: 17.88 on 4 and 68 DF, p-value: 4.493e-10

> plot(date,y)
> abline(mod6)
Warning message:
In abline(mod6) : only using the first two of 5 regression
coefficients
> annualpercentchange<--100*(1-exp(mod6$coef[2]*365))
> annualpercentchange
  date
5.108828
>
> #Sample size
>
> #change detectable with actual sample size
> nbefore<-sum((1-beforeafter)*(1-is.na(y)),na.rm=TRUE)
> nafter<-sum(beforeafter*(1-is.na(y)),na.rm=TRUE)
> pwrresult<-pwr.t2n.test(n1 = nbefore, n2= nafter,sig.level =
0.05,power=0.8)
> effsize<-pwrresult$d
> #without flow adjustment
> pcdetectable3<-100*(1-exp(-(effsize*sqrt(rms3))))
> pcdetectable3
[1] 53.22889
> #with flow adjustment
> pcdetectable5<-100*(1-exp(-(effsize*sqrt(rms5))))
> pcdetectable5
[1] 43.84184
>
> #sample size required to detect chosen change with 80%
power
> #specify change required e.g. 0.1 for 10% change, 0.3 for
30% change
> #without flow adjustment
> reqchange<-0.3
> c<-log(1-reqchange)/sqrt(rms3)
> sampsizeresult<-pwr.t.test(d=c, sig.level=0.05,
power=0.8,type=c(“two.sample”))
> sampsizeresult$n
[1] 158.5906
>
> #with flow adjustment
> reqchange<-0.3
> c<-log(1-reqchange)/sqrt(rms5)
> sampsizeresult<-pwr.t.test(d=c, sig.level=0.05,
power=0.8,type=c(“two.sample”))
> sampsizeresult$n
[1] 91.84742
>
```

Appendix 3
R code for plotting autocorrelation function and fitting a model with autocorrelated residuals

```
library(season)
library(nlme)

#Read in data
data<-read.csv(“DailyTarlandData.csv”)

#Log transformations
logP<-log(data$SRP)
y<-logP

#Set up dummy variable for before/after
date<-as.Date(data$Date,"%d/%m/%Y")

#Set up harmonic terms for seasonal effects
cos<-cos(2*pi*yrfraction(date))
sin<-sin(2*pi*yrfraction(date))

#Check for autocorrelation
plot(pacf(ts(y),na.action=na.pass))

#Fit model with autocorrelation (AR1 errors)
armod<-gls(y ~ sin+cos, correlation=corARMA(p=1,q=0),na.
action=na.omit)
summary(armod)
```

crew

Scotland's centre of expertise for waters

CREW Facilitation Team

James Hutton Institute

Craigiebuckler

Aberdeen AB15 8QH

Scotland UK

Tel: +44 (0)1224 395 395

Email: enquiries@crew.ac.uk

www.crew.ac.uk

The James Hutton Institute

Scottish Government
gov.scot

CREW is a Scottish Government funded partnership between
the James Hutton Institute and Scottish Universities.