Detection of the effects of stormwater control measure in streams using a Bayesian BACI power analysis

Dong Liang a,⁎, Lora A. Harris b, Jeremy M. Testa b, Vyacheslav Lyubchich a, Solange Filoso b

a Environmental Statistics Collaborative, Chesapeake Biological Laboratory, University of Maryland Center for Environmental Science, Solomons, MD 20688, United States of America
b Chesapeake Biological Laboratory, University of Maryland Center for Environmental Science, Solomons, MD 20688, United States of America

HIGHLIGHTS
• BACI with stratification over discharge has power to detect local BMP impacts.
• Moderate load reduction from concentration changes was not detected.
• Planning of monitoring schemes and sampling frequency are essential.

GRAPHICAL ABSTRACT

Abstract

The unpredictable timing and magnitude of precipitation events and the spatiotemporal variability of constituent concentrations are major complications to effective monitoring of watershed nutrient and sediment loads. Furthermore, detecting small changes in constituent loads in response to implementation of Stormwater control measures (SCMs) against natural variability is a challenge. Nevertheless, regulatory frameworks that direct reductions of pollutants to streams frequently depend on the ability to quantify changes in loads after management interventions. The before-after-control impact (BACI) sampling design is often used to assess the effects of an environmental change made at a known point in time. However, this approach may be complicated to apply to nutrient and sediment loads in streams as the relative impact of SCMs on nutrient concentration conditional on the long term variability of discharges has not been evaluated. Multi-scale monitoring studies that provide estimates of the natural temporal and spatial variability of discharge and concentrations could provide useful information in designing a BACI study. Here we use data from the Baltimore Long Term Ecological Research (LTER) sites and urban restoration sites to develop multiple statistical measures of the effectiveness of a given monitoring scheme in revealing the hypothesized restoration effects in terms of hydrology and nutrient loads. Stratified sampling over baseflow and stormflow and the use of multiple control streams were useful tools to detect long term cumulative reductions in concentrations due to SCMs. Moderate reductions in concentration (20%), however, were not

Keywords:
Sampling design
Markov chain Monte Carlo
Statistical power
Urban restoration
Sampling frequency
Spatial and temporal variances

https://doi.org/10.1016/j.scitotenv.2019.01.125

0048-9697/© 2019 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).
1. Introduction

There are increasing needs for reliable quantification of the effects of restoration activities to determine whether the activity has achieved its goals (Osenberg et al., 2006). However, many challenges remain for identifying a link between the restoration actions and observed effects. First, monitoring is expensive and labor intensive, thus the ability to observe the effect of restoration activity is practically constrained by the number of sample units (Harmel et al., 2006; Harmel et al., 2009; Mostaghimi et al., 2007). Secondly, the selection of environmental parameters to monitor is challenged by the unknown spatial scale in which each parameter interacts with the putative restoration activities. For example, physical-chemical parameters (e.g., nitrogen concentration) might respond at a larger spatial scale than individual biological parameters (e.g. size and condition of benthic organisms) (Osenberg et al., 1994).

Stormwater control measures (SCMs) refer to structural projects implemented to help reduce surface runoff and improve water quality. Eco-hydrological research suggests that innovative SCMs such as regenerative stream/stormwater conveyance structure have the potential to improve the capacity of urban streams to control the transport of water and materials to downstream waters (Filoso et al., 2015; Thompson et al., 2014). A key challenge for the evaluation of SCM performance, however, is to implement effective monitoring programs with sampling that is intensive enough to measure SCM impacts. For example, long term monitoring studies are expensive and there are nuances associated with applying research-driven stream sampling approaches to monitoring programs that are practical for management. A primary complication for effective monitoring is that the timing and magnitude of pollutant loads is driven by unpredictable storm events. In particular, the relative impact of SCMs on nutrient concentration conditional on the long term variabilities of discharges has not been evaluated, but this has implications for when, how, and for how long streams should be monitored to detect quantifiable changes associated with a restoration target.

The Before-After Control Impact (BACI) design is regarded as the optimal design for environmental impact assessment (Manly, 2008; Stewart-Oaten et al., 1986). The basic principle is sampling one or more sites under impact, and one or more similar sites that cannot receive any impact (i.e. controls). The assumption is that any extreme change at impact sites relative to the controls is associated with the impact. The spatial differences between impact and control sites, termed as coherence in Osenberg et al. (1994), are usually analyzed in a mixed effect or Bayesian modeling framework (Benedetti-Cecchi, 2001; Conner et al., 2016) to test the magnitude of impact. To aid in the planning of a BACI monitoring of SCMs effects in streams, it would be ideal to use data from previous BACI studies conducted for the same restoration activities and at the same spatial scale. This would enable estimation of the likely effect size and the long term variabilities of the loads. Unfortunately, logistical and financial constraints limit the availability of this information. Instead, existing monitoring studies must be synthesized in planning of a BACI study (Osenberg et al., 1994).

Two types of studies are common in monitoring restoration activities (Hurlbert, 1984; US EPA, 2002; USDA, 2003): (1) long term studies at the watershed scale documenting the natural spatiotemporal variability in loads (Kaushal et al., 2008), (2) short term local scale studies that assess the impact of a specific restoration activity (Fanelli et al., 2017). Both types of studies can provide the effect size and natural variability estimates and are potentially useful to plan a BACI monitoring study. However, the observed restoration signal from either type of study can be confounded with natural variability of loads and other processes affecting the hydrological and chemical responses. For example, inter-annual variation in precipitation patterns in temperate streams often result in a contrast of “wet” or “dry” years, with “wet” years characterized by more storms and precipitation that result in greater variability of stream discharge and groundwater inputs of non-point nutrient sources. Land use for a given watershed may also contribute to variability, with higher impervious cover resulting in flashier discharge patterns as a result of faster surface water runoff (Miller et al., 2014; Walsh et al., 2005).

The aim of this study is to demonstrate how monitoring data from both the watershed scale and restoration project specific scale can be synthesized to help plan a BACI study. Specifically, we model nutrient (e.g. total nitrogen and total phosphorus) concentrations and discharge to determine the optimal BACI design for a specified impact of SCMs, and illustrate how the design efficiency changes from watershed scale to local restoration project scale. This is a unique analysis that quantifies the impacts of SCMs solely on chemical processes using long term, high quality measurements. The hierarchical Bayesian model accounts for spatial and temporal structures in the discharge and concentration data. We begin by estimating the natural variability of discharge, concentration and load using Bayesian hierarchical analyses. We then present the resulting likely number of daily samples needed to detect the simulated SCM effects in nutrient concentration given the hydrological variability. We end with a discussion of the implication of our results in the context of environmental sampling and assessment.

2. Materials and methods

2.1. Site description

We analyzed previously-measured monitoring data for constituent concentration and loads from several watersheds on the western shore of the Chesapeake Bay in Maryland, USA. Two of our study sites are within the non-tidal portion of the Gwynns Falls watershed (Fig. 1). The Baltimore Ecosystem Study is a Long Term Ecological Research site (BES-LTER) supported by the National Science Foundation that consists of a network of four longitudinal sites along the main channel of the Gwynns Falls, including two median sized, mixed land-use watersheds, and two small watersheds with relatively homogeneous land uses (Fig. 1, Groffman et al., 2004). The Glyndon watershed represents a suburban small watershed (Table 1). The Pond Branch watershed is completely forested. The BES-LTER provides data on water and nutrient fluxes from these watersheds (BES-LTER, 2018). We also evaluated a dataset from the Coastal Plain, where the Carriage Hills study included one year of monitoring of a paired control and restored watersheds in the Severn River of Anne Arundel County, Maryland US (Fig. 1). One of the pair was restored in the summer of 2010 (Fanelli et al., 2017).

2.2. Discharge record and stream chemistry sampling

Discharge records in Gwynns Falls were maintained by the US Geological Survey (USGS) using standard technology and methods (Groffman et al., 2004). Stream chemistry sampling was conducted within a few meters of the USGS gauging stations, ensuring no additional tributary flow or seepage contribution to constituent concentrations.
Samples were collected during both baseflow and stormflow periods. Weekly water chemistry samples were collected, but the day of the week was randomly selected during the previous week. No attempt was made to avoid high precipitation events. Both filtered and unfiltered total nitrogen (TN) and total phosphorous (TP) samples were analyzed using the US Environmental Protection Agency (EPA) standard methods (Groffman et al., 2004 and reference therein). Discharge records were collected continuously in Carriage Hills sites. Fifteen event based stormflow sampling efforts were conducted for TN and total suspended solid analyses (Fanelli et al., 2017).

2.3. Yield calculations

Daily loads of TN and TP exported from watersheds were estimated based on discharge \(m^3 \text{d}^{-1}\) versus constituent concentration relationships using the LOADEST method (Runkel et al., 2004). These relationships were estimated from weekly concentration data and continuous discharge values between 1998 and 2009 for the Gwynn Fall sites, and between 2011 and 2012 for the Carriage Hills sites. Nine multiple linear regression models were applied to concentrations based on daily discharge, season and decimal days since the start of long term records. The daily loads were estimated based on approximate Maximum Likelihood Estimation. Residual analyses of the loads for estimated and measured loads were conducted to assess the modeling assumptions. Models were selected based on the Akaike Information Criterion for each parameter and watershed. Daily loads \(kg \text{d}^{-1}\) were normalized by the corresponding watershed areas to generate yield estimates.

2.4. Markov chain Monte Carlo simulation

Detection of the impacts of restoration activities requires a powerful statistical test. Simulation based approaches, which do not rely on a single error term underlying classical F-tests, could be used to improve the test's capability to detect the impacts of specified magnitude (Benedetti-Cecchi, 2001). A hierarchical model was used to simulate concentration, yield and discharge data sets over the study area. Modeling was conducted on a natural logarithm scale to study the multiplicative effect of SCMs on load reduction. Let \(y_t\) denote natural logarithm of an estimated parameter (e.g. discharge, TN concentration or yield) at watershed \(i\) and day \(t\) where \(i = 1, \ldots, n\) the total number of watersheds, and \(t = 1, \ldots, T\) the total number of days in each study. Denote spatial random effect \(\eta_i\) to capture land uses or other watershed specific characteristics. We assume a Conditional Autoregressive Prior (Besag and Kooperberg, 1995) on \(\eta_i\) to model the spatial variation of the responses. The neighborhood structured was defined based on watershed sizes and land use (Table 1). Likewise, assume temporal effects \(\theta_t\) and \(\xi_t\) to capture seasonal pattern and trends of the water responses, which were assigned non-parametric seasonal and random walk of order 1 priors (Rue and Held, 2005). Our model also includes random effects \(\delta_{it}\) to model spatio-temporal interactions across watersheds and days, which could serve as surrogates for long term changes in precipitation patterns or land use. The spatiotemporal prior for \(\delta_{it}\) was constructed using the Kronecker product between spatial random effect \(\eta_i\) and temporal random effect

![Fig. 1. Baltimore Ecosystem Study sites in Gwynn Fall watersheds in the Piedmont physiographic province, and Carriage Hills restored and control watersheds in the Coastal Plain province and land uses.](image)

<table>
<thead>
<tr>
<th>Duration</th>
<th>Name</th>
<th>Drainage (ha)</th>
<th>Landuse</th>
</tr>
</thead>
<tbody>
<tr>
<td>10/1998–03/2009</td>
<td>Pond Branch</td>
<td>32.3</td>
<td>Forest</td>
</tr>
<tr>
<td>Glyndon</td>
<td>81</td>
<td>Suburban</td>
<td></td>
</tr>
<tr>
<td>08/2011–08/2012</td>
<td>Carriage Hills Control</td>
<td>5.6</td>
<td>Suburban</td>
</tr>
<tr>
<td>Carriage Hills restored</td>
<td>5.4</td>
<td>Suburban</td>
<td></td>
</tr>
</tbody>
</table>
\( \xi \) (Rue and Held, 2005). Let \( \mu \) denote overall mean and \( \varepsilon_t \) the white noise process at day \( t \) and watershed \( i \).

\[
y_{it} = \mu + \tau_i + \theta_t + \omega_k + \varepsilon_t \quad \text{where } \varepsilon_t \sim N(0, \sigma^2) \tag{1}
\]

The Bayesian model was completed with conjugate gamma priors on the scales parameters such as \( \sigma^2 \) and parameterized to be non-informative. Detailed prior specification is given in Appendix. Sensitivity analyses were conducted to assess the robustness of sample size calculation in those prior specifications. Posterior inference was conducted using an efficient Markov chain Monte Carlo (MCMC) sampler (Liang and Kumar, 2013) and implemented in the R statistical language. Eight parallel MCMC chains were run for 7000 iterations with the initial 2000 iterations discarded as burn-in, and every 16th draw was retained for posterior inferences.

2.5. Natural variability assessment

For the modeling effort it was necessary to document the variability of the loads, and then use this characterization to generate simulated loading estimates for the analysis. Posterior predictive density for the loads, and then use this characterization to generate simulated yield ratios between each design, the sampling options include simple random sampling (SRS) or stratified random sampling (STS). The days were stratified into baseow flow and stormflow periods. The 95th concentration was detectable with 72 days of sampling using the SRS and BA study design. The sample size for BACI with single control was higher at 1216 days. Incorporating a second control reduced the sample size to 664 days. Stratification using discharge records into baseow and storm flow was efficient; fewer days were needed to detect the same SCM effect under stratification. The differences were all statistically significant given the posterior uncertainty estimates. An 80% reduction of concentration was detectable with 72 days of sampling using the SRS and BA. The sample size for BACI with single control was higher at 82 days, and the difference was statistically significant. The difference between SRS and STS was not statistically significant. None of the three simulated reductions in TP concentration was detectable using the same designs.

3. Results

3.1. LTER results

TN yields from the completely forested headwater watershed (Pond Branch) were low relative to the urban headwater watershed (Glyndon) (Fig. 2). The mean yields in urban headwaters were >10 times that of the forested watershed (data not shown). Fall yields were most variable while winter yields were least variable. The seasonal pattern was more obvious in the forested watershed with a consistent winter-spring peak, while the urban watershed yields exhibited more flashy patterns throughout the years. Peak yield episodes were temporally aligned between the paired watersheds. The yields estimates were based on measured discharge and concentration rather than model simulations.

From the Glyndon results (Table 2), a 20% reduction of TN concentration was not detectable with any of the design options, given the high long term variability of discharge. A 50% reduction of TN concentration was detectable with 1075 days of sampling using the SRS and BA study design. The sample size for BACI with single control was higher at 1216 days. Incorporating a second control reduced the sample size to 664 days. Stratification using discharge records into baseow and storm flow was efficient; fewer days were needed to detect the same SCM effect under stratification. The differences were all statistically significant given the posterior uncertainty estimates. An 80% reduction of concentration was detectable with 72 days of sampling using the SRS and BA. The sample size for BACI with single control was higher at 82 days, and the difference was statistically significant. The difference between SRS and STS was not statistically significant. None of the three simulated reductions in TP concentration was detectable using the same designs.

<table>
<thead>
<tr>
<th>Site</th>
<th>Scenario (a)</th>
<th>(b)</th>
<th>(c)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glyndon (TN)</td>
<td>SRS STS</td>
<td>SRS STS</td>
<td>SRS STS</td>
</tr>
<tr>
<td>20% n.a.</td>
<td>1075 823 1216</td>
<td>1001 664 582</td>
<td></td>
</tr>
<tr>
<td>50% (92)</td>
<td>(63) (101) (72)</td>
<td>(61) (49)</td>
<td></td>
</tr>
<tr>
<td>80% (71.2)</td>
<td>(69.1) (82.2) (83.2)</td>
<td>(42.1) (47.1)</td>
<td></td>
</tr>
<tr>
<td>Glyndon (TP)</td>
<td>SRS STS</td>
<td>SRS STS</td>
<td>SRS STS</td>
</tr>
<tr>
<td>20% n.a.</td>
<td>50% n.a.</td>
<td>80% n.a.</td>
<td></td>
</tr>
<tr>
<td>50% (51.5)</td>
<td>(46.4) (74.9) (74.9)</td>
<td>(38.4) (43.4)</td>
<td></td>
</tr>
<tr>
<td>Carriage Hills (TN)</td>
<td>SRS STS</td>
<td>SRS STS</td>
<td>SRS STS</td>
</tr>
<tr>
<td>20% 64(8)</td>
<td>55(6) 59(13) 90(11)</td>
<td>47(6) 51(6)</td>
<td></td>
</tr>
<tr>
<td>80% (42.5)</td>
<td>(39.4) (61.7) (63.6)</td>
<td>(31.4) (37.3)</td>
<td></td>
</tr>
</tbody>
</table>

Number of sampling days (posterior mean and standard deviation in brackets) to detect the simulated reductions in concentration (20%, 50%, 80%) with type I error 0.10 and power 0.90. Design options include (a) Before and After; (b) Before After Control Impact (with one control site); and (c) Before After Control Impact (with two control sites). SRS denotes random sampling of days and STS denotes stratified sampling of stormflow days (defined as the 95th percentile of discharge from the urban site) and baseow days. TN: total Nitrogen; TP: total Phosphorus. n.a.: not detectable with finite sample sizes.
3.2. Carriage Hills results

The Carriage Hills monitoring focused on TN and total suspended solids (TSS). Total phosphorus was only estimated from a sub-set of samples based on TSS. TN yields from the restored watershed in the Carriage Hills site were low relative to the control counterpart (Fig. 3). There was no obvious seasonal pattern in yields. The yields were more stable in the restored watershed, while the control watershed still

---

**Fig. 2.** Time series of LOADEST estimates of yields using measured discharges and concentrations. In Glyndon (urban) watershed and Pond Branch (forested) watershed in Gwynn Falls.

**Fig. 3.** Time series of LOADEST estimates of yields using measured discharges and concentrations in control and restored watershed in Carriage Hills sites.
exhibited the typical flashy pattern of an urban watershed. A storm event in summer 2012 resulted in elevated yields in both watersheds. The simulated effects of 20%, 50%, and 80% reductions in concentrations due to SCMs in Carriage Hills were detected (Table 2). Smaller sample sizes were needed to detect larger effects, but there was no statistically significant difference between the BA and BACI design, and the SRS and STS designs.

The MCMC approach generates dependent samples from the joint posterior distribution, thus we evaluate whether MCMC samples have sampled the joint distribution sufficiently, and whether the results are accurate enough. Selected model diagnostic plots indicate that the MCMC algorithms have converged to posterior modes (Supplemental Fig. 1), and the model fits (Supplemental Fig. 2) were not sensitive to hyper-prior specifications.

4. Discussion

At the long term Baltimore study sites, changes in nutrient loading caused by SCM effects that only involve concentration are hard to detect given the natural variability of discharge. Specifically, potential changes in loading associated with a moderate reduction (i.e. 20%) of TN and TP concentrations are masked by discharge variation and the corresponding load reduction cannot be detected with any of the simulated design options. This challenge calls for careful planning and design of future monitoring studies, and communication of expectations regarding how well we can quantify the results of restoration efforts. In particular, the Chesapeake Bay Total Maximum Daily Loads and its emphasis on “Watershed Implementation Plans” where numerous restoration activities are proposed to meet required percent reductions in pollutant loads will require thoughtful attention to monitoring to ensure expected changes are quantifiable. These considerations are especially timely, considering that future climate projections in this region call for elevated volume and seasonality of precipitation (Wagenet et al., 2018). We did identify stratification over discharge and paired watershed monitoring as useful design options to increase the probability of detecting such SCM effects. For a simulated 50% reduction of TN concentration, using two instead of one control resulted in a 40% reduction in sample size; while stratification provided 10–20% reduction in sample sizes across all detectable load reduction scenarios and number of controls (Table 2). Using BACI instead of BA resulted in a 13% increase in sample size using SRS, this result indicates the challenges of selecting an appropriate control site. Inappropriate pairing of sites could lead to lack of coherence in load time series, and loss of power from the BACI design.

Our results from the Carriage Hills sites indicate that the effects of a specific and local restoration activity can be detected with much less sampling effort than the Baltimore simulation study that represents efforts to detect change from multiple efforts across a given watershed that are implemented with an expectation of a percentage reduction in pollutant load. This was probably a consequence of the differential patterns of chemistry and hydrology between a focused local restoration activity during a short period versus a mixture of restoration activities and their cumulative improvements over a relatively long time periods (Benedetti-Cecchi, 2001; Underwood, 1994). For local and focused restorations, the volume of water conveyed in the treatment stream during smaller storms is clearly reduced, and the corresponding decreases in concentration were strong relative to the natural variability of yields (Runkel et al., 2004), and increase the needed sample size. Thus the sample size reported here may be too small. However, the LOADEST regression for discharge and load reported relatively high R² values, we do not expect qualitative changes in the ordering of these design options.

Long term monitoring before a planned restoration activity is rare in practical environmental studies due to logistical and financial constraints (Osenberg et al., 2006; Underwood and Chapman, 2003). Thus the results presented here assuming equal monitoring efforts before and after restoration activity are not realistic for practical impact assessment. It is however, reasonable to expect regular monitoring of many watersheds of a random chosen set of sites for a standard set of parameters. The randomization ensures that these watersheds are representative in term of physical and biological conditions. It is reasonable to assume that the spatial and temporal variabilities estimated from these regular monitoring data could approximate the natural variability of the impacted sites “before” restoration (Underwood, 1994). These pseudo-control data can thus be incorporated as prior information in a BACI study. Our results show that using two such pseudo-controls can reduce the sample sizes by 40%. USGS gauging stations provide a network of regular monitoring data at multiple spatial scales and order of streams. These data may serve as pseudo-controls to detect the hydrological impacts of SCM implementations at multiple spatial scales (Underwood, 1994).

5. Conclusions

- Magnitude of cumulative and long term load reduction from reduced concentration only was hard to detect, which may indicate the importance of hydrology for planning restoration activities.
• Appropriate selection of pseudo-control watersheds, with stratification over baseflow and stormflow provided the highest probability of detecting the load reduction. Results indicate the strength of BACI design principles in instances where long term monitoring data are available.

• Data indicate that despite the challenges, short term intensive restoration activities can be detectable with feasible monitoring. This may have implication for regular monitoring of a random set of focused sites to improve the utility of the monitoring data.

Acknowledgements

The authors would like to thank the reviewers for their comments, which have resulted in a clearer manuscript. The authors acknowledge data from Baltimore ecosystem study as part of the NSF LTER monitoring (P. Groffman). Carriage Hills restoration data was supported as part of the Severn Riverkeeper monitoring grant to S. Filoso and a NOAA grant to M. Williams and others. We acknowledge inputs from the modeling group at the Chesapeake Bay Program (G. Shenk) regarding the simulation designs. The project is funded by grant 07.4.31575 from the Maryland Department of Natural Resources and the Chesapeake Bay Trust, United States and by Drs. Liang and Lyubchich’s start-up funds provided by the Chesapeake Biological Laboratory, UMCES. This is UMES Publication # 5555.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.scitotenv.2019.01.125.

References


Liang, D., Kumar, N., 2013. Time-space kriging to address the spatiotemporal misalignment in the large datasets. Atmos. Environ. 72, 60–69.


10.1016/j.scitotenv.2019.01.125.