November 2016

Predictive modeling of riverine constituent concentrations and loads using historic and imposed hydrologic conditions

Mark Hagemann
University of Massachusetts - Amherst

PREDICTIVE MODELING OF RIVERINE CONSTITUENT CONCENTRATIONS AND LOADS USING HISTORIC AND IMPOSED HYDROLOGIC CONDITIONS

A Dissertation Presented

by

MARK W. HAGEMANN

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

September 2016

Civil and Environmental Engineering

Predictive Modeling of Riverine Constituent Concentrations and Loads using Historic and Imposed Hydrologic Conditions

A Dissertation Presented

by

MARK W. HAGEMANN

Approved as to style and content by:

_____
Mi-Hyun Park, Chair

_____
John E. Tobiason, Member

_____
Nicholas G. Reich, Member

_____
Richard N. Palmer, Department Head
Civil and Environmental Engineering

DEDICATION

To my brother Scott, who taught me to love to learn.

ACKNOWLEDGMENTS

Funding for this research was provided by the Massachusetts Department of Conservation and Recreation (DCR). I am privileged to have had the opportunity to contribute to a long history of research by UMass in partnership with DCR. In this I am grateful for support and input from DCR staff, namely Larry Pistrang, Pat Austin, Steve Sulprizio, and Dan Crocker.

For academic support and guidance I am indebted to a great many people. Principally I have my advisor, Dr. Mi-Hyun Park to thank for taking me through my grad-school journey. The DCR research group, led by Dr. John Tobiason, provided larger guidance and nudged my ideas into useful directions. My teammate in this project, Lily Jeznach has been a joy to work alongside. Advice in data science and statistics came from Professors Daeyoung Kim, John Staudenmayer, and Nick Reich, the latter of whom has graciously served on my dissertation committee. I must also thank the community of grad students, too many to name, whose comradery has made the past four years far more enjoyable than graduate school is supposed to be.

Finally, this work and others like it would not be possible without a legacy of collecting, measuring, archiving, and ultimately sharing data. For this I again must acknowledge the DCR staff mentioned above, and field and lab personnel throughout the years whose labors produced the monitoring data we modelers too often take for granted. Chapter 2, in particular, relied upon the decision of countless organizations to extend their data's value by sharing them with the greater scientific world.

ABSTRACT

PREDICTIVE MODELING OF RIVERINE CONSTITUENT CONCENTRATIONS AND LOADS USING HISTORIC AND IMPOSED HYDROLOGIC CONDITIONS

September 2016

MARK W. HAGEMANN, B.A., CARLETON COLLEGE

Ph.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Dr. Mi-Hyun Park

This research was principally concerned with the task of quantifying dissolved and suspended constituents carried in river water, when direct measurements are not available. This is a question of scientific and societal relevance, and one with a long history of study and a great deal of remaining difficulty. The three studies that comprise the chapters of this document sought to advance this field, primarily in the context of drinking water supplies, and with an emphasis on extreme precipitation events.

Chapter 1 investigated the impact of model form and flexibility on estimates of constituent loads to a water-supply reservoir. A series of load-estimation regression models were calibrated and used to predict nutrient (nitrate-nitrogen and total phosphorus) and organic carbon loads from three major tributaries of a water-supply reservoir. These models included traditional linear models (LMs) as well as semiparametric generalized additive models (GAMs). The relative performance of each model was determined using cross-validation. GAMs, which employ more flexible model structures, outperformed LMs in most cases, explaining an additional 2% of load variance and 5% of concentration variance in validation data on average. Resulting point

load estimates from the 1.5-year study period were similar between the two modeling approaches, yielding overlapping 95% confidence intervals in all 9 cases modeled. A novel graphical method was developed that presents relative agreement between the two model types as a function of the time interval over which the load is estimated. This revealed relative differences between the two modeling approaches in excess of 100% depending on the time interval of the load estimate. The time-dependent disagreement between similarly well-performing models highlights an aspect of model uncertainty not visible in more structurally restrictive modeling approaches.

Chapter 2 assessed the applicability of GAMs to the prediction of riverine solute concentrations during extreme high-flow hydrologic events, when such events are absent from the models' calibration data. Using a version of the differential split-sample test, and a large validation dataset (n = 6921) from sites across the US Northeast, such models showed a tendency to overpredict extreme-event concentrations, with increasing bias and variance for increasingly extreme hydrologic conditions. The validation framework in this study effectively compared model performance across disparate hydrologic regimes and constituents, yet can be used to estimate individual model performance under an unobserved extreme-flow condition, regardless of whether any extreme-flow data are available for that model. The validation procedure can further be generalized to explore model performance in an arbitrarily defined extreme condition for a broad range of model types. Despite an overall increase in uncertainty for extreme-event concentration estimates, estimates under extreme

hydrologic conditions could be improved by taking into account the observed bias in the aggregated regional database.

Chapter 3 developed and applied a methodology to generate reservoir tributary discharge and constituent concentration time-series for an imposed extreme-event scenario. In this approach, discharge is generated deterministically using historical observed storm hydrographs, while constituent concentrations are generated probabilistically using quasi-Monte Carlo sampling. A multivariate probability model was developed for constituent concentration in an arbitrary number of tributaries and water-quality constituents, conditional on time and hydrological condition. The resulting high-dimensional sampling distribution is reduced to a more manageable low-dimensional space using principal component analysis, and quasi-Monte Carlo samples are drawn from the lower-dimensional space. These samples are then used to as inputs to a process-based model of the receiving water body. As an application of the methodology, two separate historical storm events were modified using 3 extreme precipitation depths on tributaries of the Wachusett Reservoir Watershed in Massachusetts, U.S. Concentrations and loads were predicted for 5 constituents including nutrients and organic carbon. Despite several constituents having large mean-square error, this uncertainty was fully characterized and propagated through the reservoir model.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

ESTIMATING NUTRIENT AND ORGANIC CARBON LOADS TO A WATER-SUPPLY RESERVOIR USING SEMIPARAMETRIC MODELS

(A version of the paper published in the Journal of Environmental Engineering[*])

1.1 Introduction

Accurate estimation of constituent mass loading is a vital component of watershed management, and has particular bearing in watersheds of water-supply reservoirs. Nutrients and organic matter, either dissolved or suspended, are transported through tributaries to receiving waters, impacting their water quality. Nutrient loading drives biological dynamics, determining the trophic state of the receiving water body, and can lead to undesirable algae blooms. These are of special concern in water supply reservoirs, as the chemical byproducts of such blooms can include harsh taste and odor compounds as well as toxins. Excessive organic matter is also problematic in water supply reservoirs: in addition to increased chemical costs associated with its treatment, natural organic matter can combine with disinfectants to form potentially carcinogenic disinfection byproducts (DBPs).

Mass load (L, [M]) for a given time period is a time integral of the product of concentration ($c(t)$, [M L^{-3}]) and stream discharge ($q(t)$, [L^3 T^{-1}]),

$$L = \int_{T1}^{T2} q(t)\, c(t) dt \qquad \text{(Eq. 1.1)}$$

[*] Hagemann, Mark, Daeyoung Kim, and Mi Hyun Park. "Estimating Nutrient and Organic Carbon Loads to Water-Supply Reservoir Using Semiparametric Models." *Journal of Environmental Engineering*, 2016: 04016036

Discharge is often measured at sufficiently high resolution to approximate a continuous time series, but it may be practical to measure concentration only at weekly or monthly intervals. Load estimation therefore relies on methods that use a relatively sparsely sampled concentration dataset to infer its continuous behavior, or the long-run average of such behavior.

Numerous estimation methods have been applied to this task, including simple averages, ratio estimators, and regression methods, and these have been widely compared in the literature (Swistock et al., 1997; Ullrich and Volk, 2010; Verma et al., 2012). Of these, regression methods are among the most mathematically rigorous, exploiting relationships between easily measured quantities—often hydrology or time variables—and a less-easily measured variable of interest, i.e. concentration or load. Such models are empirical, with parameters determined by optimizing a function involving observed data.

"Rating-curve" regressions have been widely used in this regard, estimating the logarithm of concentration or load using the logarithm of flow and other exogenous variables relating to long-term trend and seasonal changes (e.g. Stenback et al., 2011; Huntington and Aiken, 2013; Kumar et al., 2013; Yoon and Raymond, 2012). For example, a 7-parameter model described in Cohn (1992) uses quadratic log-flow, quadratic time, and 4 seasonal harmonic variables to predict log-transformed load. This model—or some subset thereof—has become well established, and is employed in the U.S. Geological Survey (USGS) LOADEST load estimation software (Runkel et al., 2004).

Although widely applicable and mathematically lucid, regression models of this type rely on specific flow-concentration relationships and are prone to bias where these are weak or absent. For example, strong log-log relationships between flow and concentration are widespread, but not ubiquitous. Other hydrologic predictor variables can substitute for or complement log-transformed flow, and many examples of this are reported in the literature. Some authors have tried to capture antecedent conditions, for example using a simple running average of flow (Aulenbach and Hooper, 2006), 1-day differenced log flow (Brett et al., 2005), time-lagged flow (Drewry et al., 2009), or an exponential smoothing filter (Wang et al., 2011). Furthermore, it may be more appropriate to employ a model using a transformation other than the logarithm (e.g. Aulenbach and Hooper, 2006)—or no transformation at all—for concentration.

Two classes of regression models extend the flexibility of linear models in a generalized framework. Generalized linear models (GLMs) relax the normal-distribution assumption and allow a link function to be applied to the response variable (i.e. load or concentration). Rating curve models apply a log link function explicitly, requiring retransformation and bias correction following estimation, whereas GLMs incorporate these steps automatically. Despite these seeming advantages, the use of GLMs in load estimation has been limited to a few cases (Cox et al., 2007). Generalized additive models (GAMs) further expand on GLMs and allow more complicated relationships to be fitted to the data using semi-parametric functions of predictor variables. GAMs apply "smooth functions" to the predictor variables rather than describing a quantity as a linear or polynomial function of predictor variables (Wood, 2011). Smooth functions,

such as locally weighted scatterplot smoothing (LOWESS), have been used in load estimation (Helsel and Hirsch, 2002), but GAMs have not yet been extensively used for load estimation (Wang et al., 2011).

The aims of the present study were to explore impacts of model choice—including relaxing assumptions of linearity and using uncommonly considered predictor variables—on statistical measures of model performance as well as on resulting load estimates. We applied the GAM framework to three different water-quality constituents in three subbasins with disparate size and land-use in the Wachusett Reservoir watershed, Massachusetts (Figure 1.1). The GAM models were compared to linear models (LMs) to understand the range of model performance over different subbasin conditions. Finally, we investigated the hydrological conditions under which model predictions vary most widely between linear and semiparametric models, and explored the timescales on which the resulting differences in load estimates persist.

1.2 Material and methods

1.2.1 Study location

The Wachusett Reservoir Watershed (Figure 1.1) drains an area of 300 km^2, contributing approximately half of the total inflows to the Wachusett Reservoir, with the remainder arriving via aqueduct from the Quabbin Reservoir. Water withdrawn from the Wachusett Reservoir supplies approximately 2.5M users in the Boston metropolitan area. In general, tributary water quality is high, having low concentrations of nutrients and low turbidity. The Stillwater and Quinapoxet River subbasins together comprise

approximately 70% of the Wachusett watershed area, and are comparable in land-use. Although the Quinapoxet subbasin is 1.8 times the size of the Stillwater (143 km^2 versus 79 km^2), roughly one third of Quinapoxet drainage is diverted to off-basin reservoirs, giving it and the Stillwater subbasin comparable effective areas. Mean flows on the Quinapoxet and Stillwater Rivers since 1998 were 1.9 m^3/s and 1.6 m^3/s respectively at their stream gages (Figure 1.1). Land use in these subbasins is primarily forest (67% and 74% for Quinapoxet and Stillwater, respectively), and includes some urban/developed areas (15% and 11% for Quinapoxet and Stillwater, respectively). The Gates Brook subbasin is considerably smaller than the Quinapoxet and Stillwater subbasins, comprising approximately 1.5% of the reservoir watershed. However, 63% of its area is urban/developed land-use; for this reason, Gates Brook contributes certain contaminants from urban runoff and other anthropogenic sources in amounts disproportionate to its drainage area (Fiedler, 2009). Agricultural land-use area is less than 6% in all 3 subbasins.

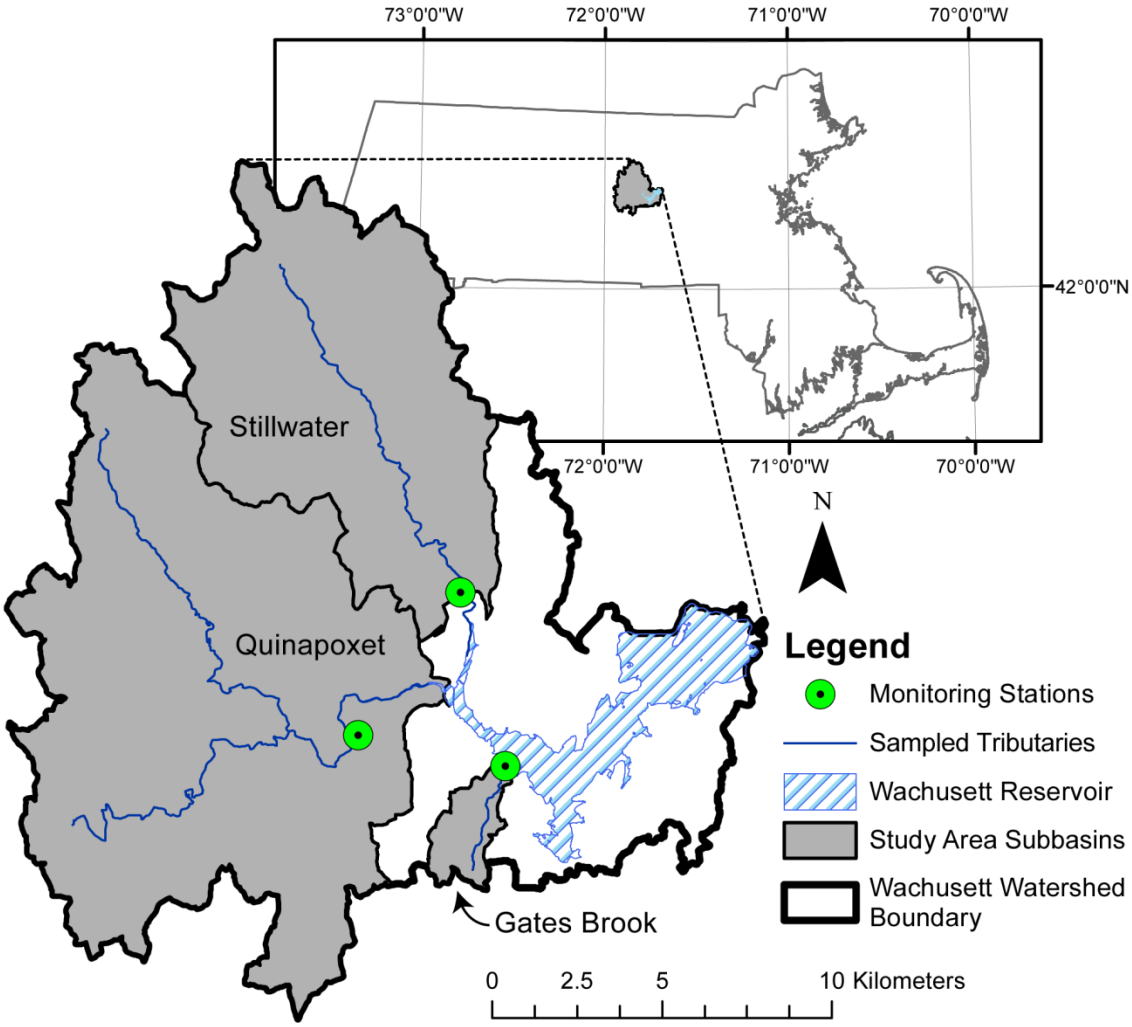

Figure 1.1: Map of Wachusett Reservoir Watershed and subbasins studied

1.2.2 Data

The Massachusetts Department of Conservation and Recreation collected nitrate-nitrogen (NO_3-N), total phosphorus (TP), and total organic carbon (TOC) concentration data from the Stillwater and Quinapoxet Rivers and Gates Brook as routine grab samples since 1999 (monthly since 2011), and additionally as flow-weighted storm-composite samples during selected events since mid-2011 (Table 1.1).

Storm event samples were collected approximately 12 times annually, with separate volumetric composite samples collected during rising and falling hydrograph limbs (MassDCR, 2013). In Gates Brook rising-limb and falling-limb samples often occurred in the same day; in such cases the mean was used to provide a single measurement for the day. All samples were analyzed for a suite of constituents including NO_3-N, TP, and TOC at the Massachusetts Water Resources Authority (MWRA) Deer Island Laboratory. TOC was analyzed using Standard Method 5310B; Nitrate-N and TP were analyzed using EPA methods 353.2 and 365.1, respectively (L. Pistrang, personal communication, 2013).

Table 1.1: Summary of concentration and flow data for sampled days. Discrepancies in mean flow in a given tributary are the result of imperfect overlap of sampled days for different solutes.

Tributary	Constituent	Flow (m^3/s)			Concentration (mg/L)			Storm Event samples	Routine samples
		max	mean	min	max	mean	min		
Gates	NO_3-N	0.54	0.18	0	1.26	0.7	0.23	28	13
	TP	0.54	0.18	0	0.32	0.1	0.01	26	13
	TOC	0.54	0.17	0	9.1	4.15	1.35	26	13
Quinapoxet	NO_3-N	5.58	1.3	0.2	0.39	0.22	0.11	23	13
	TP	5.58	1.33	0.2	0.12	0.03	0.01	24	13
	TOC	5.58	1.35	0.2	8.97	4.52	2.81	23	13
Stillwater	NO_3-N	5.78	1.58	0.1	0.28	0.15	0.03	22	13
	TP	5.78	1.58	0.1	0.13	0.03	0.01	22	13
	TOC	5.78	1.55	0.1	6.05	3.7	2.18	20	13

The USGS operates stream gages on the Stillwater and Quinapoxet Rivers and Gates Brook. These provided a continuous flow record with 15-minute resolution, aggregated to daily mean flow for this study. The Stillwater and Quinapoxet gages have been operating since 1994 and 1996, respectively, while the Gates Brook gage came on-line in December 2011. Prior to this time, flow on Gates Brook was manually measured on a weekly basis. Two missing values from the Stillwater gage were linearly interpolated from adjacent days; all other flow records were complete.

This study used data from an 18-month time period from December, 2011 to June, 2013. This time period was chosen for its consistency of sampling methods and frequency, and the availability of daily flow data from the three tributaries.

1.2.3 Model Development

A suite of 96 candidate models was developed, with models differing from each other in 3 ways: by the predictor variables used, by the response-variable transformation (i.e. log-transformed or not), and by the functional relationship between predictors and response (linear or semiparametric). One subset of models (36 of 96) was developed using traditional multiple linear regression (LM), while the remaining models (60 of 96) were developed based on the GAM framework.

A set of potential predictor variables (Table 1.2) was identified using exploratory data analysis including pairwise scatter plots between a number of hydrologic and time variables and concentration. In addition to commonly used predictor variables, (e.g. flow, time, and season), precipitation, antecedent dry days, and 1-day change in flow

were investigated as potential predictors. A further set of predictors was created by modifying flow-derived variables using an exponential smoothing "discount" function (Wang et al., 2011, Wang et al., 2013) to capture hysteresis effects. The discount function is defined as follows:

$$y_t = \frac{\sum_{i=0}^{t-1} d^i x_{t-i}}{\sum_{k=1}^{t} d^k}$$

where y_t is the discounted variable on day t, x_t is the (undiscounted) flow-derived variable on day t, and d is a discount factor between 0 and 1; this study used $d = 0.6$, reflecting the relatively short characteristic time scales of the study catchments (Wang et al., 2011).

Table 1.2: Variables used in regression models, including models not selected for load estimation

Variable	Description	Units	% of models with variable	
			LM	GAM
adry	Number of antecedent dry days	days	50	50
cj	Cosine of Julian day	(none)	100	100
ddflow	discounted (exponential smooth) 1-day change in flow	CFS	33	7
dflow	1-day change in flow	CFS	33	7
logC	log-transformed concentration	(none)	50	50
logQ	log-transformed mean-centered flow	(none)	33	7
sj	sine of Julian day	(none)	100	100
t	mean-centered time in days	days	67	40
t2	square of mean-centered time	$Days^2$	33	20
C	concentration	mg/L	50	50
s(ddflow)	nonparametric smooth function of ddflow	CFS	0	27
s(dflow)	nonparametric smooth function of dflow	CFS	0	27
s(logQ)	nonparametric smooth function of logQ	(none)	0	27
s(t)	nonparametric smooth function of t	days	0	40

All 96 candidate models were calibrated to the concentration data using maximum-likelihood estimation of model parameters. The R statistical computing platform was used for all calculations, with GAMs fitted using the mgcv package (Wood, 2011). Backward stepwise elimination was applied to all models to remove non-significant predictor variables using a significance threshold of p-value <0.05.

Once calibrated, each individual model's predictive power was evaluated using cross-validation. Validation against withheld data is useful for assessing model

performance, but withholding data from calibration degrades the accuracy of model parameter estimation. Cross-validation overcomes some of this tradeoff by iterating the validation process using collectively exhaustive subsets of the data and calculating the overall predictive power from all the individual validations. Two variations of cross-validation were used in this study: leave-one-out and 10-fold. Leave-one-out cross-validation (LOOCV) iterates through the data set, each time omitting a different single point, calibrating a model using the remaining data, predicting the omitted value and obtaining a validation residual from the prediction. For a data set of size n, each calibration data set has size n-1 and the resulting set of validation residuals has size n. 10-fold cross-validation instead performs 10 iterations, each time omitting 10% of the data and calibrating a model using the remaining data. The resulting set of validation residuals are used to calculate goodness-of-fit statistics, such as the root mean square error (RMSE), coefficient of determination (R^2), and Nash-Sutcliffe Efficiency (NSE). Due to the relatively small data sets (n approximately 35-40) used in this study, LOOCV was used as the primary validation method.

The model list included models for concentration as well as log-transformed concentration. As models using log-transformed response variables exhibit well-documented bias arising from back-transformation from log-space (Cohn et al., 1989; Ferguson, 1986), the smearing bias-correction factor (Duan, 1983) was applied to concentration and load estimates for models of this type. This adjusts back-transformed concentration estimates using an empirical adjustment factor, defined as the ratio of observed mean concentration to mean unadjusted back-transformed concentrations.

From the suite of 96 models, a single best-performing LM and GAM were selected for each location and constituent based on LOOCV (Table 1.3). Model assumptions (independence, distribution, functional relationship) were checked using residual plots and tests for serial correlation. Models that violated assumptions were discarded and replaced by the next-best-performing model. The best LM and GAM were then used to estimate a daily-resolution time series of concentration, from which mass loads were calculated as in Equation 1.1. As a comparison of LMs and GAMs against more rudimentary estimation methods, two sets of load estimates were calculated by linear interpolation of concentration measurements (Littlewood et al., 1998), one using routine data (omitting storm-event samples) only, and one using routine data and storm-event samples.

Table 1.3: Selected model forms used for load estimation

Tributary	Solute	LM	GAM
Gates	NO_3-N	logC ~ ddflow + sj	C ~ cj + sj + s(logQ)
	TP	C ~ logQ + sj	C ~ sj + s(ddflow)
Quinapoxet	TOC	C ~ dflow + cj + sj	C ~ cj + sj + s(dflow)
	NO_3-N	C ~ logQ + cj + sj	C ~ logQ + s(t)
	TP	logC ~ ddflow + adry + t + t2	logC ~ ddflow + cj + sj
Stillwater	TOC	logC ~ ddflow + sj + t	logC ~ ddflow + s(t)
	NO_3-N	C ~ logQ + cj + sj	C ~ logQ + s(t)
	TP	C ~ ddflow	C ~ ddflow
	TOC	logC ~ logQ + cj + sj	logC ~ logQ + adry + s(t)

1.2.4 Uncertainty estimation

The estimated cumulative load is a sum of estimated daily loads, (Equation 1.1), and has an associated uncertainty that is a function of the variance of each individual daily load estimate and the covariance structure of these estimates. Specifically, the variance of the cumulative load is given by the sum of the elements of the covariance matrix for the load time series:

$$var\left(\sum_{t=1}^{T} l_t\right) = \sum_{i=1}^{T}\sum_{j=1}^{T} cov(l_i, l_j)$$

$$= \sum_{t=1}^{T} var(l_t) + 2\sum_{i=2}^{T}\sum_{j=1}^{i-1} cov\left(l_i, l_j\right) \ , \qquad (Eq. 1.2)$$

where $l_t = q_t c_t$ is the load on day t, and T is the number of days over which the load is being estimated. While some studies have investigated the impact of considering flow uncertainty (e.g. Wang et al., 2011; Leisenring and Moradkhani 2012; Vogel et al., 2005), the high temporal frequency of flow measurement suggests that its contribution to load uncertainty would be minimal, and in this study its uncertainty is assumed to be zero.

Several studies have estimated load uncertainty for rating-curve models analytically, based on assumptions of the underlying model form and residual distributions (Gilroy et al., 1990; Vogel et al., 2005). This study used a less restrictive estimate of the covariance matrix based on statistics of temporal structure in the model estimates. This method was preferred for its mathematical simplicity and applicability

across various model forms (for example, using either log-transformed or untransformed concentration).

For a time series of load estimates, l_t, Equation 1.2 can be decomposed as follows:

$$var\left(\sum_{t=1}^{T} l_t\right) = \sum_{t=1}^{T} q_t^2 \, var(\mathrm{c}_t) + 2\sum_{i=2}^{T}\sum_{j=1}^{i-1} q_i q_j cov(\mathrm{c}_i, \mathrm{c}_j) \qquad (Eq. 1.3)$$

Equation 1.3 leaves two related quantities to be estimated—the variance of each concentration estimate and the autocovariance in the concentration estimates. This study used a strictly empirical quantification of these terms, derived from the autocovariance of the estimated concentrations.

$$var\left(\sum_{t=1}^{T} l_t\right) = \sum_{t=1}^{T} q_t^2 (var(\hat{c}_t) + \sigma_{res}^2) + 2\sum_{i=2}^{T}\sum_{j=1}^{i-1} q_i q_j cov(\hat{c}_i, \hat{c}_j) \qquad (Eq. 1.4)$$

where σ_{res}^2 is the residual variance.

1.3 Results and Discussion

1.3.1 Model Performance

The best model for each constituent and model type was selected based on load predictive power as determined by LOOCV R^2 (Figure 1.2; Table 1.3). For NO_3-N and TOC, R^2 values were >0.75 for both LM and GAM models, indicating that the models explained at least three quarters of total load variance in the validation data. For TP, R^2 was lower, between 0.35 and 0.77. Models with high predictive power for loads did not

necessarily predict concentrations well. For example, the LM and GAM models selected for Stillwater TP had a LOOCV R^2 of 0.77 for load, but only 0.03 for concentration (Figure 1.2), implying that in these cases load dynamics were more strongly determined by flow than by concentration. GAM models slightly outperformed LMs in load prediction for 7 of 9 cases, slightly underperformed linear models in 1 case, and were identical in the remaining case. This tendency for GAMs to outperform linear models is further reflected in the NSE (Figure 1.2). The NSE results indicate that GAMs generally explained more variance in the concentration data, although this advantage was greatly reduced after converting the estimates to loads; even models that explained a small fraction of concentration variance were able to capture well over half of the total load variance.

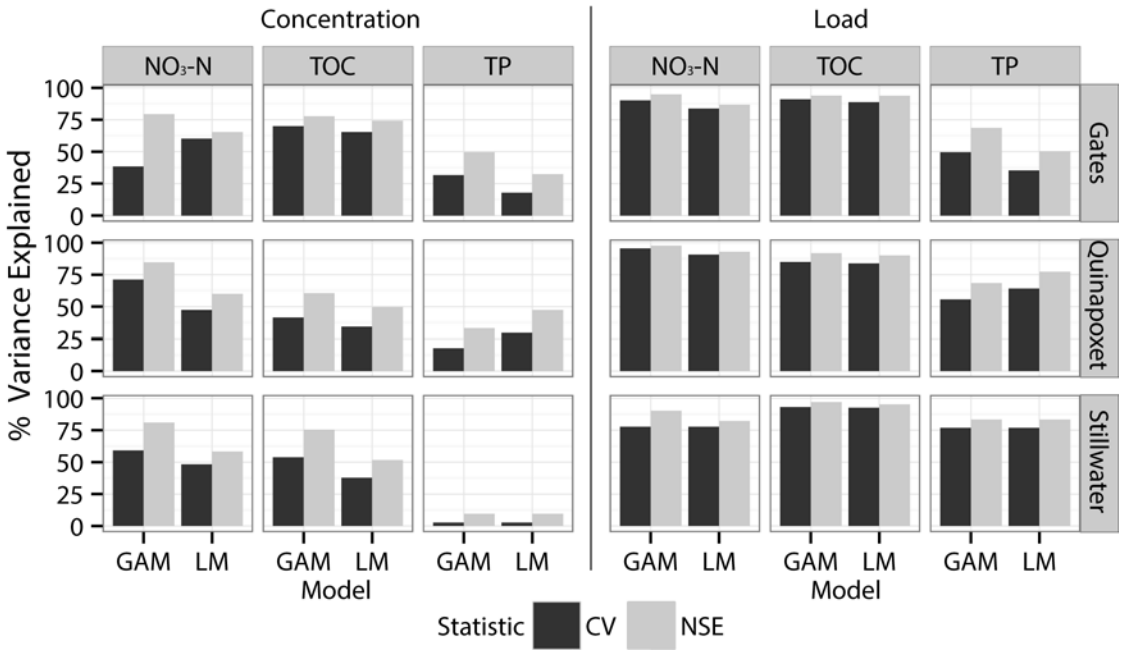

Figure 1.2: Goodness-of-fit statistics for concentrations and loads estimated by GAMs – Generalized Additive Models and LMs – Linear Models. NSE: Nash-Sutcliffe Efficiency; LOOCV: leave-one-out cross-validation R2.

Selected models differed in form between constituents and stations (Table 1.3). GAM models tended to be similar to corresponding LMs for the same constituent and station, and in one case (Stillwater TP) the model was identical to the corresponding LM, implying that no improvement was derived from introducing flexibility to the model functions. In several cases the only difference between LMs and GAMs was the application of a nonparametric smoothing function to a predictor variable (Gates TOC, Quinapoxet NO_3-N, Quinapoxet TOC, Stillwater NO_3-N). Others used different sets of predictor variables (Gates NO_3-N, Gates TP, Quinapoxet TP, Stillwater TOC), and one used different response transformations (logarithmic transformation vs. no transformation for Gates NO_3-N). In these cases introducing model flexibility using

nonparametric functions not only affected the model functional forms, but also allowed other explanatory relationships to be identified and exploited.

GAMs for Gates Brook, which has a small, highly developed catchment, tended to fit nonlinear relationships to hydrologic predictor variables, whereas GAMs for the two larger tributaries' models fit linear relationships to hydrologic variables, and nonlinear terms to time variables only (Table 1.3). In all cases except Stillwater TOC, model forms other than the traditional log-log rating-curve were selected, achieving better predictions using untransformed concentration, untransformed predictors, or both. The selected models for larger tributaries were more likely to use log-transformed concentration although half of the models employed untransformed concentration. Both discounted differenced flow and log-transformed flow proved to be important predictor variables as these variables were selected in 4 of 9 models for both GAMs and LMs. This suggests that effects of hysteresis and changes in hydrology are important indicators of constituent concentrations in the study watershed.

1.3.2 Load Estimates

Time series of model estimates show time-varying disparities between GAM and LM concentration and load estimates. Figure 1.3 shows these estimates for Gates NO_3-N and Stillwater TOC as illustrative cases. Concentration estimates had larger relative differences than load estimates, reflecting the strong dependence of loads on flow. Models with similar sets of predictor variables, such as those for Stillwater TOC (Figure 1.3 a) had greater agreement in concentration estimates than models with different

predictors, such Gates NO_3-N (Fig 3 b). Even cases with large disagreement in estimated concentration had similar load estimates between the two models, with the only sustained load disparities arising from a combination of high flow and poor concentration agreement (Figure 1.3 b).

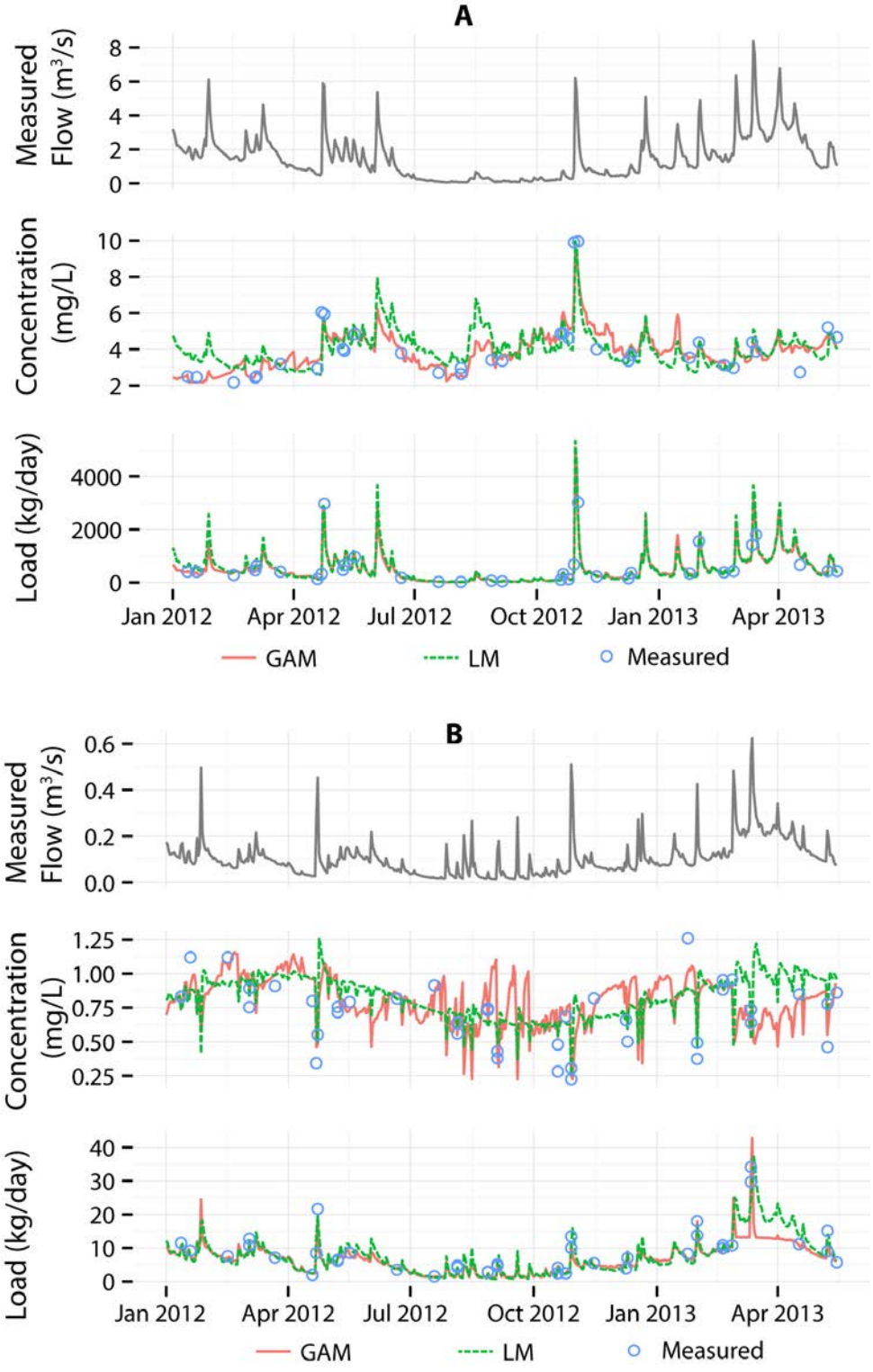

Figure 1.3: Flow, concentration, and daily load measurements and estimates from GAM – Generalized Additive Model and LM – Linear Model for (a) TOC in Stillwater River, and (b) NO_3-N in Gates Brook

On average, loads from Quinapoxet and Stillwater were substantially larger than loads from Gates Brook, reflecting their larger drainage areas (Table 1.4). Quinapoxet contributed the most NO_3-N, while Stillwater contributed slightly more TOC and TP. On a per-unit-area basis, Gates Brook contributed more NO_3-N and TP than the larger tributaries, while TOC contributions were fairly even across tributaries. The largest of these differences is in NO_3-N, with Gates Brook having export coefficients between 5 and 6 times greater than those from Stillwater and Quinapoxet. This is evidence of nutrient load contributions from sources such as urban runoff and septic systems, as Gates is both the most urbanized and has the highest density of on-site septic systems of the three subbasins.

Table 1.4: Estimated loads and export coefficients from mean of GAM and LM models

Variable	Station	Area (ha)	Load (kg/yr)	Daily Load (kg/day)	Export Coefficient (kg/ha/yr)
NO_3-N	Gates	506	2702	7.4	5.34
	Quinapoxet	9175*	9131	25.02	1.00
	Stillwater	7886	7131	19.54	0.90
TP	Gates	506	295	0.81	0.58
	Quinapoxet	9175*	1322	3.62	0.14
	Stillwater	7886	1362	3.73	0.17
TOC	Gates	506	10519	28.82	20.79
	Quinapoxet	9175*	190987	523.3	20.82
	Stillwater	7886	194693	533.4	24.69

* Effective area. Water from 36% of the 14340 ha Quinapoxet subbasin area is diverted out of the watershed as supply for the City of Worcester.

Uncertainty analysis, available for LM- and GAM-derived estimates but not the more simplistic interpolation estimates, establishes the degree of trust that can be placed in any single load estimate. This is partly a function of model goodness-of-fit (reflected by the σ_{res}^2 term in Equation 1.4), and provides some context for evaluating whether a better-fitting model yields a more precise result. The results of this analysis are presented as 95% confidence intervals (error bars in Figure 1.4), and were of similar magnitude for GAMs and LMs. Only two cases yielded noticeable differences in precision between the two models, with the GAM having lower uncertainty in Gates NO_3-N and the LM having lower uncertainty in Quinapoxet TP.

For the entire 18-month study period, load estimates ranged widely between different estimation methods (Figure 1.4). While interpolation-derived estimates often differed greatly from regression-based estimates (e.g. Gates TOC), load estimates from GAMs and LMs were generally comparable (Figure 1.4), with one typically in the 95% confidence interval of the other. The large disagreement between interpolation-derived load estimates and regression-based estimates exposes biases arising from ignoring dependencies between hydrologic condition and concentration. Interpolation estimates depended strongly on the data used, with the inclusion of storm samples producing load estimates much different than those produced from routine samples only (Figure 1.4). The amount and direction of this disagreement reflects basin- and solute-specific aspects of the hydrology-concentration relationship, with storm conditions having higher concentrations of TP and TOC, and lower concentrations of NO_3-N; these differences were especially pronounced in Gates Brook. TP had the largest differences

between estimation methods and largest relative uncertainties, reflecting its complex transport dynamics and strong dependence on hydrologic condition. Phosphorus is present in both particulate and dissolved phases, with various sources and sinks in the watershed and riparian zones. Despite this complexity, GAMs and LMs produced confidence intervals indicating general agreement on TP estimates.

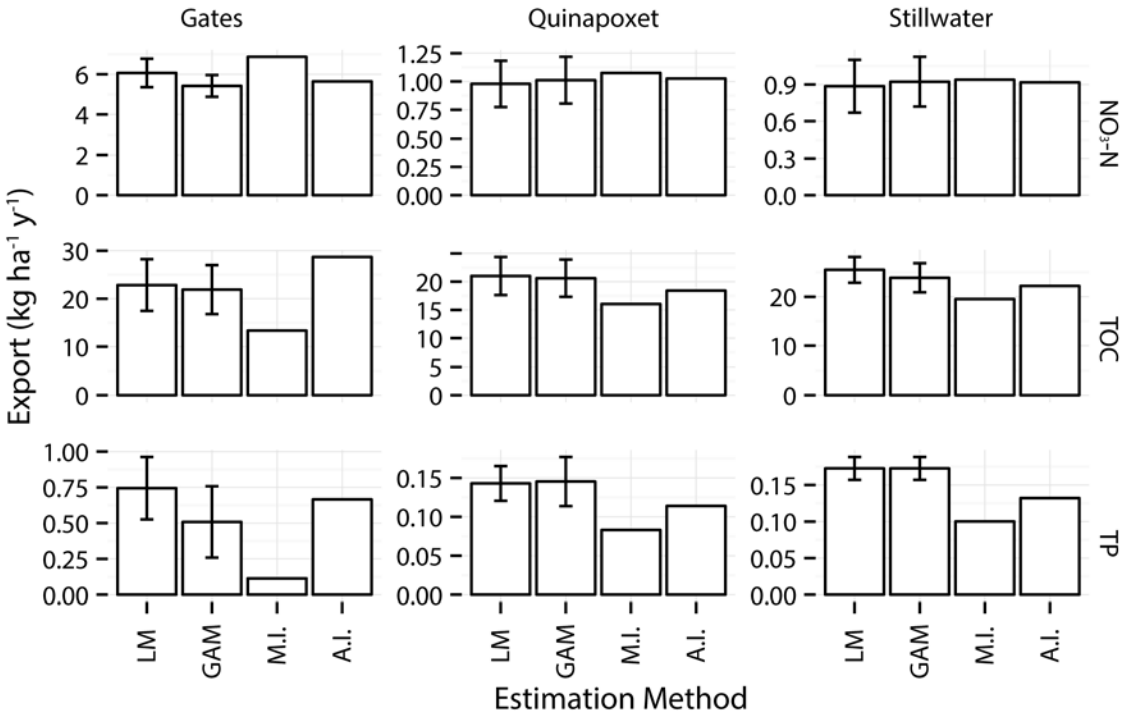

Figure 1.4: Calculated export coefficients (kg $ha^{-1}y^{-1}$) using daily concentration estimates from different methods: LM - linear model; GAM - generalized additive model; M.I. - linear Interpolation of Monthly concentration data (no storm samples); A.I. - linear Interpolation using All concentration data (including storm samples). Error bars are 95% confidence intervals. Note that the vertical scale is different for Gates vs. Quinapoxet and Stillwater

Differences in relative uncertainties between Gates Brook and the two larger tributaries (Figure 1.4) can be understood in terms of subbasin characteristics. Export

coefficients for nutrients are higher, reflecting anthropogenic sources in the highly developed Gates Brook subbasin. Gates Brook is flashier than the other tributaries, and this variability in flow, coupled with high concentration during storm events, could explain the large uncertainties in TP and TOC for this tributary. In contrast, the relatively low uncertainty in Gates Brook NO_3-N loads points to a steady supply of groundwater-sourced nitrogen during baseflow conditions, with storm events contributing a smaller fraction of the total load compared to other constituents.

To investigate differences over smaller time periods, load estimates were compared between LMs and GAMs for a series of durations ranging from single-day to annual, and spanning the study period. These comparisons are presented as two-dimensional plots, with the time at which the load estimate is centered on the x-axis, and the duration of the load interval on the y-axis, with shading corresponding to the percent difference from one method to another (Figure 1.5). For example, the 90-day period centered on 1 September 2012 (roughly 15 July to 15 October) had a discrepancy of between 10 and 20 percent (light blue) in TP load estimates at Gates Brook, while the 7-day period centered on 1 February 2012 had a discrepancy of between 90 and 100 percent (dark orange) for the same quantity (Figure 1.5 a). Here the difference is measured relative to the LM estimate. To the authors' knowledge, this is a novel approach to assessing differences between load estimates.

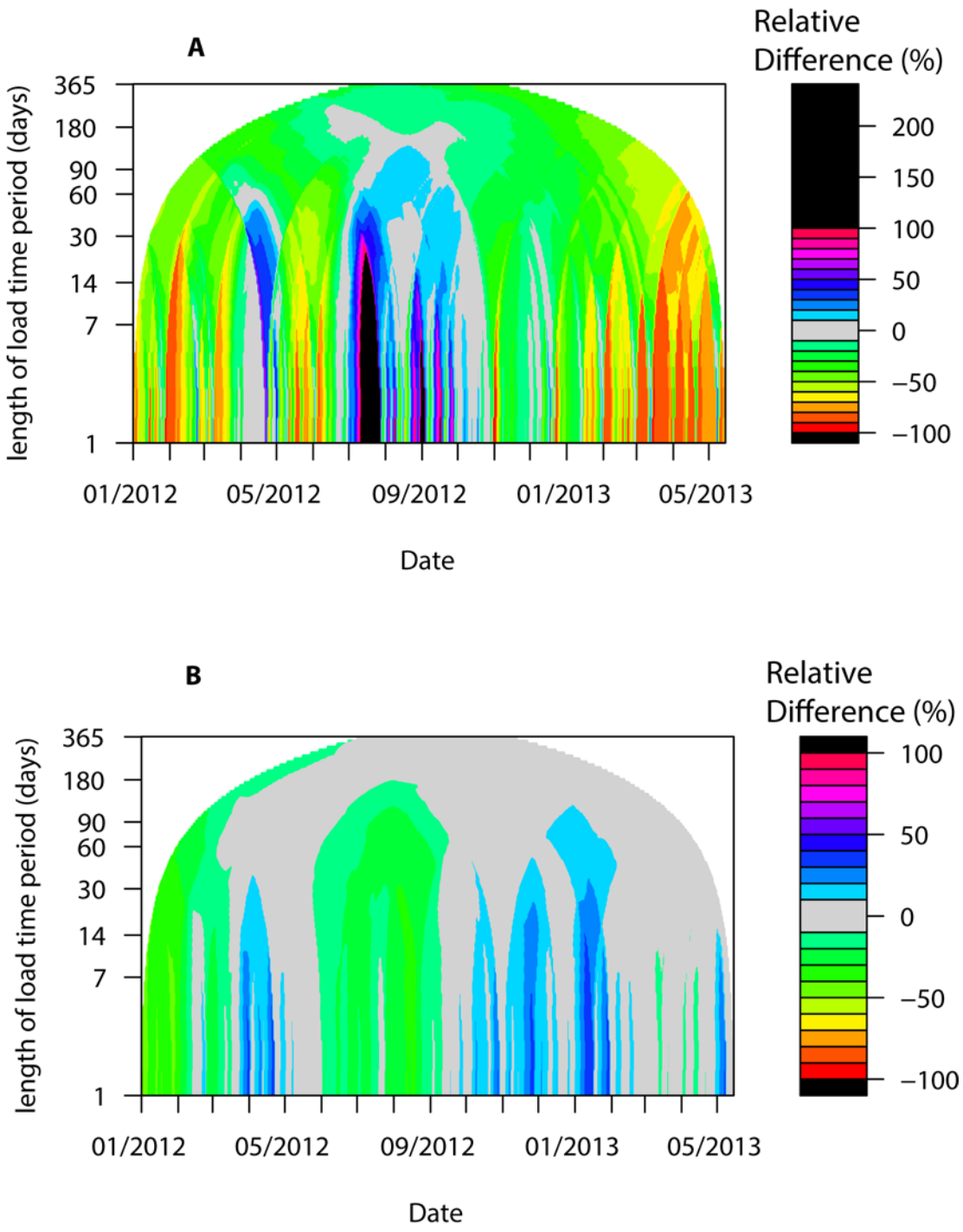

Figure 1.5: Differences in load estimates from GAMs and LMs for (a) TP in Gates Brook and (b) TOC in Stillwater River. Color indicates the relative difference between GAM and LM load estimates for a time interval centered on the date indicated on the horizontal axis, and with duration indicated on the vertical axis. Differences are relative to the LM ((GAM – LM) / LM).

The relative differences in load estimates given by the different model types are greatest for shorter time intervals. In most cases these differences dissipate with longer time intervals such that 1-year load estimates differ by less than 10% due to the averaging effect of summing loads over longer time periods (Figure 1.5 b), although in some cases differences persist over such intervals (Figure 1.5 a). Where significant, differences in load estimates tend to propagate from large-magnitude, short-duration differences in load estimates corresponding to hydrologic storm events. For example, much of the discrepancy in load 180-day Gates TP load estimates for early 2013 (yellow-green towards top right of Figure 1.5 a) stems from a series of high-flow events in spring 2013 (Figure 1.3).

Such a time-domain investigation is useful for at least two reasons. From a monitoring standpoint, the approach reveals the hydrologic conditions that are least well understood—or well captured—using empirical models. These conditions are those that coincide with large differences between the different models, and whose effects persist over longer load estimation periods. For example, in Gates Brook TP a series of summer rain events corresponded with large disagreement between models, with the GAM predicting higher loads than the LM. The same two models also disagree strongly for spring snowmelt conditions in spring 2013 (but not spring 2012, which followed a relatively snowless winter), with the LM predicting larger loads. Monitoring efforts could be focused on these conditions to improve conceptual understanding of watershed behavior. Second, this approach could motivate model improvement, revealing

conditions where the current models fail--or at least disagree with one another. This could prompt efforts at validation, or suggest adding model complexity, for example using data stratification (Guo et al., 2002) or interaction terms (Swistock et al., 1997). However, if time-domain analysis reveals that models are generally in agreement for time intervals of interest, then a simpler modeling approach may be suitable.

The results of this study highlight two advantages of using semiparametric models. The first is flexibility; GAMs can automatically fit arbitrary nonlinear relationships in the calibration data, rather than requiring specific higher-order polynomial and/or harmonic terms. While linear models are prone to bias if the assumption of linearity between predictors and response does not hold, a semiparametric model incorporates this assumption check internally, reverting to a linear model when the relationship is sufficiently linear. In the present study, linearity assumptions for hydrologic variables were found to be valid in most cases, with selected GAMs most often applying smooth functions to variables for time and season only. A semiparametric approach can be used to identify a suitable model (i.e. one whose assumptions are valid) and the resulting load estimates are less likely to be biased when using such a model. This is true for both total load estimates and estimates of uncertainty. A second advantage is ease of model selection. The inherent flexibility of GAMs obviates the complication of selecting from many models with higher order terms, simplifying the model selection process. Whereas fitting complicated long-term and seasonal fluctuations requires estimating many parameters for high-order polynomials in the LM framework, GAMs accomplish this task much more easily.

1.4 Conclusions

This study demonstrates that choices of model type are not inconsequential for load estimation and can result in large relative differences that persist at annual time scales and beyond. Even models whose estimates of long-term loads are similar showed substantial discrepancies at shorter time scales, and time-domain investigation was useful for elucidating the scales at which large differences exist, revealing gaps in understanding and potentially informing monitoring and modeling efforts. Ultimately it may be impossible or infeasible to distinguish which of multiple models is most suitable for estimation, particularly when both are comparable in quantitative measures of performance. Where observed, such a result reveals additional uncertainty in the resulting load estimate that is not captured in uncertainty estimates given by a single model. Future work should explore this source of load uncertainty.

GAMs proved to be a versatile tool for load estimation, fitting nonlinear relationships between concentration and hydrologic variables where such relationships were present, and otherwise manifesting as linear models. Such an approach to load estimation is less prone to bias associated with improperly assuming a linear (or log-linear) model form. The ability of GAMs to exploit relationships not well described by linear regression models led to models with enhanced explanatory and predictive capacity, and ultimately to more accurate load predictions, compared with those from linear models. Future work will include a systematic comparison of LMs and GAMs across a wide range of water-quality datasets on a regional scale, including Wachusett Reservoir tributary data from outside the time period of the present study.

CHAPTER 2

CAPACITY OF SEMIPARAMETRIC REGRESSION MODELS TO PREDICT EXTREME-EVENT WATER QUALITY IN THE NORTHEASTERN U.S.

2.1 Introduction

Concentration and mass flux of riverine constituents are two environmental parameters that are of high social and environmental interest, yet difficult to measure with satisfactory time resolution. This has led to widespread modeling efforts that attempt to fill in the resulting gaps in the observed time series, typically using regression models--often referred to as "rating curves"--that estimate log-transformed concentration or mass flux as a function of more easily measured variables including discharge.

The empirical (and explicit, in the case of mass flux) relationships with hydrological variables on which such models rely lead to a natural focus on moments of large variations in flow, i.e. on storm conditions. These periods constitute transport "hot moments" (*Vidon et al.*, 2010) for many riverine constituents, and are responsible for transporting over half of all mass loads in many cases (*Raymond and Saiers*, 2010). Several recent extreme high-flow events in the US Northeast, including Hurricanes Irene (August 2011) and Sandy (October 2012) have prompted increased scientific attention on the impacts such events have on transport of constituents including nutrients (*Yoon and Raymond*, 2012), organic matter (*Caverly et al.*, 2013; *Dhillon and Inamdar*, 2013, 2014), and suspended sediment (*Yellen et al.*, 2014). Several of these studies point to

disproportionately large exports during such events, exceeding model predictions. However, these studies focus primarily on quantifying exports from individual storms and do not make a systematic assessment of model performance under such extremes.

A guiding principal for hydrologic modeling was stated in *Klemeš* (1986)--"Before it is used operationally, a model must demonstrate how well it can perform the kind of task for which it is intended". As modelers increasingly seek to predict impacts of previously unobserved weather and climate conditions (e.g. *Carpenter et al.*, 2015), empirical constituent models may be used to predict water-quality responses to a hypothetical extreme storm event, or to estimate unmeasured conditions during an actual event. Other environmental modeling disciplines have attempted to establish the range of climatic conditions under which their models yield acceptable results (*Andréassian et al.*, 2009; *Coron et al.*, 2012), but to date no systematic assessment has been made of rating-curve models under extreme hydrologic conditions.

This study addresses the question of how well a rating-curve model makes predictions in extreme-flow conditions, given that such conditions are beyond the range seen in its calibration data. Before addressing this question, it is useful to lay out some of the assumptions upon which rating-curve models rest.

1. The mean (μ_{lnC}) of the random variable representing log-transformed concentration (lnC) is a function of log-transformed flow and other variables such as season and time. The mean can thus be written conditionally on a set of measurable variables X: $\mu_{lnC|X}$. As predictors in a regression model, these variables explain a portion of the variance in the modeled quantity (i.e. the "response

variable"); knowing the value of the predictors reduces the uncertainty in the response. In most cases, the functional relationships are assumed to be linear or quadratic with respect to a transformation of the predictors (e.g. logarithm for flow, harmonic for season; (*Cohn et al.*, 1992)). These assumptions can lead to model bias where they are incorrectly applied (*Hirsch*, 2014), and other functional forms have been introduced in extensions of the linear model (*Autin and Edwards*, 2010; *Hirsch et al.*, 2010; *Wang et al.*, 2011).

2. The variance (σ_{lnC}^2) and standard deviation (σ_{lnC}) of log-transformed concentration are constant for all times and flow conditions. Often concentration is assumed to follow a log-normal distribution, i.e. $lnC \sim N(\mu_{lnC}, \sigma_{lnC}^2)$ (*Esmen and Hammad*, 1977; *Helsel and Hirsch*, 2002), but this assumption is not required except when making probabilistic inferences such as confidence intervals and hypothesis tests (*Helsel and Hirsch*, 2002).

One mathematical consequence of these assumptions is that the standard deviation of concentration ($\sigma_{C|X}$) is directly proportional to the conditional mean, implying that larger estimates have larger uncertainty. In the case of log-normality, the proportionality constant grows exponentially with increasing σ_{lnC}^2, the variance of lnC, about its conditional mean: $\sigma_{C|X} = (e^{\sigma_{lnC}^2} - 1)^{\frac{1}{2}}\mu_{C|X}$. As a result, concentration estimates during large hydrologic events are inherently more uncertain than those for less extreme events, whereas estimates of log-transformed concentration have similar precision for all conditions. This study therefore sought to evaluate whether rating-

curve predictions of lnC retain their predictive capacity in extreme high-flow events, relative to their performance in less extreme conditions. We pay specific attention to the supposition of thresholds beyond which constituent behavior undergoes fundamental changes (*Dhillon and Inamdar*, 2013, 2014), rendering models inaccurate (*Yoon and Raymond*, 2012). We further make recommendations about the collection, management, and dissemination of water-quality data in order to improve large-scale data-driven studies.

2.2 Materials and Methods

2.2.1 Notation

This study used data from multiple sites, with each site potentially having multiple datasets and each dataset having multiple calibration and validation data. The following notation is used to distinguish these variables.

- M: the number of datasets, equivalent to the number of models
- n_m: the number of data used to calibrate model m, $m = 1,2,\ldots,M$
- $N = \sum_{m=1}^{M} n_m$: the total number of calibration data across all datasets
- v_m: the number of validation data from dataset m
- $V = \sum_{m=1}^{M} v_m$: the total number of validation data across all datasets
- $lnC_{i,m}$, $i = 1,\ldots,n_m$: the i^{th} observation of log-transformed concentration from the m^{th} dataset's calibration data
- $lnC_{j,m}$, $j = 1,\ldots,v_m$: the j^{th} observation of log-transformed concentration from the m^{th} dataset's validation data

The above subscript indices are used equivalently to differentiate between other model- and data-specific variables.

2.2.2 Data acquisition

Concentration data for streams in the US Northeast were obtained from the National Water Quality Monitoring Council Water Quality Portal (http://www.waterqualitydata.us/). The initial database query extracted all water-quality data for selected constituents from stream monitoring stations located in the Northeast US between 36°N and 48°N latitude and between 81°W and 66°W longitude (Figure 2.1). Daily discharge data for water-quality monitoring sites were obtained from the US Geological Survey (USGS) National Water Information System (NWIS). The data were filtered to include only the datasets with at least 30 concurrent measurements of concentration and discharge for a given station and constituent. A total of 2747 datasets were obtained from 459 monitoring stations, with each dataset representing a unique combination of constituent (nutrients such as nitrogen and phosphorus, organic carbon, and total suspended solids), fraction (dissolved, suspended, or total), and monitoring station (Table 2.1). In some cases where it was not explicitly provided the "suspended" fraction was calculated from the difference between "total" and "dissolved" fractions, while the fractions of certain dissolved constituents reported as "total" were discarded following USGS recommendations (*Ricket*, 1992). Each dataset contained between 31 and 3098 concurrent observations of flow, concentration, and date of measurement.

Table 2.1: Summary of DSST data.

Constituent	Fraction	no. stations	no. samples (entire database)	no. samples (DSST validation set)
Ammonia	Dissolved	202	29551	634
Kjeldahl nitrogen	Dissolved	118	15007	339
Kjeldahl nitrogen	Suspended	100	10975	302
Kjeldahl nitrogen	Total	173	31290	564
Nitrate	Dissolved	177	28049	471
Nitrite	Dissolved	127	16103	390
Nitrogen	Dissolved	43	4092	12
Nitrogen	Suspended	56	4178	63
Nitrogen	Total	165	10671	26
Organic Carbon	Dissolved	57	5795	106
Organic Carbon	Suspended	44	4287	87
Organic Carbon	Total	67	10508	177
Organic Nitrogen	Dissolved	109	12555	371
Organic Nitrogen	Suspended	82	8705	283
Organic Nitrogen	Total	133	23423	508
Organic phosphorus	Dissolved	2	79	0
Phosphate	Dissolved	222	35393	644
Phosphorus	Dissolved	162	23154	435
Phosphorus	Suspended	149	21088	428
Phosphorus	Total	266	43080	739
Total suspended solids	Suspended	221	20794	342
	Total:	459	358777	6921

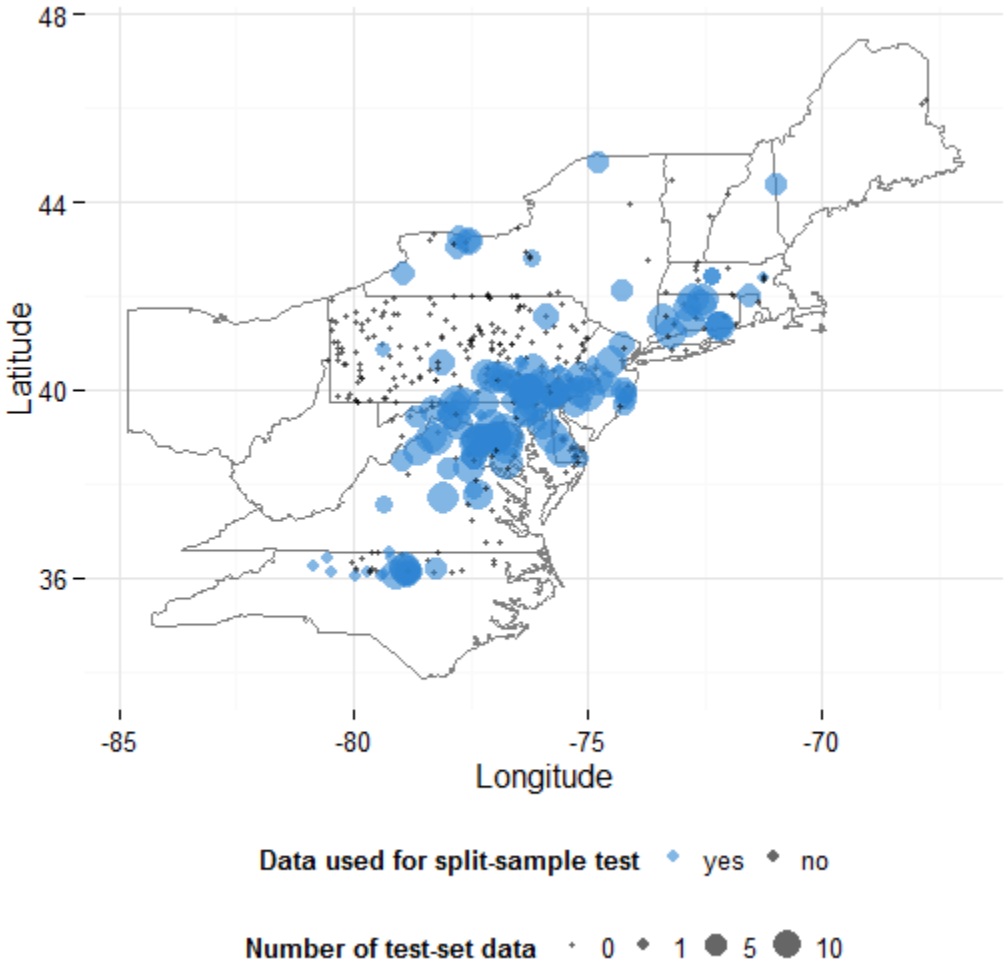

Figure 2.1: Map of dataset locations. Blue stations included concentration samples from sufficiently high flow conditions to be included in split-sample-test. Larger blue markers indicate sites that contributed more test-set data to the split-sample test.

2.2.3 Model development

A semiparametric rating-curve model (*Wang et al.*, 2011; *Kuhnert et al.*, 2012) was calibrated to each discharge-concentration dataset using the mgcv R package (*Wood*, 2011). This model is similar to a traditional rating-curve in which log-

transformed concentration is linearly regressed on log-transformed discharge and other variables representing seasonal and long-term fluctuations. The main difference is that functional relationships may be arbitrarily nonlinear. The model has the form

$$lnC = s_1(lnQ) + s_2(doy) + s_3(time) + \epsilon \qquad (Eq\ 2.1)$$

where lnC is log-transformed concentration; lnQ is log-transformed flow; doy is the numeric day of the year, (1-365 or 1-366); $time$ is the time of observation in days from the mean observation time; and ϵ is a zero-mean, constant-variance error term. The functions $s_1()$, $s_2()$, and $s_3()$ are nonparametric smooth-functions based on cubic splines (*Wood*, 2006). Although other rating-curve models have used covariates other than time, season, and discharge, these three are by far the most commonly used (*Hirsch*, 2014).

2.2.4 Differential split-sample test

In order to simulate model performance in a previously unobserved extreme hydrologic condition, a calibration-validation procedure was developed based on the differential split-sample test (DSST) proposed by *Klemeš* (1986). As originally described in the context of hydrological simulation models, the test involves dividing the available calibration data according to climatic condition, partitioning it into, for example, wetter-than-average and drier-than-average data subsets. In order to establish a model's predictive capacity under a wetter climate, the model is calibrated using the dry partition and validated using the wet partition. Our application of this approach for constituent rating-curve models entails splitting the observed data into a calibration

dataset constituting non-extreme cases and a validation dataset constituting the remaining extreme cases.

In order to partition data into "extreme" and "non-extreme" flow conditions, we defined a flow-based measure of hydrologic extremity, q_s, as the number of standard deviations of log-transformed flow from its long-term mean:

$$q_s = \frac{lnQ - \overline{lnQ}}{sd(lnQ)}$$

where $\overline{lnQ}$ and $sd(lnQ)$ are each station's observed mean and standard deviation, respectively, of log-transformed flow, calculated from the full record of daily flows. The subscript s represents "standardized", leading to a quantity that has zero mean and unit variance. The scaling of this quantity allows it to be readily compared across different stations with different flow characteristics, e.g. allowing the flow condition of a small, flashy stream to be compared to that of a large, steady river. In this case, "extreme" flow was defined as $q_s > 3$, corresponding to a high-flow condition with at least a 2-year return interval assuming log-normality of discharge. (Note that log-normality is not a requirement for considering q_s as a measure of flow extremity.) It also corresponded very nearly to the median of all stations' maximum sample-day flow ($\text{median} = 2.87$). Therefore, restricting a model's calibration data to $q_s < 3$ was equivalent to calibrating it using the full range of high-flow conditions available in a typical dataset.

Although $q_s = 3$ was used as the cutoff for defining an extreme event, much larger q_s values were present in the test set. 4 of the 8 most extreme flows, including two separate events with $q_s > 8$, occurred at a single station, USGS-01493112. This

station is known to be influenced by storm tides that can increase water levels, leading to potential overestimation of high flows (http://waterdata.usgs.gov/usa/nwis/uv?01493112). Nonetheless, all flow data used in the present study were marked as "approved" by USGS.

A model of the form given in Equation 2.1 was fit to each dataset using only non-extreme data ($q_s < 3$) for calibration. The model's simulated performance during an extreme event was then evaluated using the withheld validation data ($q_s > 3$). Because extreme events are inherently rare, each dataset contains only a small number of validation data (e.g. only one or two observations), and a single dataset seldom contains sufficient data to obtain a reliable statistic for model performance under extreme conditions. To overcome this limitation, the validation errors from all DSST validation sets were aggregated into a single error dataset. To effectively combine errors across different models, each model's prediction errors ($\epsilon_{j,m}$) were scaled prior to aggregation by dividing by the square-root of the calibrated model's generalized cross-validation score (GCV), an approximation of its mean-square prediction error (*Wood*, 2006):

$$\epsilon_{j,m} = \frac{lnC_{j,m} - \hat{lnC}_{j,m}}{\sqrt{GCV_m}}$$

where $\hat{lnC}_{j,m}$ is the model prediction of $lnC_{j,m}$. This ensured that statistics computed using the aggregated errors could be interpreted relative to each model's predictive power for less extreme cases. If the validation-set prediction errors for model m have mean and variance equal to those for the calibration set, then the scaling ensures that $\epsilon_{j,m}$ will have zero mean and unit variance. To investigate the impact of increasingly

extreme conditions, the validation-set errors were binned into increasingly extreme hydrologic conditions: q_s between 3 and 4 (n = 5729), between 4 and 5 (n = 1088) and above 5 (n = 104).

2.2.5 Metrics of model performance

Model performance on the calibration partition of the DSST data was determined using 3 related goodness-of-fit statistics (Table 2.2): Nash-Sutcliffe efficiency (NSE), coefficient of determination (R^2), and cross-validation coefficient of determination (Q^2). All three statistics estimate the amount of total variance that the model explains in the response variable. They take on values on the interval $(-\infty, 1)$, with 1 indicating a model that explains 100% of variability in the response variable, and negative values indicating a model that is outperformed by the mean of the response variable. R^2 and Q^2 are measured with respect to the log-transformed values that are modeled explicitly (Equation 2.1), whereas NSE is defined with respect to values after the log-space predictions are retransformed into their original units, e.g. mg/L (*Bennett et al.*, 2013). Whereas R^2 and NSE are measured using the same data used for model calibration, Q^2 uses prediction errors from cross-validation instead, thereby avoiding the potential to be affected by overfitting (*Quan*, 1988). Since its definition with respect to cross-validation errors makes Q^2 a true measure of model predictive capacity--rather than model *fit*--it was used as the preferred measure of model performance in this study.

Table 2.2: Statistics used in model calibration, validation, and differential split-sample test.

Description	Symbol	Mathematical Definition
Coefficient of determination	R^2	$1 - \sum_{i=1}^{n} (lc_i - \hat{lc_i})^2 / \sum_{i=1}^{n} (lc_i - \overline{lc})^2$
Cross-validation coefficient of determination	Q^2	$1 - \sum_{i=1}^{n} (lc_i - \hat{lc_{(i)}})^2 / \sum_{i=1}^{n} (lc_i - \overline{lc_{(i)}})^2$
Nash-Sutcliffe Efficiency	NSE	$1 - \sum_{i=1}^{n} (c_i - \hat{c_i})^2 / \sum_{i=1}^{n} (c_i - \overline{c})^2$
Calibration-set mean-square error of prediction	MSE_{cal}^{pred}	$\frac{1}{n} \sum_{i=1}^{n} (lc_i - \hat{lc_{(i)}})^2$
scaled validation error	ϵ_j	$(\hat{lc_j} - lc_j) / \sqrt{MSE_{cal}^{pred}}$
aggregate bias	$bias_{val}$	$\frac{1}{V} \sum_{i=1}^{V} \epsilon_j$
aggregate variance	var_{val}	$\frac{1}{V} \sum_{i=1}^{V} (\epsilon_j - bias_{val})^2$
scaled mean-square error	$SMSE$	$\frac{1}{V} \sum_{i=1}^{V} \epsilon_j^2$

Multiple statistics were used to assess model predictive capacity under extreme conditions, relative to that in less extreme conditions (Table 2.2). Aggregate bias, defined as the negative mean scaled validation error, assessed the average difference between measured and predicted log-transformed concentration. Scaled mean-squared error ($SMSE$), defined as the average ratio of mean-squared error of each model's validation predictions to that model's calibration-set mean-squared error of prediction, assessed the variability of extreme-case predictions relative to that of non-extreme predictions. The scaling of errors was performed such that an SMSE of 1 corresponded

to a prediction exactly as accurate and precise as predictions made on the calibration set.

$$\begin{aligned} SMSE &= \frac{1}{M}\sum_{m=1}^{M}\frac{MSE_m^{val}}{MSE_{cal,m}^{pred}} \\ &= \frac{1}{M}\sum_{m=1}^{M}\frac{MSE_m^{val}}{GCV_m} \\ &= \frac{1}{V}\sum_{m=1}^{M}\sum_{j=1}^{v_m}\left(\frac{\widehat{lnC}_{j,m} - lnC_{j,m}}{\sqrt{GCV_m}}\right)^2 \\ &= \frac{1}{V}\sum_{m=1}^{M}\sum_{j=1}^{v_m}\epsilon_{j,m}^2 \end{aligned}$$

where MSE denotes mean squared error.

Although statistics such as R^2 and Q^2 are specific to each individual dataset and model, the scaling of validation-set errors employed in this study allowed for a mapping of Q^2 into the extreme-event case for each dataset used in the DSST. The expected value of each model's validation-set Q^2, Q_{val}^2, was calculated using SMSE as follows:

$$\begin{aligned} E[Q_{val,m}^2] &= E[1 - \frac{SSE_{val,m}}{SST_{val,m}}] \\ &= 1 - \frac{SSE_{cal,m}^{pred}}{SST_{cal,m}}E[\frac{MSE_{val,m}}{MSE_{cal,m}^{pred}}] \end{aligned}$$

This is estimated by

$$1 - (1 - Q_{cal}^2)SMSE \qquad (Eq.\,2.2)$$

where SSE and SST are error sum of squares and total sum of squares, respectively. The Wilcoxon signed rank sum test (*Bauer*, 1972), a nonparametric test for difference in median, was applied to test errors for each of the 3 levels of extremity outlined above. This tested the hypothesis that the median error of model predictions in extreme

hydrologic conditions is nonzero, meaning that such predictions are biased. Since the DSST errors contain outliers and are not normally distributed, a nonparametric test is preferred over a parametric test such as the t-test. Separate tests were performed for errors from each constituent type and fraction, in addition to a single test on the entire set of DSST errors for each level of flow extremity.

The R statistical computing platform (version 3.2) was used to calibrate and validate all models.

2.3 Results

2.3.1 Calibration

Of the 2747 datasets obtained, 1204 contained data that met the extremity criterion of $q_s > 3$ and that were included in the split-sample test (DSST) set (Figure 2.2). The DSST validation set contained 6921 data points from 149 stations, comprising 1.9% of all data points acquired from the WQP database. The DSST stations were not uniformly distributed in space, but disproportionately came from basins in Connecticut, North Carolina, and the Chesapeake Bay watershed (Figure 2.1). DSST stations also differed widely in the number of validation-set data they contained, ranging between 1 and 389, with a median of 15. Validation sets for individual constituents and fraction types varied in size from 12 to 739 (Table 2.3). Drainage areas contributing to validation-set stations broadly reflected the size composition of the overall dataset (Figure 2.3), although this was affected by a small number of stations that disproportionately contributed validation-set data.

Table 2.3: Differential split-sample test validation-set size by constituent and fraction.

Characteristic	Dissolved	Suspended	Total
Ammonia	634	NA	NA
Kjeldahl nitrogen	339	302	564
Nitrate	471	NA	NA
Nitrite	390	NA	NA
Nitrogen	12	63	26
Organic Carbon	106	87	177
Organic Nitrogen	371	283	508
Phosphate	644	NA	NA
Phosphorus	435	428	739
TSS	NA	342	NA

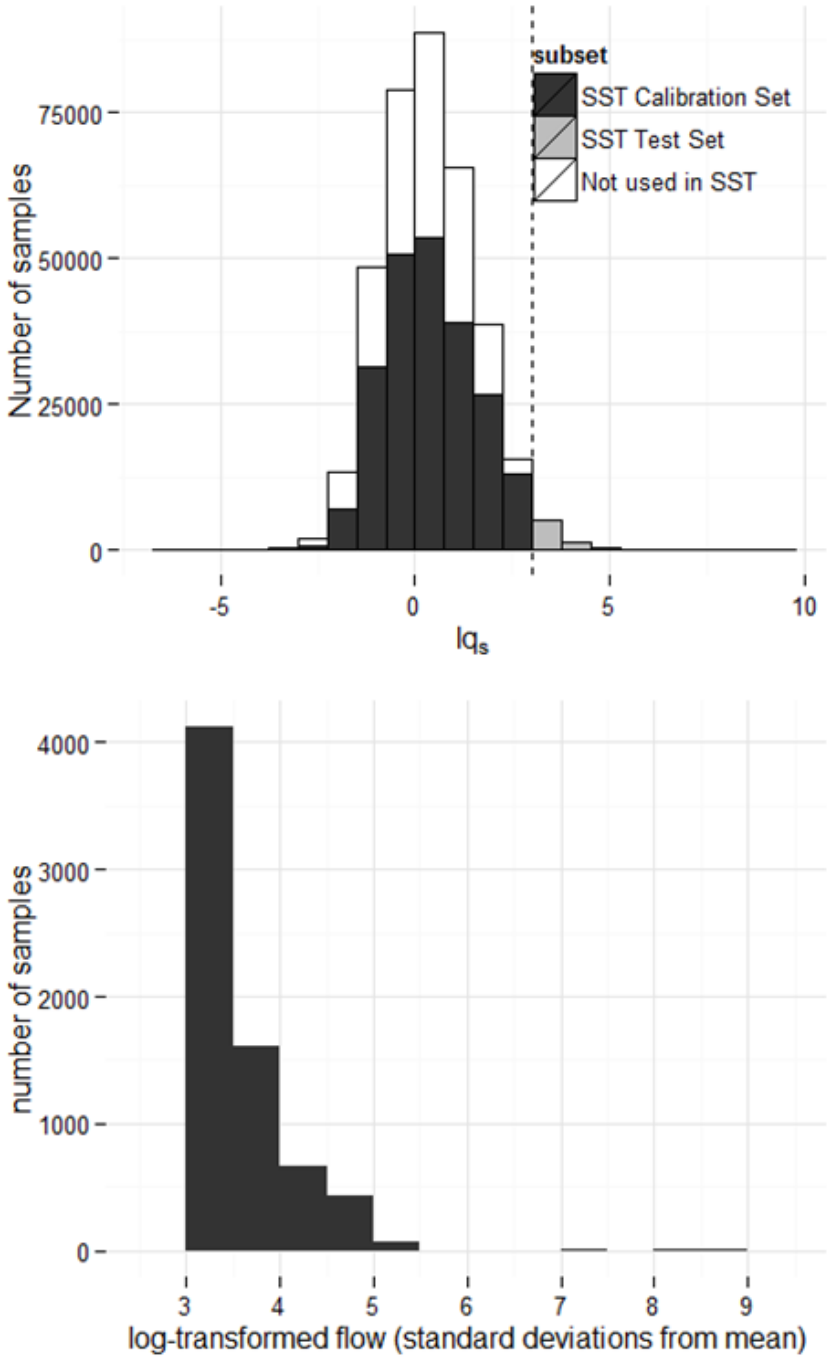

Figure 2.2: Histogram of flows associated with concentration samples in (a) entire database, and (b) split-sample-test validation set

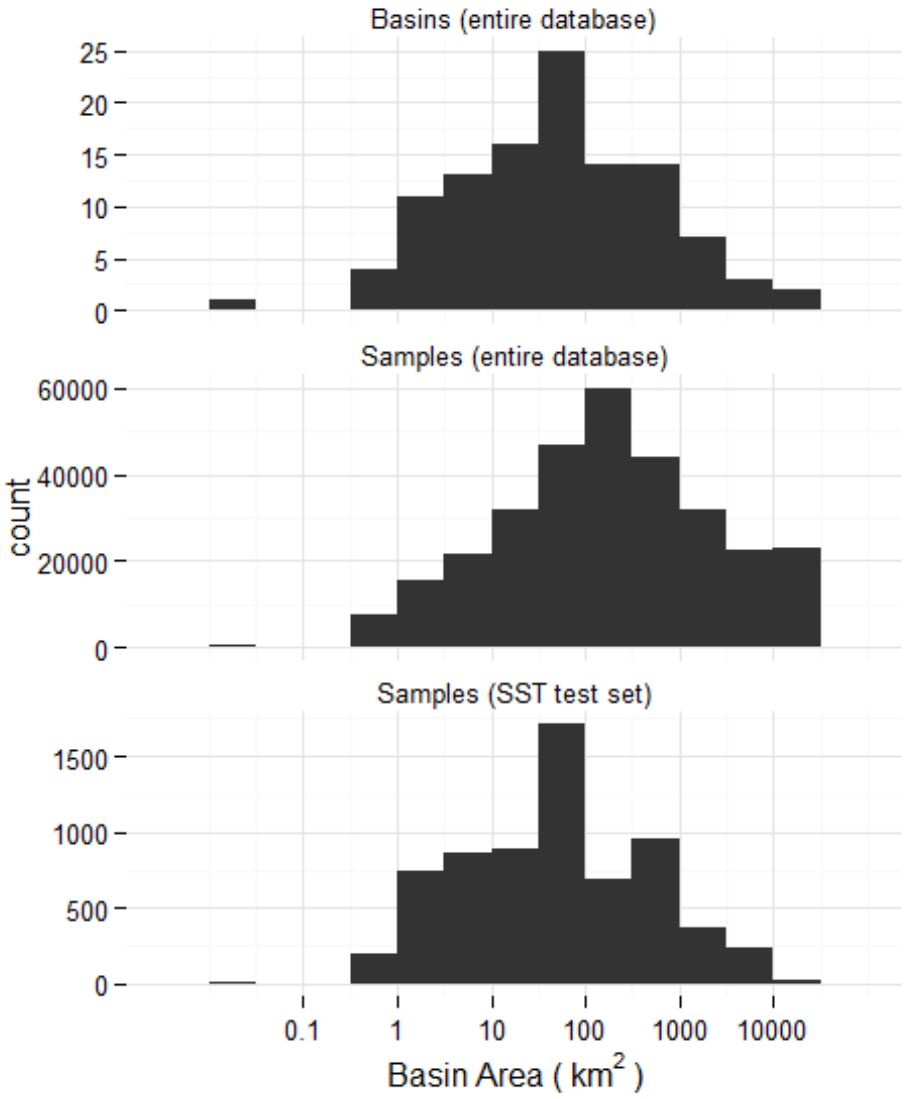

Figure 2.3: Distribution of basin sizes. Top: distribution of all basins in database. Middle: Distribution of basin size across all water-quality samples. Bottom: distribution of basin sizes across SST test-set samples.

The vast majority of calibrated models had goodness-of-fit statistics greater than zero, with most values falling between 0.2 and 0.6 (Figure 2.4). NSE values were generally lower than Q^2, which were (by definition) lower than R^2. Goodness-of-fit was

similar across fraction types, with dissolved fractions having slightly higher NSE on average, and slightly lower R^2 and Q^2 on average compared to suspended and total fractions.

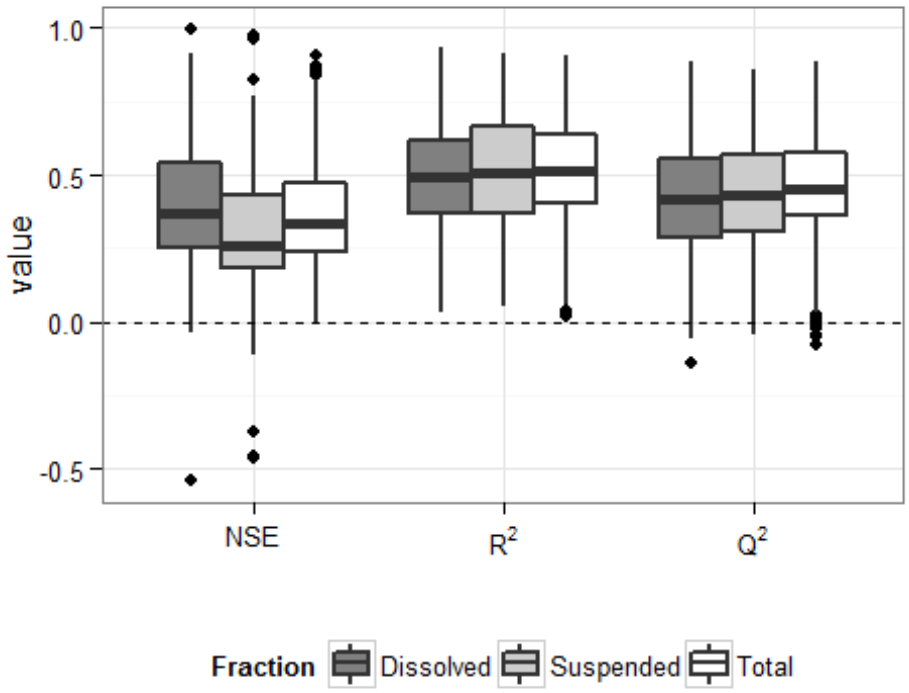

Figure 2.4: Distribution of goodness-of-fit statistics for calibrated SST models, shaded by constituent fraction.

2.3.2 Validation

Scaled DSST validation-set errors were aggregated across the 1204 calibrated models creating an error set of size V = 6921. In the aggregated data, extreme-case predictions exhibited increasing positive bias and variance with more extreme flows, reflecting a deterioration of model performance and a tendency for models to

overpredict extreme-flow concentrations. Density plots of errors in each range of extremity (Figure 2.5) show degrading model performance with more extreme flows, reflected in the shifting (increasing model bias) and flattening (decreasing model precision) of distributions as flow becomes more extreme. The amount of change was similar for all fraction types (Figure 2.5), although the "total" fraction was shifted somewhat more than the other fractions in the most extreme conditions ($q_s > 5$)

Out of 57 Wilcoxon tests on individual constituent-fraction combinations, 39 were significantly nonzero using a significance level of $\alpha = 0.05$ (Table 2.4, Figure 2.6), indicating a nonzero prediction bias in these cases. An additional 6 constituent-fraction combinations were not possible to test due to insufficiency of data for the particular extremity level (NA values in Table 2.4). Of the statistically significant tests, 16 of 18 indicated positive bias in the $3 < q_s < 4$ interval, as did 15 of 16 and 4 of 5 in the $4 < q_s < 5$ and $q_s > 5$ intervals, respectively. Only nitrate exhibited an increasing negative bias with increasing flow extremity, meaning that models for nitrate tended to underpredict nitrate concentrations during extreme flows.

Table 2.4: Results from Wilcoxon test for nonzero median error values for 3 different levels of hydrologic extremity. Bold numbers are statistically significant at α=0.05.

Fraction	Constituent	(3,4]	(4,5]	(5,8.8]
Dissolved	All	**0.202**	**0.688**	**1.736**
Suspended	All	**0.403**	**1.119**	**2.537**
Total	All	**0.408**	**1.046**	**3.359**
Dissolved	Ammonia	**0.248**	**0.576**	1.380
Dissolved	Kjeldahl nitrogen	**0.279**	**0.916**	2.875
Suspended	Kjeldahl nitrogen	**0.308**	**0.773**	0.812
Total	Kjeldahl nitrogen	**0.361**	**0.734**	2.631
Dissolved	Nitrate	**-0.271**	**-0.870**	**-2.817**
Dissolved	Nitrite	**0.134**	**0.652**	0.881
Dissolved	Nitrogen	**-1.307**	4.364	NA
Suspended	Nitrogen	**0.576**	**1.168**	5.719
Total	Nitrogen	-0.647	3.626	4.980
Dissolved	Organic Carbon	**0.382**	0.613	NA
Suspended	Organic Carbon	**0.493**	**2.546**	NA
Total	Organic Carbon	0.015	0.115	2.273
Dissolved	Organic Nitrogen	**0.226**	**0.688**	2.388
Suspended	Organic Nitrogen	**0.374**	**0.852**	2.125
Total	Organic Nitrogen	**0.466**	**1.001**	2.550
Dissolved	Phosphate	**0.312**	**1.200**	**2.716**
Dissolved	Phosphorus	**0.403**	**1.236**	**3.505**
Suspended	Phosphorus	**0.400**	**0.992**	**2.774**
Total	Phosphorus	**0.522**	**1.353**	**4.156**
Suspended	Total suspended solids	**0.468**	**1.541**	1.936

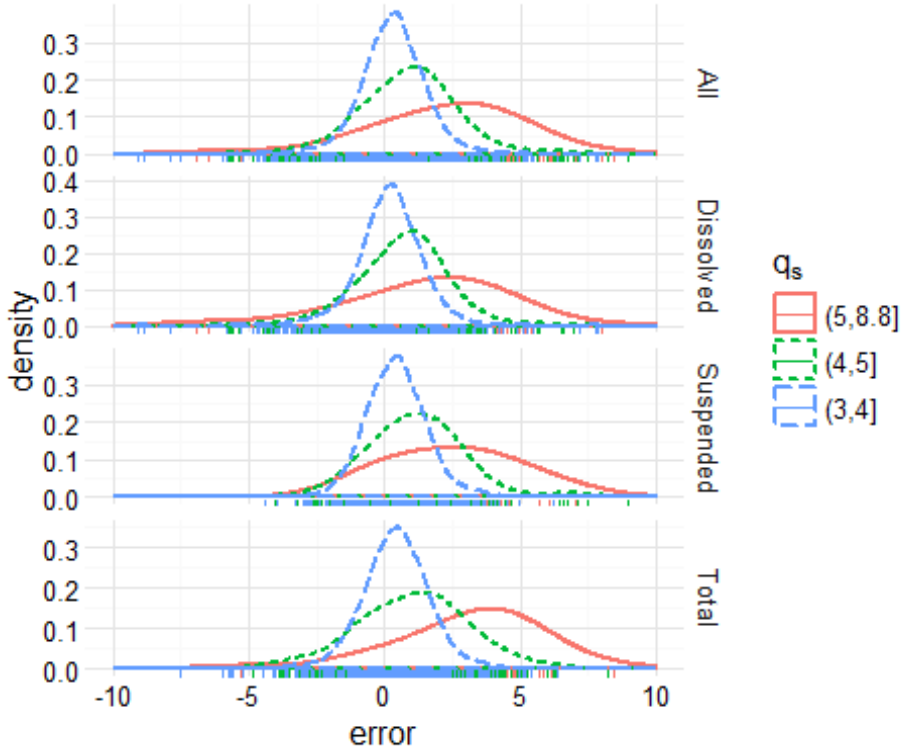

Figure 2.5: Density plots of DSST validation-set errors for 3 different levels of hydrologic extremity.

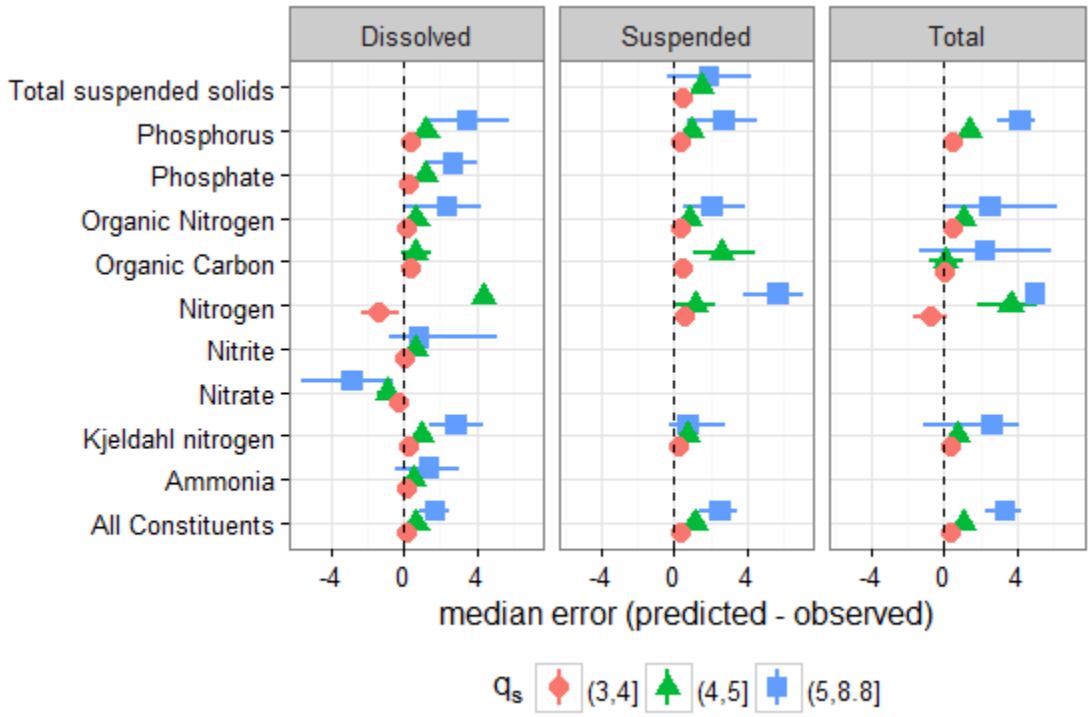

Figure 2.6: Estimates with 95% confidence bounds for median error in validation-set predictions, for 3 different levels of flow extremity.

2.3.3 Impact on model predictive capacity

The deterioration of model performance in extreme flows degraded the Q^2 scores (Equation 2.2), resulting in 55.1% of models having $Q^2_{val} > 0$ (Table 2.5), compared with 99.5% of models having $Q^2_{cal} > 0$. The extent of Q^2 deterioration increased with increasing q_s (Figure 2.7), as did its uncertainty. For models with $Q^2_{cal} > 0.75$, the associated Q^2_{val} was almost always greater than 0, and greater than 0.5 for the $3 < q_s < 4$ flow condition. Thus all but the best-performing models were outperformed by the calibration-set mean in the extreme case.

Table 2.5: Q^2 scores for models in calibration and validation predictions.

Variable	% above 0	% above 0.5
Q^2_{cal}	99.52	39.06
Q^2_{val}	55.13	8.20
$Q^2_{val,shift}$	62.53	9.71

Since Q^2_{val} decreases as a function of $SMSE$ (Equation 2.2), it can be improved by reducing $SMSE$, for example by accounting for model bias. Bias-adjusting the validation estimates--by subtracting the median error for each fraction and flow condition--resulted in higher Q^2_{val} particularly for $q_s > 5$ (Figure 2.7), but the resulting improvement in aggregate model performance was marginal (Table 2.5).

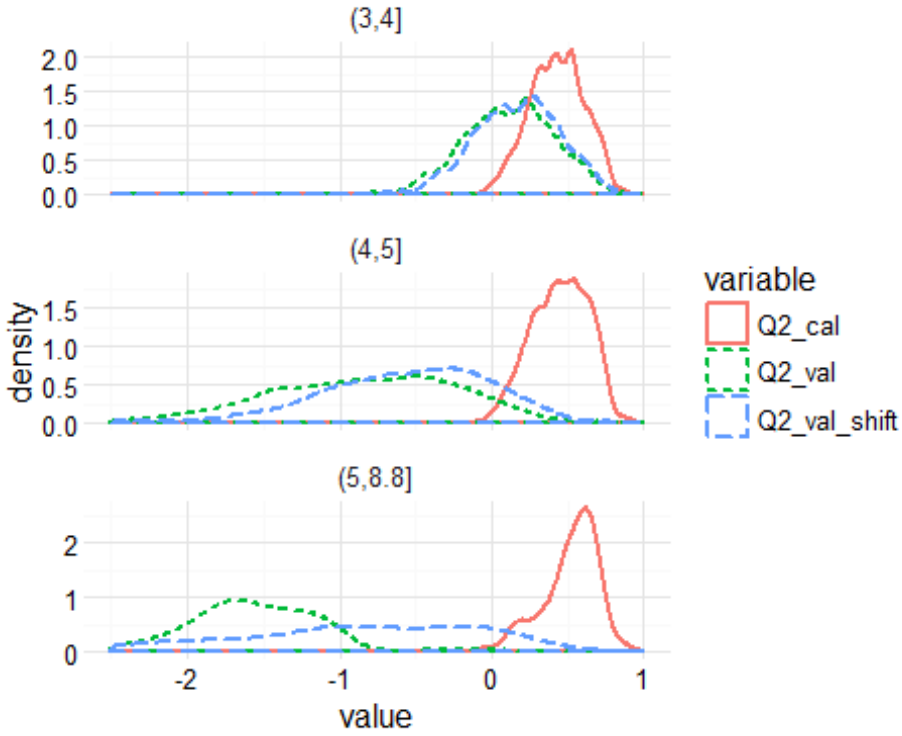

Figure 2.7: DSST calibration-set, validation-set, and bias-adjusted validation-set Q^2.

2.4 Discussion

This study was unique in that it incorporated water-quality measurements taken during many extreme-hydrologic events from a large number of monitoring stations into a single validation dataset that could be used to assess model predictive power in extreme conditions. This effectively overcame the issue of data scarcity in conditions that are by definition rare, and for which it is seldom feasible to compute performance statistics using a single dataset. The scaling of flow condition and model errors employed in this study permitted their comparison across different streams and

constituents and allowed insights obtained from the aggregated dataset to be rescaled and applied to individual site- and constituent-specific models.

Generally, model predictions degraded in quality under increasingly extreme hydrologic conditions when they had been calibrated using only data below a certain flow threshold. This result is not unexpected, as model extrapolations have additional uncertainty beyond that for calibration conditions. However, the observed errors reflected a significant bias as well as variance. While increased variance is a natural result of extrapolation, increased bias is not, and its widespread presence suggests a systematic model misspecification under extreme conditions, potentially due to changing transport processes in these conditions. Empirical models are process-agnostic, but rely on the simplicity of underlying relationships between hydrologic variables. They further expect such relationships to be consistent across the conditions to which the models are applied. Extreme events introduce several processes that could violate this consistency, including bank erosion, streambed scouring (*Yellen et al.*, 2014), changes to flow-paths, and contribution of water from condition-specific sources (*Inamdar et al.*, 2004).

More interesting is the positivity of the observed bias, reflecting a tendency for models to overpredict extreme-event concentrations on average. This is contrary to suggestions from studies in individual watersheds (*Yoon and Raymond*, 2012; *Yellen et al.*, 2014), which reported a tendency for models to underpredict extreme-event concentrations and loads. This discrepancy could be due to methodological differences, such as the modeling choice, or it could be an artifact of the small sample size in such

studies. One possibility is that dilution may play a larger role at very high flows, as constituent sources are depleted within the watershed. However, this seems unlikely in the case of particulate constituents such as total suspended solids, which would be increasingly sourced from erosion and scouring processes in extreme flows. More research will be required to confirm and determine the causes of this observed predictive bias.

A notable exception to the overall trend of positive model bias was in nitrate, which like many other constituents showed increasing severity of bias with more extreme flows, but here the progression was of increasingly negative of bias. This could reflect different source and transport dynamics of nitrate (*Inamdar et al.*, 2004), with the implication that rating-curve models tend to underestimate nitrate concentrations in extreme flow conditions. Unlike many other constituents including organic nitrogen and carbon, nitrate concentrations are often highest in baseflow conditions (*Jordan et al.*, 1997); this may explain the anomalous negative bias of nitrate predictions.

The validation results suggest opportunities for improving upon rating-curve models when applying them to extreme-event conditions. Specifically, the increasing prediction bias with increasingly extreme flow conditions implies that predictions may be improved by taking this bias into account. This finding could facilitate model-based planning in anticipation of a future climate characterized by more frequent high-intensity storms such as hurricanes (*Bender et al.*, 2010). While the relatively large contribution of variance (versus bias) to SMSE resulted in only a marginal improvement

of predictive performance in this study, further refinements to the approach may result in more substantial improvements; one such modification is described below.

Except for a small number of outlying cases, the similarity in scale between calibration- and validation-set errors points to a gradational—rather than catastrophic—deterioration of rating-curve models when applied to extreme-event conditions. The distribution of these errors offers some guidance for applying models in extreme-event conditions. Generally, predictions performed better for smaller hydrologic distance between calibration and prediction conditions, and for better fitting models (as determined using calibration-set statistics). For "mildly extreme" conditions ($3 < q_s < 4$) the validation predictions had similarly good fit to the calibration, and could provide useful concentration estimates. More extreme flows resulted in poorer validation performance, but no evidence was found for supposed threshold behavior in solute concentrations (*Dhillon and Inamdar*, 2013).

While certain choices of methodology were somewhat arbitrary, including the form of model used to estimate concentration and the definition of extreme condition, the validation framework employed in this study is applicable to a variety of model types and hydrologic conditions. Different measures of hydrologic extremity, for example using precipitation thresholds, could be used in the place of q_s. Other variations of rating-curve models could similarly be tested, such as models employed in the LOADEST FORTRAN package (*Runkel et al.*, 2004) or Weighted Regression on Time, Discharge, and Season (WRTDS) (*Hirsch et al.*, 2010) and the approach in this study could serve as a benchmarking strategy for extreme-event modeling. The framework

can further be generalized to deal with any observation-specific condition. While other hydrologic conditions—e.g. drought, snowmelt—are the most obvious application, it may be possible to extend to conditions defined by biologic or human activity.

This multi-catchment framework for assessing model performance in unobserved conditions within an individual catchment is a novel approach to the incorporation of external data when making predictions using empirical water-quality models. Historically, performance information from other catchments' models was incorporated implicitly, for example in selecting predictor variables or model hyperparameters such as exponential smoothing discount factor (*Wang et al.*, 2011) or weighting-function window width (*Hirsch et al.*, 2010). In contrast, the estimation of extreme-condition model performance and associated bias correction in this study constitutes an *explicit* and *quantitative* use of information from multiple catchments' models. This allows for further statistical approaches to be applied using the set of prediction errors. For example, a simple extension of this approach could construct an analysis of variance (ANOVA) model on the aggregated error set, using various catchment and storm characteristics to "explain away" some amount of SMSE, and thereby refining the bias-adjustment for individual predictions.

As this analysis is data-driven, it is also data-constrained, particularly in the most extreme hydrologic conditions. The increasing sparsity of data with increasingly extreme conditions limits a quantitative assessment of differences in model bias and variance as a function of q_s, especially for individual constituent types. In this analysis this was circumvented by binning analyses by fraction type. A more robust database would

facilitate more nuanced investigations into differences in extreme-event responses between different constituent types. Other data were not consistently reported across the different datasets, such as sample collection method and contributing drainage area. Sample collection method, in particular, could prove important to the interpretation of the results. Since this study exclusively concerns periods of rapid hydrologic fluctuation, grab samples are likely to give a poor estimate of event-mean concentration, and could bias the DSST validation data depending where they fall on the storm hydrograph (*Robertson and Roerish*, 1999).

Other influences beyond the scope of this analysis could be further investigated in the future. The results may benefit from a thorough investigation of storm and catchment characteristics, which were not reported in the database but which might be available from other sources. Sub-daily discharge data are often available, and might be used to differentiate between water-quality impacts at different points in the storm hydrograph. Many studies have shown effects of antecedent conditions on stormwater-quality response, such as solute build-up and depletion; these too could be incorporated into such a multi-site study.

Finally, integrative data studies such as the framework and extensions described here are only possible to the extent that data are made available. Since available datasets were not uniformly distributed throughout the region, the results could be biased toward local effects at contributing stations, such as soil type, bedrock lithology, and climatology. Inconsistent documentation of metadata—such as time of sample collection, and sample collection and analytical methods—hamper large-scale

hydrologic investigations. It is likely that finer-scale assessments will be made possible as open access to data becomes an increasingly popular and expected practice (*Hanson and Hilst*, 2014), resulting in both more accurate models and a deeper understanding of their underlying processes.

2.5 Conclusions

This study shows a widespread, systematic, and directional deterioration of rating-curve predictive performance under increasingly extreme high-flow conditions. The effect pervaded all fractions (dissolved, suspended, total), and nearly all constituent types (nutrients, organic matter, suspended solids). This could reflect a failure of models to recognize an increasing importance of dilution at higher flows. However, the large variance in prediction accuracy at such flows reflects an overall deterioration of model performance, including instances of underprediction despite an overall tendency to overpredict extreme-event concentrations. The extent of deterioration in extreme-case goodness-of-fit is not always prohibitive, and can be improved by bias-correcting the predictions. These findings are an example of what can be gleaned from open access to data, and can be further built upon as data access, documentation, and consistency of collection are improved. Although this analysis was conducted on the aggregated results from many models and locations, it can be used to calculate model-specific goodness-of-fit statistics, giving a site- and constituent-specific estimate of model performance in predicting extreme-event concentrations.

CHAPTER 3

SIMULATION OF EXTREME-EVENT IMPACTS ON RESERVOIR TRIBUTARY WATER QUALITY

3.1 Introduction

Storm events are transport hot moments (Vidon et al., 2010) in watersheds, contributing the bulk of annual mass loads for many constituents (Inamdar et al., 2006; Raymond and Saiers, 2010), despite their short duration and relative infrequency. The most extreme events, such as hurricanes and tropical storms, are especially impactful, although the magnitude of this impact is difficult to measure accurately, requiring high-frequency sampling for the duration of the storm event (Inamdar et al., 2006; Yoon and Raymond, 2012).

Several recent studies provide localized examples of extreme-event solute transport. A 210-mm summer monsoon rainfall event produced exports exceeding 60% and 20%, respectively, of total annual particulate organic carbon (POC) and dissolved organic carbon (DOC) loads in a South Korean catchment, disproportionate to the associated 9% of annual flow volume. In separate catchments in the Northeastern US, Hurricane Irene (August, 2012) produced transport events exceeding 40% and 30%, respectively, of annual DOC and dissolved organic nitrogen (DON) loads (Yoon and Raymond, 2012), 56% of annual POC loads (Dhillon and Inamdar, 2013), and more than double the average annual suspended sediment load (Yellen et al., 2014).

When the receiving water body is a drinking water reservoir, such pulses of constituents adversely impact treatment costs, finished water aesthetics, and potentially public health. Sediment transport may lead to turbidity levels exceeding US Environmental Protection Agency (EPA) guidelines in the short-term, and increasing reservoir sedimentation in the long-term (Mukundan et al., 2013; Walling, 2009). Organic matter influx can increase the formation potential of harmful carcinogenic disinfection byproducts (DBPs) such as haloacetic acids (HAAs) and trihalomethanes (THMs) (Jung et al., 2014). Nutrients (nitrogen and phosphorus) impact biological processes and can cause algal blooms that produce undesirable taste and odor compounds as well as algal toxins (Young et al., 2015).

Despite their importance, predicting the impact of extreme events is difficult for several reasons. Data for such events are typically scarce or nonexistent for a given watershed, either due to the absence of such events in the historical record or logistical difficulties in sampling during such an event. Where they do exist, extreme-event datasets typically contain only a small number of observations and are difficult to generalize across watersheds and storms. Separate from water-quality considerations, the scarcity of extreme events makes their probability of occurrence difficult to estimate, although there is increasing evidence that climate change may increase the severity and frequency of such events (Bender et al., 2010). With large uncertainty in both the probability and impacts of extreme events, risk-based frameworks are inadequate for preparing against associated water-quality degradation.

Due to these difficulties, past studies synthesizing extreme-event water-quality impacts have been qualitative. A recent report of water-quality impacts from extreme weather events (including non-hydrologic events such as earthquakes and wildfires) described 44 case-studies from water utilities in Australia and the US, including the type of event experienced, type of utility system, and observed impacts (Stanford et al., 2014). To the author's knowledge, no studies have used simulation modeling as a proactive tool to anticipate extreme-event impacts on water-quality.

This study and its companion (Jeznach et al., 2016) present a proactive modeling framework to predict water-quality impacts of extreme events in drinking water reservoirs. The framework couples two modeling approaches: process-based reservoir models and data-driven, probabilistic tributary water-quality models. While process-based models (i.e. those that numerically solve equations related to physical and chemical processes) may be well-suited to simulate reservoir processes during and following an extreme event, they require the specification of inputs including streamflow and constituent concentrations in contributing tributaries. This study focuses on the probabilistic behavior of these inputs in an imposed extreme event; reservoir modeling is presented in a separate study (Jeznach et al., 2016). Both modeling frameworks attempt to provide a full account of model predictive uncertainty in discharge of tributary and the response of receiving water body. The outcome of the modeling framework is a distribution of water-quality response at a location of interest, for example a drinking-water withdrawal point, conditional on storm parameters such as precipitation depth and date of occurrence.

Various models exist for hydrograph simulation given storm precipitation, ranging from distributed and lumped-parameter process-based models to simple transfer-function models based on existing hydrographs. Constituent concentration and load estimates are typically obtained using regression modeling, also referred to as "rating curves" (Cohn et al., 1992; Ferguson, 1986; Stenback et al., 2011). This empirical modeling approach is less well suited to prediction when data are scarce, as in the case of extreme events. Furthermore, the distributional assumptions on which such models lie often lead explicitly to a degree of uncertainty proportional to the magnitude of concentration, i.e. the water-quality estimates during extreme events are likely to be "extremely uncertain". However, the probabilistic underpinnings of such models make them well suited to deal with predictive uncertainty. This representation of uncertainty in reservoir inputs can be carried through to the process-based model via repeated Monte Carlo sampling.

3.2 Methodology

This study generated hydrologic scenarios based on deterministically imposed storm-rainfall depths. Then a simplified probabilistic model was developed to predict water-quality constituent concentrations.

3.2.1 Imposed hydrologic scenarios

Extreme storm hydrographs were generated using observed historical hydrographs and hyetographs and an imposed extreme-event precipitation depth. For

all tributaries, the observed hydrograph ($q_t, m^3/s$) was separated into baseflow ($b_t, m^3/s$) and direct runoff (f_t, a.k.a. *quickflow*, m^3/s) using the recursive digital filter method (Lyne and Hollick, 1979).

$$q_t = f_t + b_t$$

Using the resulting separated hydrograph, the depth of precipitation losses (P_L, m) was calculated as the difference between the observed total (P_{obs}) and excess precipitation depth (P_e):

$$P_L = P_{obs} - P_e$$

$$P_e = \frac{1}{A}\sum_{t=1}^{T} f_t \, \Delta t$$

where A is the basin area (m^2) and Δt is the time interval (seconds) at which the hydrograph is measured. Baseflow and losses were assumed to remain constant for a scaling-up of storm magnitude, while quickflow and excess rainfall were not.

The imposed hydrograph, Q_t was calculated by volumetrically scaling up the quickflow portion of the observed hydrograph using the imposed extreme-event precipitation (P_{ext}):

$$Q_t = b_t + \frac{P_{ext} - P_L}{P_{obs} - P_L} f_t$$

Three precipitation depths were imposed as extreme-event scenarios: 4-inch (102 mm), 6-inch (152 mm), and 8-inch (203 mm), corresponding to historic recurrence intervals of

4, 50, and 100-years, respectively. The 4-inch storm also corresponded roughly to the observed precipitation in the Wachusett Reservoir watershed resulting from Hurricane Irene. Precipitation was assumed to arrive uniformly over the entire watershed for a duration of 24 hours, and all excess precipitation was assumed to be converted to runoff within 7 days, beginning the day of the imposed rainfall.

3.2.2 Probabilistic model for extreme-event concentration behavior

For a process-based reservoir model whose inputs include concentration values of R constituents in each of S tributaries during days $t = 1, \ldots, T$, these (unknown) inputs can be represented as T realizations of a random vector with dimension $R \cdot S$. The concentration of constituent r in tributary s on day t is denoted c_{rst}, and the random vector is denoted in boldface as $\mathbf{c}_t$. The logarithm of $\mathbf{c}_t$, $\mathbf{logC}_t$ is assumed to follow a multivariate normal distribution with mean that is a function of each tributary's flow and time:

$$\mathbf{logC}_t \sim MVN(\mathbf{f}(\mathbf{q}_t, t), \Sigma) \qquad (Eq.\ 3.1)$$

where $\mathbf{q}_t$ is the vector of all tributaries' flow at time t, and Σ is the covariance matrix of model errors, assumed constant across time and flow condition.

Each element of $\mathbf{logC}_t$ is modeled as a function of flow, q_{st}, and time, t, in a regression framework.

$$logC_{rst} = f_{rs}(q_{st}, t) + \epsilon_{rst}$$

where ϵ_{rst} are normally distributed, independent errors with mean zero and variance σ_{rs}^2. Although temporally dependent at short timescales, measurements are assumed to be spaced far enough apart that this can be ignored (Helsel and Hirsch, 2002).

The functional form, $\mathbf{f}()$, of the conditional mean of $\mathbf{logC}$ is estimated using multiple regression for each constituent and station. Various forms of regression modeling have been employed in models of this type, including linear and polynomial regression using maximum-likelihood or least-absolute-deviation estimation (Runkel et al., 2004); locally weighted regression (Hirsch et al., 2010); and semiparametric models (Autin and Edwards, 2010; Kuhnert et al., 2012; Wang et al., 2011). This study employed a semiparametric generalized additive model (GAM) with the form shown below (Wang et al., 2011; Hagemann et al., 2016):

$$logC_{rst} = s_1(log(q_{st})) + s_2(d_t) + s_3(t) + \epsilon_{rst} \qquad (Eq. 3.2)$$

where d_t is Julian day (1-365 for regular years or 1-366 for leap years) corresponding to day t and the functions $s_1(), s_2(), s_3()$ are spline-based smooth functions selected using penalized maximum likelihood estimation (Wood, 2006). Due to its greater flexibility compared with linear models, this model is less susceptible to bias than fully parametric models (Hirsch et al. 2010).

The covariance matrix Σ in Equation 3.1 can be estimated using the model residuals:

$$\hat{\Sigma} = cov\left(\mathbf{logC} - \hat{\mathbf{f}}(\mathbf{q}, t)\right)$$

A full accounting of predictive uncertainty necessarily includes uncertainty about the estimate of the regression function, in this case $\hat{\mathbf{f}}()$. The predictive uncertainty is therefore larger where the value of $\hat{\mathbf{f}}()$ is less certain, for example in extrapolations into extreme-event hydrologic conditions. In order to fully reflect this uncertainty, the simulation distribution used a modified version of the covariance matrix that incorporates this functional uncertainty.

$$\dot{\Sigma}_t = \mathbf{S}_t \mathbf{P} \mathbf{S}_t$$

where $\mathbf{S}_t$ is an $R \cdot S$ by $R \cdot S$ diagonal matrix with diagonal elements corresponding to the standard error of prediction for each constituent at time t and $\mathbf{P}$ is the correlation matrix of regression residuals. This retains the assumption that the true covariance of $\mathbf{logC}$ is constant but incorporates a condition-dependent component of uncertainty arising from the error in estimating the conditional mean of this random vector.

3.2.3 Sampling Procedure

While in theory the probabilistic model described above could be used to generate a sample reflecting the uncertainty of water-quality response to each imposed scenario, and this used as input to the reservoir model, such a direct Monte Carlo approach requires a large number of samples in order to converge and can be prohibitive depending on the computational cost of each simulation (Lee and Chen, 2009; Rahman and Hu, 2004). In particular, high-dimensional sample spaces are characterized by very large distances between randomly selected points, and are

computationally expensive to generate a sample with coverage of the sample space (Aggarwal, 2001; Hastings, 1970). For the case where $R = 5,\ S = 8,\ \text{and}\ T = 7$, as in this study, this corresponds to a 280-dimensional sample space.

In order to reduce this dimensionality, two simplifications were made to the probabilistic model. First, the errors ϵ_{rst} were assumed to be perfectly correlated in time (all ϵ_{rst} equal for fixed r, s) for the duration of the simulations (7 days), reducing the sampling distribution to dimension $R \cdot S$. Concentrations are known to be highly correlated in time, although this correlation is less during periods of rapid hydrologic change, i.e. storm events (Kirchner et al., 2004). Second, principal component analysis (PCA) was applied to the rating-curve residuals, and simulation samples were drawn from the resulting lower-dimensional space defined by the first two principal components (PCs;). This process was conducted as follows:

1. Regression-model residuals, $r_{rsj} = logC_{rsj} - f_{rs}(q_{sj}, t_j), r = 1, \ldots, R; s = 1, \ldots, S; j = 1, \ldots, n$ were computed for each rating-curve model, and assembled into a matrix with $R \cdot S$ columns corresponding to variables (combinations of constituent and sampling location) and n rows corresponding to dates of observations used in model calibration.

2. The correlation matrix $\mathbf{P}$ of the residuals was estimated from the residuals matrix. Because several days were missing observations for a given constituent at a station, this matrix was estimated using pairwise observations.

3. The eigendecomposition for the correlation matrix was computed as $\mathbf{P} = \mathbf{V}\Lambda\mathbf{V}^{-1}$, and the first two eigenvectors of $\mathbf{P}$ were extracted, along with their associated eigenvalues. The original random vector of model errors ϵ was then approximated as

$$\epsilon \approx \dot{\epsilon} \equiv w_1\mathbf{v_1} + w_2\mathbf{v_2}$$

where $\mathbf{v_1}$ and $\mathbf{v_2}$ are the first and second principal components, respectively, and w_1 and w_2 are independent random variables distributed as $w_i \sim N(0, \lambda_i)$, where λ_i is the eigenvalue corresponding to the i^{th} principal component.

A quasi-random sample was then generated from the bivariate distribution of $\mathbf{w} = \binom{w_1}{w_2}$ using a Halton sequence (Morokoff and Caflisch, 1995) of length N = 100 in order to ensure low discrepancy and to optimize the coverage of the sampling distribution. This sample of w_1 and w_2 were used to generate a sample of errors $\dot{\epsilon}$. For day t, the vector of simulation residuals was calculated as

$$\dot{\epsilon}_t^{(i)} = S_t V \Lambda^{\frac{1}{2}} \mathbf{w}^{(i)}$$

where V is the $(R \cdot S) \times 2$ matrix whose columns are the first two eigenvectors of the residual correlation matrix, and $\Lambda^{\frac{1}{2}}$ is the 2×2 diagonal matrix whose diagonal entries are the square-root of the first two eigenvalues of the residual correlation matrix.

The samples $\dot{\epsilon}_t^{(i)}$ constitute a random sample in the 2-dimensional subspace of the $R \cdot S$-dimensional sample space in which the largest proportion of variance lies. A simple analog of this methodology is shown in Figure 3.1.

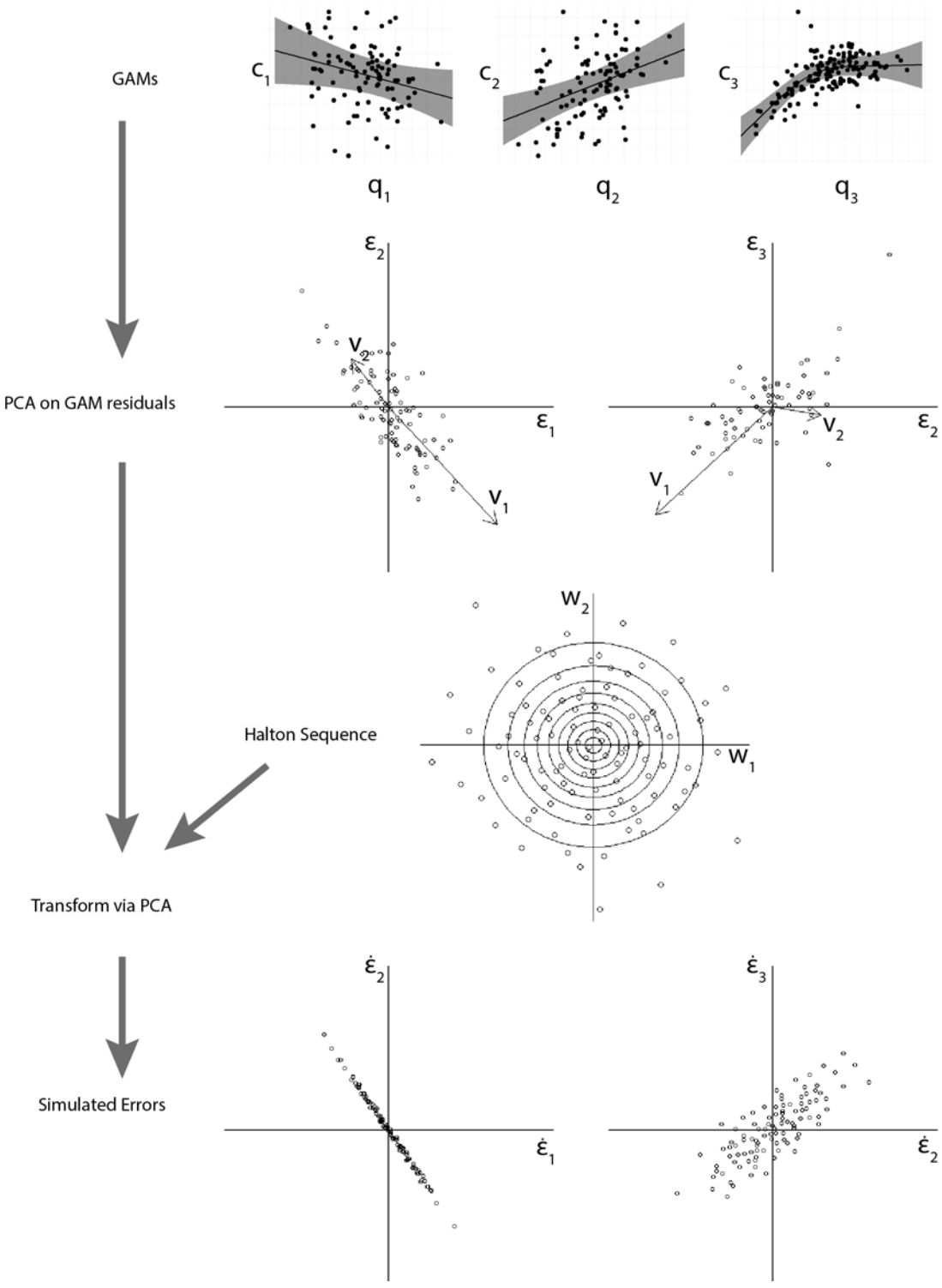

Figure 3.1: Illustration of sampling methodology for three arbitrary constituents having (respectively) log-concentrations c_1, c_2, c_3 and GAM errors ε_1, ε_2, and ε_3.

The sampling procedure is thus:

For iteration i = 1, ..., N runs:

1. Choose a values, $w_1^{(i)}, w_2^{(i)}$ from the quasi-random Halton sequence
2. Compute $\dot{\epsilon}^{(i)} = w_1^{(i)}\mathbf{v_1} + w_2^{(i)}\mathbf{v_2}$.
3. For each r, s, t, obtain $c_{rst}^{(i)} = exp(\hat{f}_{rs}(q_t, t) + \dot{\epsilon}_{rs}^{(i)})$

3.2.4 Study Area

The methodology was applied to the Wachusett Reservoir watershed in central Massachusetts. The reservoir serves as the primary drinking water supply for 51 communities in the Boston metropolitan area, and as an unfiltered water supply is subject to stringent water-quality requirements as part of the EPA's Filtration Avoidance Criteria (Austin et al., 2013; Kavanaugh, 1998). The reservoir has a volume of $250M\ m^3$, with most of its water arriving via the Stillwater and Quinapoxet Rivers and an aqueduct from the Quabbin Reservoir. In addition to the two rivers, water from the Wachusett Reservoir watershed arrives via 7 minor tributaries as well as direct runoff and precipitation (Table 3.1). Total inflows to the reservoir average $1.12\times10^6\ m^3/\text{day}$.

Table 3.1: Reservoir inflows from major tributaries.

Station	Mean flow (CFS)	Mean flow (m^3/day)
French	4.2	10309
Gates	4.8	11703
Malagasco	2.0	4880
Malden	4.6	11213
Muddy	2.2	5327
Quinapoxet	65.5	160369
Stillwater	57.0	139416
Waushacum	5.4	13218
W. Boylston	0.7	1748

Land use for the Wachusett Reservoir Watershed is primarily forest (67%), with less than 10% each of wetland, agriculture, residential, and other land use types. The watershed is divided into 9 subbasins corresponding to the major and minor tributaries. The individual subbasins vary substantially in land-use, with developed urban/residential land-use comprising over 50% of the Gates Brook and West Boylston Brook subbasins (Table 3.2). The two largest subbasins, those of the Stillwater and Quinapoxet Rivers, are similar to each other in land-cover, reflecting the composition of the entire reservoir watershed (Table 3.2).

Table 3.2: Land-use in major subbasins of the Wachusett Reservoir Watershed.

Tributary	Percent Cropland	Percent Developed	Percent Forest	Percent Pasture	Percent Water-Wetland	Area (ha)
Direct Runoff	2.7	15.9	72.7	0.6	8.0	2608
French	1.5	23.3	64.0	0.7	10.4	549
Gates	0.1	63.3	29.0	3.9	3.7	470
Malagasco	1.7	17.7	62.3	0.3	18.0	230
Malden	4.5	40.8	43.8	2.8	8.1	681
Muddy	0.3	35.8	55.0	0.1	8.8	190
Quinapoxet	4.0	15.4	67.0	1.8	11.8	14339
Stillwater	2.7	10.7	75.2	2.7	8.7	7887
Waushacum	2.3	24.1	57.1	2.7	13.8	1648
W.Boylston	5.3	55.2	31.9	1.4	6.2	111
Total	3.3	16.5	67.7	2.0	10.5	28715

The Massachusetts Department of Conservation and Recreation (Mass DCR) measured concentrations of water-quality constituents including nitrate-N (NO_3-N), ammonia-N (NH_3-N), total organic carbon (TOC), and total phosphorus (TP) on 8 out of 9 tributaries approximately monthly for the years 2005-2013. Analyses were performed at the Massachusetts Water Resources Authority (MWRA) Deer Island Laboratory, using Standard Method 5310B for TOC and EPA methods 350.1, 353.2, and 365.1, respectively, for NH_3-N NO_3-N, and TP (L. Pistrang, personal communication, 2013).

Flow for the Stillwater and Quinapoxet Rivers was measured by USGS stream gages with sub-daily resolution for the entire study period. Flow on minor tributaries was manually measured by Mass DCR with approximately weekly resolution, and interpolated to daily flow series by scaling Stillwater flow data proportional to the subbasin area.

Two rain gages operated in the watershed during the storms used in this study, located at the Stillwater River USGS gage and Mass DCR office in West Boylston, MA. A daily rainfall dataset was compiled from the average of these measurements.

3.2.5 CE-QUAL-W2 model

CE-QUAL-W2 (Cole and Wells, 2006) is a 2-dimensional, laterally averaged hydrodynamic and water quality model that simulates longitudinal and vertical hydrodynamics, in addition to chemical and biological processes. A CE-QUAL-W2 model for the Wachusett Reservoir has subsequently been extensively calibrated and updated by graduate students and faculty at the University of Massachusetts, Amherst (Jeznach et al., 2014). Model simulations are commonly used to evaluate the impact of external forcing (contaminant spills, climate change) on the water quality at the Cosgrove Intake and other locations of interest in the reservoir. More information on CE-QUAL-W2 and its applications is provided separately in the companion paper (Jeznach et al., 2016).

The Wachusett CE-QUAL-W2 model uses over 70 input files, including daily flow and concentration in each of 9 tributaries, daily precipitation, and sub-daily meteorology. Some of these inputs are assumed constant across time (e.g. bathymetry); others vary but are measured explicitly (temperature, flow on major tributaries); still others are measured infrequently (flow on minor tributaries), or not at all, and must be estimated using other measurements.

A single run of the model on a modern desktop computer requires approximately 30 minutes of CPU time for a 2-year simulation. This constrains the type

of analysis that can be performed using this model; for example, it limits the number of sample draws possible in a Monte Carlo or sensitivity-analysis study. Model outputs include time series of system parameters at various locations within the system, including constituent concentrations, flow rates, and water surface elevation. The most relevant output location in the Wachusett system is the withdrawal to the John J. Carroll Water Treatment Plant, the primary treatment facility for the water supply.

3.2.6 Baseline scenarios

Of 15 years for which reliable data exist for the reservoir and watershed, the year 2011 was selected for use in the extreme-event simulation study. The shape of the seasonal hydrograph was typical for this system, reflecting a moderate snowmelt hydrograph in the early spring and occasional rainstorms throughout the year. Additionally, two easily isolated events of moderate extremity occurred in 2011, amid typical spring and summer hydrologic conditions. A mid-April rainstorm with total storm depth of 65 mm occurred several weeks after ice-off conditions were reached on the reservoir. Stream flows prior to the event were high, reflecting high shallow groundwater storage following snowmelt conditions. While no snow cover was present in the watershed during this storm, the high antecedent levels of groundwater and surface storage led to relatively high runoff generated by this storm, calculated at 60% of rain depth. The second event, a 108 mm one-day rainfall related to Hurricane Irene, occurred in late August amidst summer low-flow conditions. This storm recorded the largest single-day rainfall on record in the Stillwater River rain gage, which began

operating in 1998. However, rainfall totals were significantly greater in other parts of the regions, reaching 200 mm in Western Massachusetts (Yellen et al., 2014).

3.3 Results and Discussion

3.3.1 Calibrated models

A total of 40 rating-curve models (Equation 3.2) were calibrated using between 53 and 156 observations, constituting 10 to 11 years of monitoring data. Calibration R^2 ranged between 0.18 and 0.84. Goodness-of-fit did not differ significantly across the different constituents, but did differ significantly across stations. For example, R^2 was greater than 0.58 for all constituents in French Brook and less than 0.45 for all constituents in Malden Brook (Table 3.3).

Table 3.3: Goodness-of-fit statistics (R2) for Wachusett-tributary GAM models.

	Tributary							
Constituent	French	Gates	Malagasco	Malden	Muddy	Quinapoxet	Stillwater	W.Boylston
NH_3-N	0.58	0.20	0.41	0.44	0.84	0.45	0.37	0.38
NO_3-N	0.66	0.64	0.30	0.23	0.56	0.67	0.41	0.58
TOC	0.78	0.37	0.49	0.24	0.48	0.64	0.62	0.18
TP	0.82	0.33	0.48	0.39	0.26	0.36	0.40	0.26
UV254	0.75	0.36	0.55	0.38	0.60	0.75	0.63	0.40

Term plots for the calibrated GAM models (Figure 3.2) reveal constituent- and site-specific linear and nonlinear relationships between (log-transformed) flow and (log-transformed) concentration. In general, nitrogen concentrations (NH_3-N and NO_3-N) decreased with increasing flow, suggesting dilution of groundwater nitrogen sources. In contrast, organic carbon concentrations (TOC and UV254) generally increased with increasing flow, whereas relationships between TP and flow were not consistently positive or negative across the different stations.

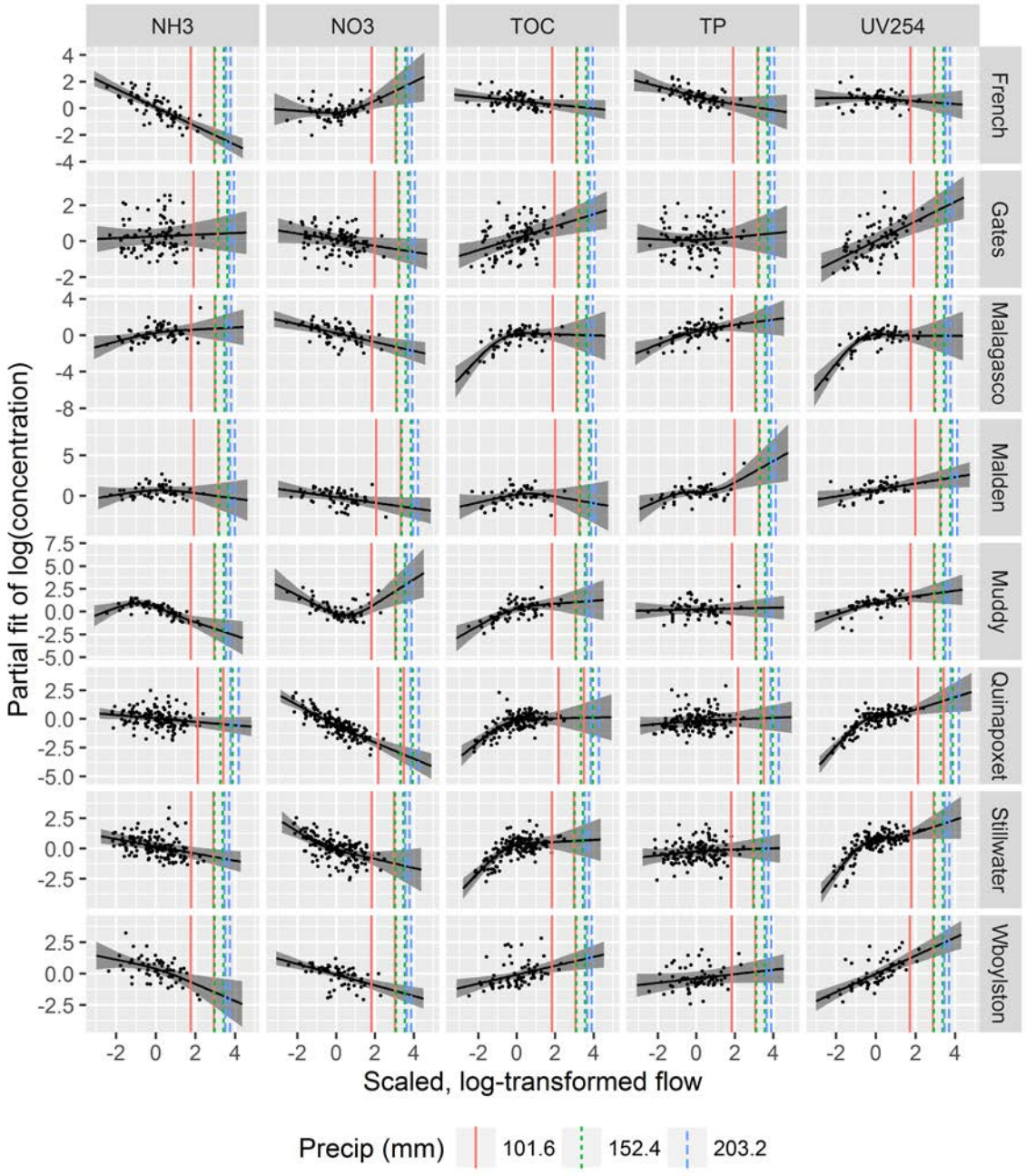

Figure 3.2: Flow-term plots for calibrated GAM models. Vertical lines indicate flows resulting from simulated precipitation events.

Although many models explained only a small fraction of total variance (low R^2), this is not problematic due to their probabilistic treatment. The unexplained variance is preserved in the covariance structure of the multivariate probability model, and is propagated through the reservoir model.

3.3.2 Imposed flow scenarios

The imposed extreme-event scenarios produced simulated flows that exceeded the original observed peak flows by as much as six-fold. In the most extreme cases (203 mm precipitation depth), the resulting flows were among the highest on record for both the April and August storm dates (Table 3.4). The middle scenario (152 mm precipitation depth) resulted in record flows for the April storm, but not the August storm, while the least-extreme scenario (102 mm precipitation depth) produced flows over twice the observed peak for the April Storm and approximately equal to the observed peak for the August storm. Flows were generally higher for the April storm, due to higher antecedent moisture that resulted in a greater fraction of rainfall converting to runoff, whereas the relatively dry antecedent conditions in the August storm attenuated the resulting peak flows. However, this effect was less pronounced in the most extreme scenario, for which infiltration accounted for a smaller fraction of the storm-total rainfall depth.

Table 3.4: Hydrologic characteristics of imposed extreme-event scenarios.

		Peak Flow (CFS)				Max Observed Flow (CFS)
Tributary	**Storm Date**	**Observed**	**101mm**	**152mm**	**203mm**	
Quinapoxet	2011-04-16	561	1488	2380	3272	1790
	2011-08-28	368	398	1277	2156	1790
Stillwater	2011-04-16	392	881	1415	1949	1380
	2011-08-28	389	277	914	1551	1380

Predicted water-quality responses to the simulated extreme-event flows differed by station and storm date, both in mean and variance. While some stations and constituents show a clearly shifting median (either increasing or decreasing) for progressively more extreme scenarios, a more consistent effect is increasing uncertainty with increasing precipitation extremity (Figure 3.3). This is evidenced by the tendency for median predictions (shape symbols in Figure 3.3) to stay relatively stationary across storm scenarios, whereas 95% confidence intervals (error bars in Figure 3.3) grow progressively wider with increasing storm depth. Further, the amount of concentration variability across stations is generally much larger than the variability from the precipitation scenarios or the storm date.

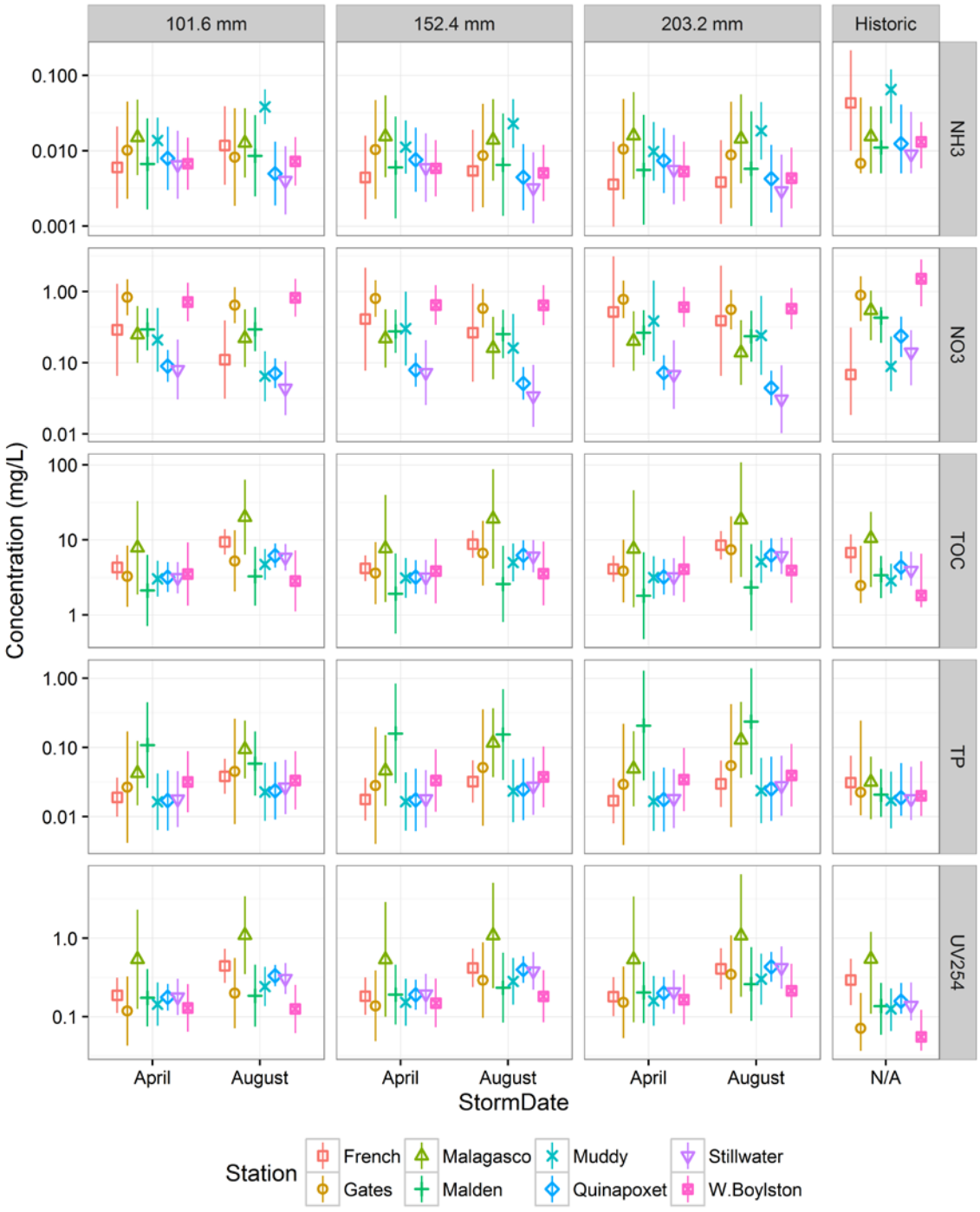

Figure 3.3: Extreme-event concentration predictions with 95-percent prediction intervals. Right panel shows observed concentrations in all historical data (5th, 50th, 95th percentiles)

3.3.3 Principal Component Analysis

The first two principal components of the concentration residuals explained 40% of the total variance, approximating a 40-dimensional probability distribution using just two dimensions (Figure 3.4). The first principal component (PC1) largely separated nitrogen variables (NH_3-N and NO_3-N) from the remaining constituents, while the second principal component (PC2) further differentiated certain constituents, particularly TP, and grouped similar stations within a given constituent (Figure 3.4). For example, PC2 separated NO_3-N observations in the highly developed West Boylston and Gates Brook tributaries from those in the primarily forested French, Quinapoxet and Stillwater tributaries, whereas PC1 did not separate these considerably.

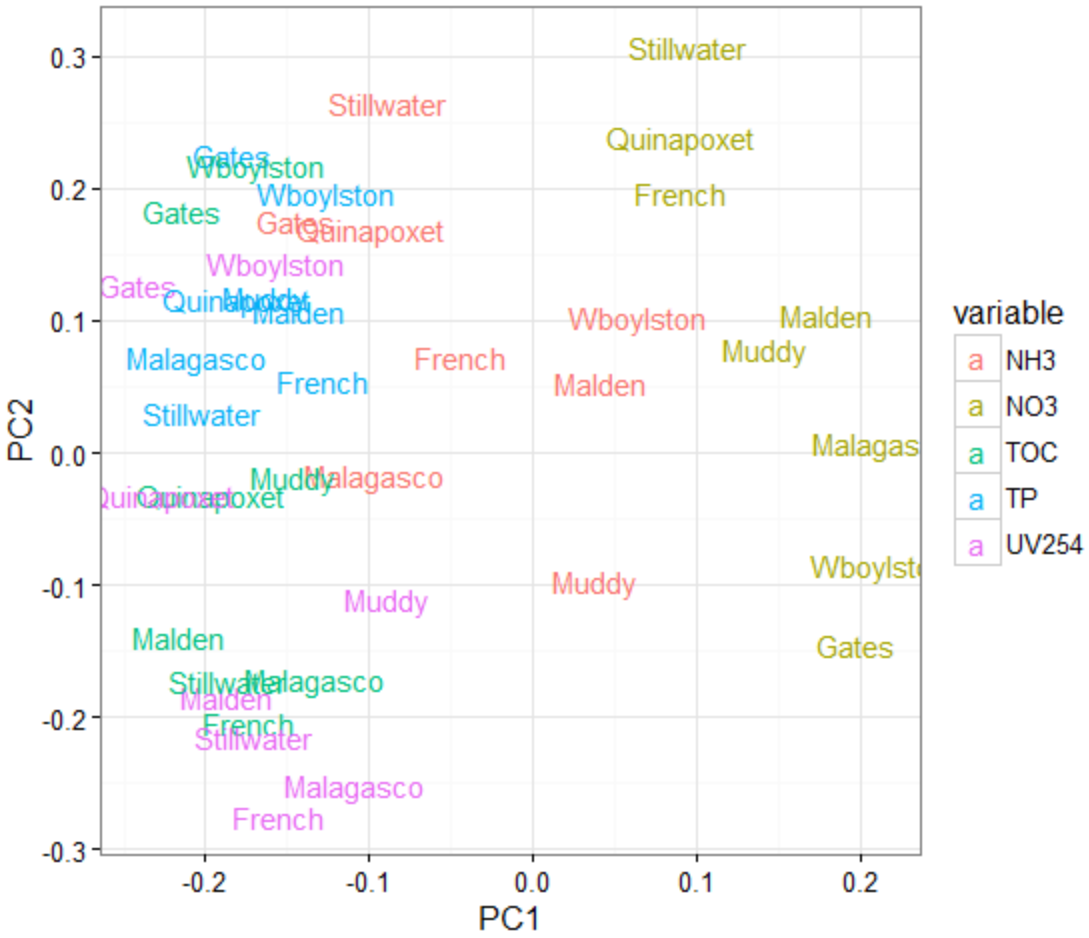

Figure 3.4: Plot of water-quality variables in first-two principal component-space (biplot)

3.3.4 Reservoir Inputs

Total inputs to the reservoir over the 7-day period beginning the day of the simulated extreme event varied by constituent and storm scenario (Figure 3.5). Storm inflow volume in April was greater than that in August for all storm depths. Median and maximum loads of nutrients (NH_3-N, NO_3-N and TP) were higher for April storms than for August storms. Organic matter loads (TOC and UV254) were higher in August for the

203-mm event, higher in April for the 101-mm event, and roughly equivalent across seasons for the 152-mm event. As was the case for the individual tributaries' concentrations, more extreme events resulted in more variance about higher mean loads, reflecting larger uncertainties at these extreme events.

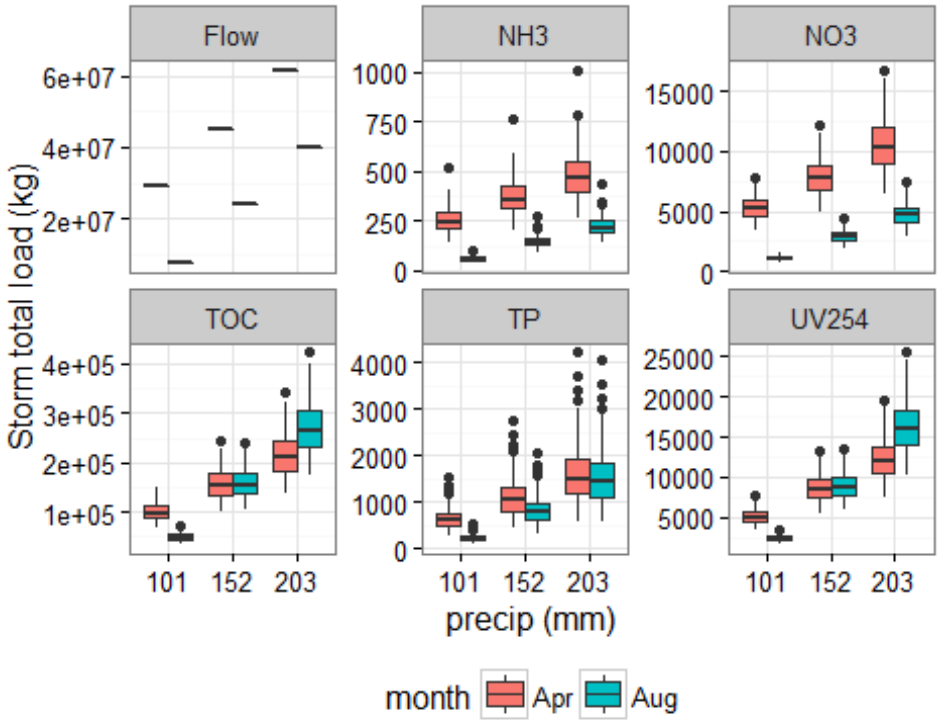

Figure 3.5: Boxplots of total reservoir inputs for week beginning at date of imposed event

A number of simplifications were required in order to couple with a computationally intensive reservoir model and a large number of uncertain reservoir inputs including flow and concentration. The role of each is discussed here in turn.

Hydrologic conditions including flow and precipitation were imposed as deterministic scenarios based on simplified hydrologic processes. These are crude representations of a complex, heterogeneous system, but give an extreme-event scenario necessary in order to apply a probabilistic simulation model. As this study was concerned primarily with predicting water-quality response to an arbitrary extreme event, this simplistic generation of streamflow was adequate. However, the methodology does not require such an approach, and can readily accommodate any method of streamflow simulation.

Generating Monte Carlo samples requires uncertainty to be quantified via a probability distribution, and the multivariate probability model used here (Equation 3.1) is only an estimate of this uncertainty based on available data. Assumptions including log-normality of concentration and the functional form of the conditional mean are simplifications required in order to make a complex reality analytically tractable.

Similarly, the GAM rating-curves (Equation 3.2) quantify the observed relationships between easily-measured variables (flow and time of year) and each constituent of interest, including their (marginal) uncertainty. However, the relationships they quantify are not necessarily causal, and are only best-estimates given available monitoring data. As seen in the assessment of model goodness-of-fit, the uncertainty about a given variable may not be well constrained using such models (i.e. models may have low R^2). This is not problematic in such a probabilistic simulation

study so long as the uncertainty is fully accounted for and carried through to the simulation step.

Principal component analysis allowed a high-dimensional space spanned by by many covarying random variables ($logC_{rs}, q = 1, \ldots, R; s = 1, \ldots, S$) to be approximated as a low-dimensional space that could of be sampled using a computationally feasible number of model runs. Further, the use of a quasi-random Halton sequence instead of a pseudo-random number generator improved the sampling efficiency in this low-dimensional space (Morokoff and Caflisch, 1995).

These assumptions and approximations provided a computationally tractable procedure to simulate realizations of watershed response to extreme-event hydrologic forcing. As with any model, they constitute an imperfect approximation of a natural system, including its associated uncertainty. Sources of uncertainty not accounted for in this approximation include uncertainty in extrapolating observed relationships in the data beyond their observed range (Figure 3.2) and error in the estimation of the covariance matrix used to compute the principal components.

3.4 Conclusion

The objective of this study was to predict, with uncertainty, the flow and concentration response of reservoir tributaries during an imposed extreme precipitation event condition. A methodology was developed that is sufficiently general to be applied to any reservoir model requiring flow and concentration time-series as inputs. Tributary flows were deterministically imposed using observed hydrographs and observed and

imposed rainfall, while concentration time series were probabilistically generated using a joint probability distribution obtained from generalized additive models and observed concentration data. In order to make the sampling procedure computationally feasible, a quasi Monte Carlo procedure was used to sample from a low-dimensional space defined by the first two principal components of the full joint probability distribution. The methodology was applied to a drinking water reservoir in central Massachusetts, and used estimate impacts from six imposed extreme-event scenarios. Flows and constituent loads were generally larger for spring scenarios than summer scenarios, resulting from high antecedent baseflow conditions. Concentration response varied by constituent, with nitrogen species generally having higher concentrations in the summer scenarios, while organic carbon species tended to have higher concentrations in the spring scenarios. Imposed precipitation depth more strongly impacted the uncertainty in constituent concentration than the estimated concentration itself, with larger events having larger concentration uncertainties. The results of this study and its companion paper (Jeznach et al., 2016) demonstrate the viability of proactively modeling extreme precipitation event impacts to a drinking-water reservoir and watershed using historical data, empirical and process-based models, and a full quantification of predictive uncertainty.

CHAPTER 4

CONCLUSIONS OF RESEARCH

Precise prediction of hydrologic variables, including concentration, remains elusive due to the large complexity and heterogeneity of watersheds. While it is unlikely that modeling efforts will overcome these issues in a process-based framework through computational brute force, this and other research show avenues for improvement. As other researchers have pointed out, traditional parametric models of concentration and load are inadequate for representing the often nonlinear relationships between hydrologic and water-quality variables. Fortunately, recent interest in predictive modeling across many disciplines has made new, more flexible methods readily available. This research demonstrates how such methods can be applied operationally. In particular, generalized additive models were found to perform as well as or better than linear models when applied to load estimation in tributaries of the Wachusett Reservoir. The differences in prediction from different methods were probed in-depth using a novel visualization method comparing load estimates across time and duration.

Although not a new concern, this research emphasized the importance of uncertainty quantification, providing several examples of its use. Chapter 2 demonstrated how errors in empirical models can be probed for systematic bias, and this used to improve their predictions. It further demonstrated how condition-specific predictive performance can be estimated a priori using aggregated errors from external catchments. While focused on extreme high-flow conditions in the U.S. Northeast using generalized additive models, the methods applied are extensible to other regions,

models, and conditions, and may be a direction for future research. Chapter 3 demonstrated that model uncertainty is not necessarily problematic for making predictions, and can be incorporated operationally into a proactive modeling framework for a water-supply reservoir. By incorporating multiple sources of uncertainty and considering its multivariate structure, the predictions generated in this methodology represent a fuller account of system understanding, and more importantly provide a means to exploit this knowledge when making management decisions.

A key contribution of this research, particularly that presented in Chapter 2, is an example how insights may be distilled from a large database containing multiple datasets from disparate catchments and monitoring organizations. This work uncovered a regionally persistent bias in predictions made during extreme high-flow events, and allowed model performance to be estimated in out-of-sample conditions using external data. As open access to historical datasets becomes increasingly widespread, methods such as these will be required to reap the full benefit of this collaboration.

BIBLIOGRAPHY

Aggarwal, Charu, Alexander Hinneburg, and Daniel A Keim. “On the Surprising Behavior of Distance Metrics in High Dimensional Space.” *Database Theory*, 2001: 420-434.

Apel, Heiko, Annegret H Thieken, Bruno Merz, and Günter Blöschl. "Flood risk assessment and associated uncertainty." *Natural Hazards and Earth System Science*, 2004: 295-308.

Aulenbach, Brent T, and Richard P Hooper. "The composite method: an improved method for stream-water solute load estimation." *Hydrological Processes*, 2006: 3029-3047.

Austin, Patricia, Lisa Gustavsen, Rebecca Budaj, Lawrence Pistrang, Kelley Freda, and Joel Zimmerman. *2013 Watershed Protection Plan Update.* Massachusetts Department of Conservation and Recreation, 2013.

Autin, Melanie A, and Don Edwards. "Nonparametric harmonic regression for estuarine water quality data." *Environmetrics*, 2010: 588-605.

Bauer, David F. "Constructing confidence sets using rank statistics." *Journal of the American Statistical Association*, 1972: 687-690.

Bender, Morris A, et al. "Modeled impact of anthropogenic warming on the frequency of intense Atlantic hurricanes." *Science*, 2010: 454-458.

Bennett, Neil D, et al. "Characterising performance of environmental models." *Environmental Modelling & Software*, 2013: 1-20.

Benotti, Mark J, Benjamin D Stanford, and Shane A Snyder. "Impact of drought on wastewater contaminants in an urban water supply." *Journal of environmental quality*, 2010: 1196-1200.

Brett, Michael T, Sara E Mueller, and George B Arhonditsis. "A daily time series analysis of streamwater phosphorus concentrations along an urban to forest gradient." *Environmental management*, 2005: 56-71.

Carpenter, Stephen R, Eric G Booth, Christopher J Kucharik, and Richard C Lathrop. "Extreme daily loads: role in annual phosphorus input to a north temperate lake." *Aquatic Sciences*, 2015: 71-79.

Caverly, Emma, James M Kaste, Gregory S Hancock, and Randolph M Chambers. "Dissolved and particulate organic carbon fluxes from an agricultural watershed during consecutive tropical storms." *Geophysical Research Letters*, 2013: 5147-5152.

Cohn, Timothy A, Dana L Caulder, Edward J Gilroy, Linda D Zynjuk, and Robert M Summers. "The validity of a simple statistical model for estimating fluvial constituent loads: An empirical study involving nutrient loads entering Chesapeake Bay." *Water Resources Research*, 1992: 2353-2363.

Cohn, Timothy A, Lewis L Delong, Edward J Gilroy, Robert M Hirsch, and Deborah K Wells. "Estimating constituent loads." *Water resources research*, 1989: 937-942.

Cole, Thomas M, and Scott A Wells. "CE-QUAL-W2: A two-dimensional, laterally averaged, hydrodynamic and water quality model, version 3.5." 2006.

Coron, Laurent, et al. "Crash testing hydrological models in contrasted climate conditions: An experiment on 216 Australian catchments." *Water Resources Research*, 2012.

Cox, Nicholas J, Jeff Warburton, Alona Armstrong, and Victoria J Holliday. "Fitting concentration and load rating curves with generalized linear models." *Earth Surface Processes and Landforms*, 2008: 25-39.

Dhillon, Gurbir Singh, and Shreeram Inamdar. "Storm event patterns of particulate organic carbon (POC) for large storms and differences with dissolved organic carbon (DOC)." *Biogeochemistry*, 2014: 61-81.

—. "Extreme storms and changes in particulate and dissolved organic carbon in runoff: Entering uncharted waters?" *Geophysical Research Letters*, 2013: 1322-1327.

Drewry, J J, L T Newham, and B F Croke. "Suspended sediment, nitrogen and phosphorus concentrations and exports during storm-events to the Tuross estuary, Australia." *Journal of environmental management*, 2009: 879-887.

Duan, Naihua. "Smearing estimate: a nonparametric retransformation method." *Journal of the American Statistical Association*, 1983: 605-610.

Esmen, Nurtan A, and Yehia Y Hammad. "Log-normality of environmental sampling data." *Journal of Environmental Science & Health Part A*, 1977: 29-41.

Ferguson, R I. "River loads underestimated by rating curves." *Water Resources Research*, 1986: 74-76.

Fiedler, Erich. "Modeling the Wachusett Reservoir, central Massachusetts, tributaries for improved watershed management." 2009.

Fienen, Michael N, and Nathaniel G Plant. "A cross-validation package driving Netica with python." *Environmental Modelling & Software*, 2015: 14-23.

Gilroy, E J, R M Hirsch, and T A Cohn. "Mean square error of regression-based constituent transport estimates." *Water Resources Research*, 1990: 2069-2077.

Hagemann, Mark, Daeyoung Kim, and Mi Hyun Park. "Estimating Nutrient and Organic Carbon Loads to Water-Supply Reservoir Using Semiparametric Models." *Journal of Environmental Engineering*, 2016: 04016036.

Hanson, Brooks, and Rob Hilst. "AGU's Data Policy: History and Context." *Eos, Transactions American Geophysical Union*, 2014: 337-337.

Hastings, W. K. "Monte Carlo Sampling Methods Using Markov Chains and Their Applications." *Biometrika*, 1970: 97-109

Helsel, Dennis R, and Robert M Hirsch. *Statistical methods in water resources.* Vol. 323. US Geological survey Reston, VA, 2002.

Hirsch, Robert M. "Large Biases in Regression-Based Constituent Flux Estimates: Causes and Diagnostic Tools." *JAWRA Journal of the American Water Resources Association*, 2014: 1401-1424.

Hirsch, Robert M, Douglas L Moyer, and Stacey A Archfield. "Weighted Regressions on Time, Discharge, and Season (WRTDS), with an Application to Chesapeake Bay River Inputs." *JAWRA Journal of the American Water Resources Association*, 2010: 857-880.

Huntington, Thomas G, and George R Aiken. "Export of dissolved organic carbon from the Penobscot River basin in north-central Maine." *Journal of Hydrology*, 2013: 244-256.

Inamdar, S P, N O'Leary, M J Mitchell, and J T Riley. "The impact of storm events on solute exports from a glaciated forested watershed in western New York, USA." *Hydrological Processes*, 2006: 3423-3439.

Inamdar, Shreeram P, Sheila F Christopher, and Myron J Mitchell. "Export mechanisms for dissolved organic carbon and nitrate during summer storm events in a glaciated forested catchment in New York, USA." *Hydrological Processes*, 2004: 2651-2661.

Jeznach, Lillian C, John E Tobiason, and David P Ahlfeld. "Modeling conservative contaminant effects on reservoir water quality." *JOURNAL AWWA*, 2014: 6.

Jeznach, L.C., Mark W. Hagemann, Mi-Hyun Park, and John E Tobiason. 2016. "Proactive modeling of water quality impacts of extreme precipitation events. Part 2: Drinking water reservoir. In preparation.

Jordan, Thomas E, David L Correll, and Donald E Weller. "Relating nutrient discharges from watersheds to land use and streamflow variability." *Water Resources Research*, 1997: 2579-2590.

Jung, B.-J., J.-K. Lee, H Kim, and J.-H. Park. "Export, biodegradation, and disinfection byproduct formation of dissolved and particulate organic carbon in a forested headwater stream during extreme rainfall events." *Biogeosciences*, 2014: 6119-6129.

Kavanaugh, James. "To filter or not to filter: A discussion and analysis of the Massachusetts filtration conflict in the context of the safe drinking water act." *BC Envtl. Aff. L. Rev.*, 1998: 809.

Kirchner, James W, Xiahong Feng, Colin Neal, and Alice J Robson. "The fine structure of water-quality dynamics: the (high-frequency) wave of the future." *Hydrological Processes*, 2004: 1353-1359.

Klemeš, Vit. "Operational testing of hydrological simulation models." *Hydrological Sciences Journal*, 1986: 13-24.

Kuhnert, Petra M, Brent L Henderson, Stephen E Lewis, Zoe T Bainbridge, Scott N Wilkinson, and Jon E Brodie. "Quantifying total suspended sediment export from the Burdekin River catchment using the loads regression estimator tool." *Water Resources Research*, 2012.

Kumar, Saurav, Adil N Godrej, and Thomas J Grizzard. "Watershed size effects on applicability of regression-based methods for fluvial loads estimation." *Water Resources Research*, 2013: 7698-7710.

Lee, Sang Hoon, and Wei Chen. "A comparative study of uncertainty propagation methods for black-box-type problems." *Structural and Multidisciplinary Optimization*, 2009: 239-253

Leisenring, Marc, and Hamid Moradkhani. "Analyzing the uncertainty of suspended sediment load prediction using sequential data assimilation." *Journal of hydrology*, 2012: 268-282.

Littlewood, I G, C D Watts, and J M Custance. "Systematic application of United Kingdom river flow and quality databases for estimating annual river mass loads (1975–1994)." *Science of the Total Environment*, 1998: 21-40.

Lyne, V, and M Hollick. "Stochastic time-variable rainfall-runoff modelling." *Institute of Engineers Australia National Conference.* 1979. 89-93.

MassDCR (Massachusetts Department of Conservation and Recreation) 2013. "Water Quality Report: 2012 Wachusett Reservoir and Sudbury Reservoir Watersheds" <http://www.mass.gov/eea/docs/dcr/watersupply/watershed/2012wachusettwqreport.pdf>

Morokoff, William J, and Russel E Caflisch. "Quasi-monte carlo integration." *Journal of computational physics*, 1995: 218-230.

Mukundan, Rajith, et al. "Factors affecting storm event turbidity in a New York City water supply stream." *Catena*, 2013: 80-88.

Quan, Nguyen T. "The prediction sum of squares as a general measure for regression diagnostics." *Journal of Business & Economic Statistics*, 1988: 501-504.

Rahman, S. and Xu, H. "A univariate dimension-reduction method for multi-dimensional integration in stochastic mechanics." *Probabilistic Engineering Mechanics*, 2004: 393-408.

Raymond, Peter A, and James E Saiers. "Event controlled DOC export from forested watersheds." *Biogeochemistry*, 2010: 197-209.

Ricket, David A. *Discontinuation of the National Water Quality Laboratory determinations for "total" nitrite, "total" nitrite plus nitrate, "total" ammonia,and "total" orthophosphate (using the four-channel analyzer).* United States Geological Survey, 1992.

Robertson, Dale M, and Eric D Roerish. "Influence of various water quality sampling strategies on load estimates for small streams." *Water Resources Research*, 1999: 3747-3759.

Runkel, Robert L, Charles G Crawford, and Timothy A Cohn. "Load Estimator (LOADEST): A FORTRAN program for estimating constituent loads in streams and rivers." 2004.

Stanford, B D, B Wright, J Routt, J Debroux, and S Khan. "Water Quality Impacts of Extreme Weather-related Events." *Water Environment Research FoundationWater Services Association of Australia*, 2014.

Stenback, Greg A, William G Crumpton, Keith E Schilling, and Matthew J Helmers. "Rating curve estimation of nutrient loads in Iowa rivers." *Journal of Hydrology*, 2011: 158-169.

Swistock, Bryan R, Pamela J Edwards, Frederica Wood, and David R Dewalle. "Comparison of methods for calculating annual solute exports from six forested Appalachian watersheds." *Hydrological Processes*, 1997: 655-669.

Tobin, James. "Estimation of relationships for limited dependent variables." *Econometrica: journal of the Econometric Society*, 1958: 24-36.

Ullrich, Antje, and Martin Volk. "Influence of different nitrate–N monitoring strategies on load estimation as a base for model calibration and evaluation." *Environmental Monitoring and Assessment*, 2010: 513-527.

Verma, Siddhartha, Momcilo Markus, and Richard A Cooke. "Development of error correction techniques for nitrate-N load estimation methods." *Journal of Hydrology*, 2012: 12-25.

Vidon, Philippe, et al. "Hot spots and hot moments in riparian zones: Potential for improved water quality management1." Wiley Online Library, 2010.

Vogel, Richard M, Beth E Rudolph, and Richard P Hooper. "Probabilistic behavior of water-quality loads." *Journal of Environmental Engineering*, 2005: 1081-1089.

Walling, Desmond E. *The impact of global change on erosion and sediment transport by rivers: current progress and future challenges.* Unesco, 2009.

Wang, You-Gan, and Ting Tian. "Sediment concentration prediction and statistical evaluation for annual load estimation." *Journal of Hydrology*, 2013: 69-78.

Wang, You-Gan, Petra Kuhnert, and Brent Henderson. "Load estimation with uncertainties from opportunistic sampling data–a semiparametric approach." *Journal of hydrology*, 2011: 148-157.

Wood, Simon. *Generalized additive models: an introduction with R.* CRC press, 2006.

Wood, Simon N. "Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2011: 3-36.

Yellen, B, J D Woodruff, L N Kratz, S B Mabee, J Morrison, and A M Martini. "Source, conveyance and fate of suspended sediments following Hurricane Irene. New England, USA." *Geomorphology*, 2014: 124-134.

Yoon, Byungman, and Peter A Raymond. "Dissolved organic matter export from a forested watershed during Hurricane Irene." *Geophysical Research Letters*, 2012.

Young, Ian, Ben A Smith, and Aamir Fazil. "A systematic review and meta-analysis of the effects of extreme weather events and other weather-related variables on Cryptosporidium and Giardia in fresh surface waters." *Journal of water and health*, 2015: 1-17.