HYDROLOGICAL PROCESSES
Hydrol. Process. **28**, 1260–1283 (2014)
Published online 4 January 2013 in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1002/hyp.9614

Application of the SPARROW model in watersheds with limited information: a Bayesian assessment of the model uncertainty and the value of additional monitoring

Christopher Wellen,[1]* George B. Arhonditsis,[1] Tanya Labencki[2] and Duncan Boyd[2]

[1] *Ecological Modelling Laboratory, Department of Physical & Environmental Sciences, University of Toronto, Toronto, ON, Canada, M1C 1A4*
[2] *Great Lakes Unit, Water Monitoring & Reporting Section, Environmental Monitoring and Reporting Branch, Ontario Ministry of the Environment, Toronto, ON, Canada, M9P 3V6*

Abstract:

How can spatially explicit nonlinear regression modelling be used for obtaining nonpoint source loading estimates in watersheds with limited information? What is the value of additional monitoring and where should future data-collection efforts focus on? In this study, we address two frequently asked questions in watershed modelling by implementing Bayesian inference techniques to parameterize SPAtially Referenced Regressions On Watershed attributes (SPARROW), a model that empirically estimates the relation between in-stream measurements of nutrient fluxes and the sources/sinks of nutrients within the watershed. Our case study is the Hamilton Harbour watershed, a mixed agricultural and urban residential area located at the western end of Lake Ontario, Canada. The proposed Bayesian approach explicitly accounts for the uncertainty associated with the existing knowledge from the system and the different types of spatial correlation typically underlying the parameter estimation of watershed models. Informative prior parameter distributions were formulated to overcome the problem of inadequate data quantity and quality, whereas the potential bias introduced from the pertinent assumptions is subsequently examined by quantifying the relative change of the posterior parameter patterns. Our modelling exercise offers the first estimates of export coefficients and delivery rates from the different subcatchments and thus generates testable hypotheses regarding the nutrient export 'hot spots' in the studied watershed. Despite substantial uncertainties characterizing our calibration dataset, ranging from 17% to nearly 400%, we arrived at an uncertainty level for the whole-basin nutrient export estimates of only 36%. Finally, we conduct modelling experiments that evaluate the potential improvement of the model parameter estimates and the decrease of the predictive uncertainty if the uncertainty associated with the current nutrient loading estimates is reduced. Copyright © 2012 John Wiley & Sons, Ltd.

Supporting information may be found in the online version of this article.

KEY WORDS SPARROW model; Bayesian inference; uncertainty analysis; value of information; nutrient loading estimation; Hamilton Harbour

Received 12 April 2012; Accepted 8 October 2012

INTRODUCTION

Despite decades of research on the nutrient export dynamics of watersheds, nonpoint sources of excess nutrients continue to impair the quality of receiving water bodies, leaving a substantial number of North American and European lakes classified as eutrophic or hypereutrophic (Schindler, 2006). In this regard, there is a pressing demand for watershed models which can support water quality management goals, such as the estimation of nonpoint source nutrient loads and the examination of alternative land use scenarios (Rode *et al*., 2010). While a suite of distributed process-based models exists to meet these needs (e.g. SWAT, Arnold *et al*., 1994; HSPF, Donigian *et al*., 1995; see also review by Borah and Bera, 2003), such models are often too complex and data-demanding to be applied in any but the most intensively monitored catchments (Borah and Bera, 2004). At the other end of the complexity spectrum, simple empirical models for estimating loads do exist (Cohn *et al*., 1989, 1992), but their application does not offer any insights into watershed functioning, and thus no ability to project future watershed response to management interventions and changing climatic conditions or land uses. The SPAtially Referenced Regressions on Watershed attributes (SPARROW) modelling approach was developed to bridge the gap between process-based and empirical models of catchment water quality (Schwarz *et al*., 2006). SPARROW expresses mean annual loads (mass year^{-1}) as nonlinear functions of watershed attributes, including nutrient sources, deliveries, stream and reservoir attenuation. Albeit a regression model, SPARROW explicitly considers a number of processes that enable the exploration of land use scenarios.

To date, the SPARROW model has been applied only in catchments which have been comparatively well studied, with study sites having at least 36 water quality monitoring stations with bi-weekly sampling (Schwarz *et al*., 2006; see their Table 1.1). Yet, many watersheds of management interest are not intensively monitored, and the development

*Correspondence to: Christopher Wellen, Ecological Modelling Laboratory, Department of Physical & Environmental Sciences, University of Toronto, Toronto, Ontario, Canada, M1C 1A4.
E-mail: christopher.wellen@utoronto.ca

of methodological frameworks to guide model implementation in such cases is often highlighted as one of the emerging imperatives of the contemporary modelling practice (Rode *et al*., 2010). In this context, uncertainty analysis must be a cornerstone feature for quantifying predictive uncertainty associated with model inputs as well as knowledge gaps from the studied catchment (Pappenberger and Beven, 2006). Formal Bayesian approaches have been proposed as a promising means to accommodate the uncertainty underlying the challenges of watershed modelling in a comprehensive and statistically defensible manner (Kuczera and Parent, 1998; Vrugt *et al*., 2005). These challenges include the spatial and temporal correlation of real-world processes and model residuals (Yang *et al*., 2007, 2008; Rode *et al*., 2010) as well as the lack of commensurability between measured variables and model inputs, e.g., point precipitation is often measured, whereas mean aerial precipitation is used for model input (Kavetski *et al*., 2006a,b; Vrugt *et al*., 2008; Balin *et al*., 2010). A major advantage of the Bayesian methods when calibration data are scarce is the incorporation of prior knowledge on model parameters, thereby improving our capacity to locate realistic areas of the parameter space associated with high model likelihood values (Omlin and Reichert, 1999; Qian *et al*., 2003). Importantly, Bayesian uncertainty analysis techniques feature statistically sound likelihood functions, the use of which provide meaningful credible intervals for model predictions (Hong *et al*., 2005). There are multiple techniques of different degrees of complexity to quantify model uncertainty due to parameters and other model inputs, model structure and data error (Wagener and Gupta, 2005; Ajami *et al*., 2007; Arhonditsis *et al*., 2008a,b; Rode *et al*., 2010).

Qian *et al*. (2005) present a Bayesian application of SPARROW which clearly demonstrated the advantages of statistical formulations characterizing the spatial structure of model residuals due to autocorrelated forcing factors, e.g., climate and soils. Wellen *et al*. (2012) demonstrated a Bayesian approach to incorporate interannual variability into the SPARROW model. Hitherto, there has not been a Bayesian application of SPARROW focused on addressing three fairly core issues, and the second two have been entirely neglected: (i) the uncertainty of model calibration data; (ii) the importance (or lack thereof) of informative prior parameter distributions in assisting model calibration; and (iii) the implications of the covariance of model parameters on the inference drawn and the posterior patterns derived. The first problem is quite critical considering that estimates of mean annual load are often obtained by rating curve models and are typically characterized by substantial uncertainty (Cohn *et al*., 1989, 1992; Alexander *et al*., 2002, 2004; Moatar and Meybeck, 2005). Despite the questionable quality of the calibration datasets, most SPARROW applications do not explicitly account for their uncertainty. Regarding the second issue, the use of information about the relative plausibility of parameter values aims to reduce the disparity between what ideally we want to learn (internal description of the system) and the data available to guide model calibration. In doing so, we can conceivably overcome the problem of poor parameter identification when basing model calibration on limited data. Evidence about the latter problem was provided by Qian *et al*. (2005), who showed that three of the SPARROW parameters were highly correlated and concentrated around a narrow banana-shaped region of the prespecified parameter space. Because this covariance pattern can undermine the search of the maximum likelihood with conventional numerical optimization algorithms, the same study also underscored the importance of selecting (i) an efficient sampling scheme for generating input vectors, which are then evaluated with regard to the model performance, and (ii) a proper statistical description of the prior parameter space and likelihood function.

The main objective of this study was to develop a Bayesian framework for applying SPARROW to watersheds which are not intensively monitored, and subsequently to assess the uncertainty of the model application while addressing the three core issues mentioned previously. After introducing the case study and describing the watershed data, we present the calculation of the mean annual loads and the associated uncertainties, relying directly on mean values of concentration and flow. We then develop data quality submodels, which quantify the uncertainty of the load estimated at each water quality monitoring station, thereby ensuring that the predictive statements made by the model also reflect the uncertainty surrounding the calibration dataset. We examine three data error characterizations and five statistical configurations of the SPARROW model, two of which aim to overcome the spatial autocorrelation of model residuals. There are also two statistical formulations designed to explicitly accommodate parameter covariance, an issue not examined before in the context of SPARROW models. Finally, we conduct modelling experiments that evaluate the potential improvement of parameter estimates and the decrease of predictive uncertainty if the precision of the currently available nutrient loading estimates is increased.

METHODOLOGY

Case study

Hamilton Harbour is a large embayment at the western end of Lake Ontario, which has a history of eutrophication problems manifested as algal blooms, low water transparency, prevalence of toxic cyanobacteria and low hypolimnetic oxygen concentrations during the late summer (Hiriart-Baer *et al*., 2009; Ramin *et al*., 2011). Although it is clear from earlier work that the sewage treatment plants play a critical role in governing total phosphorus and chlorophyll α concentrations in the Harbour, there is substantial uncertainty in the ambient conditions driven by the nutrient loadings from the drainage basin (Gudimov *et al*., 2010). Hamilton Harbour's drainage basin is about 450 km^2 in aerial extent and consists of watersheds dominated by agricultural land use (Grindstone and Spencer Creeks) and urban land use (Redhill and Indian Creeks; see Figures 1 and 2; Ontario Ministry of Natural Resources, 2005, 2008). The soils of the Harbour basin

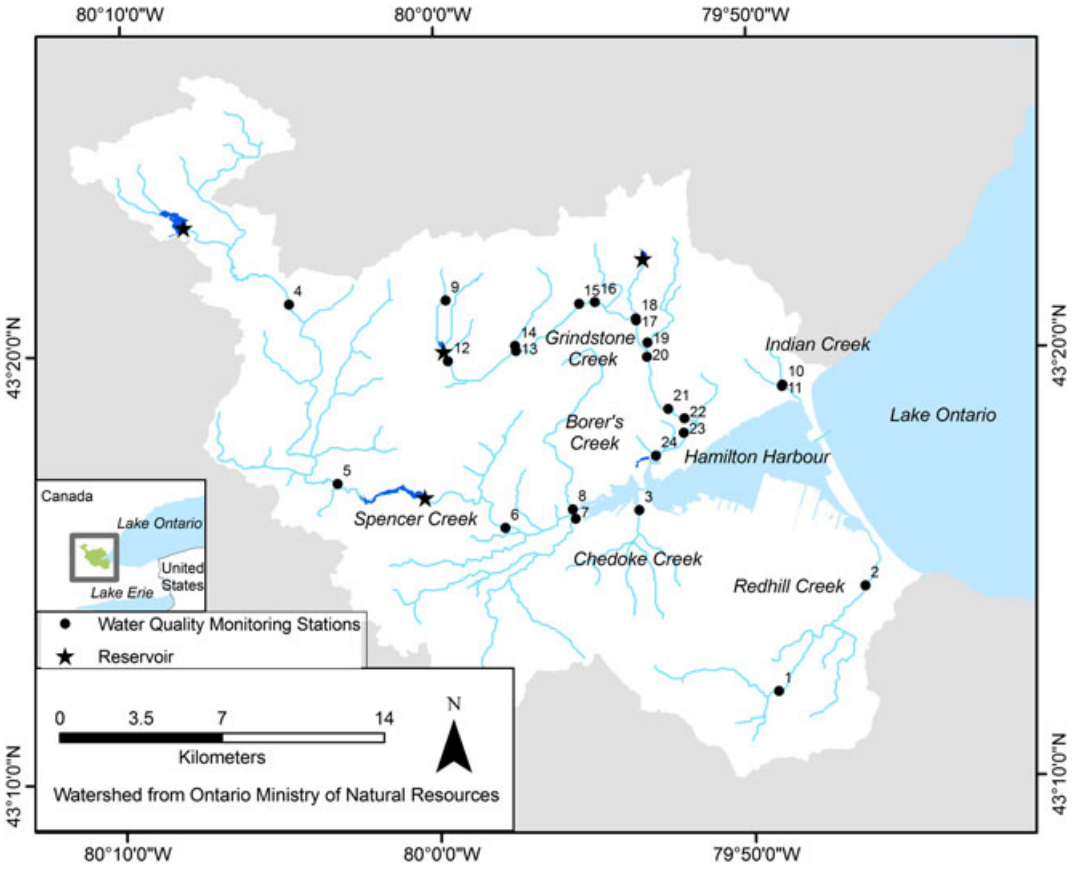


Figure 1. Map of the Hamilton Harbour watershed, western end of Lake Ontario, Ontario, Canada

are mainly loams (25%), sandy loams (28%) and silty loams (20%), whereas organic soils, silty clay loams and clay loams together make up about 10% of the basin soils, with most of the remainder composed of rocky outcroppings and ravines. Soils are spread relatively evenly between the four Natural Resources Conservation Service's soil hydrologic runoff groups – groups A and B, those least likely to generate runoff, have 23% coverage, respectively, group C has 29% coverage and group D, the group most likely to generate runoff, has 24% coverage (Ontario Ministry of Agriculture and Food, 2005). The slopes of the Harbour basin are mild, with the exception of the Niagara Escarpment. The average slope of the entire basin is 4.4%, and ignoring all slopes greater than 30% the average is 3.8% (Ontario Ministry of Natural Resources, 2005).

SPARROW model

The SPARROW model has been extensively described elsewhere (Alexander *et al.*, 2002; McMahon *et al.*, 2003; Qian *et al.*, 2005; Schwarz *et al.*, 2006), so only a basic introduction is given here. SPARROW consists of a two-level hierarchical spatial structure. Watersheds are first divided into subwatersheds, each of which drains to a water quality monitoring station. Each subwatershed is then disaggregated into reach catchments drained by a particular stream segment. In this paper, the mean annual load of total phosphorus is the response variable of SPARROW, whereas watershed attributes aggregated to the reach catchments are used as predictor variables.

The SPARROW model can be expressed as

$$\mu_i = Ln\left\{\sum_{n=1}^{N}\sum_{j=1}^{J_i}\beta_n S_{n,j} e^{(-\alpha Z_j)} H_{i,j}^S H_{i,j}^R\right\}\varepsilon_i \qquad (1)$$

where the subscripts i and j refer to subwatersheds and reach catchments, respectively; μ_i refers to the (log-transformed) mean annual total phosphorus load measured at station i in metric tons per year; n, N refers to the source index, where N is the total number of sources (diffuse and point sources) and n is an index for each source; J_i refers to the number of reaches in subwatershed i; β_n refers to the estimated source coefficient for source n; $S_{n,j}$ refers to the quantity of source n in reach j, where $\beta_n S_n$ has units metric tons per year; α refers to the vector of land to water delivery coefficients; Z_j is a vector of the land-surface characteristics associated with drainage to reach j; $H_{i,j}^S$ refers to the fraction of nutrient mass originating in reach j remaining at station i as a function of first-order loss processes in streams; $H_{i,j}^R$ refers to the fraction of nutrient mass originating in reach j remaining at station i as a function of first-order loss processes in lakes and reservoirs; and ε_i refers to a random multiplicative error term assumed to be independently and identically distributed across all subwatersheds.

Nutrient loss processes in streams (e.g., loss to sediments and biota) are modelled with a first-order loss function:

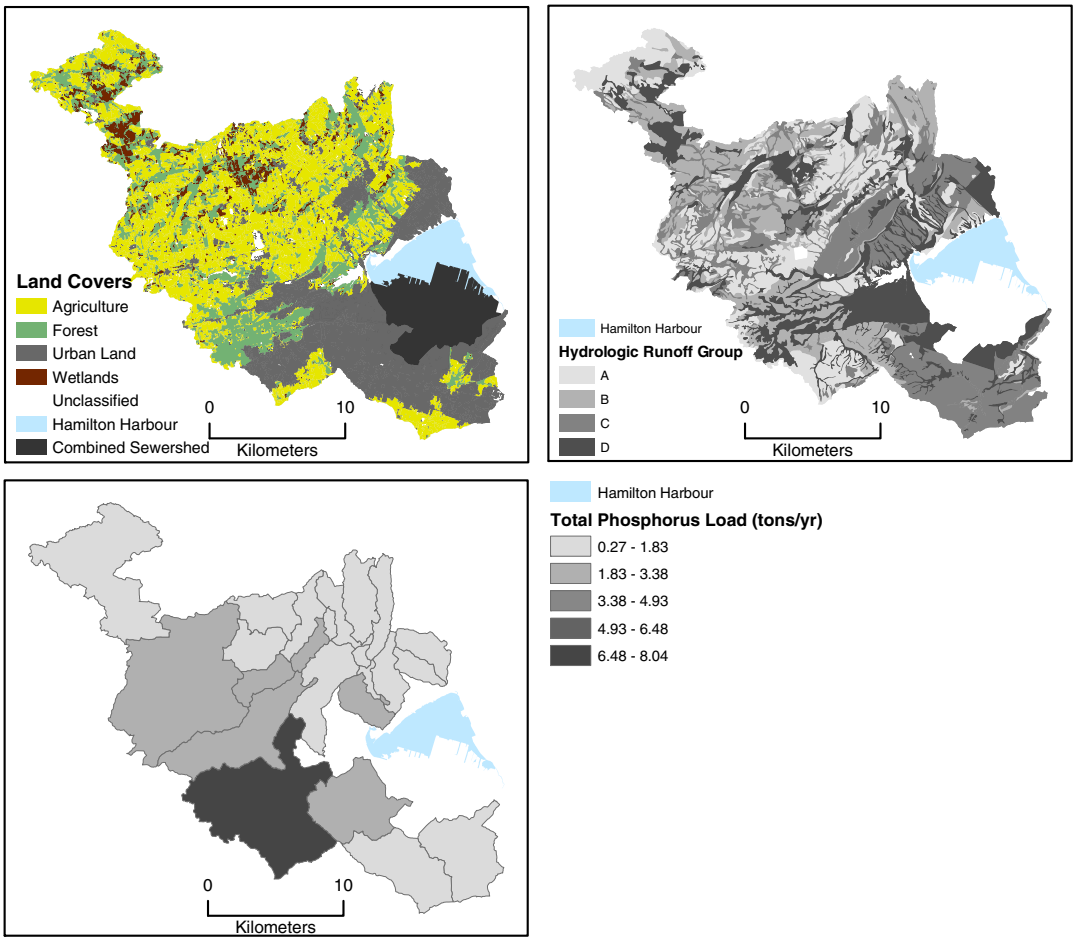


Figure 2. Land cover classification (upper left), hydrologic runoff groups (upper right) and measured total phosphorus loads derived from the E2 dataset (lower left) in the Hamilton Harbour watershed

$$H^{S}_{i,j} = \exp\left(-k_s L_{i,j}\right) \tag{2}$$

where k_s refers to the first-order loss coefficient for streams (km^{-1}) and $L_{i,j}$ refers to the stream length in kilometres between reach i and station j. To aid the reader's interpretation, a k_s value of 0.04 indicates that total phosphorus is removed from streams at a rate of about 4% per kilometre.

First-order loss processes operating in lakes and reservoirs are limited to loss to sediment, which is expressed as

$$H^{R}_{i,j} = \prod_{l} \exp\left(-k_r q_l^{-1}\right) \tag{3}$$

where l refers to any lakes or reservoirs between reach i and station j, k_r refers to the first-order loss coefficient or settling velocity (m $year^{-1}$), q_l refers to the aerial hydraulic loading of the lake/reservoir (m $year^{-1}$). Table I presents all the calibrated parameters (and other stochastic nodes) of the model.

Data sets

Spatial data sets. We provide an extensive description of the spatial datasets used as inputs to the SPARROW model in the Electronic Supplementary Material (ESM) and a brief overview here. We used a 10-m digital elevation model to delineate the subwatersheds. Our calibration dataset had 24 subwatersheds. Their areas ranged from 0.3 to 75.8 km^2, with a mean of 17.9 km^2 and an interquartile range of 25.7–6.8 = 18.9 km^2. There are a total of 175 reach catchments, and each reach catchment discharges into a confluence, reservoir, or water quality monitoring station. Reach catchment areas ranged from 0.02 to 19.3 km^2, with a mean of 2.5 km^2 and an interquartile range of 3.5–0.9 = 2.6 km^2. Each reach is drained by a single stream. The mean stream length is 2.4 km with an interquartile range of 3.2–1.2 = 2.0 km. Only one stream class is included in the model. Four reservoirs were used during the parameter estimation of the SPARROW model (Figure 1). Nonpoint nutrient sources included in the model were agricultural land and urban land, together representing 80% of the basin area. A single waste water treatment plant, the Waterdown plant, drained into one of the streams. The mean loading for this plant between 1996 and 2007 was 0.3 tons of total phosphorus per year, with an interquartile range of 0.4–0.2 = 0.2 tons per year. We assumed that the delivery of total phosphorus to streams is primarily controlled by soil runoff potential, parameterized as a function of the soil hydrologic runoff group. Following McMahon *et al*. (2003), we calculated an area-weighted average of soil hydrologic runoff group for each reach catchment. We assigned values of 1 through 4 to soil

Table I. Stochastic nodes of the different SPARROW model configurations examined

Parameter	Description	Units
α	Land to water delivery coefficient	
β_1	Export coefficient for urban land	Metric tons P km^{-2} year^{-1}
β_2	Export coefficient for agricultural land	Metric tons P km^{-2} year^{-1}
k_r	Reservoir settling velocity	m year^{-1}
k_s	Stream attenuation coefficient	km^{-1}
$s_{\theta 1 \theta 2}$	Covariance between two of the parameters α, β_1, β_2	*Ln*(tons P km^{-2} year^{-1}) or *Ln*(tons P km^{-2} year^{-1})^2
σ	Model error standard deviation	*Ln*(tons P year^{-1})
τ	Model error standard deviation specific to CAR and STSP	*Ln*(tons P year^{-1})
ψ	Model error standard deviation specific to CAR	*Ln*(tons P year^{-1})

SPARROW, SPAtially Referenced Regressions On Watershed attributes; STSP, state space; CAR, continuous autoregressive.

groups A, the most well-drained group, through D, the most poorly drained group. We then took the reciprocal of the reach-level average so that lower numbers indicate higher nutrient delivery rates. Figure 2 contains maps of land use and soil runoff groups.

Total phosphorus loads. There are many approaches to calculating annual constituent loads when using noncontinuous concentration records. However, relatively few methods exist for noncontemporaneous records of concentration and flow. Moatar and Meybeck (2005) compared the accuracy and precision of a number of different approaches to calculate annual phosphorus loads and recommended the use of the product of means of sampled concentrations and annual discharge, similar to the approach adopted herein. Our study builds upon Moatar and Meybeck's (2005) findings, and the natural logarithm of the mean annual load is expressed as follows:

$$Ln(\text{Load}_i) = Ln(\text{Flow}_i) + Ln(\text{Conc}_i) \quad (4)$$

where the subscript i refers to a subwatershed, $Ln(\text{Flow}_i)$ is estimated from a discharge-area regression presented in the ESM, and $Ln(\text{Conc}_i)$ represents the mean of the natural log-transformed in-stream total phosphorus concentrations measured by the Ontario Ministry of the Environment's Provincial Water Quality Monitoring Network (Ontario Ministry of the Environment, 2010). This program measures stream water quality during coordinated field trips with monthly or bi-weekly frequency. Only concentration measurements after the year 1987 at stations that drained at least 5.25 km^2 were used for load estimation, and thus, we developed two loading datasets: The first loading dataset (Error 1 or E1) was based on mean annual loads derived from Equation (4) at all 24 stations. The mean loading across all 24 subwatersheds of the E1 dataset is 1.62 tons per year, with an interquartile range of 1.88–0.77 = 1.11 tons per year. This method of calculating annual loads may result in an underestimation of annual loads, as it implicitly assumes that flow and concentration are independent (Preston *et al.*, 1989). In this particular study, the introduced bias is fairly minimal as the correlation coefficients between contemporaneous measurements of flow and concentration ranged from −0.06 to 0.36, with a mean of 0.1.

The second loading dataset (Error 2 or E2) computes the mean annual load with Equation (4) at the 18 stations, which lack contemporaneous data, and also uses a rating curve to estimate the mean annual load at the six water quality monitoring stations where contemporaneous data are available. All rating curve calculations were carried out with the United States Geological Survey LOADEST program (Runkel *et al.*, 2004). All available contemporaneous measurements of concentration and flow were used to parameterize a linear regression between log-transformed daily flow and log-transformed daily loading:

$$Ln(\text{Load}) = a_0 + a_1 Ln(Q) \quad (5)$$

where a_0 and a_1 are regression coefficients, and $Ln(Q)$ refers to trend-corrected stream flow. Estimated daily loading values were then averaged and aggregated to a yearly timescale to yield mean annual loading. The mean loading across all 24 subwatersheds of the E2 dataset is 1.61 tons per year, with an interquartile range of 1.88–0.77 = 1.11 tons per year. Thus, the summary statistics of the two datasets were almost identical, the low correlation between concentration and flow suggests limited bias of the E1 estimates, whereas their comparison also suggests reasonable correspondence, i.e., $Ln(\text{Load})_{\text{Eq.5}} = 1.07\ Ln(\text{Load})_{\text{Eq.4}} + 0.54$ ($r^2 = 0.92$, $n = 6$). We have included a table in the ESM detailing various attributes and summary statistics of the calibration dataset (Table ESM-1).

Data quality submodel. There are two approaches for representing measurement error in models. The classical approach assumes that the observed values of a variable, Y_i, are drawn from a probability distribution with an expected value, Load_i, the 'true' value of the variable being sampled (Carroll *et al.*, 2006). The classical approach is appropriate when the error stems from deficiencies in sampling or measurement and has been recently used to model the uncertainty of point rainfall estimates (Balin *et al.*, 2010). The Berkson model takes the opposite approach, in that the true value is assumed to be drawn from a distribution with expected value equal to the observed datum. The Berkson approach is appropriate when the uncertainty is assumed to stem from a lack of commensurability between what has been measured and what the variable one is interested in,

and has been applied to estimate mean aerial rainfall from point measurements (Kavetski *et al*., 2006a,b; Ajami *et al*., 2007). The key difference between the two strategies resides in whether the observed values vary about the true ones (classical) or the true values vary about the observed ones (Berkson). We assumed that the uncertainty in load estimates stems from a combination of sampling and analytic errors rather than a lack of commensurability, so we opted for the classical representation of measurement error for annual loads.

Our data quality submodel postulates that the log-transformed loadings are random variables drawn from normal distributions with mean values equal to the previously described estimates and variances representing the associated error and/or temporal variability at each site. Although this assumption does somewhat confound temporal variability at a site with the uncertainty of the mean load, we decided to take the most conservative (largest) estimate of the uncertainty in light of the low quality of the load estimates. To estimate the variances for the Error 1 dataset, we use the following equation from Ware and Lad (2003):

$$\delta_i^2 = \mathrm{Var}_{Ln(\mathrm{Conc}_i)} + \mathrm{Var}_{Ln(\mathrm{Flow}_i)} + 2\mathrm{Cov}_{Ln(\mathrm{Conc})Ln(\mathrm{Flow})} \tag{6}$$

where δ_i^2 refers to the variance of the measured nutrient loading, *Cov* refers to covariance, Conc refers to total phosphorus concentration and i refers to a subwatershed. $\mathrm{Var}_{Ln(\mathrm{Conc}_i)}$ was estimated as the variance of the log-transformed concentration measurements at each water quality monitoring station. $\mathrm{Var}_{Ln(\mathrm{Flow}_i)}$ was estimated by assuming a constant coefficient of variation (CV) for mean annual log-transformed flow equal to 0.023, i.e., the largest CV of the nine subwatersheds used to parameterize the discharge-area model. $\mathrm{Cov}_{Ln(\mathrm{Conc})Ln(\mathrm{Flow})}$ was estimated at each of the six stations with contemporaneous data using the log-transformed daily values of concentration and flow. The maximum of the six values (0.16) was used. The mean variance of the loadings in E1 dataset is 1.34 $(Ln(\text{tons year}^{-1}))^2$, with an interquartile range of 1.62–0.8 = 0.82 $(Ln(\text{tons year}^{-1}))^2$. In assessing the variability of log normally distributed data, Limpert *et al.* (2001) emphasized the importance of characterizing lognormal distributions with the use of multiplicative standard deviations and lognormal coefficients of variability. Expressed as percentages, our multiplicative standard deviations ranged from 130% to 380% with a median of 180%, indicating substantial uncertainty. Our coefficients of variability ranged from 0.96 to 3.3 with a median of 1.37 for the E1 dataset.

Variance estimates for the E2 dataset differed at the six stations where contemporaneous measurements were available. In keeping with our normality assumption, the width of the 95% confidence intervals of the log-transformed mean annual loading estimates, derived from the LOADEST program, was set equal to four standard deviations. Note that the loads estimated with the rating curve do not confound temporal variability with uncertainty at all, as it is the predictive interval of the mean predicted load which is used to derive the uncertainty estimates. Therefore, the Error 2 dataset places a much greater emphasis on the well-studied sites and achieves a reasonable balance between making use of all the information available and the need to emphasize the most reliable information. The mean variance of the loadings in Error 2 is 1.16 $(Ln(\text{tons year}^{-1}))^2$, with an interquartile range of 1.62–0.63 = 0.99 $(Ln(\text{tons year}^{-1}))^2$. To evaluate the impact of the data error in our analysis, we considered a third 'reference' statistical formulation founded upon the assumption that the Error 2 dataset represents error-free loading estimates (No Error or E0). In terms of multiplicative standard deviations and coefficients of variability, the six well-studied stations ranged from 17% to 106% with a median of 22%, and their coefficients of variability ranged from 0.13 to 0.86 with a median of 0.20. Overall, the E2 dataset has multiplicative standard deviations ranging from 17% to 380% with a median of 176%. Table II summarizes the loading datasets used in this paper, whereas the ESM contains a table that describes the calibration data and their uncertainty in detail (Table ESM-1).

Bayesian parameter estimation

Bayesian inference was used as a means for estimating model parameters because of its ability to include prior information in the modelling exercise and to explicitly handle model structural and parametric uncertainty (Gelman *et al*., 2004). Bayesian inference treats each parameter θ as a random variable and uses the likelihood function to express the relative plausibility of obtaining different values of this parameter when particular data have been observed:

$$\pi(\theta|\mathrm{data}) = \frac{\pi(\theta)L(\mathrm{data}|\theta)}{\int_\theta \pi(\theta)L(\mathrm{data}|\theta)d\theta} \tag{7}$$

where $\pi(\theta)$ represents our prior statements regarding the probability distribution that more objectively depicts the existing knowledge on the θ values, $L(\mathrm{data}|\theta)$ corresponds to the likelihood of observing the data given the different θ values and $\pi(\theta|\mathrm{data})$ is the posterior probability that expresses our updated beliefs on the θ values after the existing data from the system are considered. The denominator in Equation (7) is the expected value of the likelihood function and acts as a scaling constant that normalizes the integral of the area under the posterior probability distribution. Sequences of realizations from the model posterior distributions were obtained using Markov chain Monte Carlo (MCMC) simulations. Specifically, we used the general normal proposal Metropolis algorithm as implemented in the WinBUGS software (Lunn *et al*., 2000); this algorithm is based on a symmetric normal proposal distribution, whose standard deviation is adjusted over the first 4000 iterations such as the acceptance rate ranges between 20% and 40%. We collected between 50,000 and 60,000 samples from three chains for each model configuration. The first 10,000 samples were discarded, and posterior statistics were calculated using a thin of 10,

Table II. Measurement error associated with the total phosphorus loading data: Y_i refers to the log-transformed measured load, μ_i refers to the output of the SPARROW model, Load_i is a latent variable that represents the 'true' loading values when accounting for the measurement error δ_i, and σ represents the model error

Notation	Load estimation	Error sources	Model likelihood*	Total error**
Error 1 (E1)	Loads derived by multiplying average log flow by average log concentration.	Uncertainty in loads is the sum of the variances of the log concentrations and log flows plus two times their covariance.	$Y_i \sim N(\text{Load}_i, \delta_i^2)$ $\text{Load}_i \sim N(\mu_i, \sigma^2)$	$S_i = \sigma^2 + \delta_i^2$
Error 2 (E2)	Loads derived by rating curve methods where available and by multiplying average log flow by average log concentration elsewhere.	Uncertainty in loads is the variance derived from the 95% confidence interval of mean annual loading where available. Otherwise, the uncertainty is estimates as for error 1.	$Y_i \sim N(\text{Load}_i, \delta_i^2)$ $\text{Load}_i \sim N(\mu_i, \sigma^2)$	$S_i = \sigma^2 + \delta_i^2$
No Error (E0)	Loads derived by rating curve methods where available and by multiplying average log flow by average log concentration elsewhere.	Uncertainty in loads is ignored.	$Y_i \sim N(\mu_i, \sigma^2)$	$S = \sigma^2$

*The full mathematical notation of the models is provided in Equations (8)–(10).
**The correction factors τ^2 and $\tau^2 + \psi^2$are added with the STSP and CAR models, respectively.

which yielded a sample size of at least 4000 for all the model configurations considered.

Data quality submodel. Mathematically, the classical measurement error model consists of three components: (i) the (log-transformed) measurements Y_i, (ii) the (log-transformed) true values Load_i, (iii) and the measurement error ${\delta_i}^2$. These variables are included in a hierarchical framework, in which the first level defines the relation between the observed and the true loading values:

$$Y_i \sim N\left(\text{Load}_i, {\delta_i}^2\right) \tag{8}$$

Note that because we use log-transformed data, this statement postulates multiplicative measurement error. For this exercise, the values of ${\delta_i}^2$ are prespecified and are not part of the model calibration process. The second level of the hierarchy introduces a model for the 'true' log-transformed loads:

$$\text{Load}_i \sim N\left(\mu_i, \sigma^2\right). \tag{9}$$

Because the term μ_i refers to the SPARROW model prediction, this framework essentially postulates that the model is an unbiased estimator of the 'true' annual loads with structural (or process) error drawn from a normal distribution with variance σ^2. The likelihood of the loading estimate i is then the product of the likelihood of the two levels of our hierarchical configurations:

$$\begin{gathered} p(Y_i|\text{Load}_i) \times p(\text{Load}_i|\mu_i) = \\ \frac{1}{\sqrt{2\pi}\delta_i} \exp\left(-\frac{(Y_i - \text{Load}_i)^2}{2{\delta_i}^2}\right) \times \\ \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(\text{Load}_i - \mu_i)^2}{2\sigma^2}\right). \end{gathered} \tag{10}$$

To summarize, our calibration framework considers both the discrepancies between the measured and 'true' loading data as well as between the 'true' and modelled loading. To do so, we must estimate the 'true' loading as part of the model calibration. This adds an additional number of i stochastic nodes, thereby substantially increasing the complexity of the calibration exercise but realistically accommodating the measurement errors as well as the model process error.

Statistical formulations. In this study, we examined five statistical formulations comprising different combinations of model likelihood and prior parameter specification. The first three formulations were founded upon the MCMC model of Qian *et al*. (2005), which was merely a probabilistic expression of Equation (1), i.e., the log nutrient loading follows a normal distribution with a mean defined by the model and a constant model error variance. Contrary to the common practice in SPARROW modelling (Schwarz *et al*., 2006), our analysis uses the predicted (and not the observed) upstream loads to support predictions at downstream sites. Being equivalent to having the observed nutrient loading data on both sides of the nonlinear regression model, the conventional practice represents a conceptual deficiency of the SPARROW model and undermines its application for forecasting purposes. To address this problem, we introduce the data quality submodel as a means of separating the observation from the process error. By doing so, model-estimated nutrient loads are used as inputs to downstream subwatersheds while accounting for the uncertainty of the observed loads.

The notation of this approach is as follows:

$$\begin{gathered} Y_i \sim N\left(\text{Load}_i,\ {\delta_i}^2\right) \\ \text{Load}_i \sim N\left(\mu'_{(i)},\ \sigma^2\right) \\ \mu'_{(i)} = Ln\left(\sum_{n=1}^{N}\sum_{j=1}^{J_i} \beta_n S_{n,j} e^{(-\alpha Z_j)} H_{i,j}^S H_{i,j}^R\right) \\ 1/\sigma^2 \sim \text{gamma}(0.001, 0.001) \end{gathered} \tag{11}$$

where Y_i refers to the log-transformed measured load at site i, $\mu'_{(i)}$ refers to the corresponding corrected

SPARROW output, Load_i is a latent variable representing the 'true' loading values when we account for the prespecified measurement error δ_i, σ represents the model (or process) error and gamma(0.001, 0.001) is the gamma distribution with shape and scale parameters of 0.001, representing a 'noninformative' or vague prior assigned to the process error precision ($1/\sigma^2$). When we explicitly consider the data error, it is interesting to note the structural similarities between the state space (STSP) SPARROW of Qian *et al*. (2005) and our MCMC statistical characterizations. The main difference is that although the earlier work postulated a uniform precision of the log-transformed loading data throughout the watershed and then assigned a diffuse gamma prior, our analysis allots a great deal of effort to obtain informative site-specific observation error characterizations.

The normal distribution provided the basis for the likelihood function, which was then combined with three sets of priors for the first five parameters in Table I. Our first MCMC formulation, also denoted as MCMC-1 model, assumes conditional independence among the SPARROW parameters. The second statistical formulation, called MCMC-2, has one trivariate lognormal prior distribution for α, β_1 and β_2, in which the inverse of the covariance matrix (or the precision matrix) was assumed to follow a Wishart (R, r) distribution. To represent some prior knowledge on the parameters, we chose five degrees of freedom for this distribution ($r=5$), which is greater than the rank of the matrix (=3). The scale matrix R, an assessment of the order of magnitude of the covariance matrix, contains diagonal elements equal to the prior variances assigned to the three parameters with the MCMC-1 formulation, whereas the off-diagonal elements were set equal to zero. Loosely speaking, MCMC-2 approximately has the same starting point with MCMC-1 but does allow for covariance estimation among the three parameters as part of the updating process. The third statistical formulation, called MCMC-3, uses a trivariate lognormal distribution for α, $\beta_1$and β_2 with a fully specified covariance matrix. On the basis of the patterns of Qian *et al*. (2005), we assumed correlation coefficients between α, β_1 and β_2 uniformly distributed within the 0.7–0.9 range. The variances of α, β_1 and β_2 are the same as in the previous two formulations.

Qian *et al*. (2005) presented two statistical formulations, called STSP and continuous autoregressive (CAR), designed to accommodate the spatial autocorrelation of model residuals, as revealed by the original SPARROW application to the Neuse River Estuary watershed (McMahon *et al*., 2003). Although effective, both approaches entail more complexity, with STSP and CAR including one and two additional parameters, respectively. We seek to evaluate whether a system less intensively monitored than the Neuse River Estuary can support the more complex formulations of CAR and STSP, especially when we add an extra level of complexity by considering the data quality submodel (or site-specific observation error correction). The notation of the STSP approach is as follows:

$$\begin{gathered} Y_i \sim N\left(\text{Load}_i,\ \delta_i^2\right) \\ \text{Load}_i \sim N\left(\mu_i,\ \sigma^2\right) \\ \mu_i \sim N\left(\mu''_{(i)},\ \tau^2\right) \\ \mu''_{(i)} = Ln\left(\sum_{n=1}^{N}\sum_{j=1}^{J_i}\beta_n S_{n,j} e^{\left(-\alpha Z_j\right)} H_{i,j}^S H_{i,j}^R\right) \\ 1/\sigma^2, 1/\tau^2 \sim \text{gamma}(0.001, 0.001) \end{gathered} \tag{12}$$

Drawing parallels with the work by Qian *et al*. (2005), our STSP formulation considers the global process (structural) error σ, the prespecified site-specific observation error δ_i and the additional term τ that aims to correct for the residual structural and observation error that is not accounted for by the previous two error terms. The $\mu''_{(i)}$ term implies this additional correction of the total phosphorus loading entering the subwatershed i from upstream. In doing so, our intent is to address the adequacy of the first-order serial correlation postulated by the original study by Qian *et al*. (2005); that is, the serial correlation is addressed by a first-order model and what is left is mostly white noise. The addition of one more layer into the hierarchical model structure could raise issues of numerical instability, but our contention is that the extent of the instability problem is predominantly driven by the observation error assigned. In particular, the magnitude of the data precision determines the degree of confidence on the estimate of the 'true' load, which in turn represents the anchoring point along with the model structure for the estimation of the σ and τ error term and correction factor.

Building upon the STSP approach, the CAR model implements an additional spatial correction of the process error in two dimensions, whereby the random effect v_i for all subwatersheds is a correction factor modelled by a multivariate normal distribution with mean 0 and an unknown covariance matrix (Besag and Kooperberg, 1995). This correction accounts for possible inadequacies of the model structure, which result in spatially clustered model residuals, e.g., the use of a single export coefficient for all the agricultural land uses clearly overestimates the intensity of the agricultural practices in certain (neighbouring) subwatersheds and underestimates them in others (Qian *et al*., 2005). Thus, the CAR formulation is as follows:

$$\begin{gathered} Y_i \sim N\left(\text{Load}_i,\ \delta_i^2\right) \\ \text{Load}_i \sim N\left(\mu_i,\ \sigma^2\right) \\ \mu_i = \lambda_i + v_i \\ \lambda_i \sim N\left(\mu''_{(i)},\ \tau^2\right) \\ \text{v}_i | \text{v}_{-i} \sim N\left(\bar{v}_i, \psi_i^2\right) \\ \bar{v}_i = \frac{1}{n_i}\sum v_{-i} \quad \text{and} \quad \psi_i = \psi/n_i \\ 1/\sigma^2, 1/\tau^2, 1/\psi^2 \sim \text{gamma}(0.001, 0.001) \end{gathered} \tag{13}$$

where $-i$ denotes all the adjacent subwatersheds of i and ψ^2 is the conditional variance of the v_i terms, and its prior density was based on a conjugate inverse-gamma (0.001, 0.001) distribution. Equation (13) implies that the conditional mean of v_i to be the mean of the adjacent subwatershed random

effects and the conditional variance of v_i to be ψ^2 divided by the number of adjacent subwatersheds (n_i). The conditional distribution of each term v_i is determined by the neighbouring regions in the network. We here note that the v_i terms do not enter into the model likelihood function but simply act as additional parameters (Besag and Kooperberg, 1995). Table III summarizes all the model formulations examined in this study.

Prior specification. Wherever possible, we opted for informative, log normally distributed priors. The latter selection was partly due to the SPARROW parameterization of Qian *et al*. (2005), using total nitrogen loads from three large river basins in eastern North Carolina, which presented evidence that these parameters tend to be positively skewed (see their Figure 7). The median and standard deviation values assigned herein are provided in Table IV. In particular, we had no information regarding the dependence of the total phosphorus delivery to streams on runoff potential, so α was assigned a relatively flat prior. The values of the β coefficients represented literature-based estimates of total phosphorus export (Beaulac and Reckhow, 1982; Harmel *et al*., 2008). The upper limit found in both databases was specified as the 70th percentile of our distributions; thus, the corresponding priors were relatively wide, thereby allowing more of the information contained in the posterior distributions to come directly from the data. The distribution for k_r was drawn from a work by Cheng *et al*. (2010). We based the prior distribution for k_s, the stream attenuation coefficient, loosely on values from previous models; that is, we assigned a median of 0.04 along with a large standard deviation (McMahon *et al*., 2003; Alexander *et al*., 2004). To illustrate the importance of our informative priors in arriving at a reasonable model parameterization in our data-limited situation, we updated two versions of simplest Bayesian SPARROW formulation with uninformative (flat) priors (normal distributions centered at zero with variance equal to 10,000). The first approach is identical to the MCMC configuration of Qian *et al*. (2005) and uses measured upstream loads as point inputs to downstream subwatersheds (Schwarz *et al*., 2006). The second version uses modelled upstream loads as downstream inputs but does account for the loading data uncertainty with the E2 dataset.

Model assessment. We assessed the relative model performance using the Bayes factor. Originally intended to quantify the support for a scientific theory given a set of data, the Bayes factor can be used for any pairwise model comparison (Kass and Raftery, 1995). When we compare two alternative models, the Bayes factor is the posterior odds of one model over the other (assuming the prior probability on either model is 0.5). If M_A and M_B denote the two alternative models, the Bayes factor is as follows:

$$B_{AB} = \frac{\pi(Y|M_A)}{\pi(Y|M_B)} \tag{14}$$

For model comparison purposes, the model likelihood ($\pi(Y|M_k)$; $k = 1, 2$) is obtained by integrating over the parameter space:

Table III. Bayesian configurations of the SPAtially Referenced Regressions On Watershed attributes model examined

Model notation	Description
MCMC-1	All prior parameters are independent. Model residuals are not correlated.
MCMC-2	Priors for α and β parameters are based on a trivariate joint distribution. Assumes prior independence but allows for covariance estimation as part of the updating process. Model residuals are not correlated.
MCMC-3	Priors for α and β parameters are based on a trivariate joint distribution. The covariance structure is fully specified. Model residuals are not correlated.
CAR	Conditional autoregressive modelling of residuals. Model residuals are correlated in space in two dimensions. All prior parameters are independent.
STSP	State space modelling of residuals. Model residuals are correlated serially along a river network. All prior parameters are independent.

MCMC, Markov Chain Monte Carlo; CAR, continuous autoregressive; STSP, state space.

Table IV. Properties of the prior distributions for each parameter

Parameter	Median	Standard deviation	Source
α	1.0	22025	
β_1	0.1	3.51	Beaulac and Reckhow (1982)
β_2	0.07	1.25	Harmel *et al*. (2008)
k_r	12.84	4.76	Cheng *et al*. (2010) and reference therein
k_s	0.04	0.17	

All parameters are log normally distributed.

$$\pi(Y|M_k) = \int_{\theta} \pi(Y|M_k, \theta_k)\pi(\theta_k|M_k)\mathrm{d}\theta_k \qquad (15)$$

where θ_k is the parameter vector under model M_k and $\pi(\theta_k|M_k)$ is the prior density of θ_k. Using the MCMC method, we can estimate $\pi(Y|M_k)$ from posterior samples of θ_k. Letting $\theta_k^{(i)}$ be samples from the posterior density, the estimated $\pi(Y|M_k)$ is

$$\overline{\pi(Y|M_k)} = \left\{ \frac{1}{m} \sum_{i=1}^{m} \pi\left(Y|M_k, \theta_k^{(i)}\right)^{-1} \right\}^{-1} \qquad (16)$$

the harmonic mean of the likelihood values (Kass and Raftery, 1995). Although the value of this approximation converges to the value that would be obtained using analytic means, this convergence is unstable, as low outliers exert considerable influence on the value of the posterior odds (Kass and Raftery, 1995). To ensure that our results are not unduly influenced by outliers, we omitted the 20 samples of lowest likelihood from all odds calculations, corresponding to <0.5% of all the samples collected. We also note that the model performance was examined only within, and not among, data error specifications.

To illustrative importance of informative priors in situations with limited information, we evaluated the two models with flat priors together with the simplest formulation with informative priors (MCMC-1 E2, Equation (11)) using a set of frequentist and Bayesian metrics. The frequentist metrics were the root mean squared error (RMSE), calculated as follows:

$$\mathrm{RMSE} = \sqrt{\frac{\sum (Y_i - \mu_{\hat{i}})^2}{n}} \qquad (17)$$

and the weighted root mean squared error (WRMSE), calculated as follows:

$$\begin{aligned} \mathrm{WRMSE} &= \sqrt{\sum w_i \times (Y_i - \mu_i)^2} \\ w_i &= \frac{\lambda_i}{\sum_i \lambda_i} \\ \lambda_i &= \frac{1}{\delta_i^2} \end{aligned} \qquad (18)$$

i.e., the individual squared residuals were weighted by the precision (inverse of the variance) of the measured data. We also evaluated the models with two Bayesian metrics. The first was the posterior mean deviance, defined as the residual information in data Y conditional on a parameter vector θ and is calculated as $-2\log\{\mathrm{p}(Y|\theta)\}$ or $-2\log\{\text{likelihood}\}$. The second is the deviance information criterion (DIC), a Bayesian measure of parsimony, which rewards for model fit but penalizes model complexity (Spiegelhalter *et al.*, 2002). The DIC is defined as follows:

$$\mathrm{DIC} = \overline{D(\theta)} + p_D \qquad (19)$$

where $\overline{D(\theta)}$ refers to the posterior mean deviance and p_D is a measure of the effective number of model parameters. The effective number of parameters is calculated as the posterior mean deviance of the model ($\overline{D(\theta)}$) minus the estimate of the model deviance calculated when using the posterior means of the parameters ($D(\bar{\theta})$), which corresponds to the trace of the product of Fisher's information and the posterior covariance. A smaller DIC value indicates a more parsimonious and hence 'better' model.

Following Hong *et al.* (2005), we also evaluated the degree of updating between prior to posterior parameter distributions using three different criteria. First, we computed the difference between the most likely values of the prior and posterior distributions (referred to as median shift). We selected the median as the most likely value because it is less influenced by outliers than the mean, whereas the mode may not be representative of the majority of the posterior in cases of limited data availability, when identifiability issues are likely to arise. Second, we computed the difference in the width of the 95% credible intervals of the prior and posterior distributions (referred to as width shift). This comparison assesses the change in parameter uncertainty. Third, we evaluated the change in the shape of the distribution from prior to posterior using the delta index (Endres and Schindelin, 2003). The delta index measures the distance between two probability distributions:

$$\delta_{\theta i} = \sqrt{\int \left(\pi(\theta_i) \log \frac{2\pi(\theta_i)}{\pi(\theta_i) + \pi(\theta_i|Y)} + \pi(\theta_i|Y) \log \frac{2\pi(\theta_i|Y)}{\pi(\theta_i) + \pi(\theta_i|Y)} \right) \mathrm{d}\theta} \qquad (20)$$

where $\pi(\theta_i)$ and $\pi(\theta_i|Y)$ represent the marginal prior and posterior distributions of parameter θ_i, respectively. This metric is equal to zero if there is no difference between the two distributions and equal to $\sqrt{2\log 2}$ if there is no overlap between the two distributions. All delta index values are presented as percentages of this maximum value.

Post-hoc numerical experiments. We performed a post-hoc numerical experiment to help guide future sampling efforts in this relatively understudied watershed. Watershed monitoring is costly, and so it is desirable to estimate the benefits of collecting additional information on the model parameter identification and predictive uncertainty of subwatershed loads. Our intent was to evaluate the effect of both intensity of monitoring in space (i.e. number of stations) as well as intensity of monitoring at the individual sites (i.e. number of samples taken per site). First, we simulated a high precision dataset, which would exist if all the subwatersheds had a CV of their annual nutrient loading exports equal to that of the watershed

with the highest value among the six well-studied sites. As previously described, our dataset consists of six stations with contemporaneous measurements of flow and concentrations, which allowed the estimation of the mean annual load with the use of a rating curve. The error associated with the load calculation at the six stations was very similar but certainly significantly lower than the error characterizing the loading estimates at the rest of the sparsely studied 18 monitoring sites. In this regard, our numerical experiment examines how much we can learn from a SPARROW model parameterized from a dataset that comprises 24 (and not six) well-studied stations. That is, how much does the model uncertainty decline if we obtain rating curve loading estimates in every single site of the watershed and if we are able to characterize the loading with a precision (at least) equal to the lowest precision of the current well-studied sites? We used the latent variable Load_i resulting from the E2 scenario and the MCMC-1 model as our best estimate of the actual load, calculated the lognormal CV of the six well-studied sites and used the maximum CV to characterize the error associated with the rest 18 sparsely studied sites. We also used the Load_i data with the uncertainties of the E2 case as a reference dataset that reflects the precision of our current estimates. This exercise postulates that our estimates of Load_i are correct; yet, we note that even if these estimates are misleading, we will still gain a sense of how sensitive the model results are to data uncertainty. Our second experiment selected 12 of the original 24 stations and performed the same experiment with these 12 stations. We chose to omit most of the stations along the main stem of Grindstone Creek and instead concentrate on the headwater stations. We also consolidated two urban creek stations into one (Indian Creek). The ESM contains two tables (Tables ESM-2 and ESM-3) and two figures (Figures ESM-1 and ESM-2) detailing the current and higher precisions of the datasets used for this experiment as well as the locations of the 24 and 12 sites.

RESULTS

Effect of informative priors

Our results suggest that a reasonable model parameterization in our data-limited watershed can only be obtained by the consideration of informative prior parameter distributions (Table V). Parameter posteriors resulting from noninformative priors were very poorly identified, whereas the mean values of the export coefficients were unrealistic and exceeded plausible rates of total phosphorus application for intensive agriculture. Although highly uncertain, we also note that the use of noninformative priors assigned a lower export rate to urban land (β_1) than to agricultural land (β_2). Interestingly, the SPARROW model with informative priors resulted in a reversal of the relative magnitudes of the export coefficients as well as a distinct reduction of the significance of soil runoff potential in modulating total phosphorus export. When flat priors were used, the consideration of the modelled (instead of the observed) upstream nutrient inputs resulted in an increase of the delivery and export coefficients as well as a decrease of the model structural error (σ), deviance, and WRMSE, no change in the RMSE and an increase of the DIC. The decrease of the WRMSE indicates that including the data quality submodel can minimize the impact of uncertain data, whereas the DIC increase reflects that the estimation of the 'true' loading (Load_i) increases the complexity of the modelling exercise. It is also interesting to note that the inclusion of informative priors reduces the value of nearly all metrics of fit relative to the versions combined with flat priors.

Predicted total phosphorus loads

We first examined the correspondence between the predicted mean loads and the associated uncertainty for all combinations of data error characterizations and statistical formulations. The most likely values of predicted loads were highly correlated across statistical formulations within error configurations and slightly less correlated across error configurations. Correlations ranged from

Table V. Markov Chain Monte Carlo (MCMC) estimates of the stochastic nodes of the conventional SPARROW approach, where measured upstream loads are used to fit the model, against our MCMC-1 E2 formulation with noninformative and informative priors, respectively

Parameters	Conventional SPARROW		Noninformative priors		Informative priors	
	Mean	Standard deviation	Mean	Standard deviation	Mean	Standard deviation
α	6.53	3.37	8.63	4.09	0.46	0.76
β_1	2.95	3.69	6.76	7.28	0.19	0.14
β_2	10.83	13.16	16.22	15.96	0.10	0.09
k_r	17.27	19.57	17.37	19.69	13.03	4.25
k_s	0.19	0.07	0.17	0.12	0.05	0.04
σ	0.75	0.14	0.65	0.19	0.40	0.14
Deviance	56.59	5.48	52.57	5.05	47.19	4.25
DIC	46.74		64.06		55.07	
RMSE	0.67		0.68		0.60	
WRMSE	1.03		0.83		0.51	

SPARROW, SPAtially Referenced Regressions On Watershed attributes; DIC, deviance information criterion; RMSE, root mean squared error; WRMSE, weighted root mean squared error.

0.84 to 1.0, with a mean correlation coefficient of 0.96 (Table ESM-4). All correlations less than 0.90 pertained to either the CAR or STSP formulations using the error-free scenario. Similar to the most likely values, the standard deviations of the model predictions were highly correlated within and across data error specifications (Table ESM-5). The main exceptions were the standard deviations of the loads from the CAR and STSP models parameterized with the E0 approach, showing little relationship with the standard deviations of other model runs.

Plots of modelled against measured total phosphorus loads showed reasonable correspondence with the 1 : 1 line, with r^2 values ranging from 0.53 to 0.79 and slopes ranging from 0.68 to 0.83 (Figure 3). Stations with lower

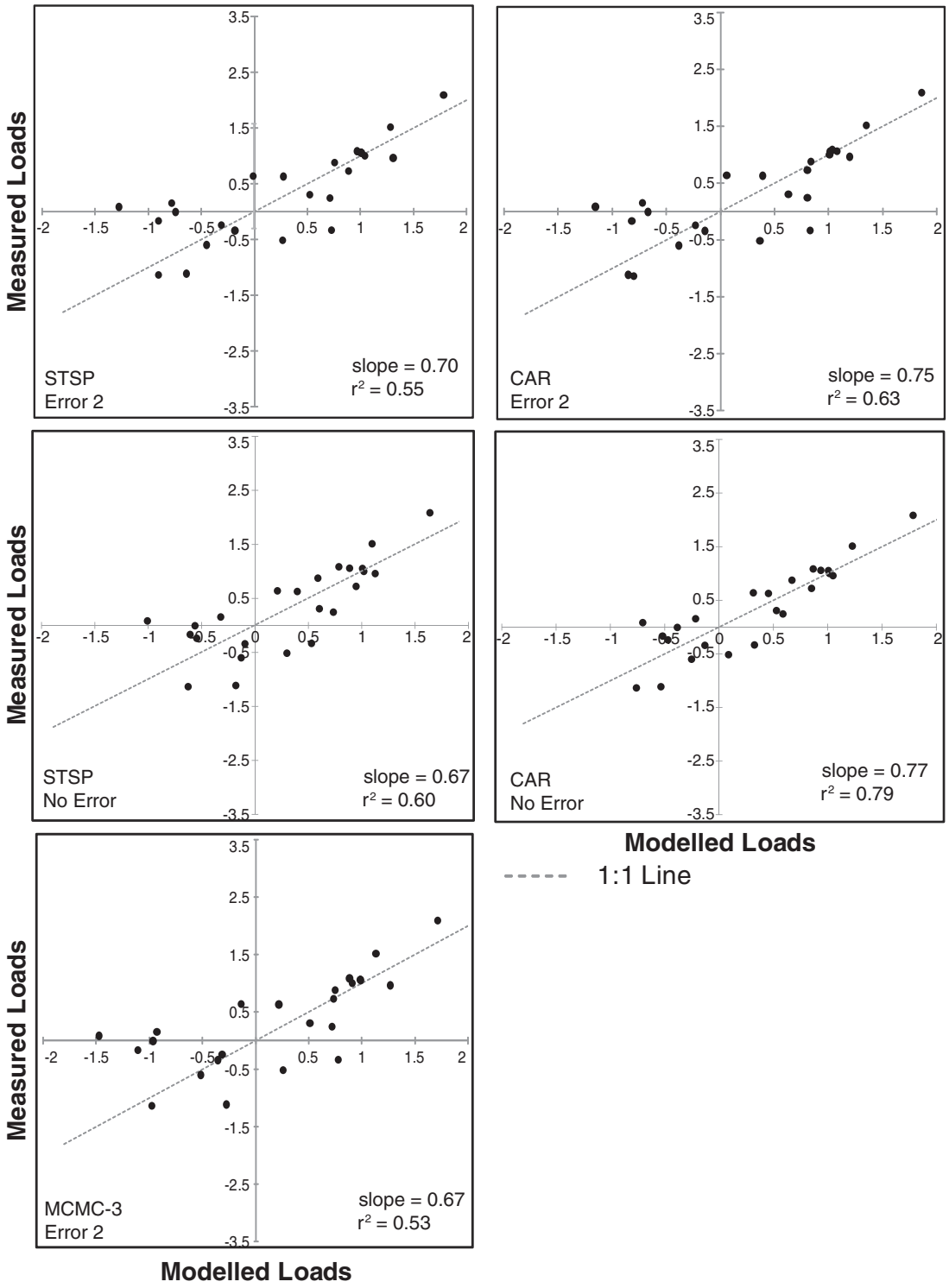


Figure 3. Measured *versus* modelled total phosphorus load (*Ln*(tons/year)) for STSP and CAR models parameterized with the E2 (top row) and the E0 data error characterizations (middle row), and MCMC-3 formulation calibrated with the E2 approach (bottom)

loading values tended to exhibit poorer model fit because of the greater uncertainties associated with the preponderance of sporadic water quality monitoring in the headwaters of Grindstone and Indian Creeks. Notably, the tight fit to the 1 : 1 line achieved by the STSP and CAR models when ignoring data uncertainty suggests that they produce better results when decoupled from the data quality submodels. That is, the STSP and CAR formulations with the explicit consideration of a prespecified site-specific data error term (δ_i) underperform relative to their counterparts that use the unconstrained global observation error term alone to capture the corresponding uncertainty.

Spatial patterns of modelled total phosphorus loads corresponded reasonably well to spatial patterns of measured loads (Figure 4). Areas of disagreement tended to coincide with areas of high measurement error. The magnitudes of model residuals displayed serial correlation, with high residuals being clustered in the headwaters of Grindstone and Indian Creeks (Figure 5). Yet, as the lower right panel of Figure 5 shows, this trend closely follows the spatial patterns of the data uncertainty. Interestingly, the residuals of the CAR and STSP statistical formulations demonstrate the same spatial structure with the residuals of the MCMC-3 formulation, a finding that is on par with the previous assertion that the predictive capacity of the CAR and STSP models is compromised when combined with the data quality submodel.

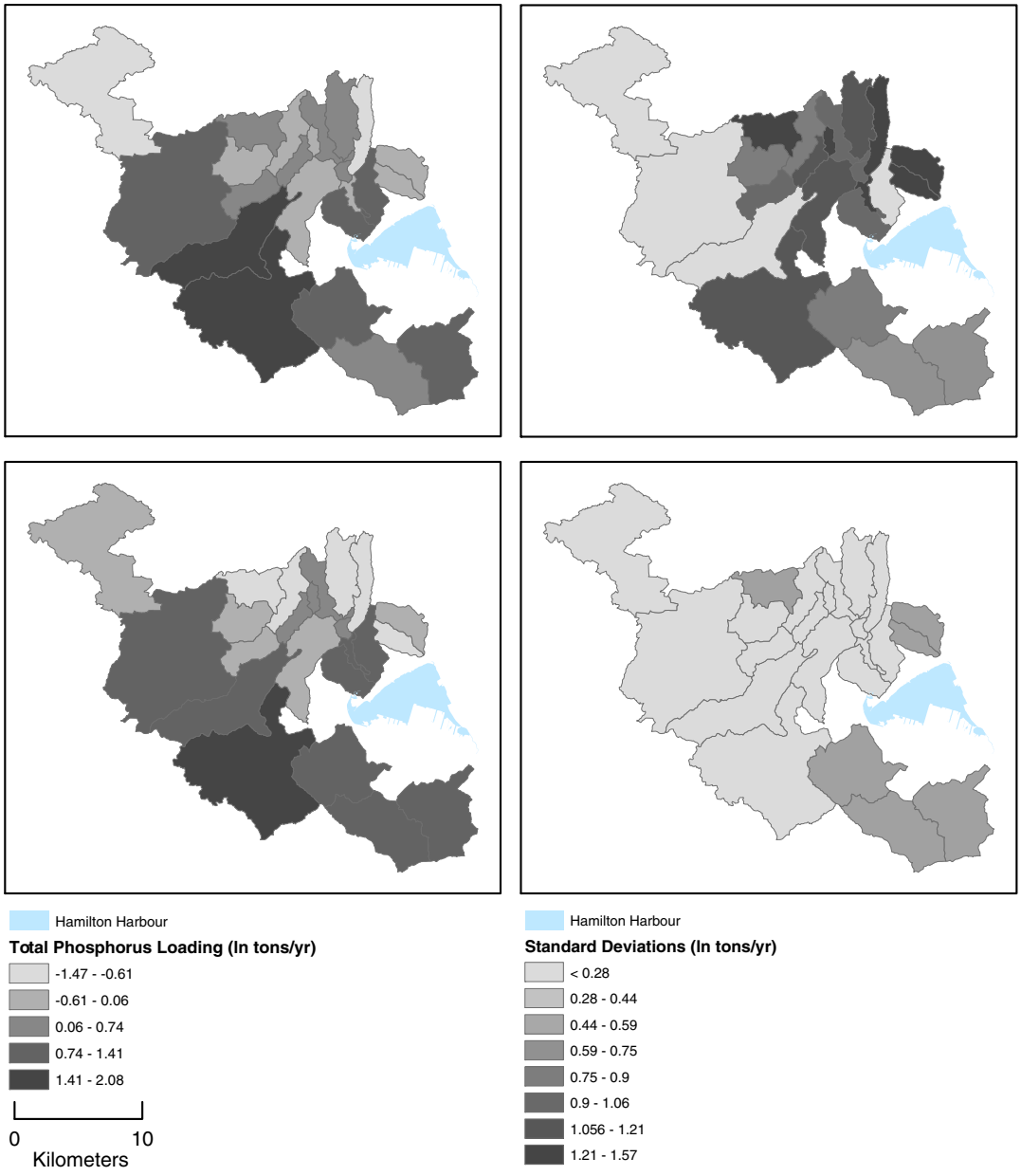


Figure 4. Measured total phosphorus loads (upper left) and associated uncertainties (upper right) against the modelled total phosphorus loads (lower left) and posterior uncertainties (lower right). Both measurements and model outputs are expressed in logarithmic scale and correspond to the E2 data error characterization and MCMC-3 configuration, respectively

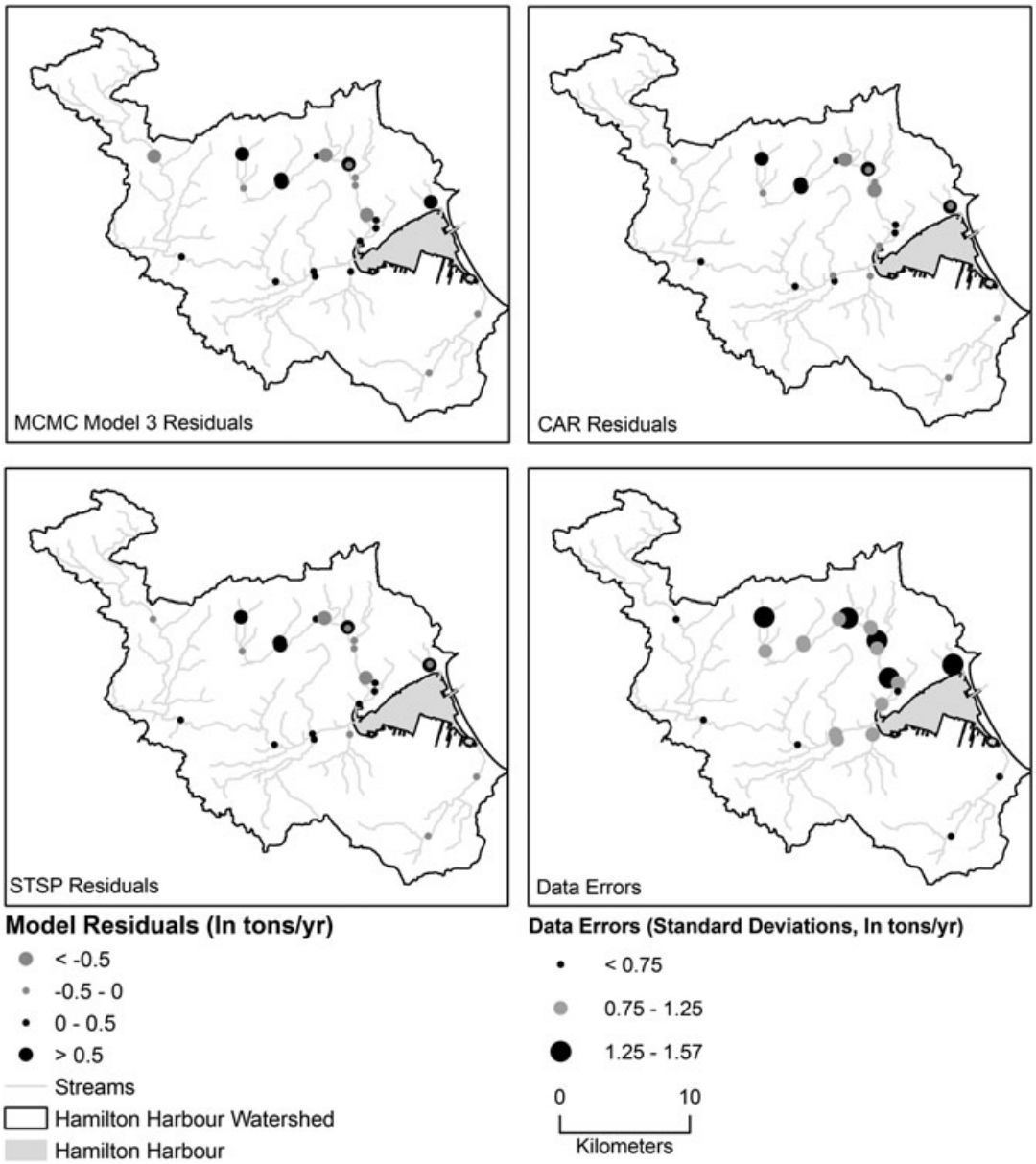


Figure 5. Model residuals for three models calibrated with the E2 data error characterization vis-à-vis the corresponding data error

Posterior parameter distributions

Table VI presents the mean and the standard deviation values for the parameter posteriors derived from the E1 case. The export coefficients of total phosphorus from urban (β_1) and agricultural (β_2) land uses are comparable and range between 0.13 and 0.17 tons P km^{-2} year^{-1}. Values of k_s indicate that between 5% and 7% of total phosphorus is attenuated per kilometre of stream length, which is a significantly lower estimate than what has been reported from previous SPARROW applications (Alexander *et al.*, 2002). Generally, we found serious identification problems when basing the parameter estimation of the five statistical formulations on the E1 dataset; especially with regard to the parameters α, β_1 and β_2. The problem was alleviated with the use of a trivariate lognormal distribution for the three parameters, i.e., MCMC-2 and MCMC-3 models. The same formulations provided somewhat lower values of k_s and distinctly higher α values relative to the rest of the models examined. In particular, the MCMC-2 model resulted in a quite high but fairly well-determined posterior estimate for the α coefficient.

Similar inferences can be drawn from the parameter posteriors obtained after the model update with the E2 approach (Table VII). Notably, aside from the CAR and STSP formulations, the process error σ is significantly higher than the corresponding values derived from the E1 scenario. Apparently, the lower measurement error assigned to the total phosphorus loadings of six stations constrains the search for the corresponding 'true' loading values, thereby exacerbating the influence of potentially erroneous observed loading estimates during the model updating. Notably, β_1 (urban export) is now higher than β_2 (agricultural export). Similar to the posteriors obtained from the E1 characterization, the MCMC-2 and MCMC-3 models produce higher values of α and lower values of k_s

Table VI. Markov Chain Monte Carlo (MCMC) estimates of the SPAtially Referenced Regressions On Watershed attributes model stochastic nodes using the E1 approach

Parameters	MCMC-1		MCMC-2		MCMC-3		CAR		STSP	
	Mean	Standard deviation	Mean	Standard deviation	Mean	Standard deviation	Mean	Standard deviation	Mean	Standard deviation
α	0.54	0.87	1.69	0.72	0.87	1.10	0.48	0.68	0.40	0.67
β_1	0.14	0.11	0.15	0.11	0.13	0.11	0.16	0.14	0.14	0.11
β_2	0.14	0.14	0.17	0.19	0.13	0.17	0.13	0.12	0.13	0.09
$s_{\alpha\beta1}$			0.05	5.73	4.01	1.36				
$s_{\alpha\beta2}$			0.20	3.74	4.02	1.38				
$s_{\beta1\beta2}$			0.10	2.41	1.89	0.49				
k_r	13.17	4.50	12.99	4.40	12.95	4.40	13.16	4.51	13.15	4.43
k_s	0.07	0.05	0.05	0.04	0.05	0.04	0.07	0.05	0.07	0.05
σ	0.16	0.14	0.17	0.15	0.16	0.14	0.18	0.15	0.17	0.15
τ							0.19	0.17	0.18	0.15
ψ							0.18	0.15		

CAR, continuous autoregressive; STSP, state space.

Table VII. Markov Chain Monte Carlo (MCMC) estimates of the SPAtially Referenced Regressions On Watershed attributes model stochastic nodes using the E2 approach

Parameters	MCMC-1		MCMC-2		MCMC-3		CAR		STSP	
	Mean	Standard deviation	Mean	Standard deviation	Mean	Standard deviation	Mean	Standard deviation	Mean	Standard deviation
α	0.46	0.77	1.63	0.65	0.75	0.85	0.60	0.99	0.45	0.74
β_1	0.19	0.14	0.20	0.13	0.16	0.11	0.24	0.30	0.20	0.15
β_2	0.10	0.10	0.13	0.10	0.10	0.08	0.12	0.17	0.10	0.09
$s_{\alpha\beta1}$			0.06	7.10	4.02	1.34				
$s_{\alpha\beta2}$			0.14	8.89	4.00	1.34				
$s_{\beta1\beta2}$			0.06	4.71	1.88	0.49				
k_r	12.99	4.30	12.81	4.20	12.89	4.20	13.16	4.40	13.00	4.33
k_s	0.05	0.04	0.03	0.03	0.04	0.03	0.05	0.04	0.05	0.04
σ	0.40	0.14	0.41	0.14	0.40	0.14	0.20	0.16	0.39	0.15
τ							0.16	0.14	0.18	0.16
ψ							0.43	0.25		

CAR, continuous autoregressive; STSP, state space.

relative to the rest of the formulations. Parameter identification was generally improved with the E2 approach, and this pattern was particularly evident with the MCMC-2 model.

Table VIII presents the parameter posteriors resulting from model calibration using the simpler E0 approach for the STSP and CAR formulations. As expected, the parameter standard deviations are generally lower than those obtained when we include the data quality submodel because of the greater amount of leverage exerted by the actual loading data on the corresponding posteriors. The model error variances (σ^2) are also higher than the posterior values obtained by the E2 characterization. Because of the absence of the intermediate latent variable Load_i, the mismatch between model predictions and measured data is entirely accounted for by the process error. We also note that the posterior means of β_1, β_2 and k_s are about the same or slightly higher than those obtained with the E1 and E2 data error specifications. Interestingly, as the uncertainty of the calibration data decreases, the STSP model error term and correction factor (σ and τ) become much better distinguished from one another. This supports our contention that any instability introduced by the addition of the data

Table VIII. Markov Chain Monte Carlo (MCMC) estimates of the SPAtially Referenced Regressions On Watershed attributes model stochastic nodes when ignoring the measurement error associated with the loading data

Parameters	CAR		STSP	
	Mean	Standard deviation	Mean	Standard deviation
α	0.37	0.54	0.39	0.62
β_1	0.23	0.13	0.22	0.10
β_2	0.17	0.10	0.19	0.11
k_r	12.51	4.28	12.39	3.99
k_s	0.10	0.05	0.11	0.04
σ	0.47	0.20	0.58	0.10
τ	0.20	0.16	0.18	0.14
ψ	0.41	0.38		

CAR, continuous autoregressive; STSP, state space.

quality submodel to the STSP configuration is mitigated through a higher degree of confidence in the data.

Evaluation of model performance and updating

Table IX presents the Bayes factor comparisons for all data error characterizations and statistical formulations. Each value represents the comparison of the model indicated in the column heading to the model shown in the rows. Following Kass and Raftery's (1995) interpretation, values between 1 and 3 are not worth more than a bare mention, values between 3 and 20 suggest positive evidence in favour of the model in the column heading, values between 20 and 150 show strong support and values greater than 150 show very strong support. Note that we only compare Bayes factors within each error configuration, as different error configurations have different assumptions inherent in their likelihood functions, and so their posterior odds are not directly comparable.

The consideration of the two data quality submodels did not result in strong support for any of the MCMC models but did provide overwhelming support in favour of all the MCMC models relative to CAR or STSP when the data quality submodel is included. STSP also had positive or strong support over CAR. Further, the MCMC-1 model slightly underperforms the two models that explicitly consider the covariance among α, β_1 and β_2. Interpreting these results, we infer that the single most important action for improving model performance in watersheds of limited information is the incorporation of an error term to accommodate the uncertainty associated with the existing datasets. Once this condition is met, the second interesting finding is that simpler statistical formulations can be more favourably supported by the data, even if the spatial correlation of the model residuals is not explicitly accounted for. In this regard, our results differ somewhat from those reported by Qian *et al.* (2005), in that neither the STSP nor the conditional autoregressive model compare well with the simpler MCMC model when we include the data quality submodel. Interestingly, when we omit the latter submodel, our analysis still does not provide strong support for the conditional autoregressive model relative to the findings of Qian *et al.* (2005) (see their Table 4). The study area is fairly small compared with other SPARROW applications, and thus, systematic spatial changes in topography, land cover and other model inputs are less profound. As a result, a model with common coefficients for all subwatersheds may be adequate. In other SPARROW applications, spatial patterns (e.g., agriculture intensity increases in the Neuse River watershed from west to east) in input variables often make a model with constant coefficients for all regions inadequate (McMahon *et al.*, 2003; Alexander *et al.*, 2004; Qian *et al.*, 2005). This probably explains why our CAR SPARROW application does not overwhelmingly outperform the other statistical models relative to what has been reported in earlier studies.

Table IX. Bayes factor comparisons for all simulations

	MCMC-1	MCMC-2	MCMC-3	CAR	STSP
			Error 1		
MCMC-1	1.00	2.04	1.61	0.00	0.10
MCMC-2	0.49	1.00	0.79	0.00	0.05
MCMC-3	0.62	1.27	1.00	0.00	0.06
CAR	90265	184358	145374	1.00	1139.70
STSP	9.68	19.77	15.59	0.00	1.00
			Error 2		
MCMC-1	1.00	0.65	1.01	0.01	0.21
MCMC-2	1.54	1.00	1.56	0.01	0.32
MCMC-3	0.99	0.64	1.00	0.01	0.20
CAR	17290	11252	17511	1.00	35.62
STSP	4.85	3.16	4.92	0.03	1.00
			No Error		
CAR				1.00	0.87
STSP				1.15	1.00

The Bayes factors are comparing models on top over models to the left. MCMC, Markov chain Monte Carlo; CAR, continuous autoregressive; STSP, state space.

Comparison of the priors with the posterior parameter distributions on the basis of the median shifts, width shifts and delta index is presented in Table X. Both the median shifts and the (nearly consistent) reduction in the width of the 95% credible intervals suggest substantial contribution of the dataset used during the model-updating process. The values of the three indices tended to increase as the uncertainty of the loading data decreased, reflecting greater leverage of the data on the posterior parameter distributions. The parameters α, β_1 and β_2 showed the largest median shifts, width shifts, and delta index values. By contrast, the parameters k_r and k_s occasionally showed large shifts, but generally their prior and posterior distributions were not drastically different. Finally, the significant reduction of the width of the credible intervals resulting from the MCMC-2 configuration (trivariate lognormal distribution for α, β_1 and β_2 with covariance matrix subject to updating) reiterate its efficiency to provide well-determined parameters, as demonstrated by a significant change of the posteriors in relation to the priors with regard to their central tendency and shape.

Identification of source areas and value of additional monitoring

The location of the major nutrient source areas is of great management interest in Hamilton Harbour. Priority subwatersheds for management intervention are those characterized by both a high total delivery of total phosphorus and a high delivery per area. For this modelling exercise, we used the posterior parameters from the MCMC-1 formulation on the basis of the E2 characterization in conjunction with combined sewer overflow (CSO – Hamilton Harbour Remedial Action Plan Technical Team, 2010) estimates to assess the total nonpoint source total phosphorus load to Hamilton Harbour. We found that the total load was 38.6 ± 6.7 tons of total phosphorus per year; 20.4 ± 1.25 tons of which originated from CSO events, leaving 18.2 ± 6.6 tons per year delivered by the local streams to the Harbour. The reported uncertainties are in units of one standard deviation and account for calibration data uncertainty, parametric uncertainty and model error (σ). It is interesting to note that even with loading data characterized

Table X. Comparison of the priors with the posterior parameter distributions in the different modelling experiments examined

Error 1

	MCMC-1			MCMC-2			MCMC-3			CAR			STSP		
	Median shift (%)	Width shift (%)	Delta index (%)	Median shift (%)	Width shift (%)	Delta index (%)	Median shift (%)	Width shift (%)	Delta index (%)	Median shift (%)	Width shift (%)	Delta index (%)	Median shift (%)	Width shift (%)	Delta index (%)
α	−83	−99	43	−62	−100	57	−55	−100	43	−81	−99	41	−80	−99	44
β_1	8	−90	45	21	−89	38	1	−82	40	21	−84	39	4	−89	37
β_2	54	−76	44	107	−71	46	29	−74	39	54	−75	27	48	−81	39
k_r	−4	−7	7	−5	−10	8	−4	−5	7	−1	−8	6	−2	−6	7
k_s	36	−57	25	−12	−66	25	−9	−67	25	32	−59	24	32	−59	25

Error 2

	MCMC-1			MCMC-2			MCMC-3			CAR			STSP		
	Median shift (%)	Width shift (%)	Delta index (%)	Median shift (%)	Width shift (%)	Delta index	Median shift (%)	Width shift (%)	Delta index (%)	Median shift (%)	Width shift (%)	Delta index (%)	Median shift (%)	Width shift (%)	Delta index (%)
α	−85	−99	44	−62	−100	57	−55	−100	45	−84	−100	47	−83	−99	42
β_1	62	−85	49	66	−84	45	36	−80	45	59	−83	44	66	−85	43
β_2	11	−86	38	63	−82	45	4	−83	46	21	−86	45	18	−80	42
k_r	−3	−12	8	−7	−10	9	−5	−14	9	−2	−7	7	−4	−10	10
k_s	−11	−67	22	−47	−82	35	−39	−78	30	−9	−66	22	−8	−64	21

No error

	CAR			STSP		
	Median shift (%)	Width shift (%)	Delta index (%)	Median shift (%)	Width shift (%)	Delta index (%)
α	−86	−100	46	−85	−100	46
β_1	90	−88	61	71	−91	67
β_2	112	−82	51	125	−82	57
k_r	−7	−13	13	−8	−15	14
k_s	128	−63	51	138	−64	54

MCMC, Markov chain Monte Carlo; CAR, continuous autoregressive; STSP, state space.

by uncertainties ranging from 17% to nearly 400%, it is possible to arrive at basin-wide predictions with a precision of about 36%. This is most likely because the well-studied sites tend to be on larger streams – about 66% of the basin area is drained by well-studied sites. In Figure 6a, we present the percentage of the total load of total phosphorus delivered to Hamilton Harbour originating from each subwatershed. The subwatersheds that are both large and close to Hamilton Harbour have the highest delivery values. Figure 6b shows the percentage of the total load of total phosphorus delivered to Hamilton Harbour normalized by the area of each subwatershed. The subwatersheds close to Hamilton Harbour have the highest delivery values per area, as the attenuation of their loads en route to the system is very low and the urban developments are more concentrated along the Harbour's shore.

To prioritize site-specific management interventions, it is desirable to estimate reach-level contributions. We estimate the precision with which a SPARROW model calculates reach-level contributions as the average width of the 95% credible intervals of all reach catchment loads to the closest monitoring station. Figure 7 shows how the uncertainties in reach-level contributions vary across statistical formulations and data error specifications. The E2 data error characterization does not consistently improve the precision of reach-level predictions relative to the E1 approach. On the other hand, the omission of the data quality submodel did result in more (but possibly misleadingly) precise predictions. It is also interesting to note that the explicit consideration of the α, β_1 and β_2 covariance resulted in more precise predictions relative to the model that assumes conditional independence (MCMC-1). The CAR and STSP formulations resulted in the most uncertain reach-level predictions, although we highlight their fairly similar average widths when updated with the 'error-free' approach and those derived from the MCMC-2 and MCMC-3 models when coupled with the data quality submodel.

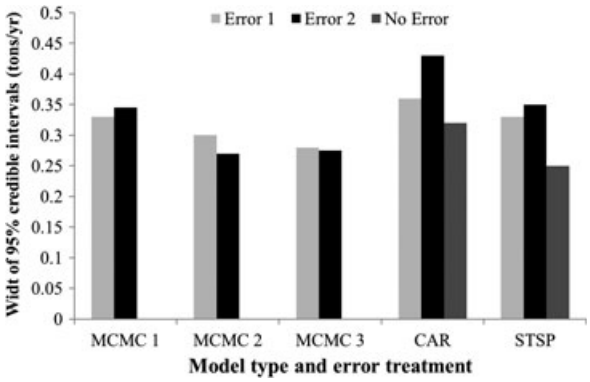


Figure 7. Average width of the posterior 95% credible intervals of reach delivery to the subsequent downstream station for each combination of statistical formulation and error treatment. Note that units are not expressed in the logarithmic scale

Finally, our post-hoc numerical experiment yielded some interesting suggestions regarding the expansion of water quality monitoring programs supporting SPARROW models. We first evaluated the degree of updating achieved by the scenarios of different data quantity and quality relative to the prior parameter distributions (Figure 8). The high precision dataset with 24 stations substantially increases the delta index values for β_1, β_2 and k_s, although there is little improvement in the widths of the 95% credible intervals of the parameter posterior distributions. Likewise, there were relatively minor differences between the prior and posterior medians derived by the two cases, with k_s and β_2 being the only notable exceptions. With only 12 stations, the improvement in parameter updating of all three metrics was lesser relative to the version of the model with all 24 stations. Further insights were gained by comparing the actual posterior statistics of the four datasets (Table XI). In particular, the high precision dataset reduced the standard deviations for all parameters, whereas only α and k_s exhibit noteworthy changes in their most likely values.

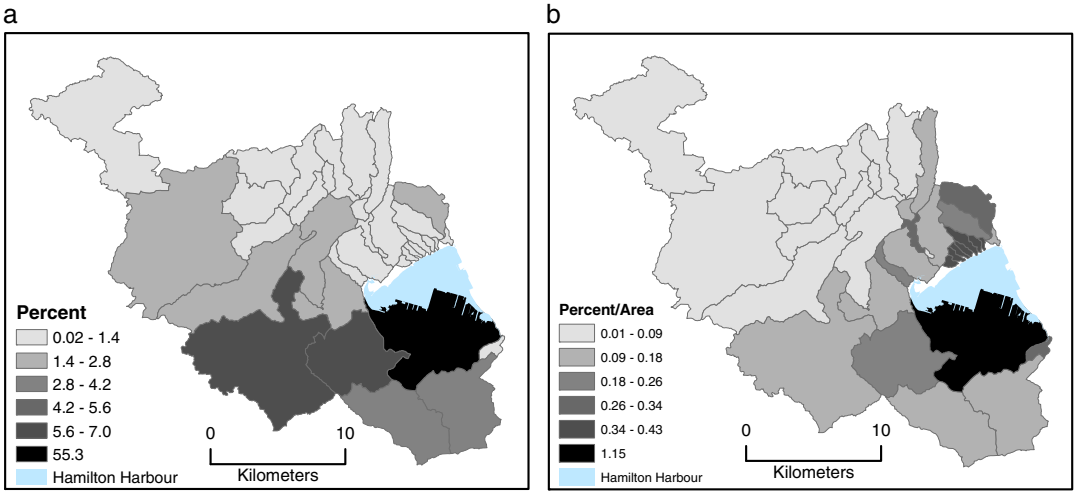


Figure 6. Estimated contribution of each subwatershed to the total phosphorus loading in Hamilton Harbour. The map on the left (a) expresses the load of each subwatershed as a percentage of the total phosphorus load, including the combined sewer overflows and taking into account attenuation en route to Hamilton Harbour. The map on the right (b) normalizes the percentage contribution by the corresponding subwatershed areas, presenting the delivered yield

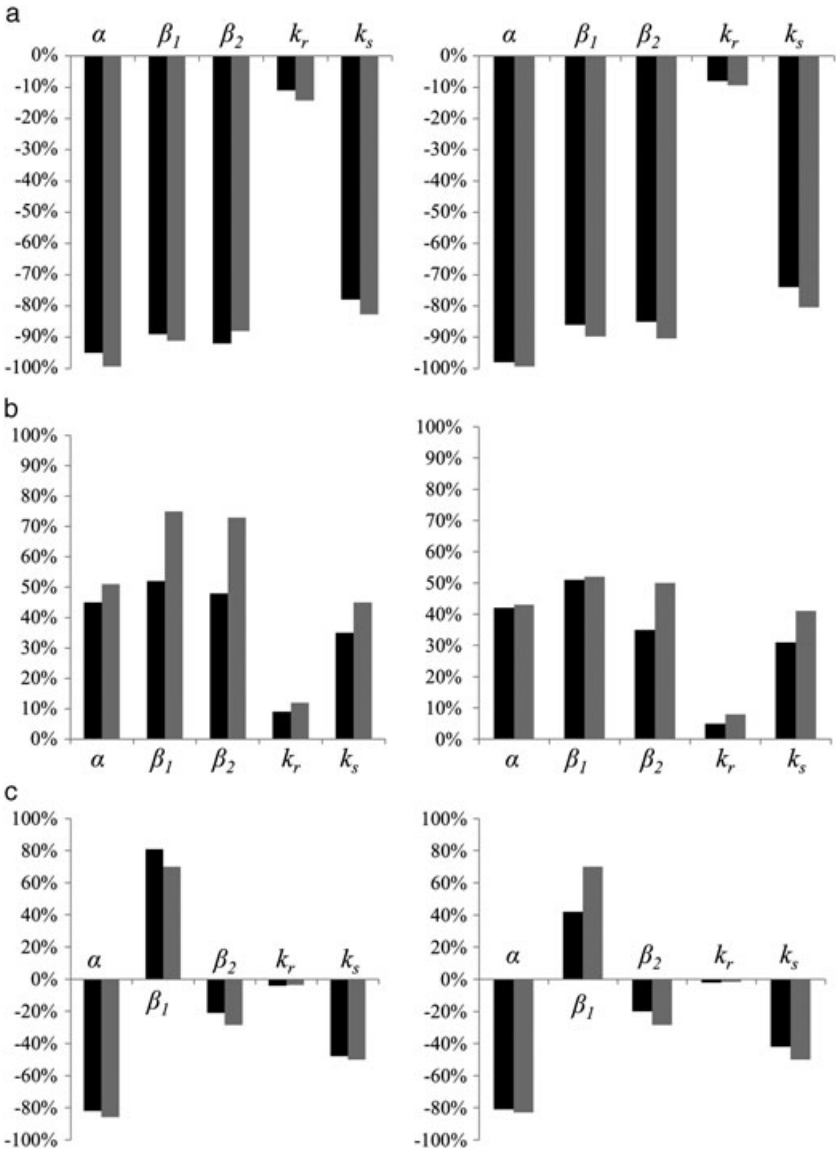


Figure 8. Value of information of additional monitoring in the Hamilton Harbour watershed. Black bars indicate current precision, whereas grey bars indicate the higher precision scenario. Left column represents sampling with all 24 stations. Right column represents sampling with a subset of 12 stations. Rows from top to bottom represent (a) the change in the width of the 95% credible interval of the informative prior and posterior parameter distributions; (b) the value of the delta index; and (c) the percentage difference between the median values of the prior and posterior parameter distributions

Table XI. Comparison of posterior parameter distributions for simulated data with the current and high precision

	Current precision, 24 stations		High precision, 24 stations		Current precision, 12 stations		High precision, 12 stations	
	Median	Standard deviation	Median	Standard deviation	Median	Standard deviation	Median	Standard deviation
α	0.14	0.76	0.13	0.50	0.16	0.87	0.17	0.86
β_1	0.17	0.12	0.16	0.08	0.17	0.14	0.17	0.12
β_2	0.05	0.06	0.05	0.03	0.05	0.08	0.05	0.06
k_r	12.54	4.25	12.32	4.13	12.59	4.45	12.63	4.39
k_s	0.02	0.03	0.02	0.02	0.02	0.03	0.02	0.03
σ	0.27	0.13	0.23	0.10	0.31	0.17	0.28	0.15

The latter result was not reflected in the prior-posterior comparison because the α and k_s priors were relatively unconstrained. When the scenario with 24 stations to that with only 12 is compared, it is clear that a greater number of points in space overwhelmingly improve parameter identification. Figure 9 presents the difference between the widths of the 95% credible intervals of the modelled log-transformed subwatershed loadings for the current

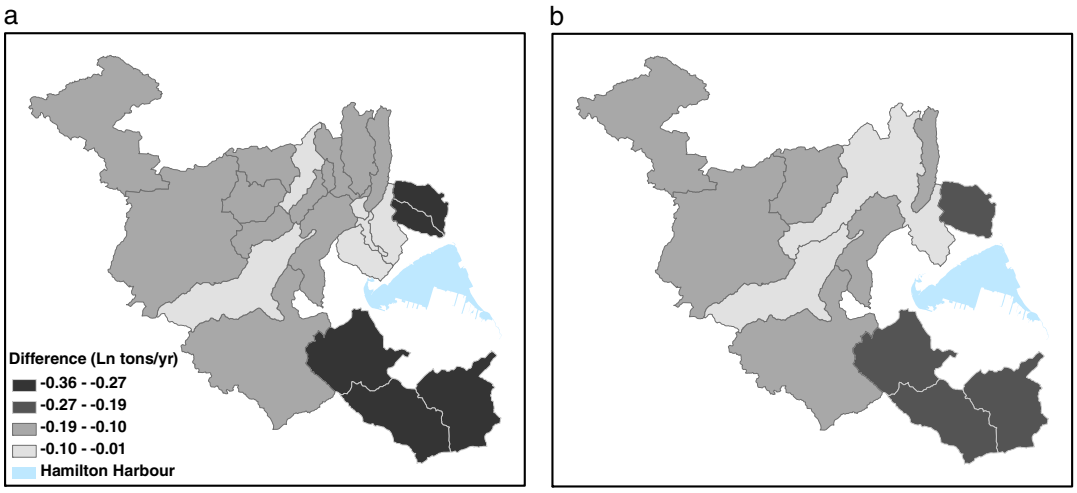


Figure 9. Value of information of additional monitoring in the Hamilton Harbour watershed. Maps show the difference between the width of the 95% credible intervals of the posterior loading estimates derived from the high and the current precision scenarios for (a) sampling with all 24 stations (right) and (b) sampling with a subset of 12 stations (left)

and high precision scenarios. Our analysis shows that an increase in the credibility of the measured loads significantly reduces the uncertainty of model predictions, even when the number of stations is halved. The '12-station' scenario did, however, moderate the benefits of the increased credibility of the observed loading data in some subwatersheds. The subwatersheds that likely to have the greatest reduction of 95% credible interval width under the '24-station' scenario are the same sites previously characterized by high absolute and areal load relative to the total nonpoint total phosphorus load entering the system.

DISCUSSION

Accounting for data uncertainty

The uncertainty in watershed modelling typically stems from errors associated with the measurements of input and response data, the parametric uncertainty, and the structural (or process) error arising from the inherent inability of a given modelling construct to reproduce the mechanisms involved in runoff generation or in biogeochemistry (Rode *et al*., 2010). Although most of the existing efforts in the literature delve into the analysis of parameter uncertainty (Beven and Freer, 2001), there is an increasing trend towards explicitly addressing the impact of other sources (e.g., Engeland and Gottschalk, 2002; Wagener and Gupta, 2005; Huard and Mailhot, 2006; Kuczera *et al*., 2006; Ajami *et al*., 2007; Liu and Gupta, 2007). In this study, one of our objectives involved the calculation of the mean annual loads and subsequently the characterization of the associated error in a watershed where limited information exists. Our Bayesian framework accounts for the data uncertainty using the classical measurement error model, i.e., the observed loading depends on a 'true' unknown value, which in turn is a random draw from a probability distribution determined by the predictive statements of the model and the associated process error (Carroll *et al*., 2006).

Measuring mean annual total phosphorus loads is not a trivial task, and so the data used to guide the parameter estimation of SPARROW models are typically *estimates* surrounded by substantial uncertainty (Cohn *et al*., 1989, 1992; Alexander *et al*., 2002, 2004; Moatar and Meybeck, 2005). Our database was likely characterized by a higher degree of uncertainty than those typically used to calibrate watershed models. However, given the lack of consideration of the calibration data uncertainty in the typical SPARROW practice, it is difficult to quantify the relative quality of our dataset. In their assessment of various annual load estimation techniques, Moatar and Meybeck (2005) report precisions (in terms of standard deviations) of between 10% and 20% for annual total phosphorus loads, depending on the method of load calculation. Of the six well-studied sites, four were within this range, whereas two (Redhill Creek) were substantially higher (around 100%). All of the sparsely studied sites were substantially higher than this level (between 150% and 400%). Yet, even with information of such quality, we were able to arrive at estimates of basin-wide nonpoint source loads with a precision of about 36%. We selected a conservative method to calculate the measurement error that disallowed the disaggregation of temporal variability from the uncertainty of the mean loads. This is defensible in a static model, such as SPARROW, where the uncertainty in model outputs can only implicitly depict the temporal variability of the system. Moatar and Meybeck (2005) found that the method employed here (our Equation (4)) resulted in the highest precision and lowest error among the methods tested with respect to total phosphorus. It is possible though that in other areas, particularly basins smaller than their case study of roughly 30,000 km^2, different methods would result in more accurate or precise estimates of annual loads. Future work could compare the results obtained herein with those obtained when using less conservative methods of estimating the uncertainty of the noncontemporaneous load estimates. For instance, the uncertainty of the

noncontemporaneous load estimates could be based on the standard errors of the sampled concentrations instead of their variances. Alternatively, daily flow estimates at ungauged sites could also be estimated using information from gauged sites and additional load and uncertainty calculation methods may then be employed (Preston *et al*., 1989; Moatar and Meybeck, 2005).

An interesting lesson learned from the consideration of data uncertainty was that it tends to inflate the uncertainty surrounding the parameter posteriors as well as the width of the predictive intervals, which in turn are a depiction of the total uncertainty of the modelling exercise, i.e., parametric, structural and data uncertainty. This increase of parametric and predictive uncertainty was a result of increasing the complexity of the modelling exercise with the data quality submodel – as we detailed in Bayesian Parameter Estimation (Methodology Section), the data quality submodel adds a number of parameters equal to the number of calibration data points in the model. We also note the somewhat counterintuitive increase of the process error when assigning lower measurement error to the total phosphorus loadings of six fairly well-monitored sites. Although this result may partly stem from the inefficient search for the corresponding 'true' loading values due to the stricter constraints imposed by the lower observation error, it may also highlight the likelihood of having several erroneous mean loading estimates in the current dataset used.

Evaluation of statistical formulations

The selection of the appropriate statistical formulation to describe the modelling problem at hand is always a critical decision and can significantly affect the posterior parameter distributions, model predictive capacity and error structure (Arhonditsis *et al*., 2008a,b). In particular, earlier work by Qian *et al*. (2005) argued that the conditional autoregressive and STSP models are effective strategies to improve SPARROW applications because of their ability to accommodate the spatial correlation of model residuals as well as to minimize the propagation of the observation error to downstream subwatersheds. Our study shows that the predominance of the CAR and STSP statistical formulations over the conventional approaches holds true only when the data uncertainty is omitted. When the CAR and STSP models are coupled with the measurement error model, the substantial complexity increase apparently becomes an impediment and tends to over-inflate the predictive uncertainty. We also pinpoint the better fit obtained by the STSP formulation without the data quality submodel vis-à-vis all the formulations that assume conditional independence of the model residuals combined with the measurement error model (e.g., see Figure 3). As both approaches are structurally identical, the former model contains two unconstrained global error term and correction factor (σ and τ) subject to updating, whereas the latter ones considers the global process error (σ) and prespecified site-specific observation errors (δ).

The advantages of accounting for the interdependencies among model parameters have been discussed in the modelling literature (Bates *et al*., 2003; Qian *et al*., 2005; Arhonditsis *et al*., 2008a,b). Yet, Hong *et al*. (2005) noted that the existing knowledge regarding the correlations among model parameters is usually insufficient to probabilistically express any prior assumptions. Further, the assumption of prior independence provides a type of robustness in the analysis, in that it allows exploring broader areas of the parameter space (Hong *et al*., 2005). When data availability is a limiting factor though, the representation of the prior parameter space by a wide hypercube, postulated by the use of several conditionally independent uniform priors, may not be adequate to elicit meaningful predictive statements from our models. Rather, we should use any prior knowledge from the literature on the relative plausibility of different values of the model parameters as well as their interdependencies, which then can be included into the 'prior–likelihood–posterior' update cycles and gradually converge towards more realistic (site-specific) values (Arhonditsis *et al*., 2007). After all, once informative data are collected from the system, the specific assumptions used in constructing the corresponding prior distributions for the different parameters will not matter. In this study, we found that the narrower hyperellipsoids implied by the MCMC-2 and MCMC-3 statistical formulations did improve parameter identification and model predictions at both the subwatershed and reach scale, although the advantages were contingent upon the type of covariance matrix specified as well as the quantification of the uncertainties of the loading data. In particular, we found that the more flexible structure of the MCMC-2 model offers fairly well-identified posteriors for α, β_1 and β_2, although the corresponding covariance estimates are poorly determined (Tables VI-VIII). On the other hand, the more rigid correlation pattern postulated by the MCMC-3 model does not appear to substantially alleviate the identification problem.

SPARROW application in watersheds of limited information and value of additional monitoring

Our analysis presented a framework for applying the SPARROW model in smaller, less intensively monitored watersheds. In any modelling framework, the results can be partly driven by the model structure selected. SPARROW's semi-distributed framework may be unable to properly resolve 'hotspot' areas where the co-occurrence of a number of favourable conditions results in disproportionately high nutrient export. Using more complex models typically requires more data, and this data may not be available for watersheds with limited information. SPARROW requires only basic information about land use, topography and nutrient loadings, and so it represents a sensible strategy when data availability is a limiting factor.

Model parameters are generally comparable to those presented by previous SPARROW models, with the exception of k_s (stream attenuation), which is generally

much higher than what we report here. This is likely due to our use of only one stream class, which assumes uniform attenuation across headwaters and large channels. Previous work has found this parameter to be much higher in smaller streams than larger ones, usually by an order of magnitude (McMahon *et al*., 2003; Alexander *et al*., 2004). Alexander *et al*. (2002) found the stream attenuation to be not significantly different from zero for large streams. It is possible that by representing small and large streams with the same coefficient, we may underestimate the attenuation in small and overestimate in large streams. Given the nature of our dataset though, it was not deemed appropriate to further inflate the uncertainty by explicitly estimating stream attenuation in both small and large streams.

The estimates of total phosphorus export in the majority of our modelling experiments suggest that urban land uses may export total phosphorus on a per area basis comparable to that of agricultural lands. This finding is somewhat contrary to the popular notion that rates the nutrient export of urban lands below that of agricultural lands due to lower nutrient subsidies (Moore *et al*., 2004; Soldat and Petrovic, 2008; Soldat *et al*., 2009). On the basis of our literature search, our prior parameter distributions assigned a slightly higher median value of total phosphorus export to urban lands than to agricultural lands, whereas the delta indices for β_1 and β_2 also showed significant influence of the dataset used on the corresponding posteriors. In particular, the six most intensively studied sites contained both agricultural and urban land covers, so it is unlikely that the higher β_1 value stems from the higher precision assigned to those sites during the model updating. It should also be noted that the same result was found ($\beta_1 \geq \beta_2$) even when identical priors were assigned to the two parameters. Further, other studies in the region of Southern Ontario have found urban total phosphorus export rates to be higher than agricultural total phosphorus export rates (Winter and Duthie, 2000). Although the water quality monitoring stations were selected to be upstream from any CSO or wastewater treatment plant effluent, there may be some connections of the sewer system to the creeks, possibly in the form of illegal connections between sanitary and storm sewers, leaky sewer pipes, or unaccounted for CSO outfalls. A more refined estimation of the export coefficients will likely require a higher quality database.

Our analysis showed the benefits of obtaining more precise estimates of total phosphorus loads within the basin, which, however, may not be possible as the resources to intensively monitor the entire watershed are not always available (Zhang and Arhonditsis, 2008). Thus, we also examined the tradeoff between increasing the spatial intensity of sampling and improving the precision of the load estimates and found that the parameter identification is predominantly driven by the appropriate sampling intensity in space. McMahon *et al*. (2003) introduced two criteria (or measures of impairment uncertainty) for allocating scarce monitoring resources and subsequently maximizing the knowledge gained. Namely, it was proposed that the additional water quality data-collection efforts should be focused on 'hot spots', characterized by either midrange likelihood of impairment (e.g., the probability of exceeding a water quality criterion was lying within the 25–75% range) or by model predictions of unacceptably high variance (McMahon *et al*., 2003). In this study, we propose two additional criteria: targeting locations where data uncertainty drives model residuals and locations where modelled loads showed the greatest reduction in the width of their 95% credible intervals when the hypothetical high precision dataset was used (Figure 9). In particular, we found that the headwater streams of Grindstone and Indian Creeks are subject to high model residuals and high load uncertainties. We also identified which subwatersheds displayed the greatest contraction in their 95% credible intervals when a hypothetical high precision dataset was used to parameterize the model. These subwatersheds include the ones at the headwater streams mentioned previously as well as those closest to the Harbour characterized by high delivery rates and urban land uses. Additional monitoring targeting the latter group of subwatersheds is underway and will considerably improve our model parameterization and load estimates.

We strongly concur with Pappenberger and Beven (2006), who called for uncertainty analysis to be an integral aspect of the environmental modelling practice. Empirical models like SPARROW are a useful scientific tool for avoiding the arbitrary selection of stringent (and often unattainable) threshold values for environmental variables (quality goals/standards) as a hedge against unknown forecast errors or risky management decisions that could result in the misallocation of the limited resources during the costly implementation of alternative environmental management schemes. Yet, the ubiquitous and often substantial uncertainty pervading any modelling exercise must be reduced or at least explicitly acknowledged and communicated in a straightforward way that can be easily used by decision makers/policy planners. In this study, the Bayesian framework illustrated the potential improvements when data uncertainty and spatial variability are accounted for as well as the ramifications that the complexity of each statistical problem description entails. On a final note, we believe that models are a worthwhile scientific activity, even when the limited knowledge from the system studied reduces their predictive ability. Rather than asking questions from a model that cannot be answered, we simply have to shift our focus on tasks that can be accomplished. After all, the development of a model is a dynamic process that is on par with the policy practice of adaptive management or 'learning while doing'. The initially uncertain model parameterization/structure can be sequentially refined as new knowledge is obtained from the system, and this gradual model evolution can provide the basis for revised (and improved) management actions.

ACKNOWLEDGEMENTS

This project has received funding support from the Ontario Ministry of the Environment (Canada Ontario Grant Agreement 120808). Such support does not indicate

endorsement by the Ministry of the contents of this material. Christopher Wellen has also received support from the Ontario Graduate Scholarships.

REFERENCES

Ajami NK, Duan QY, Sorooshian S. 2007. An integrated hydrologic Bayesian multimodel combination framework: confronting input, parameter, and model structural uncertainty in hydrologic prediction. *Water Resources Research* **43**: W01403. DOI: 10.1029/2005wr004745.

Alexander RB, Elliott AH, Shankar U, McBride GB. 2002. Estimating the sources and transport of nutrients in the Waikato River Basin, New Zealand. *Water Resources Research* **38**: 1268. DOI: 10.1029/2001wr000878.

Alexander RB, Smith RA, Schwarz GE. 2004. Estimates of diffuse total phosphorus sources in surface waters of the United States using a spatially referenced watershed model. *Water Science and Technology* **49**: 1–10.

Arhonditsis GB, Qian SS, Stow CA, Lamon EC, Reckhow KH. 2007. Eutrophication risk assessment using Bayesian calibration of process-based models: Application to a mesotrophic lake. *Ecological Modelling* **208**: 215–229.

Arhonditsis GB, Papantou D, Zhang W, Perhar G, Massos E, Shi M. 2008a. Bayesian calibration of mechanistic aquatic biogeochemical models and benefits for environmental management. *Journal of Marine Systems* **73**: 8–30. DOI: 10.1016/j.jmarsys.2007.07.004.

Arhonditsis GB, Perhar G, Zhang W, Massos E, Shi M, Das A. 2008b. Addressing equifinality and uncertainty in eutrophication models. *Water Resources Research* **44**: W01420. DOI: 10.1029/2007WR005862.

Arnold JG, Williams JR, Srinivasan R, King KW, Griggs RH. 1994. Soil and water assessment tool, USDA, Agricultural Research Service, Grassland, Soil and Water Research Laboratory, Temple, TX, 76502, USA.

Balin D, Hyosang L, Rode M. 2010. Is uncertain rainfall likely to greatly impact on distributed complex hydrological modeling. *Water Resources Research* **46**: W11520. DOI: 10.1029/2009WR007848.

Bates SC, Cullen A, Raftery AE. 2003. Bayesian uncertainty assessment in multicompartment deterministic simulation models for environmental risk assessment. *Environmetrics* **14**: 355–371.

Beaulac MN, Reckhow KH. 1982. An examination of land-use – nutrient export relationships. *Water Resources Bulletin* **18**: 1013–1024.

Besag J, Kooperberg CL. 1995. On conditional and intrinsic autoregressions. *Biometrika* **82**: 733–746.

Beven K, Freer J. 2001. Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology. *Journal of Hydrology* **249**: 11–29.

Borah DK, Bera M. 2003. Watershed-scale hydrologic and nonpoint-source pollution models: review of mathematical bases. *Transactions of the American Society of Agricultural Engineers* **46**: 1553–1566.

Borah DK, Bera M. 2004. Watershed-scale hydrologic and nonpoint-source pollution models: review of applications. *Transactions of the American Society of Agricultural Engineers* **47**: 789–803.

Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu CM. 2006. *Measurement Error in Nonlinear Models: A Modern Perspective*. Chapman & Hall/CRC Press: Boca Raton, FL; 455.

Cheng V, Arhonditsis GB, Brett MT. 2010. A revaluation of lake-total phosphorus loading models using a Bayesian hierarchical framework. *Ecological Research* **25**: 59–76. DOI: 10.1007/s11284-009-0630-5.

Cohn TA, Delong LL, Gilroy EJ, Hirsch RM, Wells DK. 1989. Estimating constituent loads. *Water Resources Research* **25**: 937–942.

Cohn TA, Caulder DL, Gilroy EJ, Zynjuk LD, Summers RM. 1992. The validity of a simple statistical-model for estimating fluvial constituent loads – an empirical study involving nutrient loads entering Chesapeake Bay. *Water Resources Research* **28**: 2353–2363.

Donigian AS, Bicknell BR, Imhoff JC. 1995. Hydrological simulation program – Fortran (HSPF). In *Computer Models of Watershed Hydrology*, Singh VP (ed.). Water Resources Publications: Highlands Ranch, CO; 395–442.

Endres DM, Schindelin JE. 2003. A new metric for probability distributions. *IEEE Transactions on Information Theory* **49**: 1858–1860. DOI: 10.1109/tit.2003.813506.

Engeland K, Gottschalk L. 2002. Bayesian estimation of parameters in a regional hydrological model. *Hydrology and Earth System Sciences* **6**: 883–898.

Gelman A, Carlin JB, Stern HS, Rubin DB. 2004. *Bayesian Data Analysis*. Chapman & Hall/CRC Press: Boca Raton, FL.

Gudimov A, Stremilov S, Ramin M, Arhonditsis GB. 2010. Eutrophication risk assessment in Hamilton Harbour: system analysis and evaluation of nutrient loading scenarios. *Journal of Great Lakes Research* **36**: 520–539. DOI: 10.1016/j.jglr.2010.04.001.

Hamilton Harbour Remedial Action Plan Technical Team. 2010. *Contaminant Loadings and Concentrations to Hamilton Harbour: 2003–2007 Update*. Hamilton Harbour Remedial Action Plan Office: Burlington, Ontario, Canada.

Harmel D, Qian S, Reckhow K, Casebolt P. 2008. The MANAGE database: nutrient load and site characteristic updates and runoff concentration data. *Journal of Environmental Quality* **37**: 2403–2406. DOI: 10.2134/jeq2008.0079.

Hiriart-Baer VP, Milne J, Charlton MN. 2009. Water quality trends in Hamilton Harbour: two decades of change in nutrients and chlorophyll a. *Journal of Great Lakes Research* **35**: 293–301. DOI: 10.1016/j.jglr.2008.12.007.

Hong BG, Strawderman RL, Swaney DP, Weinstein DA. 2005. Bayesian estimation of input parameters of a nitrogen cycle model applied to a forested reference watershed, Hubbard Brook Watershed Six. *Water Resources Research* **41**: W03007. DOI: 10.1029/2004wr003551.

Huard D, Mailhot A. 2006. A Bayesian perspective on input uncertainty in model calibration: application to hydrological model "abc". *Water Resources Research* **42**: W07416. DOI: 10.1029/2005WR004661.

Kass RE, Raftery AE. 1995. Bayes factors. *Journal of the American Statistical Association* **90**: 773–795.

Kavetski D, Kuczera G, Franks SW. 2006a. Bayesian analysis of input uncertainty in hydrological modeling: 1. Theory. *Water Resources Research* **42**: W03407. DOI: 10.1029/2005WR004368.

Kavetski D, Kuczera G, Franks SW. 2006b. Bayesian analysis of input uncertainty in hydrological modeling: 2. Application. *Water Resources Research* **42**: W03408. DOI: 10.1029/2005WR004376.

Kuczera G, Parent E. 1998. Monte Carlo assessment of parameter uncertainty in conceptual catchment models: the Metropolis algorithm. *Journal of Hydrology* **211**: 69–85.

Kuczera G, Kavetski D, Franks S, Thyer M. 2006. Towards a Bayesian total error analysis of conceptual rainfall-runoff models: Characterising model error using storm-dependent parameters. *Journal of Hydrology* **331**: 161–177.

Limpert E, Stahel WA, Abbt M. 2001. Log-normal distributions across the sciences: keys and clues. *BioScience* **51**(5): 341–352.

Liu Y, Gupta HV. 2007. Uncertainty in hydrologic modeling: toward an integrated data assimilation framework. *Water Resources Research* **43**: W07401. DOI: 10.1029/2006wr005756.

Lunn DJ, Thomas A, Best N, Spiegelhalter D. 2000. WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing* **10**: 325–337. DOI: 10.1023/A:1008929526011.

McMahon G, Alexander RB, Qian S. 2003. Support of total maximum daily load programs using spatially referenced regression models. *Journal of Water Resources Planning and Management* **129**: 315–329.

Moatar F, Meybeck M. 2005. Compared performances of different algorithms for estimating annual nutrient loads discharged by the eutrophic River Loire. *Hydrological Processes* **19**: 429–444. DOI: 10.1002/hyp.5541.

Moore RB, Johnston CM, Robinson KW, Deacon JR. 2004. *Estimation of Total Nitrogen and Total phosphorus in New England Streams using Spatially Referenced Regression Models*. United States Geological Survey: Pembroke, New Hampshire.

Omlin M, Reichert P. 1999. A comparison of techniques for the estimation of model prediction uncertainty. *Ecological Modelling* **115**: 45–59.

Ontario Ministry of Agriculture and Food. 2005. Soils of Southern Ontario. http://www.omafra.gov.on.ca/ (February 20, 2012).

Ontario Ministry of Natural Resources. 2005. Greater Toronto area digital elevation model. http://lioapp.lrc.gov.on.ca/ (July 20, 2012).

Ontario Ministry of Natural Resources. 2008. Southern Ontario Land Resource Information System (2000–2002) (SOLRIS). http://lioapp.lrc.gov.on.ca/ (July 20, 2012).

Ontario Ministry of the Environment. 2010. Provincial Water Quality Monitoring Network. http://www.ene.gov.on.ca (April 4, 2011).

Pappenberger F, Beven KJ. 2006. Ignorance is bliss: or seven reasons not to use uncertainty analysis. *Water Resources Research* **42**: W05302. DOI: 10.1029/2005wr004820.

Preston S, Bierman V, Silliman S. 1989. An evaluation of methods for the estimation of tributary mass loads. *Water Resources Research* **25**(6): 1379–1389.

Qian SS, Stow CA, Borsuk ME. 2003. On Monte Carlo methods for Bayesian inference. *Ecological Modelling* **159**: 269–277.

Qian SS, Reckhow KH, Zhai J, McMahon G. 2005. Nonlinear regression modeling of nutrient loads in streams: a Bayesian approach. *Water Resources Research* **41**: W07012. DOI: 10.1029/2005wr003986.

Ramin M, Stremilov S, Labencki T, Gudimov A, Boyd D, Arhonditsis GB. 2011. Integration of numerical modeling and Bayesian analysis for setting water quality criteria in Hamilton Harbour, Ontario, Canada.

Environmental Modelling & Software **26**: 337–353. DOI: 10.1016/j.envsoft.2010.08.006.

Rode M, Arhonditsis G, Balin D, Kebede T, Krysanova V, van Griensven A, van der Zee SEATM. 2010. New challenges in integrated water quality modelling. *Hydrological Processes* **24**: 3447–3461. DOI: 10.1002/hyp.7766.

Runkel RL, Crawford CG, Cohn TA. 2004. *Load Estimator (LOADEST): A FORTRAN Program for Estimating Constituent Loads in Streams and Rivers*. United States Geological Survey: Reston, Virginia.

Schindler DW. 2006. Recent advances in the understanding and management of eutrophication. *Limnology and Oceanography* **51**: 356–363.

Schwarz GE, Hoos AB, Alexander RB, Smith RA. 2006. *The SPARROW Surface Water-Quality Model – Theory, Applications and User Documentation*. U.S. Geological Survey: Reston, Virginia.

Soldat DJ, Petrovic AM. 2008. The fate and transport of total phosphorus in turfgrass ecosystems. *Crop Science* **48**: 2051–2065. DOI: 10.2135/cropsci2008.03.0134.

Soldat DJ, Petrovic AM, Ketterings QM. 2009. Effect of soil total phosphorus levels on total phosphorus runoff concentrations from turfgrass. *Water, Air, and Soil Pollution* **199**: 33–44.

Spiegelhalter DJ, Best NG, Carlin BR, van der Linde A. 2002. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society Series B – Statistical Methodology* **64**: 83–616. DOI: 10.1111/1467-9868.00353.

Vrugt JA, Diks CGH, Gupta HV, Bouten W, Verstraten JM. 2005. Improved treatment of uncertainty in hydrologic modeling: combining the strengths of global optimization and data assimilation. *Water Resources Research* **41**: W01017. DOI: 10.1029/2004WR003059.

Vrugt JA, ter Braak CJF, Clark MP, Hyman JM, Robinson BA. 2008. Treatment of input uncertainty in hydrologic modeling: doing hydrology backward with Markov chain Monte Carlo simulation. *Water Resources Research* **44**: W00b09. DOI: 10.1029/2007wr006720.

Wagener T, Gupta HV. 2005. Model identification for hydrological forecasting under uncertainty. *Stochastic Environmental Research and Risk Assessment* **19**: 378–387. DOI: 10.1007/s00477-005-0006-5.

Ware R, Lad F. 2003. Approximating the Distribution for Sums of Products of Normal Variables. Research Report UCDMS2003/15, Department of Mathematics and Statistics, University of Canterbury, Christchurch, NZ.. http://www.math.canterbury.ac.nz/php/research/listing/ (December 7, 2012).

Wellen C, Arhonditsis GB, Labencki T, Boyd D. 2012. A Bayesian methodological framework for accommodating interannual variability of nutrient loading with the SPARROW model. *Water Resources Research* W10505. DOI: 10.1029/2012WR011821.

Winter JG, Duthie HC. 2000. Export coefficient modeling to assess total phosphorus loading in an urban watershed. *Journal of the American Water Resources Association* **36**: 1053–1061. DOI: 10.1111/j.1752-1688.2000.tb05709.x.

Yang J, Reichert P, Abbaspour KC, Yang H. 2007. Hydrological modelling of the chaohe basin in china: statistical model formulation and Bayesian inference. *Journal of Hydrology* **340**: 167–182. DOI: 10.1016/j.jhydrol.2007.03.006.

Yang J, Reichert P, Abbaspour KC, Xia J, Yang H. 2008. Comparing uncertainty analysis techniques for a SWAT application to the Chaohe Basin in China. *Journal of Hydrology* **358**: 1–23. DOI: 10.1016/j.jhydrol.2008.05.012.

Zhang W, Arhonditsis GB. 2008. Predicting the frequency of water quality standard violations using Bayesian calibration of eutrophication models. *Journal of Great Lakes Research* **34**: 698–720.

APPLICATION OF THE SPARROW MODEL IN WATERSHEDS WITH LIMITED INFORMATION: A BAYESIAN ASSESSMENT OF THE MODEL UNCERTAINTY AND THE VALUE OF ADDITIONAL MONITORING

(ELECTRONIC SUPPLEMENTARY MATERIAL)

Christopher Wellen* **and George B. Arhonditsis**

Ecological Modelling Laboratory,

Department of Physical & Environmental Sciences, University of Toronto,

Toronto, Ontario, Canada, M1C 1A4

Tanya Labencki and Duncan Boyd

Great Lakes Unit, Water Monitoring & Reporting Section, Environmental Monitoring and Reporting Branch, Ontario Ministry of the Environment

Toronto, Ontario, Canada, M9P 3V6

* Corresponding author

e-mail: christopher.wellen@utoronto.ca, Tel.: +1 416 208 4878; Fax: +1 416 287 7279.

1. Spatial Data Sets

1.1 Topography: The delineation of subwatersheds and reach catchments is done using a digital elevation model (DEM). A stream corrected 10 meter cell size DEM generated through the application of photogrammetric methods was used for this purpose. Water quality monitoring stations were used as the discharge point for subwatersheds. More information about these stations is given in Section 2.3.5. Our calibration dataset had 24 subwatersheds. Their areas ranged from 0.3 – 75.8 km^2, with a mean of 17.9 km^2 and an interquartile range of 25.7-6.8 = 18.9 km^2. The Water Survey of Canada maintains a series of stream gauging stations which were used to develop a discharge-area (DA) model for the basin (Viessman and Lewis, 2002; Water Survey of Canada, 2011)). Flows from years 1988 – 1990, 1992 – 2002 were used to estimate the mean total annual flows. These years were chosen because they contained a complete flow record for all stations. The DA model related the natural logarithm of mean total yearly flow (m^3yr^{-1}), *Ln(Flow)*, to the subwatershed area (km^2), *Area*, with the following equation:

$$Ln(Flow) = 0.0125Area + 15.95 \ (r^2 = 0.91, n = 9) \qquad \textbf{(ESM-1)}$$

1.2 Streams, Lakes, and Reservoirs: Geographic Information System (GIS) layer files for streams, lakes, and reservoirs were obtained in two layers digitized from Natural Resource Canada's National Topographic System of maps (Natural Resources Canada, 2011). The 1:50,000 scale map series was used as the source of the dataset. The U.S. EPA River Reach File 1, used by many previous SPARROW applications (McMahon et al., 2003; Alexander et al., 2004), is digitized from source data with a scale of 1:500,000, a much coarser scale. We chose not to include all the stream segments in this SPARROW model to avoid a proliferation of miniscule reach catchments. We imposed a minimum reach catchment area of 10,000 ha as well as a minimum stream reach length of 750 m for consideration in the model. While the National Hydrographic Dataset of the United States (NHD) generally contains reaches greater than 1 mile

(1600 m) in length, we opted for 750 m due to the finer scale of our study (United States Geological Survey, 2000). The final stream layer has a mean length of 2.4 km and an interquartile range of 3.2-1.2=2.0 km. There are a total of 175 reach catchments, and each reach catchment discharges into a confluence, reservoir, or water quality monitoring station. Reach catchment areas ranged from 0.02 – 19.3 km^2, with a mean of 2.5 km^2 and an interquartile range of 3.5-0.9=2.6 km^2. We imposed two criteria that a reservoir had to fulfill in order to be included in the model. First, it had to have a minimum area of 4.05 ha, the threshold for inclusion in the NHD (United States Geological Survey, 2000). Second, it had to drain an area of at least 50,000 ha. This was roughly the x-intercept of the discharge-area model, when developed with data expressed in the original scale, and represented its application domain. Aerial hydraulic loads were calculated as the ratio of the mean total yearly flow to the reservoir area. Four reservoirs were used during the parameter estimation of the SPARROW model (Figure 1).

1.3 Nutrient Sources: Both point and non-point nutrient sources were included in the SPARROW model of Hamilton Harbour. Two point sources were considered, only one of which was used for model parameter estimation: (*i*) The Waterdown Waste Water Treatment Plant (WWTP), a small water treatment plant discharged into a tributary of Grindstone Creek during the study period, though it was taken offline in 2008. The mean loading for this plant between 1996 and 2007 was 0.3 tons of phosphorus per year, with an interquartile range of 0.4-0.2=0.2 tons per year (Hamilton Harbour Remedial Action Plan Technical Team, 2010). (*ii*) A second point source was the set of combined sewer overflows (CSOs) which result from Hamilton's combined sewer system. While holding tanks at several CSOs have been constructed, overflows can still occur when the tank capacity is reached, and CSOs without holding tanks result in more frequent overflows. Some of these combined sewers overflow directly to Hamilton Harbour,

while others flow into Cootes Paradise, a large wetland which drains directly into Hamilton Harbour. Phosphorus loads to Cootes Paradise from CSOs averaged 1.5 tons per year between 1996 and 2007, with an interquartile range of 2.1-0.6=1.5 tons per year (Hamilton Harbour Remedial Action Plan Technical Team, 2010). Phosphorus loads from CSOs directly to Hamilton Harbour averaged 21.6 tons per year between 1996 and 2007, with an interquartile range of 25.4-16=9.4 tons per year (Hamilton Harbour Remedial Action Plan Technical Team, 2010). While there is a CSO outfall upstream of the most downstream monitoring station of Redhill Creek, no information was available regarding the CSO loadings there, and so these loadings were accounted for implicitly by the model parameterization. The CSO loadings were used to estimate the total basin phosphorus load to Hamilton Harbour, but were not part of the model parameterization.

The non-point sources of total phosphorus were limited to agricultural and urban land, which included urban green space. These two land cover types were chosen because together account for 80% of the current land use of the basin; they are most likely to change in extent in the near future; and they have been found to be by far the greatest sources of phosphorus to receiving waters at the landscape scale (Beaulac and Reckhow, 1982; Alexander et al., 2004). Land uses were derived from a supervised classification of satellite imagery from 2002; Southern Ontario Land Resource Information System (Ontario Ministry of Natural Resources, 2008). Total agricultural and urban areas were estimated for each reach using GIS overlay analysis.

1.4 Landscape Characteristics: Landscape characteristics can influence the delivery of phosphorus to stream edges. Poorly drained soils have been found to deliver greater amounts of phosphorus to streams than well drained areas, as poorly drained soils tend to have a high exchange capacity and overland flow/erosion and the installation of tile drainage systems is more

common on poorly drained soils (Beaulac and Reckhow, 1982). Previous SPARROW estimations of total phosphorus have found significant delivery effects from soil permeability (Alexander et al., 2004), and therefore we assumed that the delivery of phosphorus to streams is primarily controlled by soil runoff potential, parameterized as a function of the soil hydrologic runoff group. Following McMahon et al. (2003), we calculated an area-weighted average of soil hydrologic runoff group for each reach catchment. We assigned values of 1 through 4 to soil groups A, the most well-drained group, through D, the most poorly drained group. We then took the reciprocal of the reach-level average so that lower numbers indicate higher nutrient delivery rates. The hydrologic runoff group was supplied by the Ontario Ministry of Agriculture and Food (Soils of Southern Ontario, 2005).

Table ESM-1: Aspects of the dataset used.

Station number	Creek Name	Number of concentration measurements	Mean total phosphorus concentration (mg/L)	Concentration standard deviation* (mg/L)	Drainage area (km^2)	$\overline{Q}$ (m^3/s)	Average Logged Load (tons/yr)	Variance of logged load $(tons/yr)^2$	E2 logged load (tons/yr)	E2 variance $(tons/yr)^2$	Multiplicative standard deviation, % (E1)	Multiplicative standard deviation, % (E1)
1	Redhill	75	0.10	0.08	25.53	0.249	-0.067	0.80	0.720	0.524	145	106
2	Redhill	76	0.10	0.05	51.65	0.569	0.034	1.25	0.958	0.527	206	107
3	Chedoke	117	0.25	0.11	27.80	0.277	0.998	0.66			125	
4	Spencer	48	0.02	0.01	49.05	0.537	-1.288	0.67	-1.113	0.025	127	17
5	Spencer	48	0.05	0.03	124.83	1.465	0.612	0.75	1.053	0.038	138	22
6	Spencer	47	0.06	0.05	157.16	1.861	1.126	0.80	1.509	0.037	145	21
7	Spencer	90	0.09	0.15	221.63	2.651	2.084	1.18			196	
8	Borer's	101	0.11	0.17	19.34	0.173	-0.241	1.13			190	
9	Grindstone	1	0.12	NA	11.10	0.073	0.153	2.46			380	
10	Indian	1	0.08	NA	8.06	0.035	-0.342	2.46			380	
11	Indian	1	0.11	NA	5.25	0.001	-0.007	2.46			379	
12	Grindstone	5	0.06	0.04	22.49	0.212	-0.600	0.76			140	
13	Grindstone	10	0.19	0.15	30.52	0.310	0.638	0.88			156	
14	Grindstone	12	0.10	0.05	8.66	0.043	-0.171	0.75			138	
15	Grindstone	5	0.17	0.17	43.63	0.471	0.627	1.12			188	
16	Grindstone	1	0.04	NA	45.37	0.492	-0.517	2.47			381	
17	Grindstone	6	0.09	0.08	65.59	0.740	0.306	1.03			176	
18	Grindstone	5	0.17	0.21	12.94	0.095	0.082	1.35			219	
19	Grindstone	1	0.03	NA	8.95	0.046	-1.135	2.46			380	
20	Grindstone	96	0.10	0.28	76.08	0.868	0.245	1.09			184	
21	Grindstone	1	0.03	NA	77.81	0.890	-0.335	2.47			382	
22	Grindstone	19	0.14	0.10	78.06	0.893	0.873	1.04			177	
23	Grindstone	154	0.09	0.14	86.86	1.000	0.515	1.04	1.083	0.028	177	18
24	Grindstone	88	0.15	0.30	95.49	1.106	1.057	0.98			169	

*Note that the variance of the logged concentrations at stations with only one concentration measurement were specified as 2.0, more than twice the highest variance of the rest of the stations.

Table ESM-2: Data used for the Value of Information Experiment with 24 stations. Precisions are expressed as standard deviations on the log scale.

Station Number	Posterior Mean Load (logged, tons/yr)	Measurement Error Standard Deviation - Current Precision	Measurement Error Standard Deviation - High Precision
1	0.860	0.724	0.724
2	1.298	0.726	0.725
3	1.074	0.811	0.725
4	-0.979	0.157	0.157
5	1.030	0.196	0.196
6	1.418	0.193	0.193
7	1.811	1.086	0.725
8	-0.302	1.065	0.725
9	-0.803	1.568	0.725
10	-0.186	1.568	0.725
11	-0.734	1.568	0.725
12	-0.499	0.874	0.725
13	0.000	0.940	0.725
14	-0.863	0.866	0.725
15	0.295	1.057	0.725
16	0.229	1.570	0.725
17	0.500	1.015	0.725
18	-1.243	1.160	0.725
19	-0.931	1.568	0.725
20	0.681	1.045	0.725
21	0.717	1.573	0.725
22	0.761	1.017	0.725
23	1.049	0.168	0.168
24	1.010	0.989	0.725

Table ESM-3: Data used for the Value of Information Experiment with 12 stations. Precisions are expressed as standard deviations on the log scale.

Station Number	Posterior Mean Load (logged, tons/yr)	Measurement Error Standard Deviation - Current Precision	Measurement Error Standard Deviation - High Precision
1	0.860	0.724	0.724
2	1.298	0.726	0.725
3	1.074	0.811	0.725
4	-0.979	0.157	0.157
5	1.030	0.196	0.196
6	1.418	0.193	0.193
7	1.811	1.086	0.725
8	-0.302	1.065	0.725
11	-0.734	1.568	0.725
12	-0.499	0.874	0.725
19	-0.931	1.568	0.725
23	1.049	0.168	0.168

Table ESM-4: Correlation matrix of predicted mean loads for all combinations of data error characterizations and statistical formulations. Values less than 0.90 are in bold and italics.

	MCMC-1 E1	MCMC-2 E1	MCMC-3 E1	CAR E1	STSP E1	MCMC-1 E2	MCMC-2 E2	MCMC-3 E2	CAR E2	STSP E2	CAR E0	STSP E0
MCMC-1 E1	1.00	1.00	1.00	1.00	1.00	0.98	0.98	0.99	0.97	0.98	***0.86***	***0.88***
MCMC-2 E1	1.00	1.00	1.00	0.99	1.00	0.98	0.99	0.99	0.97	0.97	***0.84***	***0.86***
MCMC-3 E1	1.00	1.00	1.00	1.00	1.00	0.98	0.99	0.99	0.97	0.98	***0.85***	***0.87***
CAR E1	1.00	0.99	1.00	1.00	1.00	0.98	0.98	0.99	0.98	0.98	***0.88***	***0.89***
STSP E1	1.00	1.00	1.00	1.00	1.00	0.98	0.98	0.99	0.97	0.98	***0.86***	***0.88***
MCMC-1 E2	0.98	0.98	0.98	0.98	0.98	1.00	1.00	1.00	0.98	0.99	***0.87***	***0.89***
MCMC-2 E2	0.98	0.99	0.99	0.98	0.98	1.00	1.00	1.00	0.98	0.99	***0.85***	***0.87***
MCMC-3 E2	0.99	0.99	0.99	0.99	0.99	1.00	1.00	1.00	0.98	0.99	***0.86***	***0.88***
CAR E2	0.97	0.97	0.97	0.98	0.97	0.98	0.98	0.98	1.00	1.00	0.91	0.92
STSP E2	0.98	0.97	0.98	0.98	0.98	0.99	0.99	0.99	1.00	1.00	0.90	0.92
CAR E0	***0.86***	***0.84***	***0.85***	***0.88***	***0.86***	***0.87***	***0.85***	***0.86***	0.91	0.90	1.00	1.00
STSP E0	***0.88***	***0.86***	***0.87***	***0.89***	***0.88***	***0.89***	***0.87***	***0.88***	0.92	0.92	1.00	1.00

Table ESM-5: Correlation matrix of standard deviations of predicted loads for all combinations of data error characterizations and statistical formulations. Values less than 0.75 are in bold and italics.

	MCMC-1 E1	MCMC-2 E1	MCMC-3 E1	CAR E1	STSP E1	MCMC-1 E2	MCMC-2 E2	MCMC-3 E2	CAR E2	STSP E2	CAR E0	STSP E0
MCMC-1 E1	1.00	0.98	0.99	0.92	0.98	0.97	0.97	0.97	0.82	0.84	***-0.07***	***-0.11***
MCMC-2 E1	0.98	1.00	0.98	0.90	0.97	0.95	0.97	0.96	0.77	0.78	***-0.10***	***-0.15***
MCMC-3 E1	0.99	0.98	1.00	0.91	0.98	0.98	0.97	0.98	0.82	0.83	***-0.07***	***-0.11***
CAR E1	0.92	0.90	0.91	1.00	0.96	0.92	0.92	0.92	0.88	0.85	***0.08***	***0.04***
STSP E1	0.98	0.97	0.98	0.96	1.00	0.95	0.96	0.96	0.87	0.87	***0.00***	***-0.04***
MCMC-1 E2	0.97	0.95	0.98	0.92	0.95	1.00	0.99	1.00	0.86	0.86	***-0.06***	***-0.11***
MCMC-2 E2	0.97	0.97	0.97	0.92	0.96	0.99	1.00	0.99	0.83	0.83	***-0.08***	***-0.13***
MCMC-3 E2	0.97	0.96	0.98	0.92	0.96	1.00	0.99	1.00	0.85	0.85	***-0.06***	***-0.12***
CAR E2	0.82	0.77	0.82	0.88	0.87	0.86	0.83	0.85	1.00	0.97	***0.16***	***0.11***
STSP E2	0.84	0.78	0.83	0.85	0.87	0.86	0.83	0.85	0.97	1.00	***0.20***	***0.15***
CAR E0	***-0.07***	***-0.10***	***-0.07***	***0.08***	***0.00***	***-0.06***	***-0.08***	***-0.06***	***0.16***	***0.20***	1.00	0.99
STSP E0	***-0.11***	***-0.15***	***-0.11***	***0.04***	***-0.04***	***-0.11***	***-0.13***	***-0.12***	***0.11***	***0.15***	0.99	1.00

Figure ESM-1: Locations of the Value of Information Experiment stations for the experiment with 24 stations.

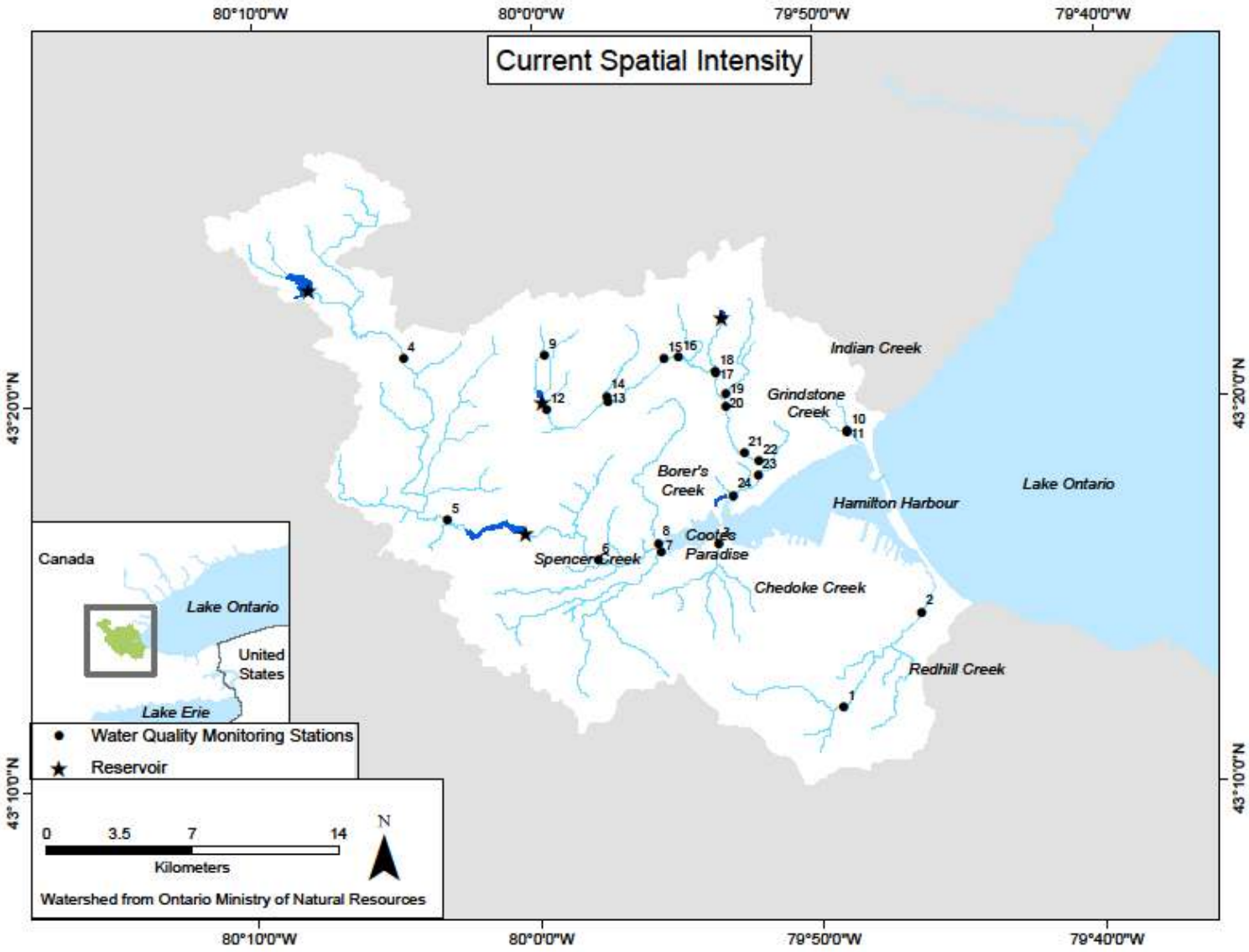
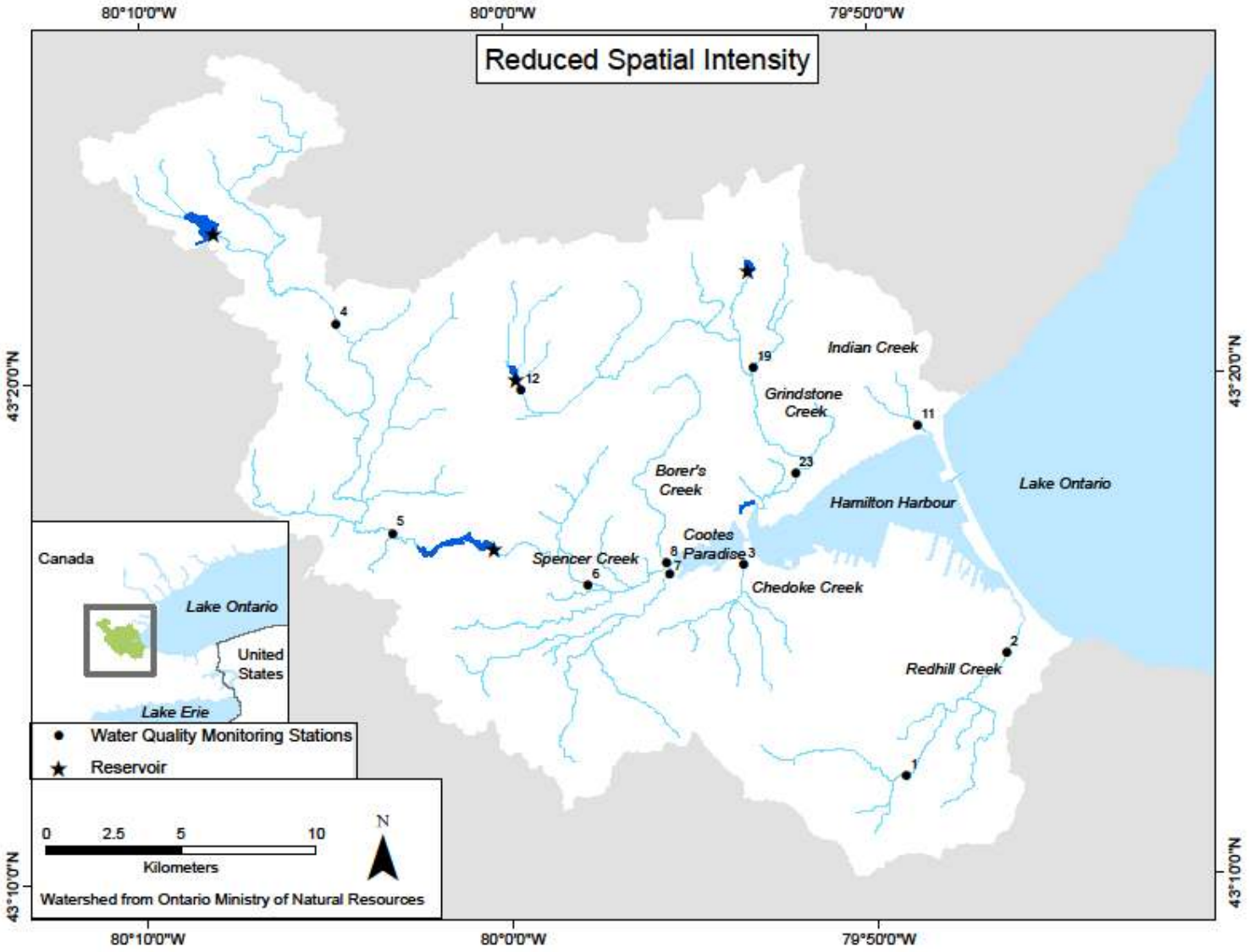


Figure ESM-2: Locations of the Value of Information Experiment stations for the experiment with 12 stations.



References

Alexander, R.B., Smith, R.A. and Schwarz, G.E., 2004. Estimates of diffuse phosphorus sources in surface waters of the United States using a spatially referenced watershed model. Water Science and Technology, 49:1-10.

Beaulac, M.N. and Reckhow, K.H., 1982. An examination of land-use - nutrient export relationships. Water Resources Bulletin, 18:1013-1024.

Hamilton Harbour Remedial Action Plan Technical Team, 2010. Contaminant Loadings and Concentrations to Hamilton Harbour: 2003-2007 Update. Hamilton Harbour Remedial Action Plan Office, Burlington, Ontario, Canada.

McMahon, G., Alexander, R.B. and Qian, S., 2003. Support of Total Maximum Daily Load Programs Using Spatially Referenced Regression Models. Journal of Water Resources Planning and Management, 129:315 – 329.

Natural Resources Canada, 2010. National Topographic System 1:50,000 Scale. Government of Canada.

Ontario Ministry of Food and Agriculture, 2005. Soils of Southern Ontario. <http://www.omafra.gov.on.ca>, April 4th 2011.

Ontario Ministry of Natural Resources, 2005. Greater Toronto Area Digital Elevation Model.

Ontario Ministry of Natural Resources, 2008. Southern Ontario Land Resource Information System (2000-2002) (SOLRIS). <http://lioapp.lrc.gov.on.ca>, April 4th 2011.

United States Geological Survey, 2000. National Hydrography Dataset: Concepts and Contents. <http://nhd.usgs.gov/>, April 4th 2011.

Viessman, W. and Lewis, G., 2002. Introduction to Hydrology (5th Edition). Prentice Hall, 612 p.

Water Survey of Canada, 2011. Archived Hydrometric Data. Environment Canada. http://www.wsc.ec.gc.ca/applications/H2O/index-eng.cfm, April 4th 2011.