A Bayesian methodological framework for accommodating interannual variability of nutrient loading with the SPARROW model

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

Received 3 January 2012; revised 17 August 2012; accepted 23 August 2012; published 3 October 2012.

[1] Regression-type, hybrid empirical/process-based models (e.g., SPARROW, PolFlow) have assumed a prominent role in efforts to estimate the sources and transport of nutrient pollution at river basin scales. However, almost no attempts have been made to explicitly accommodate interannual nutrient loading variability in their structure, despite empirical and theoretical evidence indicating that the associated source/sink processes are quite variable at annual timescales. In this study, we present two methodological approaches to accommodate interannual variability with the Spatially Referenced Regressions on Watershed attributes (SPARROW) nonlinear regression model. The first strategy uses the SPARROW model to estimate a static baseline load and climatic variables (e.g., precipitation) to drive the interannual variability. The second approach allows the source/sink processes within the SPARROW model to vary at annual timescales using dynamic parameter estimation techniques akin to those used in dynamic linear models. Model parameterization is founded upon Bayesian inference techniques that explicitly consider calibration data and model uncertainty. Our case study is the Hamilton Harbor watershed, a mixed agricultural and urban residential area located at the western end of Lake Ontario, Canada. Our analysis suggests that dynamic parameter estimation is the more parsimonious of the two strategies tested and can offer insights into the temporal structural changes associated with watershed functioning. Consistent with empirical and theoretical work, model estimated annual in-stream attenuation rates varied inversely with annual discharge. Estimated phosphorus source areas were concentrated near the receiving water body during years of high in-stream attenuation and dispersed along the main stems of the streams during years of low attenuation, suggesting that nutrient source areas are subject to interannual variability.

Citation: Wellen, C., G. B. Arhonditsis, T. Labencki, and D. Boyd (2012), A Bayesian methodological framework for accommodating interannual variability of nutrient loading with the SPARROW model, *Water Resour. Res.*, *48*, W10505, doi:10.1029/2012WR011821.

1. Introduction

[2] There is a pressing demand for watershed models that can support our efforts to effectively quantify nonpoint source pollution [*Rode et al.*, 2010]. While a suite of process-based models does exist to address this need, they have data requirements which cannot always be met (e.g., detailed subsurface properties) and are typically applied in well-studied catchments [*Borah and Bera*, 2004]. Hybrid empirical/process-based models, generally founded upon nonlinear regression equations, have been developed at large scales where a process-based model would become unwieldy and a priori knowledge about dominant biogeochemical process rates may not be available. They have been applied extensively in the United States [*Alexander et al.*, 2004], the United Kingdom [*Grizzetti et al.*, 2005], and continental Europe [*de Wit*, 2001]. The Spatially Referenced Regressions on Watershed attributes (SPARROW) model is a parsimonious hybrid empirical/process-based model developed by the United States Geological Survey to estimate nutrient loads, yields, and deliveries at landscape and regional scales. Despite its nonlinear regression structure, the inputs can be chosen according to a mechanistic understanding of nutrient source and sink dynamics. The SPARROW model has been applied at a variety of sites and scales, including New Zealand's Waikato River Basin [*Alexander et al.*, 2002], the Neuse River Estuary [*McMahon et al.*, 2003] the continental United States [*Alexander et al.*, 2004], the Mississippi River Basin [*Alexander et al.*, 2008], the Southeastern United States [*García et al.*, 2011], the United States drainage to the Laurentian Great Lakes [*Robertson and Saad*, 2011],

[1]Ecological Modeling Laboratory, Department of Physical and Environmental Sciences, University of Toronto, Toronto, Ontario, Canada.
[2]Great Lakes Unit, Water Monitoring and Reporting Section, Environmental Monitoring and Reporting Branch, Ontario Ministry of the Environment, Toronto, Ontario, Canada.

Corresponding author: C. Wellen, Ecological Modeling Laboratory, Department of Physical and Environmental Sciences, University of Toronto, Toronto, ON M1C 1A4, Canada. (christopher.wellen@utoronto.ca)

0043-1397/12/2012WR011821

the Pacific Northwest, the Missouri River Basin, the Lower Mississippi River Basin, and the New England and Mid-Atlantic drainage [*Preston et al.*, 2011]. SPARROW applications primarily focus on either nitrogen or phosphorus loadings, but models have also been developed for organic carbon [*Shih et al.*, 2010], suspended sediment [*Brakebill et al.*, 2010], and *E. coli* [*Puri et al.*, 2009]. There are several difficulties to effectively accommodate spatial and temporal variability, when using models such as SPARROW in a nested basin context. The spatial difficulties have been examined in some depth and will be briefly discussed below, followed by the lesser examined temporal difficulties, the subject of this paper.

[3] A direct ramification of SPARROW's distributed structure is the propagation of the model (process) error in space, which in turn poses a major statistical challenge. As do all distributed regression models of mass loading, the model considers upstream stations as point sources to downstream stations. This introduces a potential serial correlation of model residuals. Most SPARROW applications overcome this problem by using the measured upstream load as the input to downstream sites [e.g., *McMahon et al.*, 2003]. However, this practice is prone to essentially the same problem, as it ignores the (possibly substantial) imperfections of measured annual loads of watersheds and propagates the *measurement* error downstream. SPARROW model applications may also exhibit a spatial structure of the model residuals that does not stem from serial autocorrelation alone [*McMahon et al.*, 2003]. For the sake of parsimony, SPARROW by default assumes uniform values of the model parameters across the study watershed, an assumption that may likely be another source of residual spatial autocorrelation. That is, the use of a single export coefficient for all the agricultural land uses clearly overestimates the intensity of the agricultural practices in certain (neighboring) sites and underestimates them in others. Though some applications of SPARROW do feature some type of spatial variability of the model coefficients [*Alexander et al.*, 2004; *García et al.*, 2011], the spatial delineation of these coefficient zones is often done in an ad-hoc manner. Founded upon Bayesian inference techniques, *Qian et al.* [2005] presented a formidable framework for accommodating the serial and spatial autocorrelation of residuals in SPARROW. In addition to the classical independent model error, this study introduced an error which applies only to stations receiving loading from upstream stations (the so-called state space or *STSP* model) and error terms that account for the spatial correlation of neighboring sites regardless of the drainage network (the so-called conditional autoregressive or *CAR* model). *Qian et al.* [2005] showed that for the SPARROW application at the Neuse River Estuary watershed [*McMahon et al.*, 2003], the serially autocorrelated error contributes little to the total error, while most of the overall mismatch between model predictions and measurements could be explained by the spatially autocorrelated errors.

[4] Despite the significant progress in explicitly considering the various forms of spatial correlation, there are still no attempts in the published literature to accommodate the interannual variability of loading with either of the most commonly used hybrid empirical/process-based models (SPARROW and *PolFlow* [*de Wit*, 2001]). The typical SPARROW approach thus far has been to de-trend time series of annual loading estimates at each water quality monitoring station to a common base year [*Alexander et al.*, 2002]. This base year represents the nutrient load that would have been observed at each station if average hydrological conditions had prevailed. This strategy is a pragmatic means to focus exclusively on spatial variability, while "factoring out" both the temporal variability of loading as well as the effects of different observation periods across sampling stations. Yet, using estimated nutrient source areas generated with this approach to inform policy or target management interventions postulates that there is insignificant interannual variability of source areas, an assumption which has not been examined and most likely oversimplifies the dynamics typically experienced within the watershed context. The PolFlow model simply averages nutrient flux over a 5 year time period [*de Wit*, 2001]. While conceptually and mathematically simpler than de-trending, this approach requires a very similar data record across sites and lumps interannual variability due to both nutrient sources and climate forcing into nutrient flux estimate uncertainty.

[5] The aim of this paper is to present a methodological framework for incorporating temporal nutrient loading variability into the SPARROW model. We subsequently apply this approach to the Hamilton Harbor drainage basin, a mesoscale catchment of about 450 km^2, much smaller than those typically represented with the SPARROW model. While we focus exclusively on phosphorus in this study, our methods could be applied to the modeling of any mass flux. We employ a repeated measures approach—that is, the loading at a station for a year is treated as a datum in the regression. This time for space substitution allows us to estimate source areas and loads for each year. We adapt Bayesian configurations to accommodate the temporal correlation of model residuals and the uncertainty of the calibration data and conduct a number of numerical experiments to test two methodological approaches of incorporating temporal variability. The first approach postulates that the watershed characteristics as modeled by SPARROW represent a static, baseline level of nutrient loading associated with average conditions, while climatic predictors (e.g., precipitation) are used to describe the temporal variability around that mean. The second strategy assumes that in addition to the temporal variability associated with climatic forcing factors, there is also year-to-year variation in the source and sink processes modeled by SPARROW. We adopt methods of estimating time-varying parameters used with dynamic linear models (*DLMs*) to the nonlinear context of SPARROW. Our presentation will examine model realizations that incorporate a number of temporal predictors and different assumptions about the temporal distribution of model residuals.

2. Methodology

2.1. Description of the SPARROW Model

[6] The SPARROW model has been extensively described elsewhere [*Alexander et al.*, 2002; *McMahon et al.*, 2003; *Qian et al.*, 2005; *García et al.*, 2011], so only a basic introduction is given here. SPARROW is a hybrid empirical/process-based model designed to be applied to a network of water quality monitoring stations. 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 draining to a particular stream segment. Mean annual watershed export of any constituent is expressed as a function of watershed attributes.

[7] The model considers source and sink processes over annual timescales. Source processes, described with export coefficients, which predict constituent mobilization; delivery factors predict how landscape attributes modulate the delivery of the mobilized constituent to streams; and attenuation coefficients predict the amount of the delivered constituent remaining in transit per length of stream or per reservoir. The SPARROW model is typically calibrated to a particular base year to describe the transport of nutrient inputs occurring in that particular time frame, while incorporating the interannual variability in hydrology that occurs over a series of years. The SPARROW model is formulated as:

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

where the subscripts i and j refer to subwatersheds and reach catchments, respectively; μ_i refers to the natural logarithm of the 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 (tons P km^{-2} yr^{-1} for nonpoint sources); $S_{n,j}$ refers to the quantity of source n in reach j in units of km^2 of agricultural or urban land use for nonpoint sources, and metric tons yr^{-1} for point sources; α refers to a vector of land to water delivery coefficients; Z_j is a vector of the land-surface characteristics associated with drainage in 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; and $H_{i,j}^R$ 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.

[8] First-order loss processes in streams are expressed as

$$H_{i,j}^S = \prod_m \exp\left(-k_{s,m} L_{i,j,m}\right) \quad (2)$$

where $k_{s,m}$ refers to the first-order loss coefficient for stream class m (km^{-1}), and $L_{i,j,m}$ refers to the class m stream length in kilometers between reach i and station j. First-order loss processes in lakes and reservoirs are expressed as:

$$H_{i,j}^R = \prod_l \exp\left(-k_r q_l^{-1}\right) \quad (3)$$

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 yr^{-1}), and q_l refers to the aerial hydraulic loading of the lake/reservoir (m yr^{-1}). Table 1 contains a list of all parameters included in the SPARROW model.

2.2. Introduction of Temporal Variability to the SPARROW Model

[9] Our framework introduces temporal variability to the SPARROW model by applying a repeated measures approach to a network of water quality monitoring stations. Rather than selecting a single year to phase out the variability in time and subsequently focusing on the spatial variability, we calibrate the model to annual loads measured repeatedly at a subset of intensively monitored sites in the studied watershed. Henceforth, we will be referring to this temporal augmentation as the SPARROW with annual loads of watersheds (SWALLOW) model. With this statistical configuration, the SPARROW model is used to estimate a static baseline level of nutrient loading (μ_i) over the study period and forcing factors are being employed to explain the temporal variability around that baseline:

$$\begin{aligned} Y_{i,t} &= \mu_i + W_{v,t}\gamma_v + \varepsilon_{i,t} \\ \varepsilon_{i,t} &\sim N(0, \sigma^2) \end{aligned} \quad (4)$$

where $Y_{i,t}$ refers to the natural logarithm of the measured annual load at subwatershed monitoring station i during year t, μ_i refers to a prediction of the natural logarithm of a baseline annual load at monitoring station i estimated by the SPARROW equation, $W_{v,t}$ denotes a matrix of v, temporal forcing factors across years t, γ_v denotes the corresponding vector of coefficients, and $\varepsilon_{i,t}$ represents an independent spatiotemporal error. All errors are assumed independent, normally

Table 1. Stochastic Nodes of the Different Model Configurations Examined

Parameter	Description	Units
α	Land to water delivery coefficient.	–
β_1	Export coefficient for agricultural land.	tons P km^{-2} yr^{-1}
β_2	Export coefficient for urban land.	tons P km^{-2} yr^{-1}
k_r	Reservoir settling velocity.	m yr^{-1}
k_{s1}	Stream attenuation coefficient for first and second-order streams.	km^{-1}
k_{s2}	Stream attenuation coefficient for third and higher-order streams.	km^{-1}
γ_v	Temporal coefficient for predictor v.	–
σ	Standard model error.	Ln[tons P yr^{-1}]
ψ	Standard model error specific to *WALK*.	Ln[tons P yr^{-1}]
α_σ	Initial SD of the prior for the α parameter for the dynamic parameter estimation framework.	–
$\beta_{1\sigma}$	Initial SD of the prior for the β_1 parameter for the dynamic parameter estimation framework.	tons P km^{-2} yr^{-1}
$\beta_{2\sigma}$	Initial SD of the prior for the β_2 parameter for the dynamic parameter estimation framework.	tons P km^{-2} yr^{-1}
$k_{s1\sigma}$	Initial SD of the prior for the k_{s1} parameter for the dynamic parameter estimation framework.	km^{-1}
$k_{s2\sigma}$	Initial SD of the prior for the k_{s2} parameter for the dynamic parameter estimation framework.	km^{-1}

distributed, and with zero mean. We refer to equation (4) as the SWALLOW I model throughout this paper.

[10] We specified a second version of the SWALLOW model designed to accommodate the variability of watershed functioning in time:

$$\begin{aligned} Y_{i,t} &= \mu_{i,t} + W_{v,t}\gamma_v + \varepsilon_{i,t} \\ \varepsilon_{i,t} &\sim N(0,\sigma^2) \\ \mu_{i,t} &= Ln\left(\left\{\sum_{n=1}^{N}\sum_{j=1}^{J_i}\beta_{n,t}S_{n,j,t}e^{(-\alpha_t Z_{j,t})}H^S_{i,j,t}H^R_{i,j,t}\right\}\right) \end{aligned} \tag{5}$$

where $\mu_{i,t}$ refers to a prediction of the natural logarithm of the annual load at monitoring station i for year t independent of the effects of the temporal covariates represented by the $W_{v,t}$ matrix. All of the other variables in equation (5) are identical to their counterparts in equation (4). We refer to equation (5) as the SWALLOW II model throughout this paper. The SWALLOW model makes use of nonspatial temporal forcing factors to accommodate interannual variability of watershed loads. While any time series data could be included as forcing factor, we focused on local climatic characteristics due to their importance and availability. Table 1 presents all model parameters examined.

2.3. Bayesian Inference Framework

[11] Bayesian inference was used as a means of model calibration due to its ability to include prior information in the modeling exercise and to explicitly handle uncertainties stemming from what we assumed were the main sources of uncertainty in this modeling exercise: model parameters, calibration data, and model structure. From the Bayesian perspective, statistical inference is treated as a quantitative update of prior beliefs after taking measurements into account. Beliefs are expressed as probability distributions (i.e., random variables), with the central tendency of these distributions corresponding to the degree of certainty that the expected value of the distribution is correct [*Gelman et al.*, 2004]. Mathematically, Bayesian inference is founded upon Bayes' Theorem, expressed as

$$\pi(\theta|data) = \frac{\pi(\theta)L(data|\theta)}{\int_{\theta}\pi(\theta)L(data|\theta)d\theta} \tag{6}$$

where $\pi(\theta)$ represents our prior statements regarding the probability distribution that depicts the existing knowledge of the model parameters (θ), $L(data|\theta)$ corresponds to the likelihood of observing the data given the different θ values, and $\pi(\theta|data)$ 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 (6) 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. 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 40,000 samples each from two chains for each model realization. The first 10,000 samples were discarded and posterior statistics were calculated using a thin of 10, yielding a sample size of 6000 for all the model realizations considered. We assessed convergence qualitatively by visually inspecting plots of the posterior Markov chains for mixing and stationarity and by inspecting density plots of the pooled posterior Markov chains for unimodality. We also assessed convergence quantitatively using the modified Gelman–Rubin convergence statistic [*Brooks and Gelman*, 1998]. The accuracy of the posterior parameter values was inspected by assuring that the Monte Carlo error for all parameters was less than 5% of the sample standard deviation.

[12] Wherever possible we opted for informative priors. Priors for the export coefficients, settling velocity, and instream attenuation were log-normally distributed, owing 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 values of the β coefficients represented literature-based estimates of total phosphorus export [*Beaulac and Reckhow*, 1982]. The upper limit found for total phosphorus in the database 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 work by *Cheng et al.* [2010]. We based the prior distributions for k_{s1} and k_{s2}, the stream attenuation coefficients, loosely on values from previous models; that is, we assigned a higher median to k_{s1} than to k_{s2} along with standard deviations that are fairly large compared to the range of k_s between 0 and 1 [*Alexander et al.*, 2004]. The priors are presented in Table S1 of the auxiliary material.[1]

[13] The typical SPARROW practice uses the measured upstream load for input to downstream subwatersheds during calibration, which conceptually undermines the usefulness of the model for predictive purposes [*McMahon et al.*, 2003; *Qian et al.*, 2005]. Using the measured loads as inputs into downstream watersheds has two major problems. First, it overestimates the confidence in the loading data. Being mere estimates of the actual nutrient fluxes, the so-called "measured" loads are associated with a substantial error and failure to account for their uncertainty can result in a misleading model calibration. Second, relying on the measured loads as upstream input means that predictions at stations with most of their watershed area monitored by an upstream station may strongly depend on the measured inputs, which in turn results in a very optimistic assessment of the model error. All of the statistical formulations explored in this paper use the modeled load as input to downstream stations. As established by *Qian et al.* [2005], some representation of the uncertainty of the calibration data when using modeled loads as inputs to downstream subwatersheds is necessary to avoid a misleading

[1] Auxiliary materials are available in the HTML. doi:10.1029/2012WR011821.

model calibration. We describe our approach to do so in section 2.3.1.

2.3.1. Calibration Data Uncertainty

[14] The importance of explicitly accommodating calibration data uncertainty has been acknowledged in the literature [e.g., *Renard et al.*, 2010], though it is typically ignored in the context of SPARROW-type models. This is a significant omission, considering that annual loads are typically estimated using rating curve models or estimation approaches applied to measurements collected bi-weekly or less frequently, and are subject to substantial uncertainty [*Richards and Holloway*, 1987; *Cohn et al.*, 1992]. Two approaches have been discussed for representing measurement error in models. In the context of estimates of annual loads, the classical approach assumes that the observed values of a variable $Y_{i,t}$ are drawn from a distribution which has as its expected value $Load_{i,t}$, the "true" value of the variable being sampled [*Carroll et al.*, 2006]. The classical approach is appropriate when the uncertainty is assumed to come from deficiencies in sampling or measurement and has been used to model the uncertainty of point rainfall estimates [*Balin et al.*, 2010]. The Berkson model takes the opposite approach, assuming that the true value is drawn from a distribution with expected value equal to the observed value. The Berkson approach is appropriate when the uncertainty is assumed to stem from a lack of commensurability between what has been measured and the variable one is interested in, and has been applied to estimate mean aerial rainfall from point measurements [*Ajami et al.*, 2007]. Mathematically, the key difference between the two resides in whether the observed values vary about the true values (classical) or the true values vary about the observed (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.

[15] In our case, the classical measurement error model consists of three components: (1) the (log-transformed) measurements $Y_{i,t}$, (2) the (log-transformed) true values $Load_{i,t}$, and the measurement error $\delta^2_{i,t}$. These variables are arranged in a hierarchical framework, which has as its first level the relation of the observed to true loading values:

$$Y_{i,t} \sim N(Load_{i,t}, \delta^2_{i,t}). \tag{7}$$

Note that because we are working with log-transformed data this postulates multiplicative measurement error. For this paper, the values of $\delta^2_{i,t}$ are prespecified and are not part of the model calibration process. In section 2.5 we detail how they are calculated. The second level of the hierarchy introduces a model for the "true" log transformed loads:

$$Load_{i,t} \sim N(\mu_{i,t} + W_{v,t}\gamma_v, \sigma^2). \tag{8}$$

Because the term $\mu_{i,t} + W_{v,t}\gamma_v$ is equal to the SWALLOW 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 in year t, given the model, is then the product of the likelihood of the two levels of our hierarchical configurations:

$$\begin{aligned} &p(Y_{i,t}|Load_{i,t}) \times p(Load_{i,t}|\mu_{i,t} + W_{v,t}\gamma_v) \\ &= \frac{1}{\sqrt{2\pi}\delta_{i,t}} \exp\left(-\frac{(Y_{i,t} - Load_{i,t})^2}{2\delta^2_{i,t}}\right) \\ &\times \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\left(Load_{i,t} - (\mu_{i,t} + W_{v,t}\gamma_v)\right)^2}{2\sigma^2}\right) \end{aligned}. \tag{9}$$

To summarize, our calibration error framework seeks to minimize both the differences between the measured and "true" loading data as well as between the "true" and modeled loading. To do so, we must estimate the "true" loading as part of the model calibration. This adds an additional $i \times t$ stochastic nodes, considerably increasing the complexity of the calibration exercise but realistically accommodating the measurement errors as well as the model process error.

2.3.2. Introducing Interannual Variability With Climatic Predictors—SWALLOW I

[16] The simplest approach we investigated used the classical SPARROW model in conjunction with climatic forcing factors to estimate annual loads (equation (4)). This formulation, called Markov-Chain Monte Carlo (*MCMC*), assumes that the annual log-transformed nutrient loading is a draw from a normal distribution with a mean defined by the model and a constant model (process) error variance. Model residuals are assumed to be independent both in space and time. When we also consider the calibration data uncertainty, we can express this approach mathematically as follows:

$$\begin{aligned} &Y_{i,t} \sim N(Load_{i,t}, \delta^2_{i,t}) \\ &Load_{i,t} \sim N(\mu_i + W_{v,t}\gamma_v, \sigma^2) \\ &\mu_i = \ln\left(\sum_{n=1}^{N}\sum_{j=1}^{J_i}\beta_n S_{n,j} e^{(-\alpha Z_j)} H^S_{i,j} H^R_{i,j}\right) \\ &\sigma^{-2} \sim gamma(0.001, 0.001) \end{aligned} \tag{10}$$

where $Y_{i,t}$ refers to the log-transformed measured load of subwatershed i at time t, $Load_{i,t}$ is a latent variable that represents the "true" loading values when accounting for the measurement error $\delta^2_{i,t}$, μ_i refers to the base loading calculated from the SPARROW equation, σ represents the model (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 error precision (the inverse of variance).

[17] Even at annual timescales, watershed processes are dynamic, yet the parameters used in equation (4) are static. This may cause a temporal autocorrelation of model residuals. Temporal autocorrelation could stem from systematic trends in either the nutrient export dynamics or the spatial patterns of the different land uses. We use a temporal first-order random-walk function (*WALK*) to account for the

temporal correlation of residuals [*Arhonditsis et al.*, 2008a, 2008b; *Sadraddini et al.*, 2011a]. We posit a random effect v_t for each year t represented by a first-order random walk prior [*Shaddick and Wakefield*, 2002; *Arhonditsis et al.*, 2008a, 2008b]. When we also account for the calibration data uncertainty, the *WALK* formulation is as follows:

$$
\begin{aligned}
&Y_{i,t} \sim N(Load_{i,t}, \delta^2_{i,t})\\
&Load_{i,t} \sim N(\mu_i + W_{v,t}\gamma_v + v_t, \sigma^2)\\
&\mu_i = \ln\left(\sum_{n=1}^{N}\sum_{j=1}^{J_i}\beta_n S_{n,j} e^{(-\alpha Z_j)} H^S_{i,j} H^R_{i,j}\right)\\
&v_t|v_{-t} \sim \begin{cases} N(v_{t+1}, \psi^2) & \text{for} \quad t = 1 \\ N\left(\dfrac{v_{t-1}+v_{t+1}}{2}, \dfrac{\psi^2}{2}\right) & \text{for} \quad t = 2, \ldots, T-1 \\ N(v_{t-1}, \psi^2) & \text{for} \quad t = T \end{cases}\\
&\sigma^{-2}, \psi^{-2} \sim gamma(0.001, 0.001)
\end{aligned}
\tag{11}
$$

where $-t$ denotes the previous and subsequent years of t, T denotes the total number of years of the study period, and ψ^2 is the conditional variance of the v_t terms and its prior density was based on a conjugate inverse-gamma (0.001, 0.001) distribution. Our statistical approach reflects prior beliefs that these systematic trends in the watershed functioning are smooth and that sudden jumps between consecutive years are unlikely to occur.

2.3.3. Introducing Interannual Variability With Time Varying Parameters—SWALLOW II

[18] The rest of the formulations, referred to as SWALLOW II, considered time-variant watershed parameters and differed in the nature of the parameters allowed to vary. It is very unlikely that the landscape processes inherent in nonlinear regression models of nutrient loading (e.g., export rates, stream attenuation) operate identically from year to year, and therefore allowing these parameters to vary can be an effective means to accommodate interannual variability. The use of time variant parameters to overcome model structural deficiencies has been investigated for some time [e.g., *Beck and Young*, 1976]. Perhaps the most widely known approach is the Kalman filter [*Kalman*, 1960], a sequential model estimation approach that uses a gain function to combine model predictions with system measurements at each time step, inversely weighting each by their uncertainties. This method also postulates an error covariance structure that is estimated along with the rest of the model parameters. Kalman-type approaches to fusing model predictions and measurements online are ubiquitous in many disciplines, including watershed modeling, and have been used as a technique to estimate the values of time-varying parameters [e.g., *Moradkhani et al.*, 2005; *Lin and Beck*, 2007].

[19] Other approaches in the watershed modeling literature reproduce the temporal variability of parameter values with some type of stochastic process. *Reichert and Mieleitner* [2009], for instance, used the Ornstein-Uhlenbeck process to accommodate parametric variability in time. Although the underlying assumption of stationarity is a valuable way to account for structural uncertainty while constraining the added complexity of temporally varying parameters, we may not expect a parameter to be stationary with respect to the (always somewhat arbitrary) study time frame, e.g., the intensity of in-stream attenuation would be expected to vary with dry and wet years. *Lin and Beck* [2007] used a first-order random walk, a nonstationary process, in a dissolved oxygen model of a managed pond. Their analysis shows how time varying parameters can be used to identify structural improvements to models of environmental systems.

[20] For this paper we adapted approaches often used in the context of dynamic linear models (*DLMs*), which recognize the temporal structure in the data time series with the assumption that the level of the response variable at each time step is influenced by past levels [*West and Harrison*, 1989; *Prado and West*, 2010]. Two key points distinguish the *DLM* approach we employ in this paper from standard regression approaches. First, the *DLM* approach posits that some or all model parameters vary with time, and that their time series is autocorrelated – the closer in time, the more similar are parameter values. Second, in contrast with regression analysis, where parameters are conditioned on the entire time series, the dynamic parameter estimation is influenced only by prior and current information, not by subsequent data [*Stow et al.*, 2004; *Sadraddini et al.*, 2011a, 2011b]. In principle, the *DLM* approach is equivalent to the Kalman-type strategies, although the focus here is on a full probabilistic treatment of the underlying uncertainty, instead of a sequential updating of the mean prediction. Further, the sampling of parameter values is not done sequentially through the time series, but rather follows the standard *MCMC* approach of sampling a proposal point in the parameter space for the entire time series, evaluating that point, and applying the Metropolis Rule in deciding whether to add that point to the Markov chain [*Lunn et al.*, 2000].

[21] In this study, we introduce nonconstant and data-driven variances (with respect to time) using a discount factor on the prior of the first year [*Congdon*, 2001]. Based on experience from recent work [*Azim et al.*, 2011; *Sadraddini et al.*, 2011a, 2011b] and preliminary trials, we used values ranging from 0.95 to 0.98 and thus our dynamic parameter estimation framework is

$$
\begin{aligned}
&\theta_t = \theta_{t-1} + \varphi_t\\
&\varphi_t \sim N(0, \Phi^2_t)\\
&\Phi^{-2}_t = \zeta^{t-1} \times \Phi^{-2}_1\\
&\theta_1 \sim N(\theta_{mean}, \Phi^2_1) I(\theta_{min}, \theta_{max})\\
&\Phi^{-2}_1 \sim gamma(\alpha, \beta)
\end{aligned}
\tag{12}
$$

where θ_t represents any of the SPARROW parameters at time t, φ_t is the corresponding error term for year t sampled from normal distributions with zero mean and variance Φ^2_t, and ζ represents the discount factor. θ_{mean}, θ_{min}, and θ_{max} correspondingly represent the mean value, minimum and maximum of the literature priors used with the SWALLOW I formulations. The values of θ_1 as well as subsequent values θ_t were constrained within the range θ_{min} to θ_{max}. The gamma distribution assigned to the parameter Φ^{-2}_1 was constructed such that its mean was equal to the

variance of the SWALLOW I informative priors, while the uncertainty of the same distribution reflects our confidence of that mean estimate. To achieve commensurability between the SWALLOW I and SWALLOW II formulations, we assumed a very high level of confidence (i.e., coefficient of variation <5%). Our statistical configuration essentially postulates that between 95% and 98% of the information is carried forward from time t to $t + 1$; thus, the influence of the original priors decreases as time progresses and is gradually superseded by the influence of the data [*Azim et al.*, 2011, *Sadraddini et al.*, 2011a]. In this study, we examined different combinations of time variant parameters to identify the most parsimonious structure. In particular, we selected the following four parameter combinations: (1) delivery coefficient (α) alone; (2) the two export coefficients (β_1, β_2); (3) the two stream attenuation coefficients (k_{s1}, k_{s2}); and (4) the two export coefficients (β_1, β_2) along with the two stream attenuation coefficients (k_{s1}, k_{s2}). Table 2 presents all statistical formulations examined.

2.4. Case Study

[22] Hamilton Harbor is a large embayment at the western end of Lake Ontario. The Harbor is designated as one of 17 Canadian Areas of Concern in the Great Lakes Basin under the International Joint Commission due to a history of eutrophication problems manifested as nuisance algal blooms, turbid water, prevalence of toxic cyanobacteria, and low hypolimnetic oxygen concentrations toward the end of the summer stratified period [*Hiriart-Baer et al.*, 2009; *Ramin et al.*, 2011]. The Hamilton Harbor Remedial Action Plan (*RAP*), a consortium of government, private sector, and community actors, is mandated with restoring and protecting environmental quality and beneficial uses. RAP consultations with local stakeholders have identified a warm water fishery as a priority use for the Harbor [*Charlton*, 2001]. While earlier work highlighted the critical role of the sewage treatment plants in governing total phosphorus and chlorophyll α concentrations in the Harbor, substantial uncertainty regarding the water quality conditions exists due to the poorly defined nutrient loadings from the drainage basin [*Gudimov et al.*, 2010, 2011].

[23] Hamilton Harbor's drainage basin is about 450 km^2 in aerial extent and consists of watersheds dominated by agricultural (Grindstone and Spencer Creeks) or urban land use (Redhill and Indian Creeks; see Figure 1). Urban and agricultural land together account for 80% of the watershed's surface area. Population in Hamilton has been increasing and urban areas have been expanding, largely at the expense of agricultural land uses (Southern Ontario Land Resource Information System (SOLRIS), Ontario Ministry of Natural Resources, 2008, available from http://lioapp.lrc.gov.on.ca). The soils of the Harbor basin are mainly loams (73%), while organic soils, silty clay loams, and clay loams together make up about 10% of the basin soils. Most of the remainder is composed of rocky outcroppings and ravines. Soils are spread relatively evenly between the four soil hydrologic runoff groups – groups A and B, those least runoff prone, each have 23% coverage, group C has 29% coverage, and group D, the most runoff prone, has 24% coverage. The slopes of the Harbor 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%.

Table 2. Bayesian Statistical Formulations of the SWALLOW Models Examined

Model Notation	Description
MCMC	Model residuals are assumed independent. All parameters are static through time.
WALK	Random walk of model residuals through time. All prior parameters are independent and model residuals in space are assumed independent. All parameters are static through time.
MCMC - α_{DYN}	Model residuals are assumed independent. The α parameter (delivery to streams) varies each year, while all other parameters are static through time.
MCMC - β_{DYN}	Model residuals are assumed independent. The β parameters (export coefficients) vary each year, while all other parameters are static through time.
MCMC - k_{sDYN}	Model residuals are assumed independent. The k_s parameters (stream attenuation) vary each year, while all other parameters are static through time.
MCMC - β,k_{sDYN}	Model residuals are assumed independent. The β (export coefficients) and k_s parameters (stream attenuation) vary each year, while all other parameters are static through time.

2.5. Data Sets

2.5.1. Spatial Data Sets

[24] We provide an extensive description of the spatial datasets used as inputs to the SPARROW model in the auxiliary material and a brief overview here. We used a 10-m digital elevation model to delineate the subwatersheds. Our calibration data set had 6 subwatersheds. Their areas ranged from 25.5 – 75.8 km^2, with a mean of 49.3 km^2 and a standard deviation of 24.1 km^2. There are a total of 118 reach catchments, and each reach catchment discharges into a confluence, reservoir, or water quality monitoring station. Reach catchment areas ranged from 0.02 – 12.3 km^2, with a mean of 2.5 km^2 and an interquartile range of 3.5–1.3 = 2.2 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. Two stream classes are included in the model, one for streams of Strahler order 1 or 2, and one class for streams of Strahler order 3 or higher [*Strahler*, 1952]. Four reservoirs were used during the parameter estimation of the SWALLOW models (Figure 1). Nonpoint nutrient sources included in the model were agricultural land and urban land, together representing 80% of the basin area. A single wastewater 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 phosphorus per year, with an interquartile range of 0.4–0.2 = 0.2 tons per year (*Hamilton Harbor Remedial Action Plan Technical Team*, Contaminant Loadings and Concentrations to Hamilton Harbor: 2003–2007 Update, Hamilton Harbor Remedial Action Plan Office, Burlington, Ontario, Canada). Nutrient delivery was parameterized as a function of the proportion of each reach covered by wetlands due to their role in moderating nutrient fluxes to receiving waterbodies [*Krieger*, 2003]. Proportions of wetland ranged from 0 to 1 with a mean of 0.06 and an interquartile range of 0.06 – 0 = 0.06.

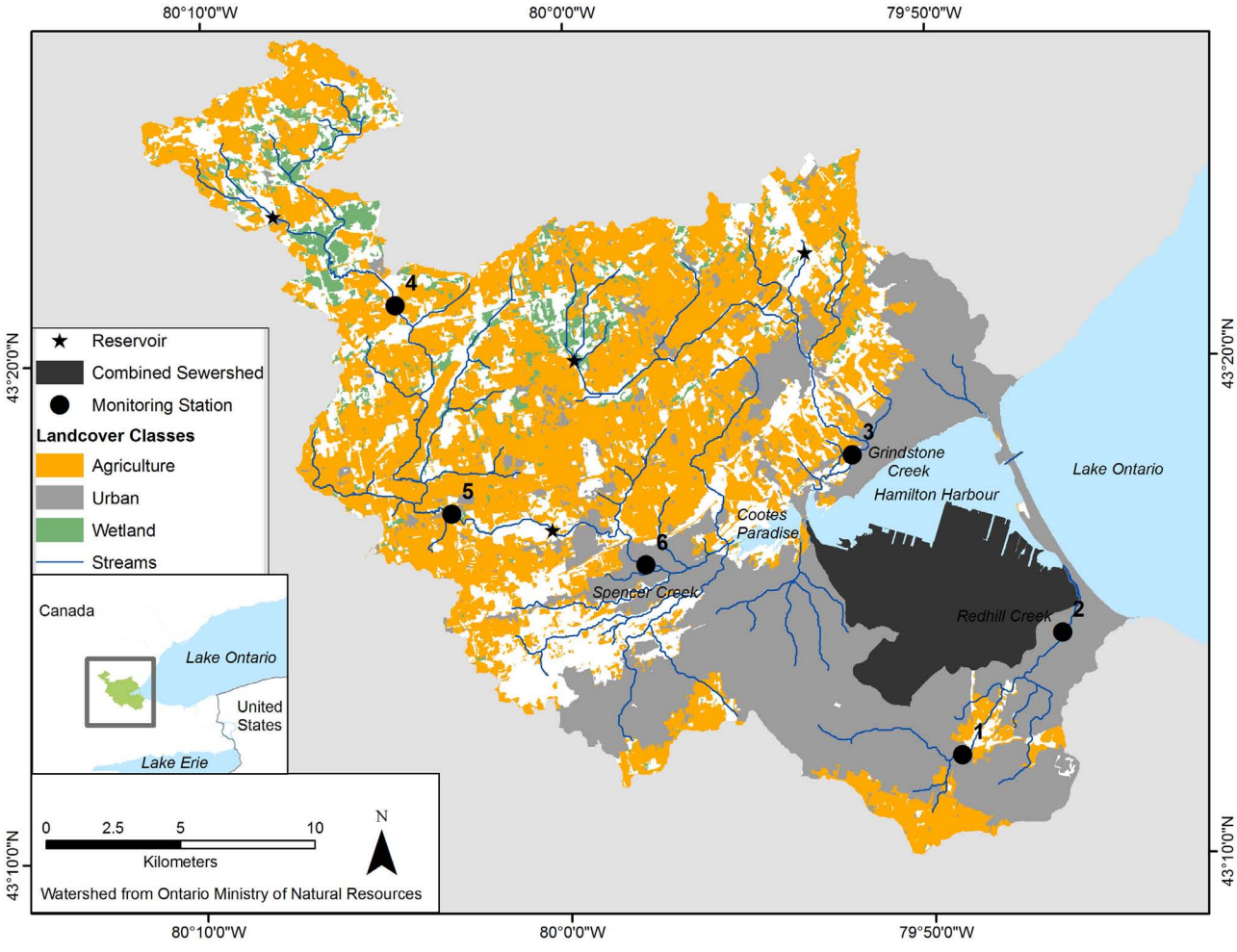


Figure 1. Map of the Hamilton Harbor watershed, western end of Lake Ontario, Ontario, Canada.

2.5.2. Nutrient Loads

[25] We estimated phosphorus loads for each year between 1988–2009 at each station using station-specific rating curves, each of which expressed log-transformed daily nutrient loading as a function of log-transformed daily flow:

$$Ln(Load) = a_0 + a_1 Ln(Q) \tag{13}$$

where a_0 and a_1 are regression coefficients, and $Ln(Q)$ refers to log transformed daily streamflow. All concentration measurements available for each station between 1988 and 2009 were employed in fitting the rating curve. We should stress that this approach to estimating loads accommodates the annual variability in loading associated with variations in the hydrograph, while variations of annual loading due to other factors such as changes in fertilizer application intensity are not captured. By assuming a single response of loading to flow throughout the study time period we may be underestimating the true temporal variability of annual loading, though the estimated log-transformed loads did show significant interannual variability, with coefficients of variability ranging from 0.27 to 0.34. The number of concentration measurements employed at each station ranged from 23 to 161 with a mean of 58. The r^2 values for the rating curves ranged from 0.71 to 0.92 with a mean of 0.82. Each rating curve was used in conjunction with daily flow records for each year to estimate average daily loading, which was multiplied by 365 to yield total annual loads for each year from 1988 to 2009. We included between 13 and 22 load estimates for each station for a total of 102 load estimates. One annual loading estimate was based on 346 days of estimated loads, while the rest were based on 365 (or 366 for leap years). Annual loads ranged from 0.2 to 6.3 tons yr^{-1} with a mean of 2.4 tons yr^{-1}. Log transformed total phosphorus values ranged from −1.6 to 1.8 Ln(t yr^{-1}) with a mean of 0.6 Ln(t yr^{-1}). All rating curve calculations were carried out with the U.S. Geological Survey's LOADEST program [*Runkel et al.*, 2004]. The concentration measurements were supplied by the Ontario Ministry of the Environment's Provincial Water Quality Monitoring Network, while the daily flows were supplied by the Water Survey of Canada (Ontario Provincial Water Quality Monitoring Network, 2011, unpublished data available from: http://www.ene.gov.on.ca/environment/; Water Survey of Canada, 2011, unpublished data available from: http://www.wsc.ec.gc.ca/applications/H2O/).

[26] Our data quality submodel postulates that the log transformed loadings are random variables drawn from normal distributions with mean values equal to the unknown true load and variances ($\delta^2_{i,t}$) representing the associated

uncertainty at each site for each year. We derive the $\delta^2_{i,t}$ terms from the 95% confidence intervals of the calculated mean daily loads at each station for each year as provided by the LOADEST program [*Runkel et al.*, 2004], which estimates the variance of the mean predicted load as the sum of the covariance of all daily load estimates [*Gilroy et al.*, 1990, equation (16)]:

$$MSE(X) = \sum_{i,j} Cov(L(i), L(j)) \tag{14}$$

where *MSE (X)* is the variance about the mean load prediction for a particular station at a particular year, *i* and *j* denote arbitrary days, and *L(i)* and *L(j)* denote the loads on days *i* and *j* predicted by the rating curve model. The covariance terms are estimated using equations given by *Gilroy et al.* [1990, equations (17)–(25)], and do account for the residual variance of the rating curve model in addition to its parametric uncertainty. The 95% confidence intervals of the mean daily load are calculated using *MSE(X)*, which we then multiplied by 365 and log-transformed to obtain the width of the 95% confidence interval on the log scale. In keeping with our assumption of log-normality, the values of $\delta_{i,t}$ were estimated as one quarter of this width. Values of $\delta_{i,t}$ ranged from 0.09 to 0.55 *Ln*(tons yr^{-1}) with a mean of 0.19 *Ln*(tons yr^{-1}).

2.5.3. Temporal Forcing Factors

[27] We use two climatic forcing factors in this paper: annual precipitation and annual potential evapotranspiration. All forcing factors were calculated from data collected at Environment Canada's Hamilton Airport station (WMO ID 71,263) between the years 1988 and 2009. Total annual precipitation ranged from 677 mm to 1115 mm, with a mean of 901 mm and an interquartile range of 1023–786 = 237 mm.

[28] Potential evapotranspiration serves as an estimate of the annual variability of atmospheric flux of water out of the basin. Evaporation is a pathway for precipitation to exit the basin without contributing to nutrient loading. We estimated daily potential evapotranspiration with the FAO's Penman-Monteith method and then summed to yearly intervals [*Allen et al.*, 1998]. While using potential evapotranspiration measured as a surrogate implicitly assumes that the main limitations to atmospheric-water fluxes are related to atmospheric conditions and energy supply and not related to water supply at the surface or stomatal/soil resistance to evapotranspiration, the inclusion of potential evapotranspiration as a temporal forcing factor may nonetheless offer some insights into the annual functioning of the Hamilton Harbor basin. All the details of the calculation are presented in the auxiliary material. Both temporal forcing factors were subjected to a nonparametric standardization ((*value-median*)/*interquartile range*) before their inclusion into the model.

2.6. Overview of Numerical Experiments and Model Evaluation

[29] The flexible framework provided by an empirical model allows many possible realizations, or combinations of inputs and statistical formulations. We here conceptualize the model realization space of SWALLOW as two dimensional, where the dimensions correspond to the statistical formulation, and temporal (climate) forcing complexity. We employed two *W* matrices: one that included only annual precipitation and one that included annual precipitation and total potential evapotranspiration. We also examined model realizations that omitted a *W* matrix altogether, which was only possible with the SWALLOW II formulation.

[30] We used two measures to evaluate the different model realizations examined. First, we used the deviance information criterion (*DIC*), a Bayesian measure of parsimony that rewards for model fit but penalizes model complexity [*Spiegelhalter et al.*, 2002]. The *DIC* is the Bayesian analog of Akaike's Information Criterion [*Akaike*, 1974]. The *DIC* is defined as follows:

$$DIC = \overline{D(\theta)} + p_D \tag{15}$$

where $\overline{D(\theta)}$ refers to the posterior mean of the deviance and p_D is a measure of the effective number of model parameters. The deviance is defined as the residual information in data *Y* conditional on a parameter vector θ and is calculated as -2 log{p($Y|\theta$)} or -2 log{likelihood}. 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. Model realizations were also evaluated for fit alone using two metrics: (1) the Root Mean Squared Error (*RMSE*), calculated using the medians of the posterior predictive distributions of the yearly log-transformed loads:

$$RMSE = \sqrt{\frac{\sum \left(Y_{i,t} - [\mu_{i,t} + W_{v,t}\gamma_v]\right)^2}{n}} \tag{16}$$

and (2) a Weighted Root Mean Squared Error (*WRMSE*) calculated using as weights the precision (inverse of variance) of the loading estimates:

$$\begin{aligned} WRMSE &= \sqrt{\sum w_{i,t} \times \left(Y_{i,t} - [\mu_{i,t} + W_{v,t}\gamma_v]\right)^2} \\ w_{i,t} &= \frac{\lambda_{i,t}}{\sum \lambda_{i,t}} \\ \lambda_{i,t} &= \frac{1}{\delta^2_{i,t}} \end{aligned} \tag{17}$$

where $Y_{i,t}$ refers to the measured log-transformed load for subwatershed *i* at year *t*, $\mu_{i,t} + W_{\nu,t}\gamma_\nu$ refers to the median of the posterior predictive distribution of the log-transformed loads from subwatershed *i* at year *t*, and *n* represents the number of total nutrient loading measurements.

3. Results

3.1. Evaluation of Model Performance

[31] The *DIC* values of the different models parameterized with the total phosphorus data are presented in Table 3. The corresponding *RMSE* and *WRMSE* values are also

Table 3. Deviance Information Criterion for All Statistical Formulations Used to Model Total Phosphorus Loading[a]

Formulation	*Pred0*	*Pred1*	*Pred2*
	SWALLOW I		
MCMC	–	−26.1	−25.8
WALK	–	−80.0	−79.7
	SWALLOW II		
MCMC - α_{DYN}	−18.5	−29.9	−30.2
MCMC - β_{DYN}	−44.6	−45.2	−45.7
MCMC - k_{sDYN}	−78.4	−80.3	−79.9
MCMC - β,k_{sDYN}	−38.1	−37.8	−37.1

[a]*Pred0* refers to the sole use of SPARROW to accommodate interannual loading variability; *Pred1* refers to the use of SPARROW along with the total annual precipitation; *Pred2* refers to the use of SPARROW along with the total precipitation and the total annual potential evapotranspiration.

presented in the auxiliary material. The highly favorable *DIC* values of the *WALK* formulations suggest systematic changes in the phosphorus exports from the watershed unaccounted for by the SPARROW model and the climatic covariates considered herein. For any number of climatic predictors though, the *RMSE* value of *WALK* is usually higher than the corresponding *RMSE* of the SWALLOW II formulations, indicating that the favorable parsimony score of *WALK* is likely driven by its lower number of stochastic nodes more than its goodness of fit.

[32] Despite the complexity entailed by the use of time-variant export and/or attenuation coefficients, the SWALLOW II formulations were generally found to be more parsimonious than the *MCMC* SWALLOW I configuration and were comparable to the *WALK* SWALLOW I configuration. For total phosphorus, when the export (β_1, β_2) or stream attenuation coefficients (k_{s1}, k_{s2}) were allowed to vary with time (SWALLOW II), they provided comparatively better results over the static *MCMC* SWALLOW I configuration. The statistical formulation that allowed both the export and stream attenuation coefficients to vary was not supported by the *DIC*. While our data set consists of 102 measurements, the SWALLOW II statistical formulation sequentially fits each year with six or fewer points. Yet, although the parameterization of a particular year is conditional upon the information contained in the preceding ones (see equation (12)), it is still likely that allowing four parameters to vary in time is simply too complex for our data set, despite the likelihood that the processes represented by both export coefficients and stream attenuation coefficients vary annually. SWALLOW II model realizations that included the precipitation variability or potential evapotranspiration as temporal predictors were typically characterized by minor improvement of their *DIC* values relative to those derived from the consideration of total precipitation alone, suggesting that temporal variability of the SPARROW model parameters may be sufficient to describe the interannual nutrient loading variability.

[33] We include quantile-quantile and autocorrelation plots of posterior mean residuals in Figures 2 and 3 in order to assess the likelihood assumptions of normality and temporal independence of residuals made by all formulations except *WALK*. We present residuals from the formulations *MCMC - Pred1* and *MCMC - Pred1 - k_{sDYN}*, the most parsimonious SWALLOW I and SWALLOW II formulations. In accordance with our likelihood assumptions, the two components of the likelihood function (equation (9)) are assessed individually. Generally, the quantile-quantile plots show that the residual distributions were centered around the 1:1 line, although the residuals of measured from estimated "true" loads are characterized by somewhat leptokurtic patterns. Interestingly, the latter deviation patterns from the normality assumption were mainly associated with the substantial uncertainty characterizing the loading estimates from Redhill Creek (see auxiliary material Figures S1–S5). The two significant deviations at the lower range of the residuals of "true" from modeled load for SWALLOW I formulation *MCMC - Pred1* both correspond to the year 1999, which as we detail in section 3.2.2 was characterized by significantly different parameter values by the SWALLOW II framework. The spatially averaged residuals of the SWALLOW I formulation are relatively independent in time, while the SWALLOW II spatially averaged residuals between "true" and modeled loads manifest some dependence on time (Figure 3, bottom-right). Interestingly, the negative correlation coefficient suggests an oscillatory pattern of the residuals, instead of the expected grouping of over and under predictions with each other, which a positive correlation would indicate. A time series plot showed oscillatory behavior of the residuals in the final 3 years of the study. We omitted these 3 years and calculated a lag-1 correlation coefficient of only −0.35, which was well below the critical correlation coefficient of $\pm 2/\sqrt{22} = 0.43$, expected from a random process generating 22 time steps of data. The three omitted years correspond to a time period when only three of the stations were active, the sparsest period of our data record.

3.2. Posterior Parameter Distributions

[34] Table 4 shows the differences of the posterior parameter means and standard deviations among the various models when considering the total precipitation as the sole temporal predictor. The reported values of the time-varying parameters are averages of the mean and standard deviation values across all years examined. The parameter distributions are generally consistent across the formulations and a careful inspection of their values offers insights into the watershed functioning. We found that the consideration of the proportion of each reach covered by wetlands led to well-identified delivery coefficients (α). For a reach with aerial wetland coverage of 6%, the mean of our data set, the delivery coefficient values predict stream deliveries of about 57% of the total phosphorus export predicted by the corresponding coefficients. The total phosphorus export coefficients from agriculture (β_1) and urban land (β_2) were well-identified and broadly in agreement with previous SPARROW applications [*Alexander et al.*, 2002; *García et al.*, 2011].

[35] The reservoir (k_r) and stream (k_s) attenuation coefficients are generally in agreement with previous SPARROW applications [*Alexander et al.*, 2002; *García et al.*, 2011]. The stream attenuation rates were higher for smaller (first and second order) streams than for larger (third and higher order) streams, reflecting the greater contact of water and streambed as well as the longer hydraulic residence time in smaller streams [*Stream Solute Workshop*, 1990]. Our parameter results indicate that on average around 16% of phosphorus is lost per kilometer of small

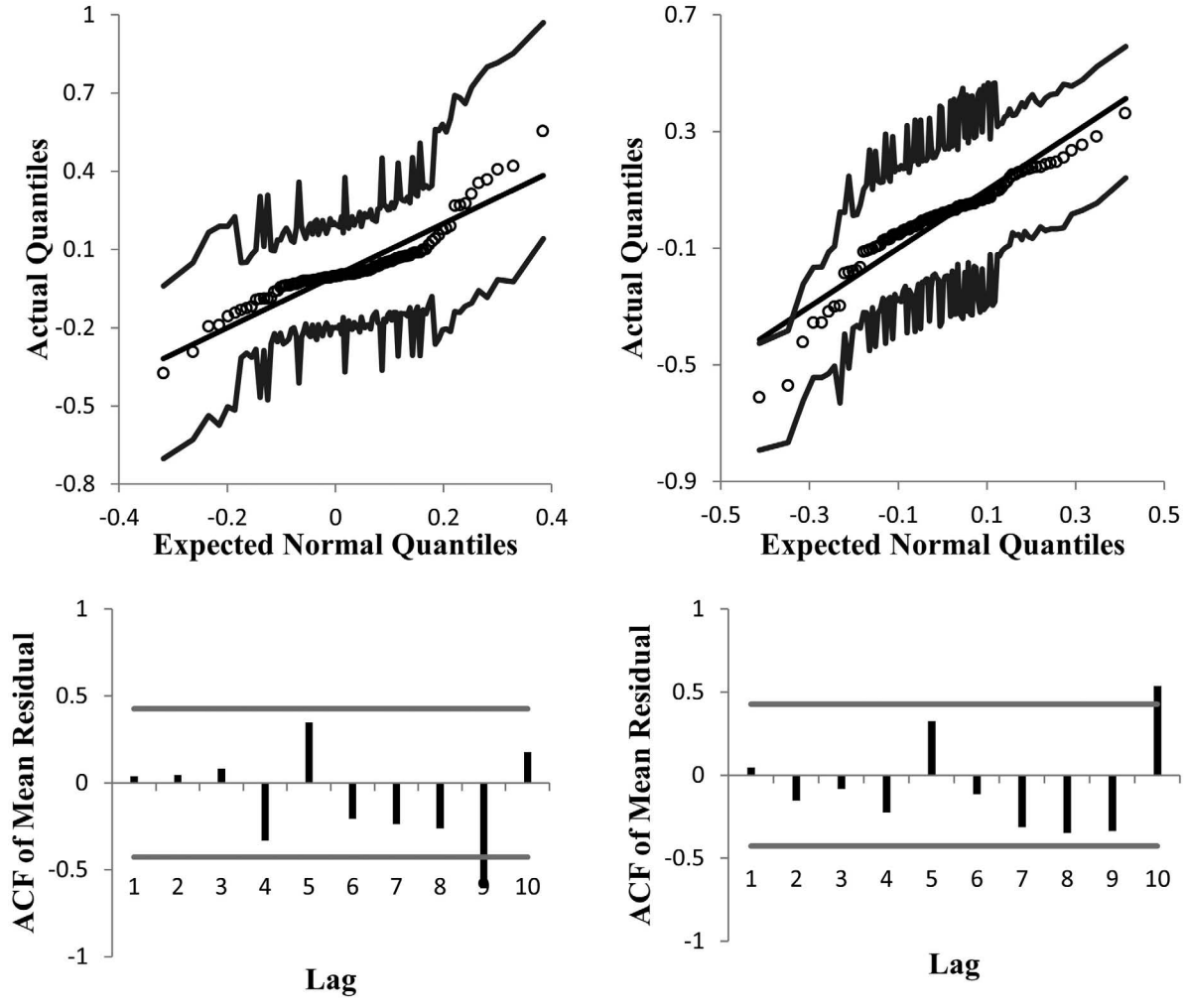


Figure 2. (top) Quantile-quantile plots and (bottom) autocorrelation function plots for SWALLOW I formulation *MCMC - Pred1*. Autocorrelation plots show average of residuals across stations. (left) Residuals of measured from estimated "true" load ($Y_{i,t} - Load_{i,t}$, see equation (7)) and (right) residuals of "true" from modeled load ($Load_{i,t} - \mu_{i,t} + W_{v,t}\gamma_v$, see equation (8)). Circles represent posterior mean residuals, gray lines the 95% credible interval, and black lines the 1:1 line. Gray lines in the autocorrelation plots represent 95% confidence interval for correlation given sample size.

stream transit and only 4% per kilometer of large stream transit. The precipitation coefficient (γ_1) has a value of roughly 0.3 and is well identified for most of the formulations examined. A positive value of this coefficient indicates that greater precipitation results in greater loading. We also note that no significant relationships were found between the time series of the dynamic parameters and the corresponding annual precipitation inputs when the precipitation was used as a covariate ($r^2 < 0.2$). On a final note, the SWALLOW II formulations were generally characterized by much lower model structural error (σ) than their SWALLOW I counterparts, reinforcing the model improvement with the dynamic watershed parameters.

[36] An interesting systematic effect was observed with the parameter estimates of the SWALLOW II formulation that were allowed to vary with time. While the posterior parameter distributions were fairly consistent across the different models examined, the SWALLOW II formulations resulted in posterior mean values for the dynamic parameters that could differ substantially compared to their static counterparts. One plausible explanation for this discrepancy may be the nature of the parameter estimation process along with the functional role of the priors with the two strategies. Namely, the SWALLOW I formulations use the literature-derived priors to update our knowledge about the average value of the different parameters for the entire time period, while the SWALLOW II formulations with sequential parameter estimation use the prior information solely for the first year, after which the estimate for the previous year supplies the most likely value for the next year's prior.

3.2.1. Effects of Temporal Predictors on Parameter Values

[37] The posterior parameter means and standard deviations for the formulation that considers dynamic stream attenuation coefficients (k_{s1}, k_{s2}) as well as the data uncertainty are provided in Table 5. The total potential evapotranspiration is a poor predictor of loading and its

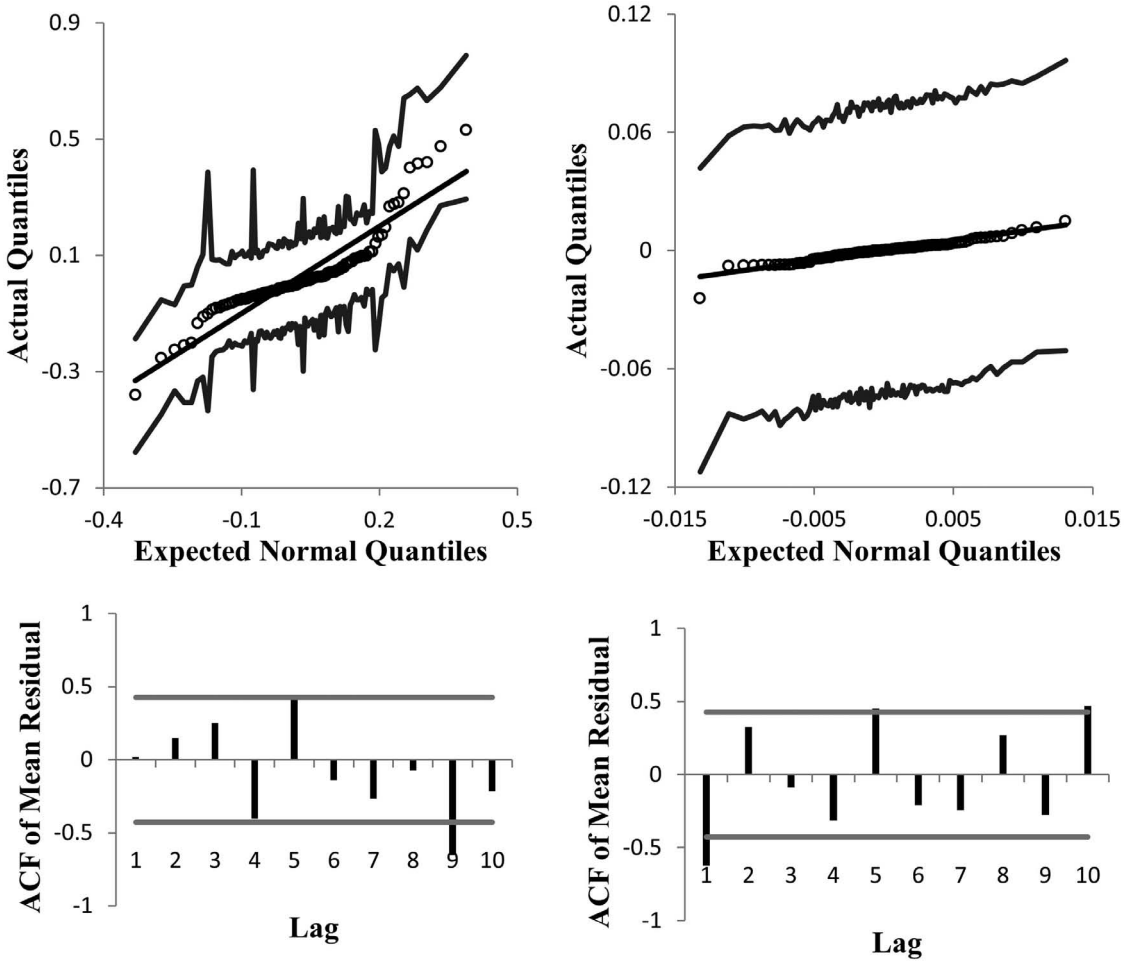


Figure 3. (top) Quantile-quantile plots and (bottom) autocorrelation function plots for SWALLOW II *MCMC - k_{sDYN}* formulation with the *Pred1* climate forcing complexity. Autocorrelation plots show average of residuals across stations. (left) Residuals of measured from estimated "true" load ($Y_{i,t} - Load_{i,t}$, see equation (7)) and right panels represents residuals of "true" from modeled load ($Load_{i,t} - \mu_{i,t} + W_{v,t}\gamma_v$, see equation (8)). Circles represent posterior mean residuals, gray lines the 95% credible interval, and black lines the 1:1 line. Gray lines in the autocorrelation plots represent 95% confidence interval for correlation given sample size.

coefficient was not well identified. This result was observed even with the much simpler SWALLOW I formulations. Yet, it is unclear whether the lack of support for potential evapotranspiration as a predictor of annual loading stems from its weak causal link with the nutrient export in Hamilton Harbor or its inability to appreciably capture the dynamics of actual evapotranspiration. While the parameter estimates are fairly consistent among the three levels of climate forcing complexity, the largest differences were found between *Pred0* and the rest of the realizations that considered temporal predictors. Adding temporal predictors tended to increase the importance of wetlands in modulating delivery to streams and decreased both the export coefficients and the small-stream attenuation coefficient (k_{s1}). We also note that no significant relationship exists between the annual k_s estimates and annual precipitation for the *Pred1* realizations ($r^2 < 0.05$), whereas the annual precipitation appears to be a significant predictor of the small stream attenuation coefficient ($r^2 = 0.32$, slope $= -0.6$, $p < 0.01$) with the *Pred0* realization. The latter finding highlights the tradeoffs when using forcing factors and time-varying parameters, in that the inclusion of significant forcing factors may reduce the variability of the time-varying watershed parameters.

3.2.2. Temporal Variability of Walk Errors and SWALLOW II Parameters

[38] We found plausible mechanisms to explain the interannual variability of the *WALK* correlated errors (v_t terms) as well as the posterior medians of the SWALLOW II parameters from various formulations. We first consider the v_t terms, as these empirical autocorrelated error terms encapsulate the "missing signal" from the static parameterization. Figure 4 shows that annual streamflow explains most of the variability of the v_t autocorrelated error terms of the *WALK E1 - Pred1* realization. Likewise, the v_t terms appear to covary with the large stream attenuation parameter estimates of the most parsimonious SWALLOW II - *E1* realization, *MCMC - k_{sDYN} - E1 - Pred1*. Further, the v_t terms explain the majority of the variability in the agricultural export terms in the *MCMC - k_{sDYN} - E1 - Pred0*

Table 4. Markov Chain Monte Carlo Estimates of the SWALLOW Models Parameterized With Total Phosphorus Data[a]

	SWALLOW I				SWALLOW II							
	MCMC		*WALK*		*MCMC - α_{DYN}*		*MCMC - β_{DYN}*		*MCMC - k_{sDYN}*		*MCMC - β,k_{sDYN}*	
Parameters	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
α	8.32	1.41	9.45	1.25	10.59	1.77	9.81	1.07	9.23	1.32	10.50	1.36
β_1	0.17	0.04	0.17	0.04	0.18	0.04	0.20	0.05	0.13	0.02	0.16	0.07
β_2	0.08	0.02	0.08	0.03	0.07	0.02	0.10	0.07	0.06	0.02	0.07	0.06
k_r	14.08	4.13	15.07	4.16	14.52	4.40	14.80	3.68	15.39	3.65	15.58	3.70
k_{s1}	0.19	0.07	0.18	0.07	0.17	0.06	0.19	0.05	0.13	0.07	0.13	0.11
k_{s2}	0.04	0.02	0.04	0.01	0.04	0.02	0.05	0.01	0.03	0.01	0.03	0.02
γ_1	0.35	0.05	0.29	0.08	0.35	0.05	0.39	0.31	0.29	0.08	0.12	0.31
σ	0.21	0.02	0.05	0.02	0.17	0.02	0.04	0.01	0.03	0.01	0.03	0.01
ψ			0.31	0.06								
α_σ					3.30	0.17						
$\beta_{1\sigma}$							1.64	0.08			1.64	0.08
$\beta_{2\sigma}$							1.82	0.09			1.82	0.09
$k_{s1\sigma}$									0.71	0.01	0.71	0.01
$k_{s2\sigma}$									1.00	0.02	1.00	0.02
DIC	−26.1		−80.0		−29.9		−45.2		−80.3		−37.8	
RMSE	0.26		0.14		0.23		0.10		0.14		0.11	
WRMSE	0.25		0.09		0.19		0.07		0.08		0.05	

[a]All statistical formulations refer to the *Pred1* level of climate forcing complexity. Reported values of β and k_s are averages of the mean and SD across all years of the study period. Units of β are tons P km^{-2} yr^{-1}. Units of k_s are km^{-1}. Units of k_r are m yr^{-1}.

realization ($r^2 = 0.60$) and the delivery terms in the *MCMC - α_{DYN} - E1 - Pred1* realization ($r^2 = 0.79$). On the other hand, there were no significant relationships of the v_t terms with the runoff ratio (total runoff/total flow; $r^2 < 0.15$), nor with the annual population of the Hamilton Census Metropolitan Area ($r^2 < 0.01$). Interestingly, precipitation was also not correlated with the v_t terms ($r^2 < 0.02$), so some aspect of the system related to flow other than total precipitation is being captured by the v_t terms.

[39] Figure 5 shows the stream attenuation coefficients as a function of average annual streamflow measured at Grindstone Creek, the largest watershed with an unmanaged flow regime. Little of the variability of the small-stream attenuation estimates can be explained by the flows at Grindstone Creek. This counterintuitive result is likely due to the deficiency of the calibration data set in headwater sites, as further described in section 3.4. More than half of the variability of the large-stream attenuation is explained by the average streamflow. This suggests that the attenuation parameter values could partially compensate for the lack of information about the rainfall-runoff process in the model. We also note that the lower attenuation values during periods of higher flow are plausible and in agreement with previous theoretical and empirical work on stream ecology, as the biotic (uptake) and abiotic (settling) processes responsible for attenuation have much less time to exert control on the nutrient load en route to the receiving water body when the streamflow rate is higher [*Stream Solute Workshop*, 1990; *Donner et al.*, 2004; *Basu et al.*, 2011]. The emergence of this pattern from an empirical model is a very interesting result. Figure 6 presents time series plots of the posterior values of the stream attenuation coefficients for one SWALLOW I realization (*MCMC - Pred1 - E1*) and one SWALLOW II realization (*MCMC - k_{sDYN} Pred1 - E1*). The spike in large-stream attenuation in 1999 corresponds with the year of the lowest average and maximum flows during the study for Grindstone (station 3) and Spencer Creeks (station 6) and the third (fourth) lowest average (maximum) streamflow values for Redhill Creek (station 2). As previously mentioned, these two formulations have different relationships to the literature prior, and therefore it is unlikely to obtain complete agreement of the resulting parameterizations. However, it is clear that there is significant interannual variability of the stream attenuation coefficients. This variability is important to take into account when a temporally static SPARROW implementation is used to estimate the locations of nutrient source areas, as discussed in section 3.3.

Table 5. Markov Chain Monte Carlo Estimates of the *MCMC - k_{sDYN}* Formulation Parameterized With Total Phosphorus Data Across Different Levels of Climate Forcing Complexity[a]

	Pred0		*Pred1*		*Pred2*	
Parameters	Mean	SD	Mean	SD	Mean	SD
α	8.73	0.88	9.23	1.32	9.17	1.28
β_1	0.16	0.02	0.13	0.02	0.13	0.02
β_2	0.08	0.02	0.06	0.02	0.06	0.02
k_r	15.94	3.77	15.39	3.65	15.32	3.56
k_{s1}	0.19	0.07	0.13	0.07	0.13	0.07
k_{s2}	0.03	0.02	0.03	0.01	0.03	0.02
γ_1			0.29	0.08	0.32	0.11
γ_2					0.04	0.10
σ	0.03	0.01	0.03	0.01	0.04	0.01
$k_{s1\sigma}$	0.71	0.01	0.71	0.01	0.71	0.01
$k_{s2\sigma}$	1.00	0.02	1.00	0.02	1.00	0.02
DIC	−75.0		−77.0		−76.6	
RMSE	0.14		0.14		0.15	
WRMSE	0.07		0.08		0.08	

[a]Reported values of k_s are averages of the mean and SD across all years of the study period.

3.3. Spatio-Temporal Identification of Source Areas

[40] The spatially distributed nature of the SPARROW model offers estimates of the sources and movement of contaminant masses within the basin, and the dynamic

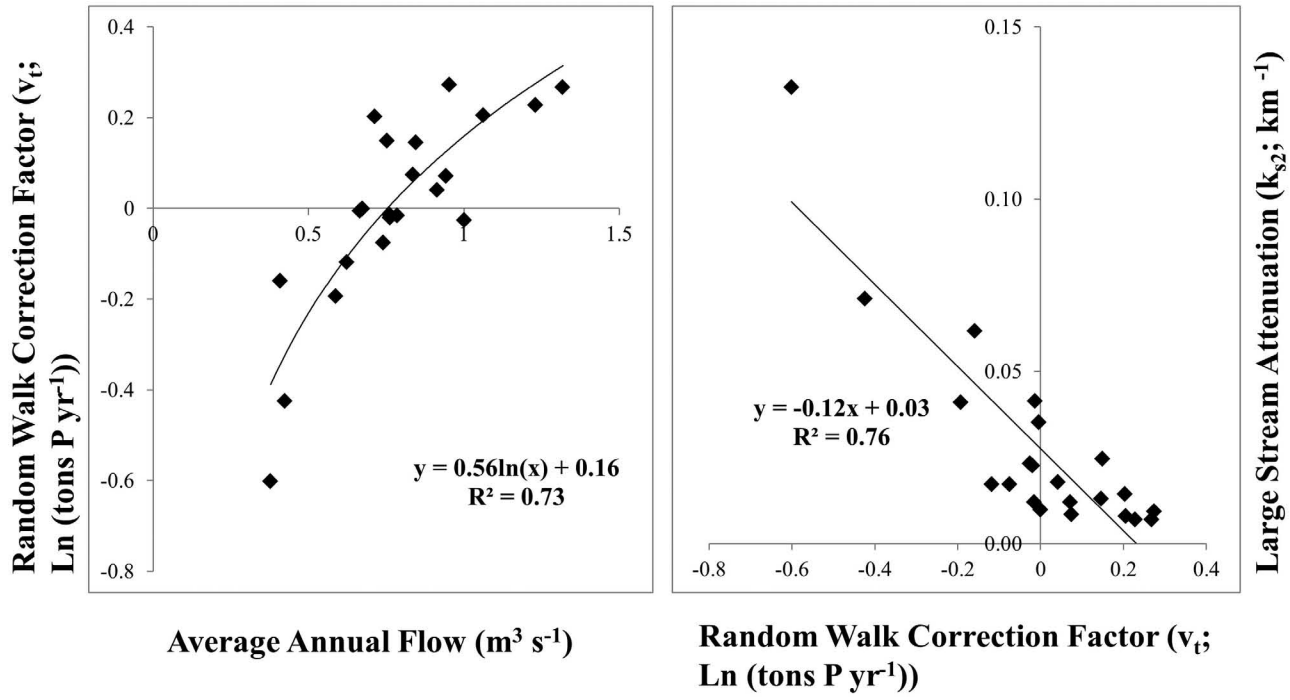


Figure 4. Scatterplots of random walk correction factor (v_t), average annual flow, and large stream attenuation (k_{s2}) for total phosphorus. Calculations were carried out using the *MCMC - k_{sDYN}* formulation with the *Pred1* climate forcing complexity.

augmentation allows this analysis to be extended in time to gain an understanding of how contaminant source and sink processes and source areas change through time. Of course, the inference drawn regarding the temporal variability of source areas in this particular exercise is subject to the ability of our calibration data set to represent temporal trends in loading. As we document in section 2.5.2, we may be underestimating the true interannual variability of our load estimates, and thus an exploration into how the annual estimates of loading propagate through the model to estimate the year-to-year contributions of source areas can offer insights into the credibility of any modeling exercise that aims to accommodate the interannual variability in a watershed context. We used the posterior parameter distributions from the *MCMC - k_{sDYN} - E1 - Pred1* realizations to estimate annual basin loads and source areas. Estimated basin load of total phosphorus ranged from 6.6 ± 2.2 to 18.1 ± 4.7 tons per year with a mean of 11.0 ± 3.3 t (the errors are in units of 1 standard deviation). While we recognize that these whole-basin load estimates are subject to the caveat of applying the model coefficients to areas smaller than the calibration subwatersheds, it should be noted that less than 15% of the total basin area falls into this category. The annual precipitation alone explained a substantial portion of the temporal variability of whole basin estimates of total phosphorus ($r^2 = 0.61, p < 0.01$).

[41] Our year-specific estimates of watershed parameters offer insights into the nutrient delivery in the Harbor for each year in addition to the static estimates typically made with SPARROW. Figure 7 shows the spatial and temporal variability of total phosphorus yield delivered to Hamilton Harbor at both the subwatershed and the reach scale. The subwatershed scale maps show the importance of proximity to the Harbor as an important factor in determining the load levels, but the reach scale maps reveal that proximity to the large (third order and higher) streams is also a significant predictor of high areal delivery, likely because the small-stream attenuation coefficients were consistently higher than the large-stream attenuation coefficients. The coefficient of variability of interannual phosphorus delivery appears to increase upstream from the Harbor, where the effect of the variability of the stream coefficients is the highest. Figure 8 presents the estimated per area deliveries at the subwatershed and reach scale for the years 1999 and 2006, i.e., the years of the highest and lowest values of large stream attenuation (see k_{s2} values in Figure 6). It is clear that the temporal variability of the watershed parameters affects the spatial variability of estimated watershed per area deliveries for total phosphorus. The estimated whole-basin delivery of total phosphorus in 1999 was 6.7 ± 2.1 tons and in 2006 was 15.0 ± 4.1 tons.

3.4. Jack-Knife Model Evaluation

[42] While the time for space substitution allowed us to parameterize the model, the spatial sampling intensity of the calibration data set was admittedly low. To evaluate whether our data contain sufficient information to impartially draw inferences about the relative contribution of different source areas as well as the interplay between temporal and spatial variability, we performed a jackknife experiment in which the most parsimonious model realization (*MCMC - k_{sDYN} - Pred1 - E1*) was parameterized against a set of data without the load measurements from one of the six stations. The same exercise was repeated six times, each time omitting a different station. Our hypothesis was that if the calibration data set does not have enough

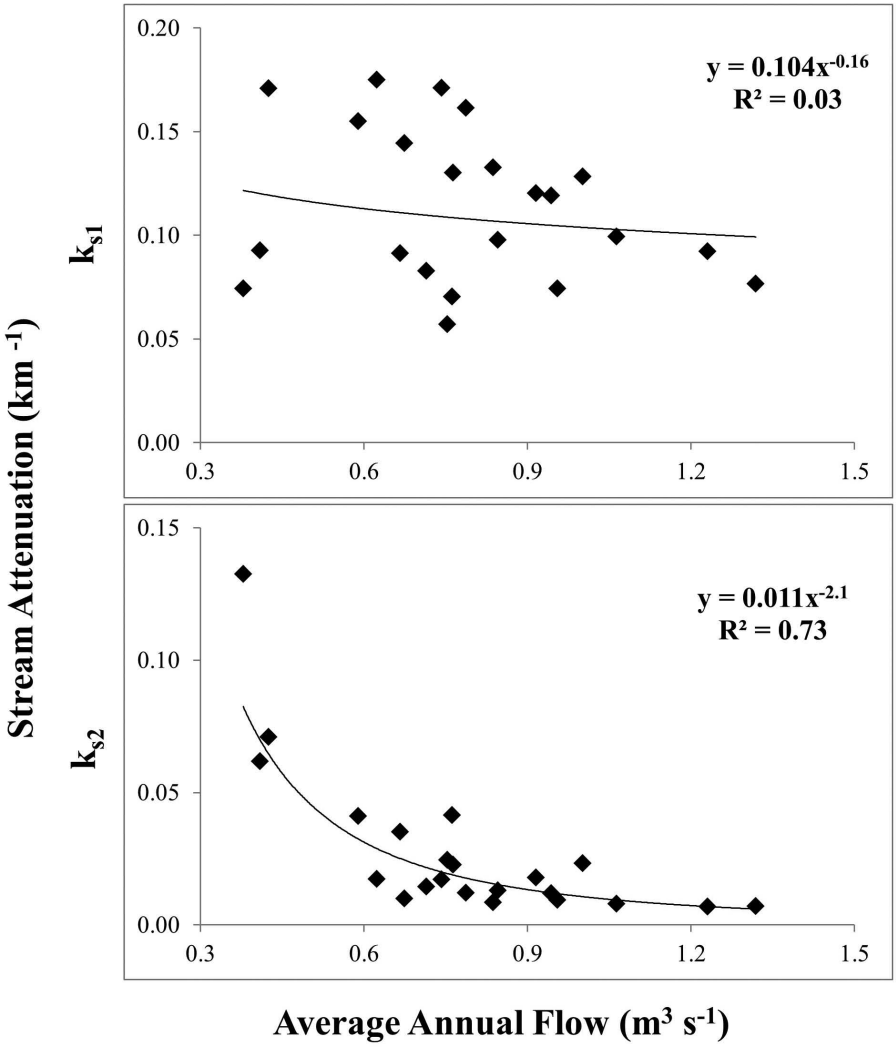


Figure 5. Scatterplots of yearly stream attenuation coefficient (k_{s1} refers to attenuation in first and second-order streams, k_{s2} to attenuation in third and higher order streams) values Total Phosphorus model against average streamflow. These results correspond to the *MCMC* - k_{sDYN} formulation with the *Pred1* climate forcing complexity.

spatial detail, the parameter values should change significantly when the data from any particular station are omitted. This is of course subject to the caveat that we do not learn how well the model performs in areas which are not explicitly represented in the calibration data set, e.g., small streams with drainage basins less than 25 km^2, which include headwater areas and areas along the shore of the Harbor. The total phosphorus parameter posteriors were fairly consistent across the moving window (Table 6). The station omission realization with the least correspondence to the posteriors obtained with the full data set is the one without the headwater station for Spencer Creek (station 4), which is the only headwater station of the entire study catchment (Figure 1). The largest discrepancy of phosphorus export coefficients occurs when one of the two urban stations is omitted (station 1).

[43] We also used the jackknife experiment to gauge the strength of the space for time substitution. We wanted to ascertain whether the model was able to reproduce the values of the omitted stations. More specifically, from column 1 in Table 6, we took the predictions of the (logged) load at station 1. From column 2, we took the loads predicted at station 2, and so on. These were used as the independent variables in a regression with the measured (logged) loading data. This regression was significant ($p < 0.001$, $r^2 = 0.91$, slope = 0.85). Even with less information in space, the model is able to reasonably predict loads at locations not used in calibration, provided those locations are comparable to those included in the calibration data set.

4. Discussion

[44] In this paper, we presented a methodological framework that aims to facilitate SPARROW application on scales of relevance to local management and to study the relative contribution of loading source areas over time. Our analysis offers a new perspective into the SPARROW modeling practice by shifting the focus toward an examination of the interplay between time and space. We adopted a repeated measures approach that enables the model to be

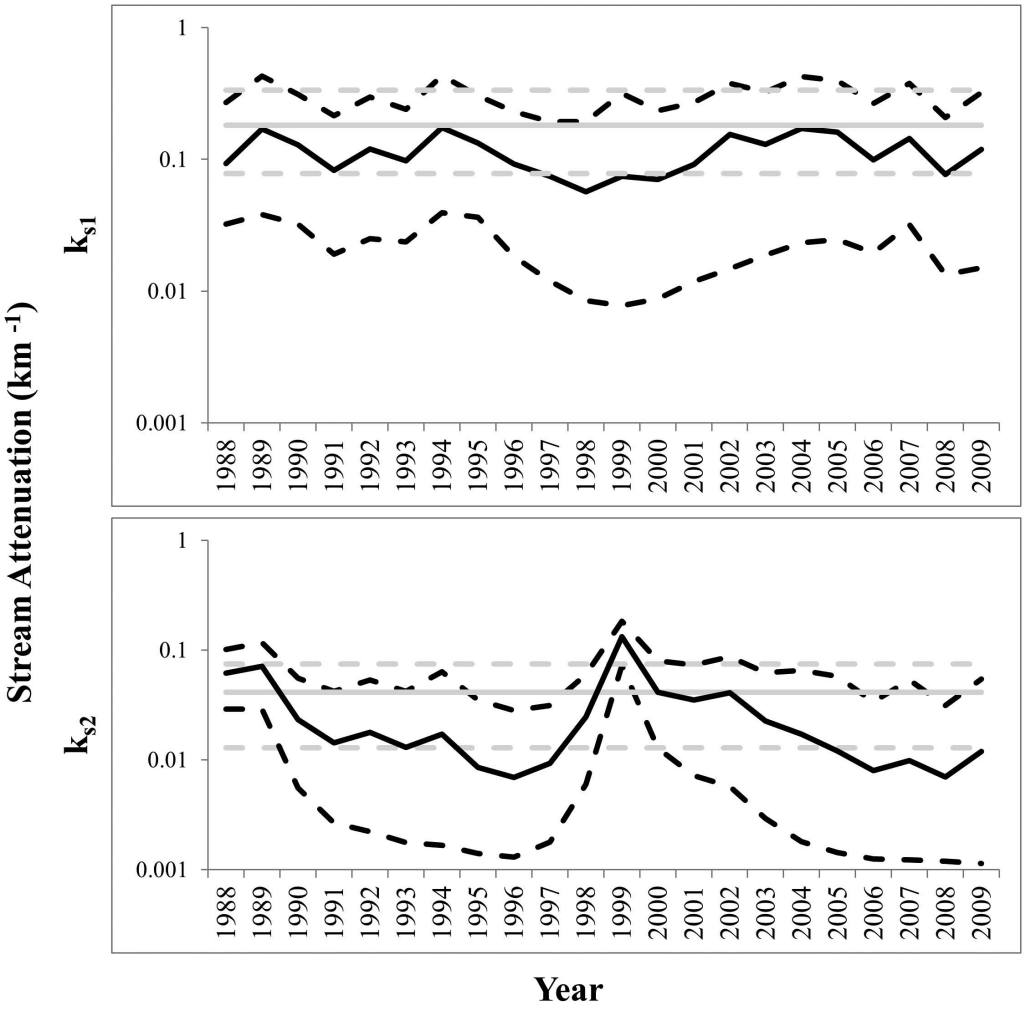


Figure 6. Time series plots of k_s for the total phosphorus model. Black and gray lines refer to parameters from the *MCMC* - k_{sDYN} and *MCMC* formulation with the *Pred1* climate forcing complexity. Dashed lines indicate upper and lower limits of the 95% credible interval, solid lines indicate the medians of the distributions.

parameterized in relatively small areas with comparatively few monitoring sites, and subsequently examined two strategies to accommodate temporal variability of the nutrient loading estimates. The first approach (SWALLOW I) assumes that the SPARROW model provides a time-invariant baseline estimate of watershed loading, while weather-related forcing factors describe the temporal variability. The second one (SWALLOW II) assumes that the processes described by the SPARROW model are dynamic and are further modulated by temporal predictors. We integrated this framework with Bayesian calibration schemes, founded upon informative prior parameter distributions and statistical formulations that can explicitly consider the data uncertainty and/or the temporal structure of model residuals. Our results show that the SWALLOW framework is able to accommodate the interannual variability of the nutrient loading estimates. Importantly, the dynamic SWALLOW II approach appears to effectively balance between performance and complexity. We also found that the temporal changes of SPARROW model parameters can be significant, thereby driving year-to-year variability of model-estimated total phosphorus source areas. The remainder of the discussion is structured to address the factors comprising the study design (statistical formulation and temporal predictors), the role of the spatial sampling protocol, and a final section examines the plausibility of the model parameterization obtained.

4.1. Role of Statistical Formulations and Temporal Predictors

[45] Previous research has considered time-varying parameters in the context of conceptual rainfall-runoff models [e.g., *Reichert and Mieleitner*, 2009] as well as models of other environmental systems (e.g., a managed pond, *Lin and Beck* [2007]). While some of these efforts have significantly improved our predictive capacity, the resulting time series of parameter values does not always give clear ideas about the structural model deficiencies. In this study, we provided two pieces of evidence that corroborate the mechanistic basis of our time-varying stream attenuation coefficients. First, we showed that the annual stream attenuation estimates of phosphorus are inversely proportional to the error terms of our *WALK* formulation, suggesting that the

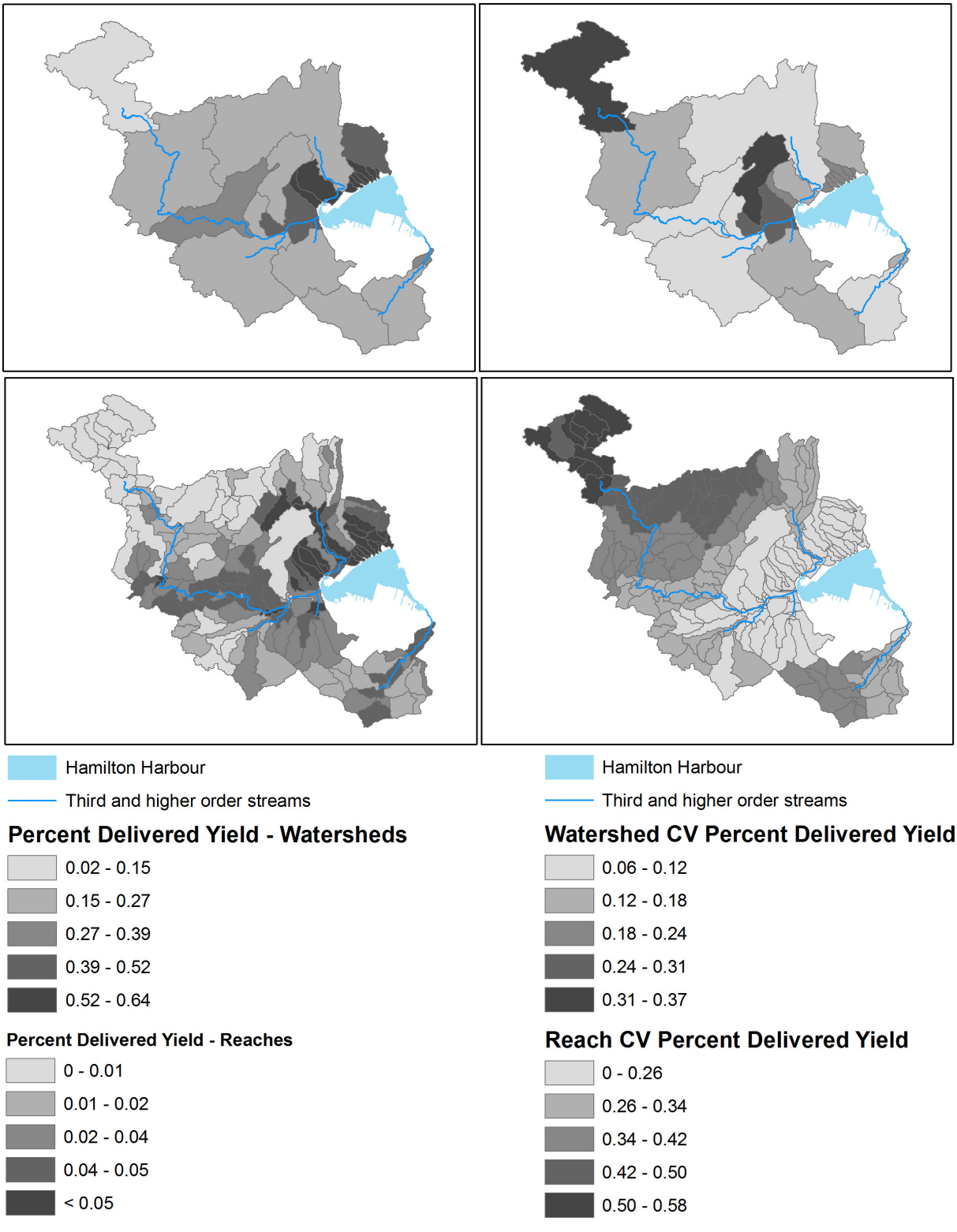


Figure 7. Spatial variability of total phosphorus delivered yield at the (top) watershed and (bottom) reach scales. (left) Mean percent contribution of total load to the Harbor for all years per square kilometer. (right) The coefficients of variability of mean percent contribution across years. These results correspond to the *MCMC* - k_{sDYN} formulation with the *Pred1* climate forcing complexity.

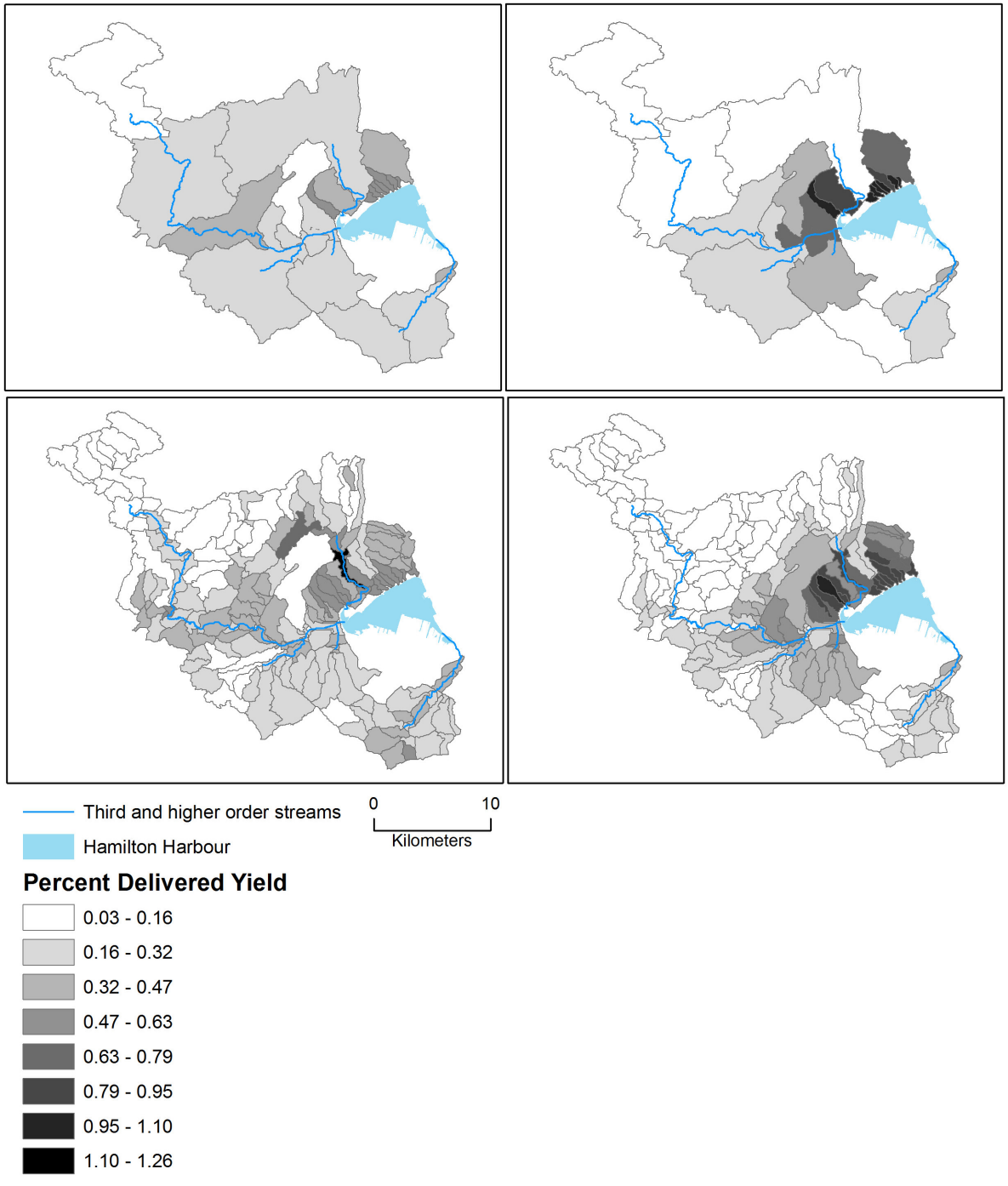


Figure 8. Spatial-temporal variability of total phosphorus delivered yield at the (top) watershed and (bottom) reach scales. (left) The percent contribution of total load to the Harbor per square kilometer for 2006, the year with the lowest value of k_{s2}. (right) The percent contribution of total load to the Harbor per square kilometer for 1999, the year with the highest value of k_{s2}. These results correspond to the *MCMC* - k_{sDYN} formulation with the *Pred1* climate forcing complexity.

assumption of a static attenuation parameter may be responsible for much of the error variability [*Sadraddini et al.*, 2011a]. Second, consistent with empirical findings and ecological theory [*Stream Solute Workshop*, 1990], model estimated (log transformed) stream attenuation is inversely proportional to the (log transformed) mean annual flow. The latter finding may partly indicate that the values of in-stream attenuation compensate for the structural inadequacy of the SWALLOW I model in describing the transformation of precipitation into runoff. Earlier work postulated a resemblance between time-varying parameters and mean-reverting statistical processes [*Riechert and Mieleitner*, 2009; *Tomassini et al.*, 2009], whereas we here adopt a formulation akin to that used in dynamic linear modeling

Table 6. Jackknife Experiment-Markov Chain Monte Carlo Estimates of the *MCMC - k_{sDYN}* Formulation With the *Pred1* Climate Forcing Complexity Parameterized With Total Phosphorus Data[a]

	Station Omitted													
	0		1		2		3		4		5		6	
Parameters	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
α	9.23	1.32	8.33	1.14	8.74	1.21	9.72	1.50	6.12	2.18	9.04	1.47	9.40	1.27
β_1	0.13	0.02	0.14	0.02	0.13	0.02	0.12	0.02	0.09	0.02	0.13	0.02	0.13	0.02
β_2	0.06	0.02	0.06	0.02	0.08	0.03	0.05	0.02	0.05	0.01	0.06	0.02	0.06	0.02
k_r	15.39	3.65	15.97	3.79	15.84	3.69	15.43	3.65	10.90	3.14	13.95	3.99	16.32	4.96
k_{s1}	0.13	0.07	0.18	0.10	0.15	0.08	0.10	0.07	0.07	0.05	0.12	0.08	0.11	0.07
k_{s2}	0.03	0.01	0.02	0.01	0.03	0.01	0.03	0.02	0.03	0.01	0.03	0.02	0.03	0.02
γ_1	0.29	0.08	0.30	0.09	0.29	0.08	0.30	0.07	0.29	0.06	0.27	0.08	0.26	0.08
σ	0.03	0.01	0.03	0.01	0.04	0.01	0.04	0.01	0.04	0.01	0.04	0.02	0.04	0.02
$k_{s1\sigma}$	0.71	0.01	0.71	0.01	0.71	0.01	0.71	0.01	0.71	0.01	0.71	0.01	0.71	0.01
$k_{s2\sigma}$	1.00	0.02	1.00	0.02	1.00	0.03	1.00	0.02	1.00	0.02	1.00	0.02	1.00	0.02

[a]Each column indicates which station was omitted. The first column is taken from Table 4 and is intended for comparison purposes.

[*Sadraddini et al.*, 2011a, 2011b]. By contrast to a mean-reverting process, intended to control the uncertainty of posterior parameter values, our approach led to a minor broadening of the 95% credible intervals of the posterior distributions of the dynamic parameters (Figure 6). Yet, our dynamic approach represents a way to relax the assumption of stationarity that a mean-reverting process assumes, and therefore depicts systematic trends that cannot be otherwise accommodated, such as the effect of in-stream nutrient attenuation on the interannual variability of the nutrient source areas of the Harbor.

[46] Our study identified the total annual precipitation as the key predictor variable to accommodate the interannual variability of the total nutrient loading into Hamilton Harbor. In particular, a preliminary exploratory analysis showed that precipitation accounts for a substantial portion of the variability of the log-transformed phosphorus loads of Redhill, Grindstone, and Spencer Creeks ($r^2 = 0.41$, $p < 0.001$ for phosphorus). Yet, should the SWALLOW model be applied elsewhere, we recommend that a variety of temporal predictors be examined. In urban areas, predictors related to population or population density could augment the land cover data typically used to infer impervious surface cover. Categorical variables related to local management practices, such as upgrades to storm water management systems or passage of stricter land use controls in agricultural systems, could also be incorporated to model their effects on watershed functioning. Further, if the SWALLOW framework is applied to broader spatial scales, the spatial variability of the temporal predictors may also need to be taken into account, i.e., our W matrix could vary in space as well as in time and each entry would correspond to the value of a predictor at a specific subwatershed or reach for a particular year.

4.2. Role of Watershed Spatial Sampling Protocol on the Model Parameterization

[47] The modeling of phosphorus was resilient to the station omissions of the jackknife experiment. By far the greatest discrepancy occurred when station 4 was omitted, which was the only nonurban station draining a predominantly headwater catchment. Station 4 also drains the subwatershed with the largest density of wetlands, so it is no surprise that the delivery coefficient and the small-stream attenuation parameter vary the most with respect to the full data set when station 4 is omitted. Also surprisingly, the jackknife experiment showed that when information from one station is omitted, the model is able to reasonably reproduce the posterior median loads from the omitted stations. In other words, the model is able to borrow enough strength from the included sites to model the load at the omitted site. It is not clear that this would be the case if more than one station were to be omitted, but this result does bolster the strength of the argument that the information in space we have is able to produce some (albeit uncertain) inference. Of course, the loads at the "missing stations" as estimated by the model when calibrated to only five stations are subject to more uncertainty than when the same loads are estimated over the entire data set. The average value of the posterior standard deviation of the log-transformed loads was 0.11, while for the loads estimated by the model calibrated to only five stations of data it was 0.14. Another caveat of the present exercise is that we were not able to condition model parameterization upon areas not explicitly represented in the calibration data set, such as small streams with drainage basins less than 25 km^2, which include many headwater areas and areas along the shore of the Harbor.

4.3. Plausibility of the Parameterization

[48] We compare the understanding of the functioning of the Hamilton Harbor watershed obtained from our modeling work with results from the SPARROW literature as well as other empirical evidence from the study area. This comparison will allow us to gauge the plausibility of the model parameterization, while enriching our understanding of the functioning of this site of intense management interest. Our model parameterization suggests that agricultural land uses result in higher phosphorus export than urban land uses. This is consistent with some previous SPARROW studies [*Moore et al.*, 2004] and empirical work [*Law et al.*, 2004; *Soldat and Petrovic*, 2008; *Soldat et al.*, 2009]. However, other SPARROW applications [*Alexander et al.*, 2004; *García et al.*, 2011] and empirical literature [*Beaulac and Reckhow*, 1982] have found the opposite – urban land exports more phosphorus than agricultural land. Some studies in Southern Ontario tend to agree with the latter assertion [e.g., *Winter and Duthie*, 2000]. Both agricultural and urban nutrient export fluxes are highly variable

and contingent upon a number of regulatory factors, including soil type, urban storm water management, agricultural intensity and conservation practices [*Beaulac and Reckhow*, 1982]. Our estimates of urban phosphorus export are slightly higher than estimates obtained from the Great Lakes region in the United States [*Robertson and Saad*, 2011], but comparable to those from the American Southeast [*García et al.*, 2011].

[49] Following empirical work, previous SPARROW applications have generally used soil parameters as delivery variables for phosphorus [*Beaulac and Reckhow*, 1982]. Wetlands have been shown to attenuate the loadings of phosphorus through processes such as particle settling, denitrification, and biotic uptake [*Reddy et al.*, 1999; *Krieger*, 2003], but have not been explicitly included in SPARROW models as a delivery factor. While some SPARROW model applications have considered soil properties that would implicitly address wetlands, such as soil organic matter and soil pH [*García et al.*, 2011], these factors would describe delivery from both wetland and upland areas and may not reflect processes unique to wetlands [*Reddy et al.*, 1999; *Krieger* 2003]. Working in the Laurentian Great Lakes, *Robertson and Saad* [2011] included a land use class intended to describe phosphorus from background sources, into which they combined wetlands, forest and scrubland. Our present results suggest that wetlands do not necessarily act as a source but may also as a sink for phosphorus at the landscape scale. In this regard, one of the lessons learned from our analysis is that SPARROW applications should consider wetland coverage as a candidate delivery variable.

[50] While we did not allow reservoir attenuation processes to vary through time, they were nonetheless an important aspect of the spatial variability of phosphorus delivery to the Harbor. The posterior mean settling velocity (k_r) for total phosphorus was about 15.4 m yr^{-1} with a 95% credible interval of 9.2–23.4 m yr^{-1}. The total phosphorus settling velocity is close to that of 14.3 m yr^{-1} obtained by *Alexander et al.* [2004] for the continental United States, but substantially greater than the value of 4.8 m yr^{-1} obtained by *Robertson and Saad* [2011] for the United States' Laurentian Great Lakes and Midwest. Notably, empirical research conducted in Cootes Paradise, a coastal wetland draining about half of Hamilton Harbor's basin, corroborates our phosphorus settling velocity results. *Prescott and Tsanis* [1997] review the net settling velocity estimates for Cootes Paradise and report values ranging from 10 to 16 m yr^{-1}. We used our posterior settling velocities to estimate 95% credible intervals for the retention of phosphorus for Cootes Paradise. Total phosphorus retention ranged from 16% to 36%, with a median of 25%. These values are in agreement with those reported by *Krieger* [2003] for a coastal wetland in the Lake Erie basin. This implies that Cootes Paradise plays a major part in reducing nutrient loading to Hamilton Harbor and, not surprisingly, the water quality in Cootes Paradise itself is degraded [*Prescott and Tsanis*, 1997].

[51] The posterior means of small-stream phosphorus attenuation were somewhat lower than previous SPARROW work in New Zealand [*Alexander et al.*, 2002], but nonetheless reasonably commensurate. Although our separation of stream classes was based on *Strahler*'s [1952] stream order and not the discharge or travel time, our results are consistent with other SPARROW studies in that the values of small stream attenuation (k_{s1}) were smaller than those for large stream attenuation (k_{s2}), reflecting the higher attenuation rates of smaller streams [*Stream Solute Workshop*, 1990; *Alexander et al.*, 2002, Figure 7]. It should again be stressed that our database is deficient in headwater sampling sites, so our estimates of small-stream attenuation and its variation in time are subject to substantial uncertainty. Nonetheless, the SWALLOW II framework is a promising one for accommodating interannual variability into SPARROW models.

[52] Estimated large stream attenuation coefficients proved to be quite variable in time for most statistical formulations. The mechanisms that modulate the variability of nutrient attenuation across stream size are fairly well established in the literature. They generally refer to the tighter coupling of smaller streams with their streambeds, whereby biological and chemical removal processes in the sediments have greater access to the nutrients in the water column [*Stream Solute Workshop*, 1990; *Alexander et al.*, 2002; *Alexander et al.*, 2004]. The longer hydraulic residence time of smaller streams also allows these processes to operate for longer times. Recent work suggests that stream stage explains the interannual variation of nutrient attenuation at a particular site over time [*Basu et al.*, 2011], implying that the coupling between the streambed and water column changes from year to year. Consistent with *Basu et al.*'s [2011] findings, we here show that the interannual variability of the average discharge, a function of stream stage, can explain more than half of the variability of stream attenuation estimates from the SPARROW model.

[53] An interesting implication of this study is that for Hamilton Harbor's basin, the interannual variability of the contribution of phosphorus source areas may be strongly affected by the capacity of stream reaches to attenuate nutrient loads (Figure 8). Empirical studies of nutrient uptake in rivers indicate significant variability of nutrient attenuation rates at annual timescales for phosphorus [*Doyle et al.*, 2003] and nitrogen [*Claessens et al.*, 2009]. *Donner et al.* [2004] found that nutrient attenuation rates varied nearly two-fold between wet and dry years in the Mississippi River, with wet years exhibiting lower attenuation. *Basu et al.* [2011] also showed an inverse relationship between stream stage and nutrient attenuation that was consistently manifested across spatial and temporal scales. This finding implies that fluctuations in stage (and discharge) may indeed affect the spatial location of significant nutrient source areas at a variety of scales and is not an artifact of the present analysis. While previous research has documented the variability of in-stream attenuation at annual timescales, the SWALLOW framework allows us to estimate how this variability impacts basin-scale nutrient source areas.

[54] In conclusion, SPARROW is a spatially distributed, empirical model that can be used to identify areas of unusually high delivery of nutrient loads to water bodies and prioritize the allocation of scarce management resources accordingly. Yet, nutrient loads, source/sink processes, and source areas are subjected to significant interannual variability, and thus a temporally static approach can oversimplify the broad range of dynamics typically experienced in a watershed context. As an alternative to employing complex, process-based models to understand the mechanisms of this variability, our SWALLOW modeling framework offers a parsimonious representation of watershed functioning through time that builds upon

the SPARROW foundation. Consistent with empirical and theoretical work, our model parameterization suggests that in-stream attenuation rates varied inversely with streamflow, which also affects the location of nutrient source areas. While we found little support for the use of time-varying export coefficients and stream attenuation coefficients, it is most likely that nutrient export and delivery to streams varies at annual timescales as well as in-stream attenuation processes. The SWALLOW II framework we present in this paper is a promising approach to arrive at a balanced depiction of the interaction of nutrient export, landscape delivery, attenuation, and climate when applied to larger datasets. By quantifying the interannual variability of nutrient delivery to the receiving water body, we believe that the modeling framework proposed can meaningfully assist long-term watershed management planning. The Bayesian nature of our approach allows the estimation of critical nutrient loads that could result in acceptable probabilities of compliance with different water quality criteria, while accounting for the different sources of uncertainty (model structure imperfection, measurement error, model input uncertainty) as well as natural system variability.

[55] On a final note, we believe that models are a worthwhile scientific activity and a sound basis for the policy-making process only if the underlying assumptions are acknowledged and impartially communicated [*Zhang and Arhonditsis*, 2008]. For example, our jackknife experiment showed that the watershed sampling protocol is deficient in headwater sampling sites. Model development is a dynamic, iterative process similar to the policy practice of adaptive management. The model parameterization/structure can be sequentially refined as new knowledge is obtained from the system, and this gradual model evolution should provide the basis for revised (and improved) management actions.

[56] **Acknowledgments.** This project has received funding support from the Ontario Ministry of the Environment (Canada Ontario grant Agreement 120,808). 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, N. K., Q. Y. Duan, and S. Sorooshian (2007), An integrated hydrologic Bayesian multimodel combination framework: Confronting input, parameter, and model structural uncertainty in hydrologic prediction, *Water Resour. Res.*, *43*, W01403, doi:10.1029/2005WR004745.

Akaike, H. (1974), A new look at the statistical model identification, *IEEE Trans. Autom. Control*, *19*(6), 716–723, doi:10.1109/TAC.1974.1100705.

Alexander, R. B., A. H. Elliott, U. Shankar, and G. B. McBride (2002), Estimating the sources and transport of nutrients in the Waikato River Basin, New Zealand, *Water Resour. Res.*, *38*(12), 1268, doi:10.1029/2001WR000878.

Alexander, R. B., R. A. Smith, and G. E. Schwarz (2004), Estimates of diffuse phosphorus sources in surface waters of the United States using a spatially referenced watershed model, *Water Sci. Technol.*, *49*(3), 1–10.

Alexander, R. B., R. A. Smith, G. E. Schwarz, E. W. Boyer, J. V. Nolan, and J. W. Brakebill (2008), Differences in phosphorus and nitrogen delivery to the Gulf of Mexico from the Mississippi River Basin, *Environ. Sci. Technol.*, *42*(3), 822–830, doi:10.1021/es0716103.

Allen, R. G., L. S. Pereira, D. Raes, and M. Smith (1998), Crop evapotranspiration: Guidelines for computing crop water requirements, *FAO Irrigation and drainage, paper 56*, Food and Agric. Org. of the United Nations, Rome, Italy.

Arhonditsis, G. B., D. Papantou, W. Zhang, G. Perhar, E. Massos, and M. Shi (2008a), Bayesian calibration of mechanistic aquatic biogeochemical models and benefits for environmental management, *J. Mar. Syst.*, *73*, 8–30, doi:10.1016/j.jmarsys.2007.07.004.

Arhonditsis, G. B., G. Perhar, W. Zhang, E. Massos, M. Shi, and A. Das (2008b), Addressing equifinality and uncertainty in eutrophication models, *Water Resour. Res.*, *44*, W01420, doi:10.1029/2007WR005862.

Azim, M. E., M. Letchumanan, A. Abu Rayash, Y. Shimoda, S. P. Bhavsar, and G. B. Arhonditsis (2011), Detection of temporal trends of alpha and gamma chlordane in Lake Erie fish communities using dynamic linear modeling, *Ecotoxicol. Environ. Saf.*, *74*(5), 1107–1121.

Balin, D., L. Hyosang, and M. Rode (2010), Is uncertain rainfall likely to greatly impact on distributed complex hydrological modeling?, *Water Resour. Res.*, *46*, W11520, doi:10.1029/2009WR007848.

Basu, N. B., P. S. C. Rao, S. E. Thompson, N. V. Loukinova, S. D. Donner, S. Ye, and M. Sivapalan (2011), Spatiotemporal averaging of in-stream solute removal dynamics, *Water Resour. Res.*, *47*, W00J06, doi:10.1029/2010WR010196.

Beaulac, M. N., and K. H. Reckhow (1982), An examination of land-use–nutrient export relationships, *Water Resour. Bull.*, *18*(6), 1013–1024.

Beck, M. B., and P. C. Young (1976), Systematic identification of DO-BOD model structure, *J. Environ. Eng. Div. Am. Soc. Civ. Eng.*, *102*, 909–927.

Borah, D. K., and M. Bera (2004), Watershed-scale hydrologic and nonpoint-source pollution models: Review of applications, *Trans. ASAE*, *47*(3), 789–803.

Brakebill, J. W., S. W. Ator, and G. E. Schwarz (2010), Sources of suspended-sediment flux in streams of the Chesapeake Bay watershed: A regional application of the SPARROW model, *J. Am. Water Resour. Assoc.*, *46*(4), 757–776, doi:10.1111/j.1752-1688.2010.00450.x.

Brooks, S. P., and A. Gelman (1998), General methods for monitoring convergence of iterative Simulations, *J. Comput. Graph. Stat.*, *7*, 434–455.

Carroll, R. J., D. Ruppert, L. A. Stefanski, and C. M. Crainiceanu (2006), *Measurement Error in Nonlinear Models: A Modern Perspective*, 455 pp., Chapman & Hall/CRC Press, Boca Raton, Fla.

Charlton, M. N., (2001), The Hamilton Harbour remedial action plan: Eutrophication, *Verh. Int. Ver. Theor. Angew. Limnol.*, *27*, 4069–4072.

Cheng, V., G. B. Arhonditsis, and M. T. Brett (2010), A revaluation of lake-phosphorus loading models using a Bayesian hierarchical framework, *Ecol. Res.*, *25*(1), 59–76, doi:10.1007/s11284-009-0630-5.

Claessens, L., C. L. Tague, L. E. Band, P. M. Groffman, and S. T. Kenworthy (2009), Hydro-ecological linkages in urbanizing watersheds: An empirical assessment of in-stream nitrate loss and evidence of saturation kinetics, *J. Geophys. Res.*, *114*, G04016, doi:10.1029/2009JG001017.

Cohn, T. A., D. L. Caulder, E. J. Gilroy, L. D. Zynjuk, and R. M. Summers (1992), The validity of a simple statistical-model for estimating fluvial constituent loads—An empirical study involving nutrient loads entering Chesapeake Bay, *Water Resour. Res.*, *28*(9), 2353–2363.

Congdon, P. (2001), *Bayesian Statistical Modeling*, Wiley, New York.

de Wit, M. J. M. (2001), Nutrient fluxes at the river basin scale. I: The PolFlow model, *Hydrol. Processes*, *15*, 743–759, doi:10.1002/hyp.175.

Donner, S. D., C. J. Kucharik, and M. Oppenheimer (2004), The influence of climate on in-stream removal of nitrogen, *Geophys. Res. Lett.*, *31*, L20509, doi:10.1029/2004GL020477.

Doyle, M. W., E. H. Stanley, and J. M. Harbor (2003), Hydrogeomorphic controls on phosphorus retention in streams, *Water Resour. Res.*, *39*(6), 1147, doi:10.1029/2003WR002038.

García, A. M., A. B. Hoos, and S. Terziotti (2011), A regional modeling framework of phosphorus sources and transport in streams of the southeastern United States, *J. Am. Water Resour. Assoc.*, *47*, 991–1010, doi:10.1111/j.1752-1688.2010.00517.x.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (2004), *Bayesian Data Analysis*, Chapman and Hall, Boca Raton, Fla.

Gilroy, E. J., R. M. Hirsch, and T. A. Cohn (1990), Mean square error of regression-based constituent transport estimates, *Water Resour. Res.*, *26*(9), 2069–2077, doi:10.1029/90WR00240.

Grizzetti, B., F. Bouraoui, G. de Marsily, and G. Bidoglio (2005), A statistical approach to estimate nitrogen sectorial contribution to total load, *Water Sci. Technol.*, *51*(3–4), 83–90.

Gudimov, A., S. Stremilov, M. Ramin, and G. B. Arhonditsis (2010), Eutrophication risk assessment in Hamilton Harbour: System analysis and evaluation of nutrient loading scenarios, *J. Great Lakes Res.*, *36*(3), 520–539, doi:10.1016/j.jglr.2010.04.001.

Gudimov, A., M. Ramin, T. Labencki, C. Wellen, M. Shelar, Y. Shimoda, D. Boyd, and G. B. Arhonditsis (2011), Predicting the response of Hamilton Harbour to the nutrient loading reductions: A modeling analysis of the "ecological unknowns," *J. Great Lakes Res.*, *37*(3), 494–506, doi:10.1016/j.jglr.2011.06.006.

Hiriart-Baer, V. P., J. Milne, and M. N. Charlton (2009), Water quality trends in Hamilton Harbour: Two decades of change in nutrients and chlorophyll *a*, *J. Great Lakes Res.*, *35*(2), 293–301, doi:10.1016/j.jglr.2008.12.007.

Kalman, R. E., (1960), A new approach to linear filtering and prediction problems, *Trans. ASME*, *82*, 35–45.

Krieger, K. A. (2003), Effectiveness of a coastal wetland in reducing pollution of a Laurentian Great Lake: Hydrology, sediment, and nutrients, *Wetlands*, *23*(4), 778–791.

Law, N., L. Band, and M. Grove (2004), Nitrogen input from residential lawn care practices in suburban watersheds in Baltimore county, MD, *J. Environ. Plann. Manage.*, *47*(5), 737–755.

Lin, Z., and M. B. Beck (2007), On the identification of model structure in hydrological and environmental systems, *Water Resour. Res.*, *43*, W02402, doi:10.1029/2005WR004796.

Lunn, D. J., A. Thomas, N. Best, and D. Spiegelhalter (2000), WinBUGS—A Bayesian modeling framework: Concepts, structure, and extensibility, *Stat. Comput.*, *10*(4), 325–337, doi:10.1023/A:1008929526011.

McMahon, G., R. B. Alexander, and S. Qian (2003), Support of total maximum daily load programs using spatially referenced regression models, *J. Water Res.*, *129*, 315–329.

Moore, R. B., C. M. Johnston, K. W. Robinson, and J. R. Deacon (2004), Estimation of total nitrogen and phosphorus in New England streams using spatially referenced regression models, *Sci. Invest. Rep. 2004-5012*, U.S. Geol. Surv., Pembroke, N.H.

Moradkhani, H., K.-L. Hsu, H. V. Gupta, and S. Sorooshian (2005), Uncertainty assessment of hydrologic model states and parameters: Sequential data assimilation using the particle filter, *Water Resour. Res.*, *41*, W05012, doi:10.1029/2004WR003604.

Prado, R., and M. West (2010), *Time Series: Modeling, Computation, and Inference*, 353 pp., CRC Press, Boca Raton, Fla.

Prescott, K. L., and I. K. Tsanis (1997), Mass balance modeling and wetland restoration, *Ecol. Eng.*, *9*(1–2), 1–18, doi:10.1016/s0925-8574(97)00015-3.

Preston, S. D., R. B. Alexander, G. E. Schwarz, and C. G. Crawford (2011), Factors affecting stream nutrient loads: A synthesis of regional SPARROW model results for the continental United States, *J. Am. Water Resour. Assoc.*, *47*(5), 891–915, doi:10.1111/j.1752-1688.2011.00577.x.

Puri, D., R. Karthikeyan, and M. Babbar-Sebens (2009), Predicting the fate and transport of *E. coli* in two Texas river basins using a spatially referenced regression model, *J. Am. Water Resour. Assoc.*, *45*(4), 928–944, doi:10.1111/j.1752-1688.2009.00337.x.

Qian, S. S., K. H. Reckhow, J. Zhai, and G. McMahon (2005), Nonlinear regression modeling of nutrient loads in streams: A Bayesian approach, *Water Resour. Res.*, *41*, W07012, doi:10.1029/2005WR003986.

Ramin, M., S. Stremilov, T. Labencki, A. Gudimov, D. Boyd, and G. B. Arhonditsis (2011), Integration of numerical modeling and Bayesian analysis for setting water quality criteria in Hamilton Harbour, Ontario, Canada, *Environ. Modell. Softw.*, *26*(4), 337–353, doi:10.1016/j.envsoft.2010.08.006.

Reddy, K. R., R. H. Kadlec, E. Flaig, and P. M. Gale (1999), Phosphorus retention in streams and wetlands: A review, *Crit. Rev. Environ. Sci. Technol.*, *29*(1), 83–146.

Reichert, P., and J. Mieleitner (2009), Analyzing input and structural uncertainty of nonlinear dynamic models with stochastic, time-dependent parameters, *Water Resour. Res.*, *45*, W10402, doi:10.1029/2009WR007814.

Renard, B., D. Kavetski, G. Kuczera, M. Thyer, and S. W. Franks (2010), Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors, *Water Resour. Res.*, *46*, W05521, doi:10.1029/2009WR008328.

Richards, R. P., and J. Holloway (1987), Monte Carlo studies of sampling strategies for estimating tributary loads, *Water Resour. Res.*, *23*(10), 1939–1948, doi:10.1029/WR023i010p01939.

Robertson, D. M., and D. A. Saad (2011), Nutrient inputs to the Laurentian Great Lakes by source and watershed estimated using SPARROW watershed models, *J. Am. Water Resour. Assoc.*, *47*, 1011–1033, doi:10.1111/j.1752-1688.2011.00574.x.

Rode, M., G. Arhonditsis, D. Balin, T. Kebede, V. Krysanova, A. van Griensven, and S. E. A. T. M. van der Zee (2010), New challenges in integrated water quality modeling, *Hydrol. Processes*, *24*(24), 3447–3461, doi:10.1002/hyp.7766.

Sadraddini, S., M. E. Azim, Y. Shimoda, S. P. Bhavsar, K. G. Drouillard, S. M. Backus, and G. B. Arhonditsis (2011a), A Bayesian assessment of the PCB temporal trends in Lake Erie fish communities, *J. Great Lakes Res.*, *37*(3), 507–520, doi:10.1016/j.jglr.2011.06.005.

Sadraddini, S., M. E. Azim, Y. Shimoda, S. Bhavsar, S. M. Backus, and G. B. Arhonditsis (2011b), Temporal contaminant trends in Lake Erie fish: A dynamic linear modeling analysis, *Ecotox. Environ. Safe.*, *74*, 2203–2214, doi:10.1016/j.physletb.2003.10.071.

Shaddick, G., and J. Wakefield (2002), Modeling daily multivariate pollutant data at multiple sites, *J. R. Stat. Soc. C-App.*, *51*(3), 351–372.

Shih, J.-S., R. B. Alexander, R. A. Smith, E. W. Boyer, G. E. Schwarz, and S. Chung (2010), An initial SPARROW model of land use and in-stream controls on total organic carbon in Streams of the conterminous United States, *U.S. Geol. Surv. Open-File Rep.*, *2010-1276*, 22 pp.

Soldat, D. J., and A. M. Petrovic (2008), The fate and transport of phosphorus in turfgrass ecosystems, *Crop Sci.*, *48*(6), 2051–2065, doi:10.2135/cropsci2008.03.0134.

Soldat, D. J., A. M. Petrovic, and Q. M. Ketterings (2009), Effect of soil phosphorus levels on phosphorus runoff concentrations from turfgrass, *Water Air Soil Pollut.*, *199*, 33–34.

Spiegelhalter, D. J., N. G. Best, B. R. Carlin, and A. van der Linde (2002), Bayesian measures of model complexity and fit, *J. R. Stat. Soc. Ser. B*, *64*, 583–616, doi:10.1111/1467-9868.00353.

Stow, C. A., E. C. Lamon, S. S. Qian, and C. S. Schrank (2004), Will Lake Michigan lake trout meet the Great Lakes strategy 2002 PCB reduction goal?, *Environ. Sci. Technol.*, *38*(2), 359–363, doi:10.1021/es034610l.

Strahler, A. N. (1952), Hypsometric (area-altitude) analysis of erosional topology, *Geol. Soc. Am. Bull.*, *63*(11), 1117–1142, doi:10.1130/0016-7606(1952)63[1117:HAAOET]2.0.CO;2.

Stream Solute Workshop (1990), Concepts and methods for assessing solute dynamics in stream ecosystems, *J. North Am. Benthol. Soc.*, *9*(2), 95–119.

Tomassini, L., P. Reichert, C. Buser, H.-R. Künsch, R. Knutti, and M. E. Borsuk (2009), A smoothing algorithm for estimating stochastic continuous-time model parameters and its application to a simple climate model, *J. R. Stat. Soc. C*, *58*, 679–704.

West, M., and J. Harrison (1989), *Bayesian Forecasting and Dynamic Models*, Springer, New York.

Winter, J. G., and H. C. Duthie (2000), Export coefficient modeling to assess phosphorus loading in an urban watershed, *J. Am. Water Resour. Assoc.*, *36*(5), 1053–1061, doi:10.1111/j.1752-1688.2000.tb05709.x.

Zhang, W., and G. B. Arhonditsis (2008), Predicting the frequency of water quality standard violations using Bayesian calibration of eutrophication models, *J. Great Lakes Res.*, *34*(4), 698–720.

A BAYESIAN METHODOLOGICAL FRAMEWORK FOR ACCOMODATING INTERANNUAL VARIABILITY OF NUTRIENT LOADING WITH THE SPARROW MODEL

(Electronic Supplementary Material)

Christopher Wellen[1*] and George B. Arhonditsis[1]

[1] Ecological Modeling Laboratory,

Department of Physical & Environmental Sciences, University of Toronto,

Toronto, Ontario, Canada, M1C 1A4

Tanya Labencki[2] and Duncan Boyd[2]

[2] 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) CALCULATION OF POTENTIAL EVAPOTRANSPIRATION

1: Master Equation: We employed the Food & Agriculture Organization (FAO) implementation of the Penman-Monteith approach [*Allen et al.*, 1998] to calculate daily evapotranspration from a reference surface with properties similar to an extended area of actively-growing, well-watered, green grass of uniform height, that is, a height of 0.12 m, a surface resistance of 70 s m^{-1}, and an albedo of 0.23:

$$ET_0 = \frac{0.408\Delta(R_n - G) + \gamma \frac{900}{T+273} u_2 (e_s - e_a)}{\Delta + \gamma(1 + 0.34u_2)} \quad \textbf{(ESM 1)}$$

where ET_0 is the reference evapotranspiration (mm day^{-1}), R_n is the net radiation at the vegetative surface (MJ m^{-2} day^{-1}), G is the soil heat flux density (MJ m^{-2} day^{-1}), T is the mean daily air temperature at 2 m height (°C), u_2 is the wind speed at 2 m height (m s^{-1}), e_s is the saturation vapor pressure (kPa), e_a is the actual vapor pressure (kPa), Δ is the slope of the vapor pressure curve (kPa $°C^{-1}$), and γ is the psychrometric constant (kPa $°C^{-1}$).

2: Non-radiation terms: This subsection deals with the terms of ESM 1 unconnected with R_n and G. These latter two terms are significantly more complex and are dealt with in subsection 3.

2.1: Psychrometric Constant and Slope of the Vapor Pressure Curve: We assumed the psychrometric constant is a function of pressure:

$$\gamma = \frac{c_p P}{\varepsilon \lambda} = 0.665 \times 10^{-3} P \quad \textbf{(ESM 2)}$$

which in turn was calculated as a function of elevation above sea level:

$$P = 101.3 \frac{(293 - 0.0065z)^{5.26}}{293} \quad \textbf{(ESM 3)}$$

where c_p refers to the specific heat of air at a constant pressure (1.013 10^{-3} MJ kg^{-1} $°C^{-1}$), P refers to atmospheric pressure (kPa), ε is the ratio of molecular weight of water vapor to dry air (0.622), λ refers to the latent heat of vaporization of water (2.45 MJ kg^{-1}), and z refers to height above sea level (m).

The slope of the saturation vapor pressure curve (Δ) was calculated as a function of daily mean air temperature T (°C):

$$\Delta = \frac{4098\left(0.6108\exp\left[\frac{17.27T}{T+237.3}\right]\right)}{(T+237.3)^2}. \quad \textbf{(ESM 4)}$$

2.2: Daily Air Temperatures: Mean daily air temperatures (°C) were calculated as the average between the maximum and the minimum daily temperatures:

$$T = \frac{T_{\max} - T_{\min}}{2} \quad \textbf{(ESM 5)}$$

2.3: Vapor Pressure Terms: The saturation vapor pressure refers to the partial vapor pressure of water in air at which the rates of condensation and evaporation are equal, a measurement of the capacity of a parcel of air to 'hold' water vapor. It is a function of temperature alone:

$$e°(T) = 0.6108\exp\left(\frac{17.27T}{T+273.3}\right) \quad \textbf{(ESM 6)}$$

where *e°(T)* refers to the saturation vapor pressure at the air temperature T (kPa, °C).

The saturation vapor pressure was calculated using the daily maximum and minimum air temperatures to correct for the non-linear response of saturation vapor pressure to air temperature:

$$e_s = \frac{e°(T_{\max}) - e°(T_{\min})}{2} \quad \textbf{(ESM 7)}$$

where *e°(T)* refers to the saturation vapor pressure at the air temperature T (kPa °C), and e_s refers to the daily saturation vapor pressure.

The actual vapor pressure (e_a), which refers to the partial pressure of the water vapor in the air at any one time, was derived from relative humidity data:

$$e_a = \frac{e°(T_{\min})\frac{RH_{\max}}{100} + e°(T_{\max})\frac{RH_{\min}}{100}}{2} \quad \textbf{(ESM 8)}$$

where RH_{max} and RH_{min} are the daily relative maximum and mimimum relative humidity, respectively, in percent, and T_{max} and T_{min} are the daily maximum and minimum temperature (°C), respectively.

2.4: Wind Speed: We transformed the wind speeds measured at Environment Canada's Hamilton Airport station (Climate ID: 6153194; WMO ID: 71263) at 10 m from the surface to wind speed estimates at 2 meters from the surface using the following equation:

$$u_2 = u_z \frac{4.87}{\ln(67.8z - 5.42)} \quad \textbf{(ESM 9)}$$

where z refers to the height above the surface of the wind measurements (10 m).

3: *Radiation terms*: This subsection offers the details of the calculation of R_n, the net radiation at the crop surface (MJ m^{-2} day^{-1}). Our radiation calculations assumed that the effects of the net daily flux of radiation into the soil (G) were negligible, a defensible assumption when constructing an annual time series.

3.1: Net Radiation (R_n): We estimated R_n as the difference between incoming net shortwave radiation (R_{ns}, MJ m^{-2} day^{-1}) and the outgoing net longwave radiation (R_{nl}, MJ m^{-2} day^{-1}):

$$R_n = R_{ns} - R_{nl} \quad \textbf{(ESM 10)}$$

3.2: Net Incoming Shortwave Radiation: The term R_{ns} is an estimate of the amount of shortwave radiation absorbed by the surface assuming a fraction is reflected back into space. This fraction is a function of the surface albedo (α), assumed to be 0.23 for the grass reference crop:

$$R_{ns} = (1 - \alpha)R_s \quad \textbf{(ESM 11)}$$

where R_s refers to the incoming solar radiation (MJ m^{-2} day^{-1}), which was estimated using Hargreaves' radiation formula:

$$R_s = k_{R_s}(\sqrt{T_{\max} - T_{\min}})R_a \quad \textbf{(ESM 12)}$$

where R_a refers to the extraterrestrial radiation (MJ m^{-2} day^{-1}), T_{max} and T_{min} refer to the daily maximum and minimum air temperature (°C), and k_{Rs} is an adjustment coefficient. The adjustment coefficient differs for coastal and interior regions. Values in the vicinity of 0.16 are suggested for continental regions, while values close to 0.19 are suggested for maritime areas. Hamilton's climate is influenced by its proximity to the Great Lakes, so we used a value of 0.18, assuming near-maritime conditions.

3.3: Extraterrestrial Radiation: We calculated R_a, the solar radiation at the top of the atmosphere, as:

$$R_a = \frac{24(60)}{\pi} G_{sc} d_r \left[\omega_s \sin(\varphi)\sin(\delta) + \cos(\varphi)\cos(\delta)\cos(\omega_s)\right] \quad \textbf{(ESM 13)}$$

where G_{sc} refers to the solar constant (0.082 MJ $m^{-2}min^{-1}$), d_r refers to the inverse relative earth-sun distance, ω_s refers to the sunset hour ansgle, δ refers to the solar decimation, and φ refers to the site latitude (43.171687 decimal degrees) expressed in radians (0.753488 radians).

We calculated the inverse relative earth-sun distance (d_r) as:

$$d_r = 1 + 0.033\cos\left(\frac{2\pi}{365} J\right) \quad \textbf{(ESM 14)}$$

where J refers to the day of the year (1 – 365). The sunset hour angle, ω_s, was calculated as:

$$\omega_s = \arccos\left[-\tan(\varphi)\tan(\delta)\right] \quad \textbf{(ESM 15)}$$

and the solar decimation (δ) as:

$$\delta = 0.409\sin\left(\frac{2\pi}{365} J - 1.39\right) \quad \textbf{(ESM 16)}$$

3.4: Net Outgoing Radiaiton: We calculated R_{nl}, the net outgoing longwave radiation (MJ m^{-2} day^{-1}), as

$$R_{nl} = \sigma\left(\frac{T_{\max,K}{}^4 + T_{\min,K}{}^4}{2}\right)\left(0.34 - 0.14\sqrt{e_a}\right)\left(1.35\frac{R_s}{R_{so}} - 0.35\right) \quad \textbf{(ESM 17)}$$

where σ refers to the Stefan-Boltzmann constant (4.903 10^{-9} MJ K^{-4} m^{-2} day^{-1}), $T_{max,K}$ and $T_{min,K}$ refer to the maximum and minimum daily temperatures (degrees Kelvin), e_a refers to the actual vapor pressure (kPa), R_s refers to the solar radiation (MJ m^{-2} day^{-1}), and R_{so} refers to the clear sky radiation (MJ m^{-2} day^{-1}), calculated as:

$$R_{so} = \left(0.75 + 2\times10^{-5} z\right)R_a \quad \textbf{(ESM 18)}$$

where z refers to the station elevation above sea level (237.7 m).

2) SPATIAL DATA

2.1 Topography: The delineation of subwatersheds and reach catchments is done using a digital elevation model (*DEM*). A stream-corrected 10 m cell size *DEM* generated through the application of photogrammetric methods was used for this purpose (Ontario Ministry of Natural Resources, Greater Toronto Area Digital Elevation Model). Water quality monitoring stations were used as the discharge point for the subwatersheds. Our calibration dataset consisted of 6 subwatersheds. Their areas ranged from 25.5 – 75.8 km^2, with a mean of 49.3 km^2 and a standard deviation of 24.1 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]. All flows on record from all the Water Survey of Canada stations in the basin were used (Water Survey of Canada, 2011, available from http://www.wsc.ec.gc.ca/applications/H2O/). The *DA* model related the mean total yearly flow (m^3yr^{-1}, *Flow)* to the subwatershed area (km^2, *Area*) with the following equation:

$$Flow = 390431 \times Area - 2393505 \ (r^2 = 0.93, n = 9) \qquad \textbf{(ESM 19)}$$

2.2 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. The 1:50,000 scale map series was used as the source of the dataset. To avoid the proliferation of miniscule reach catchments, we imposed a minimum reach catchment area of 10,000 m^2 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 a lower bound of 750 m due to the finer scale of our study as compared to the NHD's national scope (United States Geological Survey, National Hydrography Dataset: Concepts and Contents, available from nhd.usgs.gov). 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 118 reach

catchments, and each reach catchment discharges into a confluence, reservoir, or water quality monitoring station. Reach catchment areas ranged from 0.02 – 12.3 km^2, with a mean of 2.5 km^2 and an interquartile range of 3.5-1.3=2.2 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, National Hydrography Dataset: Concepts and Contents, available from nhd.usgs.gov). Second, it had to drain an area of at least 500 ha. This was roughly the x-intercept of Equation 6, and represented the limit of our confidence in its application. 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.

2.3 Nutrient Sources: Both point and non-point nutrient sources were included in the *SPARROW* model of Hamilton Harbour. While there is a combined sewer overflow (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. We also explicitly considered the Waterdown Waste Water Treatment Plant (*WWTP*), a small water treatment plant which discharged into a tributary of Grindstone Creek during the study period. 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*, Contaminant Loadings and Concentrations to Hamilton Harbour: 2003-2007 Update, Hamilton Harbour Remedial Action Plan Office, Burlington, Ontario, Canada).

The non-point sources of total phosphorus were limited to agricultural and urban land, which included paved areas and urban green space. These two land use types were selected because they have been found to be by far the greatest sources of nitrogen and phosphorus to receiving waters at the landscape scale [*Beaulac and Reckhow*, 1982; *Alexander et al.*, 2004] and because they together

comprise about 80% of the study area. Land uses were derived from a supervised classification of satellite imagery from 2002 (SOLRIS, Ontario Ministry of Natural Resources, 2008, available from http://lioapp.lrc.gov.on.ca). Total agricultural and urban areas were estimated for each reach using GIS overlay analysis.

2.4 Landscape Characteristics: Landscape characteristics can influence the delivery of phosphorus to stream edges. While most *SPARROW* applications have focused on soil properties as controlling delivery factors, we found during preliminary model applications that soil properties were not an effective way to parameterize phosphorus delivery. Nutrient delivery was parameterized as a function of the proportion of each reach covered by wetlands, due to their role in moderating nutrient fluxes to receiving waterbodies [*Krieger,* 2003]. Wetlands were included in the Southern Ontario Land Resource Information System (SOLRIS, Ontario Ministry of Natural Resources, 2008, available from http://lioapp.lrc.gov.on.ca). The proportion of wetlands covering each reach was estimated using GIS overlay analysis. Proportions of wetland ranged from 0 to 1 with a mean of 0.06 and an interquartile range of 0.06 – 0=0.06.

References

Alexander, R. B., R. A. Smith, and G. E. Schwarz (2004), Estimates of diffuse phosphorus sources in surface waters of the United States using a spatially referenced watershed model, *Water Sci. Technol.*, *49*(3), 1-10.

Beaulac, M. N., and K. H. Reckhow (1982), An examination of land-use - nutrient export relationships, *Water Resour. Bull.*, *18*(6), 1013-1024.

Krieger, K. A. (2003), Effectiveness of a coastal wetland in reducing pollution of a Laurentian Great Lake: Hydrology, sediment, and nutrients, *Wetlands*, *23*(4), 778-791.

Limpert, E., W. A. Stahel, M. Abbt, (2001), Log-normal Distributions across the Sciences: Keys and Clues, *BioScience* 51(5):341 – 352.

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

3) TABLES AND FIGURES

Table ESM3.1: Prior parameter distributions for all models.

Parameter	*Median*	*Standard Deviation*	*Shape*	*Source*
α	0	3.17	Normal	-
β_1	0.07	1.25	Log normal	*Beaulac and Reckhow*, (1982)
β_2	0.10	3.5	Log normal	*Beaulac and Reckhow*, (1982)
k_r	12.84	4.76	Log normal	*Cheng et al.*, (2010) and reference therein
k_{s1}	0.22	0.23	Log normal	-
k_{s2}	0.05	0.1	Log normal	-
γ_v	0	31.62	Normal	-

Table ESM3.2: Root Mean Squared Error (*RMSE*) values for all statistical formulations used to model total phosphorus loading. *Pred0* refers to the sole use of *SPARROW* to accommodate interannual loading variability; *Pred1* refers to the use of *SPARROW* along with the total annual precipitation; *Pred2* refers to the use of *SPARROW* along with the total precipitation and the total annual potential evapotranspiration.

Formulation	*Pred 0*	*Pred 1*	*Pred 2*
MCMC	-	0.26	0.26
WALK	-	0.14	0.14
MCMC - α_{DYN}	0.28	0.23	0.23
MCMC - β_{DYN}	0.10	0.10	0.10
MCMC - k_{sDYN}	0.14	0.14	0.15
MCMC - β, k_{sDYN}	0.11	0.11	0.11

Table ESM3.3: Weighted Root Mean Squared Error (*WRMSE*) values for all statistical formulations used to model total phosphorus loading. *Pred0* refers to the sole use of *SPARROW* to accommodate interannual loading variability; *Pred1* refers to the use of *SPARROW* along with the total annual precipitation; *Pred2* refers to the use of *SPARROW* along with the total precipitation and the total annual potential evapotranspiration.

Formulation	*Pred 0*	*Pred 1*	*Pred 2*
MCMC	-	0.25	0.24
WALK	-	0.09	0.09
MCMC - α_{DYN}	0.23	0.19	0.18
MCMC - β_{DYN}	0.08	0.07	0.07
MCMC - k_{sDYN}	0.07	0.08	0.08
MCMC - β, k_{sDYN}	0.05	0.05	0.05

Figure ESM 3.1: Coefficient of variability (CV) associated with the loadings from different Creeks. CVs were calculated as $CV = \sqrt{\exp(\sigma^2) - 1}$ [*Limpert et al.*, 2001].

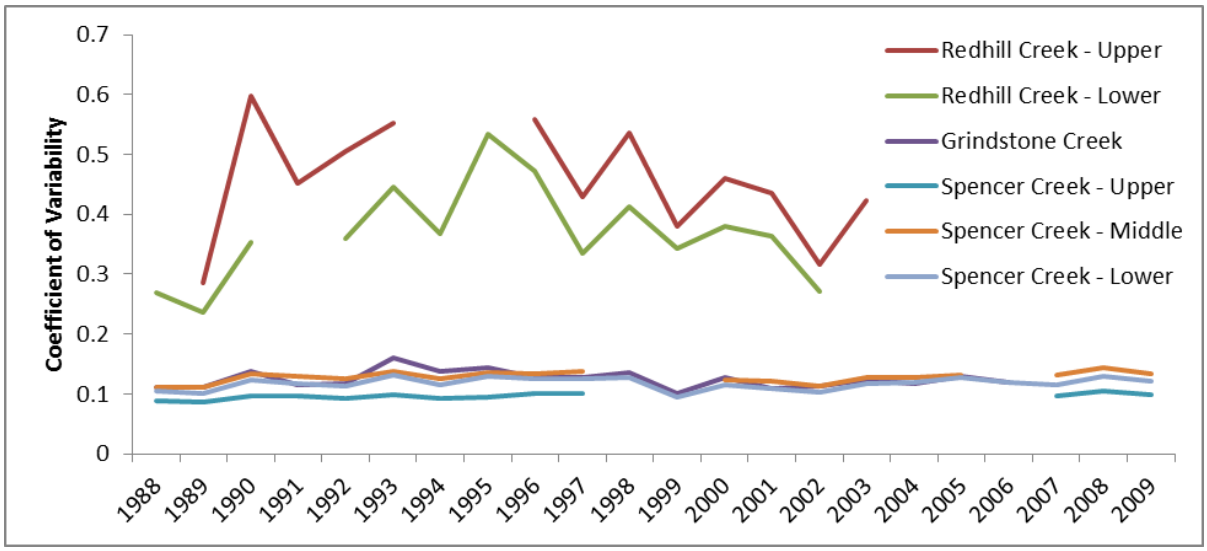


Figure ESM 3.2: Quantile-quantile plot residuals of measured from estimated 'true' load ($Y_{i,t} - Load_{i,t}$, see Equation 7) for *MCMC-Pred1*. Grey lines indicate 95% credible intervals.

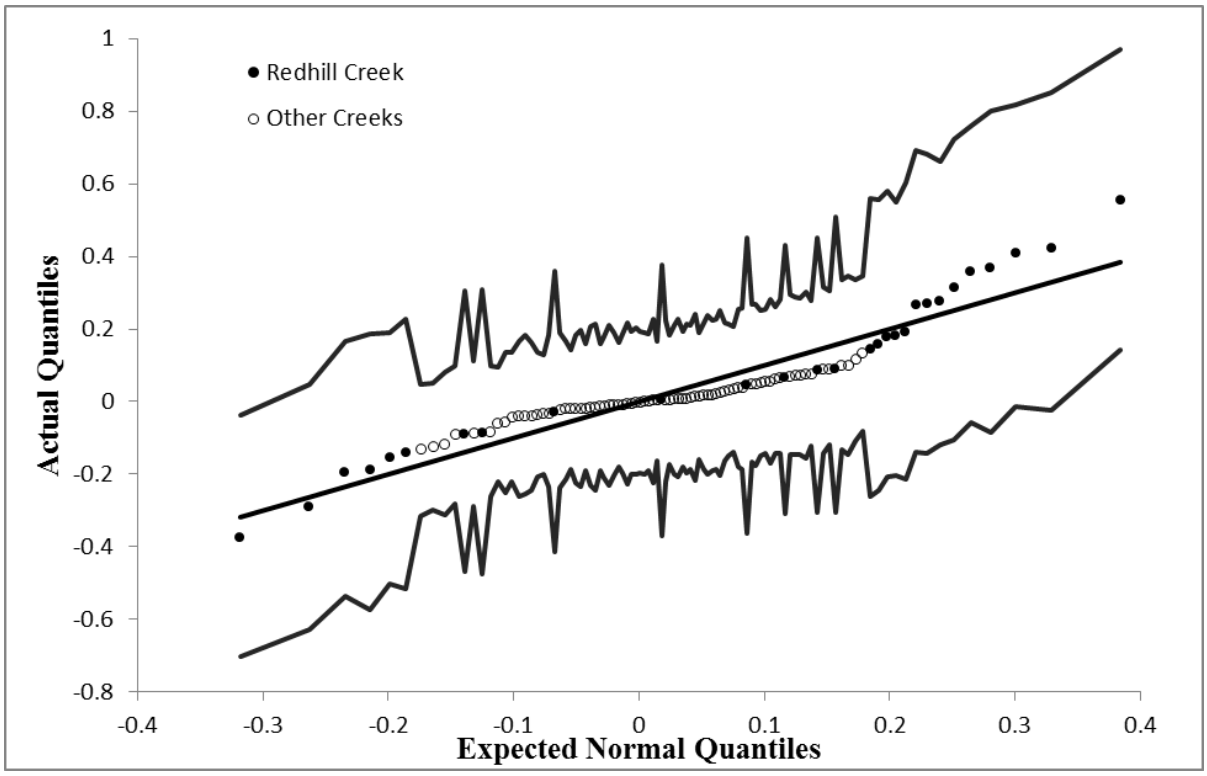


Figure ESM 3.3: Quantile-quantile plot of residuals of 'true' from modeled load ($Load_{i,t} - \mu_{i,t} + W_{v,t}\gamma_v$, see Equation 8) for *MCMC-Pred1*. Grey lines indicate 95% credible intervals.

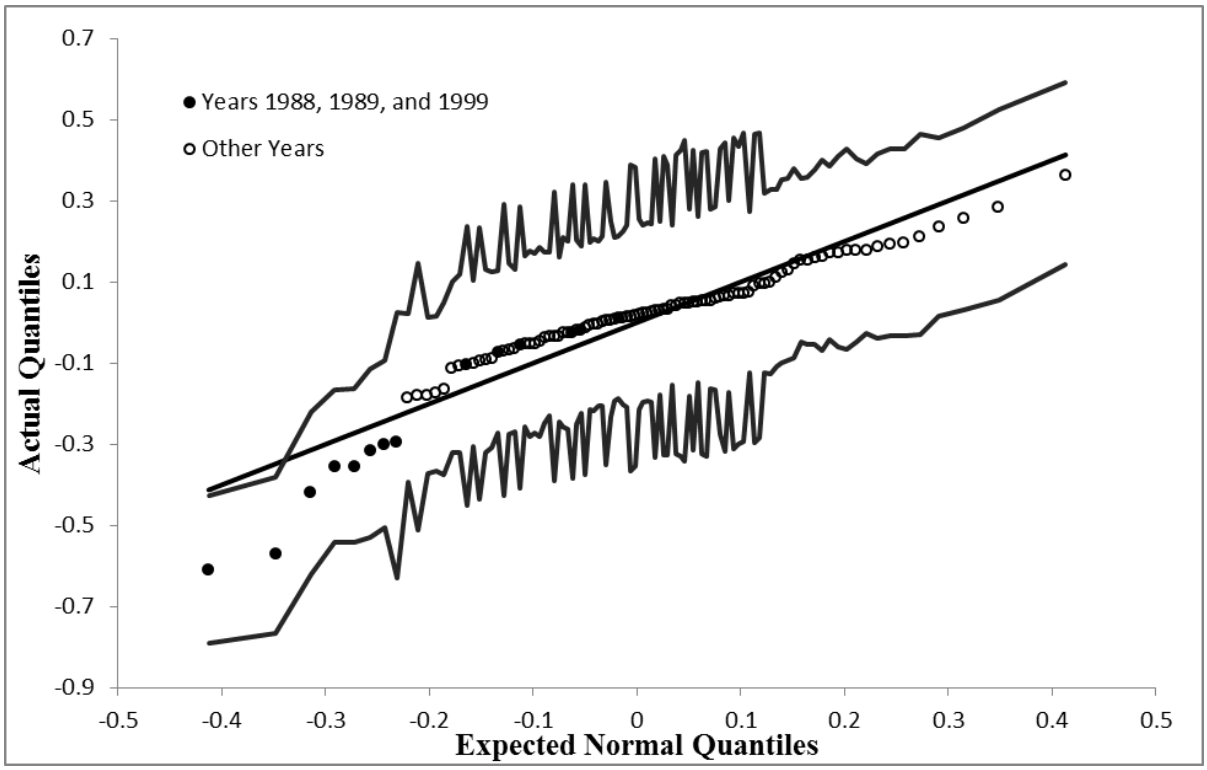


Figure ESM 3.4: Quantile-quantile plot of residuals of measured from estimated 'true' load ($Y_{i,t} - Load_{i,t}$, see Equation 7) for *MCMC* - k_{sDYN} formulation with the *Pred1* climate forcing complexity. Grey lines indicate 95% credible intervals.

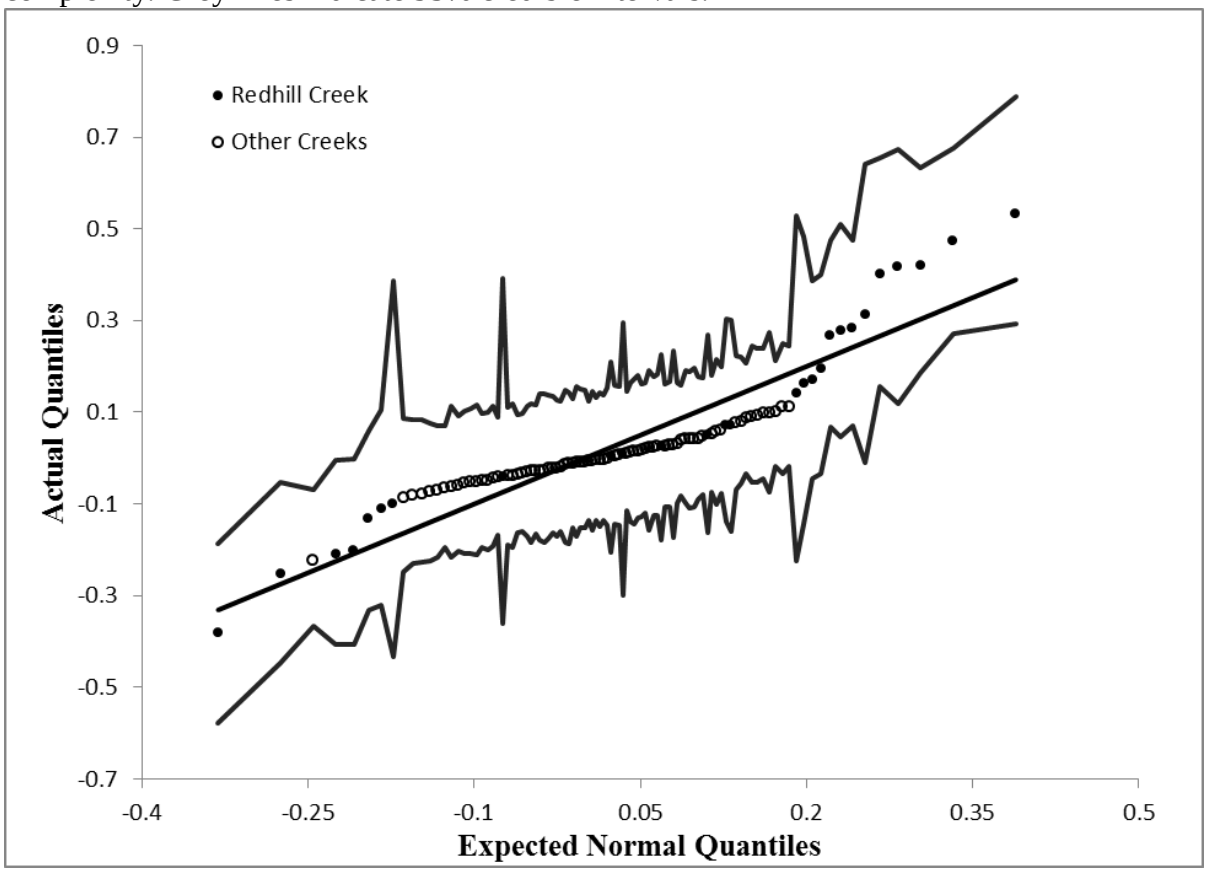
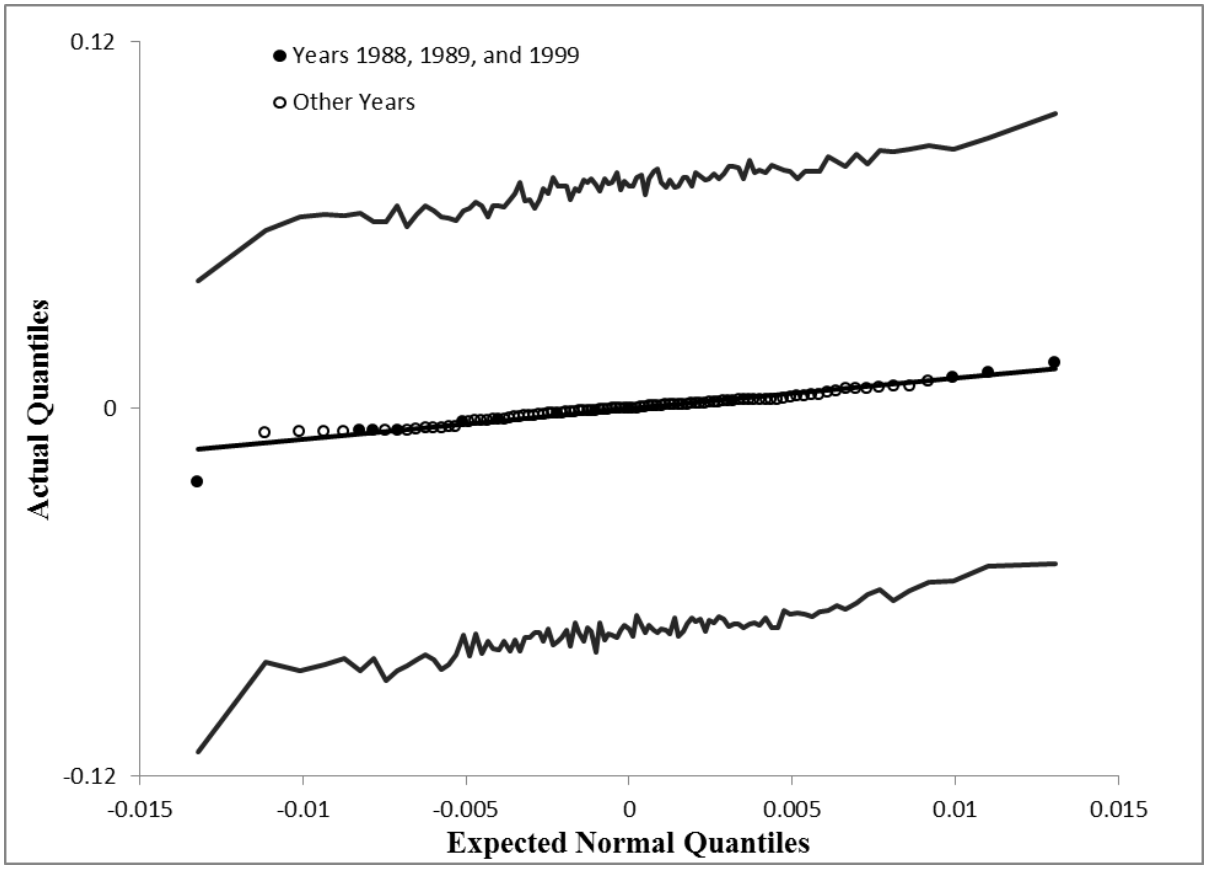


Figure ESM 3.5: Quantile-quantile plot of residuals of 'true' from modeled load ($Load_{i,t} - \mu_{i,t} + W_{v,t}\gamma_v$, see Equation 8) for *MCMC* - k_{sDYN} formulation with the *Pred1* climate forcing complexity. Grey lines indicate 95% credible intervals.


0.12
0
-0.12
Actual Quantiles
Years 1988, 1989, and 1999
Other Years
-0.015
-0.01
-0.005
0
0.005
0.01
0.015
Expected Normal Quantiles