Environment International 130 (2019) 104821

Contents lists available at ScienceDirect

Environment International

journal homepage: www.elsevier.com/locate/envint

When can we declare a success? A Bayesian framework to assess the recovery rate of impaired freshwater ecosystems

George B. Arhonditsis[a,*], Alex Neumann[a], Yuko Shimoda[a], Aisha Javed[a], Agnes Blukacz-Richards[a,b], Shan Mugalingam[c]

[a] Department of Physical and Environmental Sciences, University of Toronto, Toronto, Ontario M1C 1A4, Canada
[b] Environment and Climate Change Canada, Canada Centre for Inland Waters, Burlington, Ontario L7S 1A1, Canada
[c] Lower Trent Conservation Authority, Trenton, Ontario K8V 5P4, Canada

ARTICLE INFO

Handling Editor: Yong-Guan Zhu

Keywords:
Bayesian inference
Water quality
Ecosystem resilience
Probabilistic criteria
Bay of Quinte

ABSTRACT

Evaluating the degree of improvement of an *impaired* freshwater ecosystem resembles the statistical null-hypothesis testing through which the prevailing conditions are compared against a *reference* state. The pillars of this process involve the robust delineation of what constitutes an achievable reference state; the establishment of threshold values for key environmental variables that act as proxies of the degree of system impairment; and the development of an iterative decision-making process that takes advantage of monitoring data to assess the system-restoration progress and revisit management actions accordingly. Drawing the dichotomy between impaired and non-impaired conditions is a challenging exercise that is surrounded by considerable uncertainty stemming from the variability that natural systems display over time and space, the presence of ecosystem feedback loops (e.g., internal loading) that actively influence the degree of recovery, and our knowledge gaps about biogeochemical processes directly connected to the environmental problem at hand. In this context, we reappraise the idea of probabilistic water quality criteria, whereby the compliance rule stipulates that no more than a stated number of pre-specified water quality extremes should occur within a given number of samples collected over a compliance assessment domain. Our case study is the Bay of Quinte, Ontario, Canada; an embayment lying on the northeastern end of Lake Ontario with a long history of eutrophication problems. Our study explicitly accounts for the covariance among multiple water quality variables and illustrates how we can assess the degree of improvement for a given number of violations of environmental goals and samples collected from the system. The present framework offers a robust way to impartially characterize the degree of restoration success and minimize the influence of the conflicting perspectives among decision makers/stakeholders and conscious (or unconscious) biases pertaining to water quality management.

1. Introduction

In the Great Lakes Basin, the environmental management paradigm has been based on three fundamental processes: the designation of Areas of Concern (AOCs), the restoration of Beneficial Use Impairments (BUIs), and the setting of environmental standards (EC-USEPA, 2013). The AOCs are 43 designated geographic areas (26 in the United States, 12 in Canada, and 5 binational) that show severe environmental degradation, primarily modulated by anthropogenic activities at the local level (IJC, 2003). The BUIs[1] refer to the prevalence of undesirable conditions in terms of the physical, chemical, or biological integrity of a water body, such as poor water and sediment quality, contamination, loss of habitat/biodiversity, and other impairments that may have adverse effects on aquatic food web and/or human health (Reckhow et al., 2005; George and Boyd, 2007). The establishment of environmental standards involves a dynamic interactive process between scientific/professional knowledge and stakeholder/public opinions, whereby threshold values of measurable indicator variables are used to quantify the degree of impairment for each BUI and (most importantly) to

* Corresponding author.
E-mail address: georgea@utsc.utoronto.ca (G.B. Arhonditsis).

[1] An impairment of beneficial uses means a change in the chemical, physical or biological integrity of the Great Lakes system sufficient to affect any of the following fourteen (14) facets of the ecosystem structure and functioning: Restrictions on Fish and Wildlife Consumption; Tainting of Fish and Wildlife Flavor; Degraded Fish and Wildlife Populations; Fish Tumors or Other Deformities; Bird or Animal Deformities or Reproductive Problems; Degradation of Benthos; Restrictions on Dredging Activities; Eutrophication or Undesirable Algae; Restrictions on Drinking Water Consumption or Taste and Odor Problems; Beach Closings; Degradation of Aesthetics; Added Costs to Agriculture or Industry; Degradation of Phytoplankton and Zooplankton Populations; Loss of Fish and Wildlife Habitat.

https://doi.org/10.1016/j.envint.2019.05.015
Received 12 February 2019; Received in revised form 6 May 2019; Accepted 7 May 2019

delineate what constitutes a non-impaired or reference state (Zhang and Arhonditsis, 2008; Gudimov et al., 2011; Arhonditsis et al., 2016). These environmental standards form the basis to characterize an AOC as non-impaired or in recovery, which entails that all "*specific, measurable, achievable and scientifically defensible*" remedial actions have been undertaken and the environment is now either comparable to conditions similar to those prevailing at reference sites or requires more time to recover naturally (George and Boyd, 2007).

A critical facet of the aforementioned framework is that the freshwater ecosystem management often involves policy analysis and decision making in the face of considerable uncertainty stemming from a multitude of sources, such as the spatiotemporal variability that natural systems often display; the inherent randomness or indeterminacy, which often arises from our incomplete knowledge of the world; and the approximation uncertainty/subjective judgment, reflecting the assumptions made and imperfect knowledge used to understand the structure and inputs of the impaired environmental system (Arhonditsis et al., 2018). In view of this uncertainty, it is often argued that the water quality goals should be pragmatic by explicitly accommodating the idea that the prevailing conditions may not always be favourable in space and time, even if the central tendency of the system is on par with what is defined as a non-impaired state. A compliance rule for such a standard requires that no more than a *pre-specified number of violations* of the *targeted threshold* should occur within a *given number of samples* collected over the *compliance assessment spatiotemporal domain* (McBride and Ellis, 2001; Borsuk et al., 2002; Shabman and Smith, 2003; Zhang and Arhonditsis, 2008; Mahmood et al., 2014; Smith and Canale, 2015). Similar to any statistical hypothesis test though, there is potential for an error in the inference drawn regarding the compliance or breach of a probabilistic standard, as well as the nature of the error –Type I (falsely inferring a breach of standard) or II (falsely inferring compliance)– that might occur. To address the latter problem, McBride and Ellis (2001) presented a Bayesian approach that facilitates compliance assessment without the need to consider significance levels, Type I and Type II error risks. The same study also argued that the likelihood of bias from the influence of prior assumptions on the confidence assessment results, which is historically one of the main criticisms of Bayesian inference techniques, can be overcome either by introducing non-informative priors or through the formulation of informative distributions based on empirical evidence from the studied system (McBride and Ellis, 2001).

In this context, the overarching goal of the present study is concerned with the technical challenges and uncertainties surrounding the binary comparison between the conditions typically prevailing in impaired systems and those reflective of an "idealized" reference state. Our main objective is to introduce a Bayesian modelling framework that is designed to accommodate the covariance among multiple water quality variables, as well as the role of different sources of variability in time and space. A central concept of our framework revolves around the idea of probabilistic water quality criteria, whereby the compliance rule permits a pre-specified number of violations (water quality extremes) to occur within a given number of samples collected in time and space. Our case study is the Bay of Quinte, Ontario, Canada; an embayment lying on the northeastern end of Lake Ontario that is on the verge of being delisted as an Area of Concern, after the successful implementation of remedial measures that brought about significant water quality improvements. Nonetheless, the system occasionally experiences high ambient nutrient levels and harmful algal blooms, presumably driven by internal nutrient regeneration mechanisms that are not directly controlled by the on-going restoration practices. The fact that the prevalence of desirable conditions is intermittently interrupted by such water quality extremes poses semantic and operational challenges in unequivocally declaring that the Bay of Quinte is restored. Our framework is specifically designed to address these issues and rigorously characterize the degree of system improvement in the face of uncertainty.

2. Materials and methods

2.1. Case study

The Bay of Quinte is a Z-shaped embayment located on the northeastern shore of Lake Ontario (Fig. 1). Owing to the long history of eutrophication problems, the system has been listed as one of the AOCs

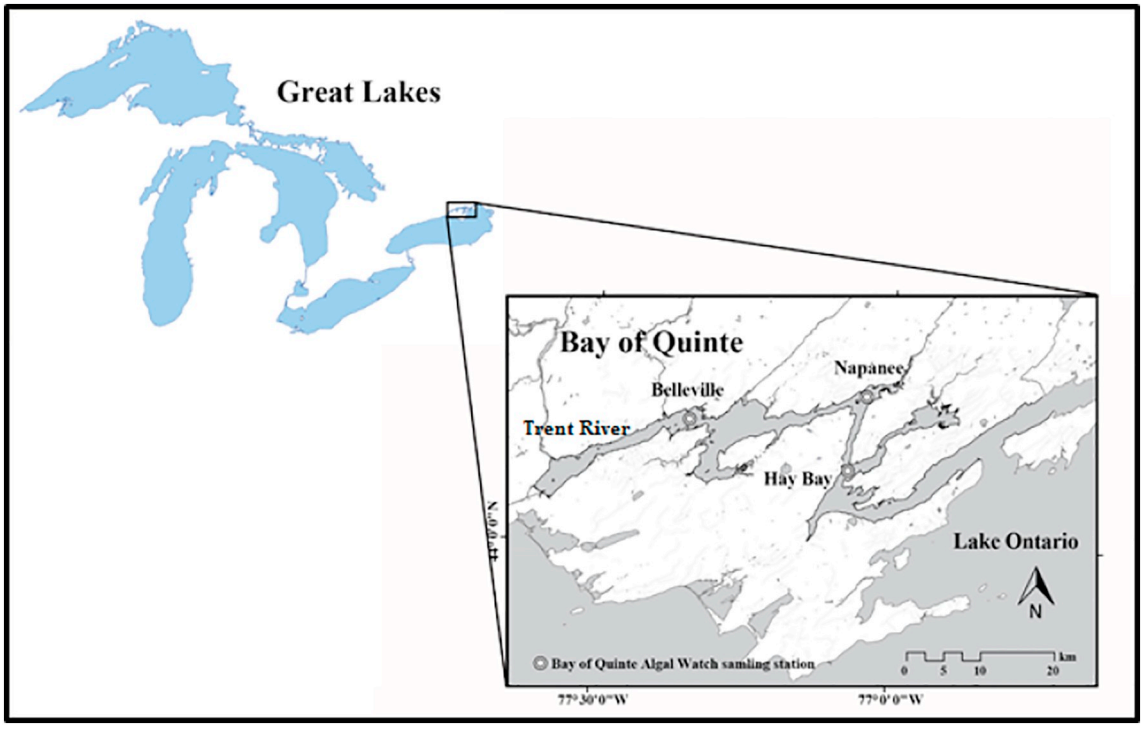

Fig. 1. Map of the Bay of Quinte, Lake Ontario (Ontario, Canada). The three sampling sites of the Project of Quinte program are also shown in the right map. Detailed description of the sampling program and analyses related to physical, chemical, and biological properties of the Bay of Quinte can be found in Nicholls (1999), Nicholls et al. (2002), and references therein.

by the International Joint Commission since 1986 (Nicholls et al., 2002). Reduction of phosphorus in detergents along with upgrades of the local waste water treatment plants resulted in a substantial decline of point-source loading during the 1970s, thereby triggering a distinct decline of ambient nutrient and phytoplankton biomass levels (Minns et al., 2011; Munawar et al., 2012). Notwithstanding the significant water quality improvement, the invasion of zebra (*Dreissena polymorpha*) and quagga (*D. bugensis*) mussels in the mid-1990s has been causally associated with the elevated end-of-summer total phosphorus (TP) concentrations as well as the dominance of the microcystin-producing cyanobacterium *Microcystis aeruginosa* (Nicholls et al., 2002; Shimoda et al., 2016). The mechanisms of sediment diagenesis have also been highlighted as another key factor that shapes phosphorus dynamics in the Bay of Quinte (Zhang et al., 2013; Kim et al., 2013; Doan et al., 2018). In particular, while the inflowing nutrient masses from Trent River predominantly influence the hydraulic regime and water quality in the innermost area of the system (Fig. 1), the sediments in the upper bay release a significant amount of phosphorus and the fluxes are likely modulated by macrophyte and dreissenid activity (Kim et al., 2013). From a management perspective, the presence of an active feedback (nutrient regeneration) loop and the occurrence of harmful algal blooms suggest that the prevalence of a non-impaired steady state in the Bay of Quinte may not be feasible in the foreseeable future (i.e., 5–10 years). Thus, additional reductions of the external point- and non-point source loading could necessitate in order to eliminate these intermittent shifts to undesirable water quality conditions and facilitate the long-term resilience of a distinctly improved state (Janse et al., 2010).

Water quality targets to delist the upper Bay of Quinte for the BUI "Eutrophication or Undesirable Algae" are based on an aggregated spatiotemporal (i.e., seasonal and system-wide) scale and are currently set at 30 μg L^{-1} and 10 μg L^{-1} for TP and chlorophyll *a* concentrations, respectively (Arhonditsis et al., 2016). However, the granularity of these water quality standards has been challenged, as it neither accommodates the significant seasonal variability in the upper bay, nor does it represent the dynamics in nearshore areas of high public exposure, e.g., beaches (Kim et al., 2013, 2018; Arhonditsis et al., 2016; Ramin et al., 2018). A single-value water quality standard, derived from averaging data that were collected from a few offshore sampling stations, is unlikely to reflect the entire range of conditions currently experienced in the system, including episodic events such as the excessively high end-of-summer ambient TP levels or harmful algal blooms (Kim et al., 2013). Striving for a more robust assessment of the prevailing water quality conditions, it has been proposed that the targets should revolve around extreme (or upper-limit) values of variables pertaining to management interest and must explicitly account for all the sources of uncertainty by permitting a realistic frequency of standard violations. In particular, Arhonditsis et al. (2016) have proposed the critical threshold TP level should be set at a value of 40 μg L^{-1}, which cannot be exceeded by > 10% in time and space. Given that the TP concentrations in the Bay of Quinte follow a log-normal distribution and values < 15 μg TP L^{-1} typically occur only 10% of the time during the growing season, then 10% exceedances of the 40 μg TP L^{-1} level are approximately equivalent to a targeted seasonal median of 26.5 μg TP L^{-1}. In a similar manner, recognizing that TP represents a "means to an end" and not "the end itself", the proposed probabilistic criterion for chlorophyll *a* must allow for no > 10% exceedances of the 12 μg chl*a* L^{-1} level, which if we follow the same reasoning as with the TP concentrations corresponds to a targeted seasonal average of 8 μg chl*a* L^{-1}.

2.2. Bayesian modelling framework

Multilevel modelling is suitable to analyze the empirical information routinely collected from the typical designs of monitoring programs, where data are organized at multiple levels (i.e., nested data). The units of analysis are usually individual measurements (at a first level) which are nested within contextual/aggregate units (at a second level), such as sampling sites, months, or years, which are themselves collectively used to infer about the contemporary ecosystem state (top level). Recognizing the need to comprehensively assess the recovery rate of the studied system, our statistical framework also accommodates the covariance among multiple water quality variables of interest, thereby allowing to draw inference regarding the joint probability of exceedance of the corresponding critical thresholds. Specifically, the statistical model of the present exercise was based on a bivariate normal likelihood in which the two means were provided by an additive (ANOVA-like) model to account for the different sources of variation of the log-transformed TP and chlorophyll *a* concentrations in time and space. We used Bayesian inference to estimate model parameters because of its ability to include prior information (e.g., literature reviews, expert knowledge, metadata, past parameter estimates) in the modelling analysis and to explicitly deal with model structural/parametric uncertainty as well as missing data and measurement errors (Gelman et al., 2013). Bayesian inference treats each parameter θ as a random variable and uses the likelihood function to express the relative plausibility of different parameter values given the available data from the system:

$$P(\theta \mid data) = \frac{P(\theta)P(data \mid \theta)}{\int_{\theta} P(\theta)P(data \mid \theta)d\theta}$$

where $P(\theta)$ represents the prior distribution of the model parameter θ, $P(data|\theta)$ indicates the likelihood of the data observation given the different θ values, and $P(\theta|data)$ is the posterior probability representing our updated beliefs on the θ values, contingent upon empirical knowledge from the system. The denominator is often referred to as the marginal distribution of the available data and acts as a scaling constant that normalizes the integral of the area under the posterior probability distribution (Gelman et al., 2013). The mathematical expression of our Bayesian multilevel modelling framework can be summarized as follows:

$$\begin{bmatrix} ln(TP_{ijtk}) \\ ln(Chla_{ijtk}) \end{bmatrix} \sim N\left(\begin{bmatrix} ln(\widehat{TP}_{ijtk}) \\ ln(\widehat{Chla}_{ijtk}) \end{bmatrix}, \begin{bmatrix} \sigma^2_{TP} & \sigma_{TP}\sigma_{Chla} \\ \sigma_{TP}\sigma_{Chla} & \sigma^2_{Chla} \end{bmatrix} \right)$$

$$ln(\widehat{TP}_{ijtk}) = \beta_{TP0} + \beta_{TP1i} + \beta_{TP2t} + \beta_{TP3j}$$

$$ln(\widehat{Chla}_{ijtk}) = \beta_{Chla0} + \beta_{Chla1i} + \beta_{Chla2t} + \beta_{Chla3j}$$

$$\beta_{TP0} \sim N(0, 10000) \quad \beta_{Chla0} \sim N(0, 10000)$$

$$\beta_{TP1i} \sim N(\mu_{TP1}, \sigma^2_{TP1}) \quad \sum_{i=1}^{I} \beta_{TP1i} = 0 \quad \beta_{Chla1i} \sim N(\mu_{chla1}, \sigma^2_{chla1}) \quad \sum_{i=1}^{I} \beta_{Chla1i} = 0$$

$$\mu_{TP1} \sim N(0, 10000) \quad \sigma^2_{TP1} \sim IG(0.001, 0.001) \quad \mu_{Chla1} \sim N(0, 10000) \quad \sigma^2_{Chla1} \sim IG(0.001, 0.001)$$

$$\beta_{TP2t} \sim N(\mu_{TP2}, \sigma^2_{TP2}) \quad \sum_{t=1}^{T} \beta_{TP2t} = 0 \quad \beta_{Chla2t} \sim N(\mu_{chla2}, \sigma^2_{chla2}) \quad \sum_{t=1}^{T} \beta_{Chla2t} = 0$$

$$\mu_{TP2} \sim N(0, 10000) \quad \sigma^2_{TP2} \sim IG(0.001, 0.001) \quad \mu_{Chla2} \sim N(0, 10000) \quad \sigma^2_{Chla2} \sim IG(0.001, 0.001)$$

$$j > 1 \qquad \begin{cases} \beta_{TP3j} \sim N(\beta_{TP3j-1}, \sigma^2_{TP3j}) \quad \sigma^{-2}_{TP3j} = \zeta^{j-1} \cdot \sigma^{-2}_{TP31} \\ \beta_{Chla3j} \sim N(\beta_{Chla3j-1}, \sigma^2_{Chla3j}) \quad \sigma^{-2}_{Chla3j} = \zeta^{j-1} \cdot \sigma^{-2}_{Chla31} \end{cases}$$

$$j = 1 \qquad \begin{cases} \beta_{TP31} \sim N(0, 10000) \quad \sigma^2_{TP31} \sim IG(0.001, 0.001) \\ \beta_{Chla31} \sim N(0, 10000) \; \sigma^2_{Chla31} \sim IG(0.001, 0.001) \end{cases}$$

$$\sum_{j=1}^{J} \beta_{TP3j} = 0 \quad \sum_{j=1}^{J} \beta_{Chla3j} = 0$$

$$\Sigma_M = \begin{bmatrix} \sigma^2_{TP} & \sigma_{TPChla} \\ \sigma_{TPChla} & \sigma^2_{Chla} \end{bmatrix} \sim IW(R,2)\ R = \begin{bmatrix} 100 & 0 \\ 0 & 100 \end{bmatrix}$$

where $ln(TP_{ijtk})$/ $ln(Chla_{ijtk})$ and $ln(\widehat{TP}_{ijtk})/ln(\widehat{Chla}_{ijtk})$ represent the log-transformed measured and modelled *TP/Chla* concentrations in sample *k*, collected from site *i*, semimonthly period *j* of the growing season, and year *t*; $\beta_{TP0}/\beta_{Chla0}$ are the overall mean concentrations for the entire system; $\beta_{TP1i}/\beta_{Chla1i}$ are the site-effect terms; $\beta_{TP2t}/\beta_{Chla2t}$ are the among-year effect terms; $\beta_{TP3j}/\beta_{Chla3j}$ are the within-year effect terms; μ_{TP1}, μ_{chla1}, μ_{TP2}, μ_{chla2} are the means of the hyperparameters of the site- and among-year effect terms; ${\sigma_{TP1}}^2$, ${\sigma_{Chla1}}^2$, ${\sigma_{TP2}}^2$, ${\sigma_{Chla2}}^2$ are the respective variances of the hyperparameters. The within-year effects are specified in terms of a sequence of normal distributions, in which the mean values for a given semimonthly period *j*, $\beta_{TP3j}/\beta_{Chla3j}$, are conditioned upon the corresponding terms for the previous time interval *j*-1, $\beta_{TP3j-1}/\beta_{Chla3j-1}$, while their precisions, ${\sigma_{TP3j}}^{-2}/{\sigma_{Chla3j}}^{-2}$, are connected with the corresponding terms for the first semimonthly time period (May 1*st*–15*th*) of the growing season, ${\sigma_{TP31}}^{-2}/{\sigma_{Chla31}}^{-2}$, through a discount factor ζ (=0.95). The latter term represents the aging of information with the passage of time, whereby we postulate that the parameter estimates assigned to the semi-monthly periods are more closely connected to their counterparts representing the preceding time intervals rather than those derived for earlier time spans of the seasonal cycle. *I, T, J* are the number of locations *i(=3)*, the total number of years *t* (=22); and the number of semimonthly periods *j* (=13) of the growing season, respectively; $N(0, 10000)$ is the normal distribution with mean 0 and variance 10,000, and $IG(0.001, 0.001)$ is the inverse gamma distribution with shape and scale parameters of 0.001. These prior distributions are considered "non-informative" or vague. The structural error covariance matrix Σ_M is assigned an inverse Wishart prior, in which the scale matrix R represents an assessment of the order of magnitude of the covariance matrix between the standard error terms σ_{TP} and σ_{Chla} (Arhonditsis et al., 2006; Gelman et al., 2013). To represent lack of confidence in the existing information, we chose two degrees of freedom for this distribution (n = 2), which is equal to the rank of the matrix. Site-, within year- and among-year effect terms are constrained to have a zero sum to make the model identifiable.

Consistent with our objective to establish probabilistic criteria that allow for a pre-specified frequency of exceedances (10%) of two water quality extremes (40 μg TP L^{-1} and 12 μg chl*a* L^{-1}), we created "confidence-of-compliance" graphs to assess the degree of compliance of the system for a given number of violations and total samples collected during the growing season. Similar to the modelling framework originally presented by McBride and Ellis (2001), we used beta-distributed prior information and a binomial likelihood to produce these graphs that are intended to provide a reference guide for tracking the degree of recovery of the Bay of Quinte:

$$CC = P(10\% > p_{crit})$$

$$n_{viol} \sim B(p_{crit}, N)$$

$$p_{crit} \sim Beta(1,1) \vee p_{crit} \sim Beta(b_1, b_2)$$

$$b_1 = \overline{x}\left[\frac{\overline{x}(1-\overline{x})}{s^2} - 1\right] \text{ and } b_2 = b_1\left[\frac{1}{\overline{x}} - 1\right]$$

where *CC* refers to the confidence of compliance which is the area below the 10% cutoff point, representing the probability that the true exceedance frequency is below the 10% allowable frequency of violations of the targeted threshold values; n_{viol} and *N* are the number of violations and the total number of samples collected during the growing season, respectively; p_{crit} is the prior probability to experience violations of the water quality criterion each time we collect a sample, which is in turn specified by a beta distribution with shape parameters either set equal to 1 (i.e., flat priors) or specified in terms of the empirical (data-based) evidence of the mean exceedance rate, $\overline{x}$, and associated standard deviation, s^2.

The statistical model used to reproduce the joint distribution of the log-transformed TP and chlorophyll *a* concentrations was parameterized with data collected from three offshore sites of the Bay of Quinte from 1996 to 2017 (Fig. 1). Even though no dramatic reduction of the point-source nutrient loading has occurred since the late 1980s/early 1990s, this period was selected as it represents the "*post-dreissenid era*", when the presence of dreissenid mussels is deemed responsible for fundamental changes in the ecosystem functioning; namely (*i*) the significant increase of light penetration, associated with the water filtration by dreissenids, is likely to have triggered the growth of submerged macrophytes and rapid proliferation of shallow-water beds into deeper water (Kim et al., 2013); (*ii*) the compositional shifts of the algal assemblage (e.g., disappearance of the late-spring diatom-dominated bloom, decline of *Aphanizomenon* and *Oscillatoria*, increase of *Microcystis*) after the dreissenid colonization that may have profound implications for the trophic efficiency and food-web integrity (Shimoda et al., 2016); and (*iii*) the substantial internal nutrient subsidies from the activity of macrophytes and dreissenids that can conceivably accentuate the fluxes emanating from the sediment diagenesis mechanisms (Arhonditsis et al., 2016), e.g., pseudofeces production, "nutrient-pump" effect (sensu Howard-Williams and Allanson, 1981).

3. Results-Discussion

The multilevel model used to characterize the variability of total phosphorus and chlorophyll *a* concentrations matched closely the measured concentrations and the corresponding standard error terms σ_{TP} and σ_{Chla} were 0.331 ± 0.005 and 0.482 ± 0.013, or when expressed in the original scale, the median error terms were 1.39 μg TP L^{-1} and 1.61 μg chl*a* L^{-1} bracketed by 95% credible intervals of 1.37–1.41 μg TP L^{-1} and 1.58–1.66 μg chl*a* L^{-1}, respectively (Table 1). To put these error estimates into perspective, the corresponding posteriors for the parameters β_{TP0} and β_{Chla0} were 3.338 ± 0.015 and 2.199 ± 0.025, i.e., the predicted median concentrations for the entire system during our 22-yr study period were 28.2 μg TP L^{-1} and 9.01 μg chl*a* L^{-1} bracketed by 95% credible intervals of 27.3–29.1 μg TP L^{-1} and 8.58–9.46 μg chl*a* L^{-1}, respectively. Among the three sampling sites considered, the innermost station at Belleville was characterized by distinctly higher TP and chlorophyll *a* concentrations, as depicted by the posterior estimates of the site-specific intercept $\beta_{TP1(Belleville)}$ = 0.077 ± 0.012 and $\beta_{Chla1(Belleville)}$ = 0.126 ± 0.024 (Table 1). Strong seasonal patterns characterize the ambient TP levels with the highest values registered during the end of summer ($\beta_{TP3(Aug16-31)}$ = 0.403 ± 0.029) and early fall ($\beta_{TP3(Sep1-15)}$ = 0.395 ± 0.028). The same pattern held true for the phytoplankton biomass with the annual maxima typically occurring in August, i.e., $\beta_{Chla3(Aug1-15)}$ = 0.666 ± 0.056 and $\beta_{Chla3(Aug16-31)}$ = 0.679 ± 0.059. Interestingly, while the system exhibits significant year-to-year variability, our model did not provide any evidence of a discernible long-term trend for either TP or chlorophyll *a* concentrations. In particular, the highest posterior β_{TP2} values were registered in 1999 (0.135 ± 0.034), 2005 (0.122 ± 0.033), and 2016 (0.119 ± 0.037). By contrast, the lowest estimates were derived for 1997 (−0.153 ± 0.035), 2009 (−0.174 ± 0.037) and 2013 (−0.174 ± 0.041). Likewise, the highest posterior β_{Chla2} estimates corresponded to 2005 (0.248 ± 0.073), 2006 (0.225 ± 0.073), and 2012 (0.284 ± 0.071), whereas the lowest values were found immediately after the invasion of dreissenids in 1996 (−0.304 ± 0.073), and 1997 (−0.271 ± 0.072), as well as during the last year of our study period, i.e., 2017 (−0.320 ± 0.083).

The distribution of the TP concentrations recorded from 1996 to 2017 was right skewed (Fig. 2a), and it is also interesting to note that the ambient TP values > 80 μg L^{-1} represented < 1% of the existing records. Collectively, 27.2% of the measured TP concentrations exceeded the proposed threshold level of 40 μg L^{-1}, while 37.1% and

Table 1
Posterior statistics of the stochastic nodes of the multilevel model used to characterize the different sources of variation of the joint total phosphorus-chlorophyll *a* distribution.

Total phosphorus				Chlorophyll *a*			
Parameters	Values	Parameters	Values	Parameters	Values	Parameters	Values
$\beta_{TP1(Belleville)}$	0.077 ± 0.012	$\beta_{TP3(May1–15)}$	−0.636 ± 0.025	$\beta_{Chla1(Belleville)}$	0.126 ± 0.024	$\beta_{Chla3(May1–15)}$	−0.899 ± 0.049
$\beta_{TP1(Napanee)}$	−0.082 ± 0.013	$\beta_{TP3(May16–31)}$	−0.445 ± 0.028	$\beta_{Chla1(Napanee)}$	−0.062 ± 0.024	$\beta_{Chla3(May16–31)}$	−0.760 ± 0.053
$\beta_{TP1(Hay\ Bay)}$	0.006 ± 0.013	$\beta_{TP3(Jun1–15)}$	−0.275 ± 0.027	$\beta_{Chla1(Hay\ Bay)}$	−0.064 ± 0.025	$\beta_{Chla3(Jun1–15)}$	−0.514 ± 0.054
$\beta_{TP2(1996)}$	−0.011 ± 0.034	$\beta_{TP3(Jun16–30)}$	−0.051 ± 0.030	$\beta_{Chla2(1996)}$	−0.304 ± 0.073	$\beta_{Chla3(Jun16–30)}$	−0.244 ± 0.057
$\beta_{TP2(1997)}$	−0.153 ± 0.035	$\beta_{TP3(Jul1–15)}$	0.078 ± 0.029	$\beta_{Chla2(1997)}$	−0.271 ± 0.072	$\beta_{Chla3(Jul1–15)}$	0.099 ± 0.055
$\beta_{TP2(1998)}$	0.106 ± 0.034	$\beta_{TP3(Jul16–31)}$	0.165 ± 0.029	$\beta_{Chla2(1998)}$	−0.004 ± 0.071	$\beta_{Chla3(Jul16–31)}$	0.427 ± 0.059
$\beta_{TP2(1999)}$	0.135 ± 0.034	$\beta_{TP3(Aug1–15)}$	0.331 ± 0.029	$\beta_{Chla2(1999)}$	−0.178 ± 0.071	$\beta_{Chla3(Aug1–15)}$	0.666 ± 0.056
$\beta_{TP2(2000)}$	0.052 ± 0.034	$\beta_{TP3(Aug16–31)}$	0.403 ± 0.029	$\beta_{Chla2(2000)}$	−0.142 ± 0.073	$\beta_{Chla3(Aug16–31)}$	0.679 ± 0.059
$\beta_{TP2(2001)}$	0.109 ± 0.034	$\beta_{TP3(Sep1–15)}$	0.395 ± 0.028	$\beta_{Chla2(2001)}$	0.067 ± 0.070	$\beta_{Chla3(Sep1–15)}$	0.587 ± 0.054
$\beta_{TP2(2002)}$	−0.054 ± 0.034	$\beta_{TP3(Sep16–30)}$	0.343 ± 0.029	$\beta_{Chla2(2002)}$	−0.076 ± 0.070	$\beta_{Chla3(Sep16–30)}$	0.583 ± 0.055
$\beta_{TP2(2003)}$	0.031 ± 0.034	$\beta_{TP3(Oct1–15)}$	0.124 ± 0.029	$\beta_{Chla2(2003)}$	0.231 ± 0.074	$\beta_{Chla3(Oct1–15)}$	0.344 ± 0.058
$\beta_{TP2(2004)}$	−0.052 ± 0.035	$\beta_{TP3(Oct16–31)}$	−0.021 ± 0.035	$\beta_{Chla2(2004)}$	0.014 ± 0.071	$\beta_{Chla3(Oct16–31)}$	−0.089 ± 0.069
$\beta_{TP2(2005)}$	0.122 ± 0.033	$\beta_{TP3(Nov1–15)}$	−0.411 ± 0.138	$\beta_{Chla2(2005)}$	0.248 ± 0.073	$\beta_{Chla3(Nov1–15)}$	−0.879 ± 0.199
$\beta_{TP2(2006)}$	0.058 ± 0.034	$\sigma_{TP3(May1–15)}$	0.147 ± 0.037	$\beta_{Chla2(2006)}$	0.225 ± 0.073	$\sigma_{Chla3(May1–15)}$	0.241 ± 0.058
$\beta_{TP2(2007)}$	0.036 ± 0.033	$\sigma_{TP3(May16–31)}$	0.151 ± 0.037	$\beta_{Chla2(2007)}$	0.047 ± 0.071	$\sigma_{Chla3(May16–31)}$	0.247 ± 0.059
$\beta_{TP2(2008)}$	−0.021 ± 0.034	$\sigma_{TP3(Jun1–15)}$	0.155 ± 0.038	$\beta_{Chla2(2008)}$	0.002 ± 0.073	$\sigma_{Chla3(Jun1–15)}$	0.254 ± 0.061
$\beta_{TP2(2009)}$	−0.174 ± 0.037	$\sigma_{TP3(Jun16–30)}$	0.159 ± 0.039	$\beta_{Chla2(2009)}$	−0.078 ± 0.082	$\sigma_{Chla3(Jun16–30)}$	0.260 ± 0.063
$\beta_{TP2(2010)}$	−0.074 ± 0.036	$\sigma_{TP3(Jul1–15)}$	0.163 ± 0.040	$\beta_{Chla2(2010)}$	0.001 ± 0.083	$\sigma_{Chla3(Jul1–15)}$	0.267 ± 0.065
$\beta_{TP2(2011)}$	−0.055 ± 0.035	$\sigma_{TP3(Jul16–31)}$	0.167 ± 0.041	$\beta_{Chla2(2011)}$	0.207 ± 0.071	$\sigma_{Chla3(Jul16–31)}$	0.274 ± 0.066
$\beta_{TP2(2012)}$	0.030 ± 0.036	$\sigma_{TP3(Aug1–15)}$	0.172 ± 0.043	$\beta_{Chla2(2012)}$	0.284 ± 0.071	$\sigma_{Chla3(Aug1–15)}$	0.281 ± 0.068
$\beta_{TP2(2013)}$	−0.174 ± 0.041	$\sigma_{TP3(Aug16–31)}$	0.176 ± 0.044	$\beta_{Chla2(2013)}$	0.042 ± 0.073	$\sigma_{Chla3(Aug16–31)}$	0.289 ± 0.071
$\beta_{TP2(2014)}$	0.002 ± 0.110	$\sigma_{TP3(Sep1–15)}$	0.181 ± 0.045	$\beta_{Chla2(2014)}$	−0.205 ± 0.069	$\sigma_{Chla3(Sep1–15)}$	0.296 ± 0.072
$\beta_{TP2(2015)}$	0.003 ± 0.108	$\sigma_{TP3(Sep16–30)}$	0.185 ± 0.046	$\beta_{Chla2(2015)}$	0.008 ± 0.196	$\sigma_{Chla3(Sep16–30)}$	0.304 ± 0.074
$\beta_{TP2(2016)}$	0.119 ± 0.037	$\sigma_{TP3(Oct1–15)}$	0.190 ± 0.047	$\beta_{Chla2(2016)}$	0.204 ± 0.057	$\sigma_{Chla3(Oct1–15)}$	0.312 ± 0.075
$\beta_{TP2(2017)}$	−0.031 ± 0.058	$\sigma_{TP3(Oct16–31)}$	0.195 ± 0.048	$\beta_{Chla2(2017)}$	−0.320 ± 0.083	$\sigma_{Chla3(Oct16–31)}$	0.319 ± 0.077
σ_{TP}	0.331 ± 0.005	β_{TP0}	3.338 ± 0.015	σ_{Chla}	0.482 ± 0.013	β_{Chla0}	2.199 ± 0.025
σ_{TPChla}	−0.268 ± 0.028						

47.5% were the exceedances for the TP values of 35 and 30 μg L^{-1}, respectively. To understand the distance of the system from the "ideal" conditions stipulated by the proposed TP criterion (solid black line in Fig. 2a), the same frequencies of exceedance for the 40, 35, and 30 μg L^{-1} levels should approximately be 10%, 18.5%, and 35%, respectively. The relevance of our modelling exercise for risk management decisions can be shown in the histograms, where we plot the frequencies of exceedance of the same threshold values as predicted by our model, or the "*probability distribution of the predicted exceedance probabilities*" (Fig. 2b–d). In these histograms, we included a benchmark value of 10% representing the maximum acceptable frequency to violate the levels of 40, 35, and 30 μg TP L^{-1} for each sample collected from the Bay of Quinte. According to the conditions currently prevailing in the system, only 35.9% of the mean predicted probabilities of exceedance of the 40 μg TP L^{-1} were below the 10% benchmark. Simply put, based on the present state of the Bay of Quinte, the graph illustrated in Fig. 2b suggests a lower than 40% confidence that the system will not exceed the specified water quality extreme (i.e., 40 μg TP L^{-1}) > 10% of the time and space considered. For the sake of comparison, the degree of confidence that the likelihood of exceedance for 35 and 30 μg TP L^{-1} will be < 10% was 28.6 and 20.5%, respectively. The onus now is on the local water quality managers to decide what is a realistic level of confidence (i.e., probability below the 10% benchmark) in order to infer that the system is resilient and therefore the likelihood of undesirable water quality shifts is kept to an acceptably low level.

The frequency histogram for the chlorophyll *a* concentrations similarly suggests a right-skewed distribution (Fig. 3a) in which 41.3% of the measured values exceeded the proposed threshold level of 12 μg L^{-1}. In a similar manner, 49.2% and 57.3% were the recorded exceedances for the chlorophyll *a* values of 10 and 8 μg L^{-1}, respectively. These exceedance frequencies clearly suggest that the system in its current state differs significantly from the reference conditions prescribed by the proposed chlorophyll *a* criterion (solid black line in Fig. 3a), whereby the same levels of exceedance of 12, 10, and 8 μg chl*a* L^{-1} should approximately be 10%, 23.9%, and 50%, respectively. According to the frequency histogram of the exceedance probabilities of 12 μg chl*a* L^{-1}, our confidence that this threshold value will not be exceeded by > 10% in time and space is 31.7% (Fig. 3b). In other words, the "*probability distribution of the predicted exceedance probabilities*" suggests that for each growing season when we monitor the three offshore stations in the Bay of Quinte, our confidence that the targeted value of 12 μg chl*a* L^{-1} will not be exceeded by > 10% of the collected samples is just over 30%. Likewise, the degree of confidence that the likelihood of exceedance for 10 and 8 μg chl*a* L^{-1} will be lower than 10% is 25.8 and 15.2%, respectively (Fig. 3c–d). These estimates of our degree of confidence reinforce our previous assertion that the Bay of Quinte still has a long way to go until the prevailing conditions resemble those stipulated by the chlorophyll *a* criterion. Given that the two water quality variables typically exhibit strong positive covariance (Zhang et al., 2013), the consideration of their joint TP-chl*a* distribution offers the complete picture with respect to the compliance of the Bay of Quinte with the two delisting criteria for the BUI "Eutrophication or Undesirable Algae" during the post-dreissenid era (Fig. 4a). In particular, our analysis suggests that only 27% of the samples (snapshots from the system) available met the expectation that the frequency of violations of the two water quality extremes should simultaneously be lower than 10% (Fig. 4b).

From an operational standpoint, the number of samples required during the growing season (end of May-early October) in order to assess system compliance with the targeted TP and chlorophyll *a* thresholds can vary from 30 to 162, depending on the sampling frequency and duration of the period examined (Table 2). For illustration purposes, we created "confidence-of-compliance" graphs to evaluate the likelihood of compliance against the TP criterion that the 40 μg L^{-1} level will not be exceeded by > 10% when different sample sizes are collected. This exercise was first conducted by assuming no prior knowledge about the prevailing conditions in the Bay of Quinte and then by informing our prior beliefs from water quality data collected during the *post-dreissenid* (1996–2017) and *eutrophic* (1975–1982) periods (Fig. 5 & Table 2).

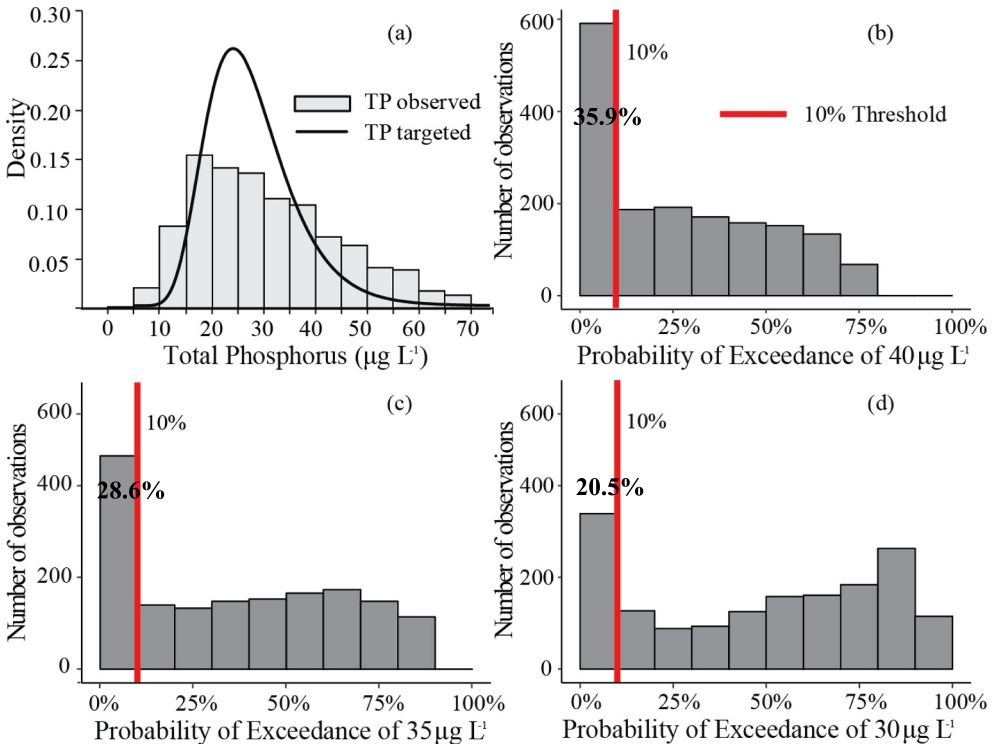

Fig. 2. Frequency histograms depicting (a) the measured TP concentrations during the 1996–2017 period against the "ideal" distribution targeted by the established TP criterion in the Bay of Quinte, (b) the mean probability values for TP concentrations to exceed the 40 μg L^{-1}, (c) 35 μg L^{-1} and (d) 30 μg L^{-1} threshold levels.[2] The vertical red lines in panels (b)–(d) represent a tentative 10% probability benchmark of TP violations. The probability below that cut-off point quantifies our degree of confidence that the TP threshold concentration(s) will not be exceeded by > 10% within the samples collected in time and space. It represents the probability that the true exceedance frequency is below the 10% guideline and is termed the confidence of compliance (CC). The mean values of the distributions in panels (b–c) are termed the expected exceedance of a targeted threshold, and were 26.5 ± 23.7%, 36.4 ± 28.3%, and 48.5 ± 31.7%, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

According to these graphs, when samples are collected from the system on a biweekly basis over the course of two years (n = 60), assuming no prior knowledge of its water quality status, then 4 samples with measured TP concentrations higher than 40 μg L^{-1} suggest that our confidence of compliance with the targeted criterion will be 74.6% (Point A in Fig. 5a). If the number of samples with > 40 μg TP L^{-1} is 6, then the degree of our confidence drops down to 40.9% (Point B in Fig. 5a). With 8 (Point C) and 10 (Point D) violations, our confidence of compliance becomes distinctly lower, 15.6% and 3.9%, and thus the resilience of the system to the potential pressure exerted by external or internal stressors will be under question. Another way to construct the same graph is to revisit our assumption of no prior knowledge of the water quality of the system, and instead base our prior beliefs on the empirical evidence gained over the course of a 22-yr study period. According to our model predictions, after the colonization of the Bay of Quinte from dreissenids, the exceedance rate of the 40 μg TP L^{-1} threshold level has been 26.5 ± 23.7%, and if we use this estimate to formulate our *Beta* prior, the resulting "confidence-of-compliance" graph paints a more optimistic picture (Table 2). For example, 4 violations of the TP criterion out of 60 collected samples would suggest a greater confidence of compliance (80.3%), and so will 6 (47.9%), 8 (21.5%), and 10 (4.8%) violations. Simply put, the consideration of the recent history of the Bay of Quinte provides a higher degree of confidence about the prevailing conditions in the system, relative to a water quality assessment exercise that will be entirely based on present data without any consideration of past information (Fig. 5b). In stark contrast, if the formulation of our prior beliefs is based on data from the 1970s/early 1980s when the system was eutrophic, the "confidence-of-compliance" predictions offer a more pessimistic/conservative perspective about the degree of recovery (Table 2). In particular, 4 violations of the TP criterion in samples collected over the course of two years are suggestive of a distinctly lower confidence of compliance (48.3%), and the same will be true if 6 (19.6%), 8 (5.8%), and 10 (1.3%) violations are registered. Another way to view this result is that our modelling framework formalizes the tendency to be more conservative in our assessment, when the history of the studied system has frequent incidences of water quality extremes.

The latter illustration highlights the conceptual advantages of the proposed framework in that the water quality assessment neither comes at the "cost" of having to state one's subjective prior belief as to likely exceedance rates, which is a usual criticism of the Bayesian inference (Dennis, 1996), nor does it disconnect the history of the system from the decision-making process, as depicted by the available empirical knowledge and monitoring data. Rather than "rolling the dice", each time we base our assessment exercise solely on a new dataset and implicitly assume no prior knowledge of the system, the "prior–likelihood–posterior" update cycles of our framework solidify the continuity of the information used to characterize the ecological conditions. Even if skeptical views counter-argue that there are times when the history of the system may not be relevant in assessing future responses, it is important to note that a well-known property of Bayesian analysis is that the results from classical and Bayesian statistics become more similar with larger sample sizes (see also confidence-of-compliance estimates in Table 2), as the information in the data increasingly dominates over that in the chosen prior distributions (McBride and Ellis, 2001; Arhonditsis et al., 2008). Our "confidence-of-compliance" graphs can determine the optimal number of samples or the duration of the compliance period that will allow to draw robust inference about the prevailing ambient conditions, and address any criticism about the potential subjectivity with the methodological practices followed. For example, a system, like the Bay of Quinte, that is still susceptible to the intermittent occurrence of water quality extremes (e.g., harmful algal blooms), inherently unpredictable ecological stressors (e.g., invasive species), and not directly controlled biogeochemical mechanisms (e.g., internal loading) may require a longer assessment (2- or 3-yr) period and a higher number of samples in order to determine its recovery rate with a greater degree of confidence.

[2] The former value represents the criterion (upper-limit threshold) proposed to local stakeholders (Shoreline Municipalities, Ministry of the Environment, Conservation and Parks, Ministry of Agriculture, Food and Rural Affairs, Environment and Climate Change Canada), while the latter two concentrations are used to more comprehensively characterize the prevailing conditions in the system.

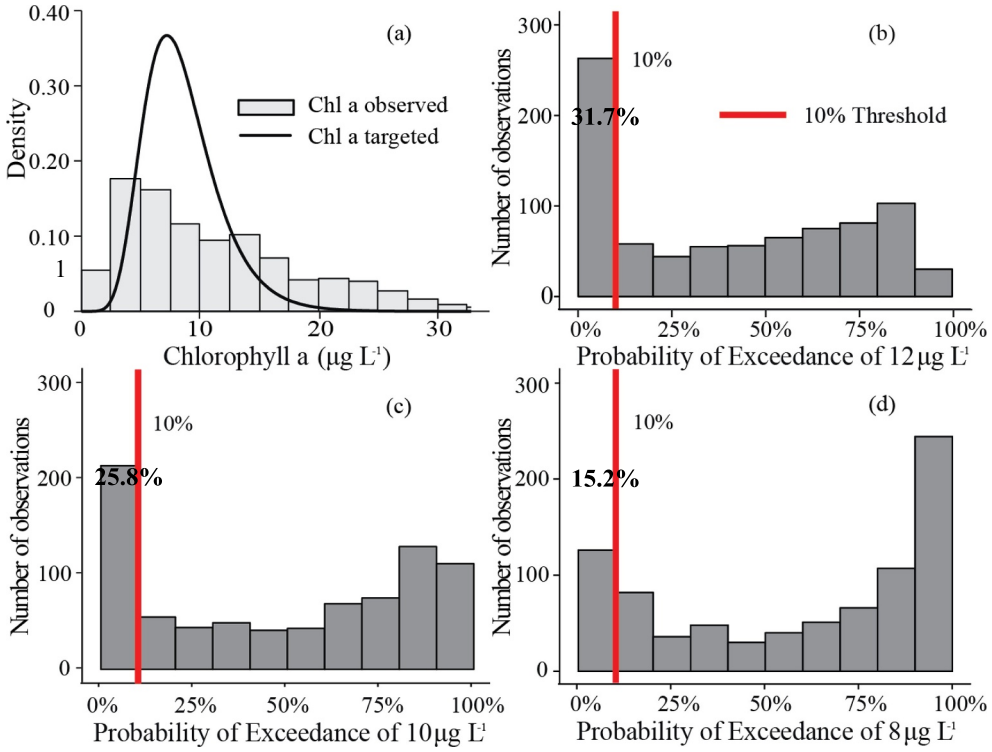

Fig. 3. Frequency histograms depicting (a) the measured chlorophyll *a* concentrations during the 1996–2017 period and the "ideal" distribution targeted by the established chlorophyll *a* criterion in the Bay of Quinte, (b) the mean probability values for chlorophyll *a* concentrations to exceed the 12 μg L^{-1}, (c) 10 μg L^{-1} and (d) 8 μg L^{-1} threshold levels[2]. The vertical red lines in panels (b)–(d) represent a tentative 10% probability benchmark of chlorophyll *a* violations. The expected (mean) exceedance in panels (b–c) were 39.8 ± 32.7%, 48.4 ± 34.9%, and 58.4 ± 35.3%, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

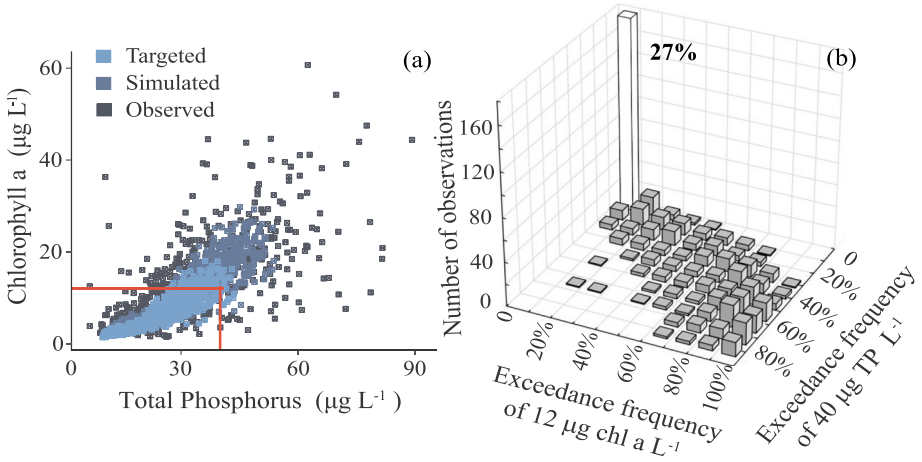

Fig. 4. (a) Scatterplot illustrating the joint distribution of chlorophyll *a* and TP concentrations based on the observed data (dark blue), simulated current (navy blue), and "ideal" conditions (light blue). The two red lines delineate the space of the "acceptable" system realizations, confined below the 12 μg chl*a* L^{-1} and 40 μg TP L^{-1} threshold values. (b) Frequency histogram of the joint likelihood of violations of the two water quality criteria. The white bar represents the snapshots from the system in which the frequency of violations of the two water quality extremes was lower than 10%. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 2
Assessment of our confidence that the system will comply with the targeted TP criterion of < 10% exceedance frequency of 40 μg L^{-1}, when a given number of violations occurs and no prior knowledge about the status of the system is assumed. Numbers in parentheses provide the same assessment of confidence of compliance, but with prior beliefs informed by data collected from the system during the *post-dreissenid* (1996–2017) and *eutrophic* (1975–1982) periods. The first row (numbers in italics) provides the sample size required during the growing season (early May-end of October) in order to assess compliance with the targeted water quality criterion at three locations in the Bay of Quinte, with a weekly or biweekly sampling frequency over the course of 1-, 2-, or 3-year periods.

Number of violations	1-Year period		2-Year period		3-Year period	
	Biweekly	Weekly	Biweekly	Weekly	Biweekly	Weekly
	30	*54*	*60*	*108*	*90*	*162*
4	19.4% (25.2% vs 4.7%)	65.1% (72.7% vs 38.1%)	74.6% (80.3% vs 48.3)	98.6% (99.3% vs 94.1)	95.8% (97.0% vs 85.2%)	100% (100% vs 99.9%)
6	3.3% (4.7% vs 0.1%)	30.9% (36.5% vs 12.6%)	40.9% (47.9% vs 19.6%)	92.6% (94.2% vs 81.4)	82.1% (84.7% vs 63.2%)	99.7% (99.9% vs 98.8)
8	0.2% (0.5% vs < 0.1%)	9.4% (12.6% vs 2.7%)	15.6% (21.5% vs 5.8%)	77.1% (81.9% vs 59.0%)	57.0% (62.1% vs 36.0%)	98.6% (98.8% vs 95.7)
10	< 0.1% (< 0.1% vs < 0.1%)	1.7% (2.8% vs < 0.1%)	3.9% (4.8% vs 1.3%)	53.1% (59.1% vs 35.6%)	30.3% (36.0% vs 15.5%)	94.2% (95.4% vs 85.7)

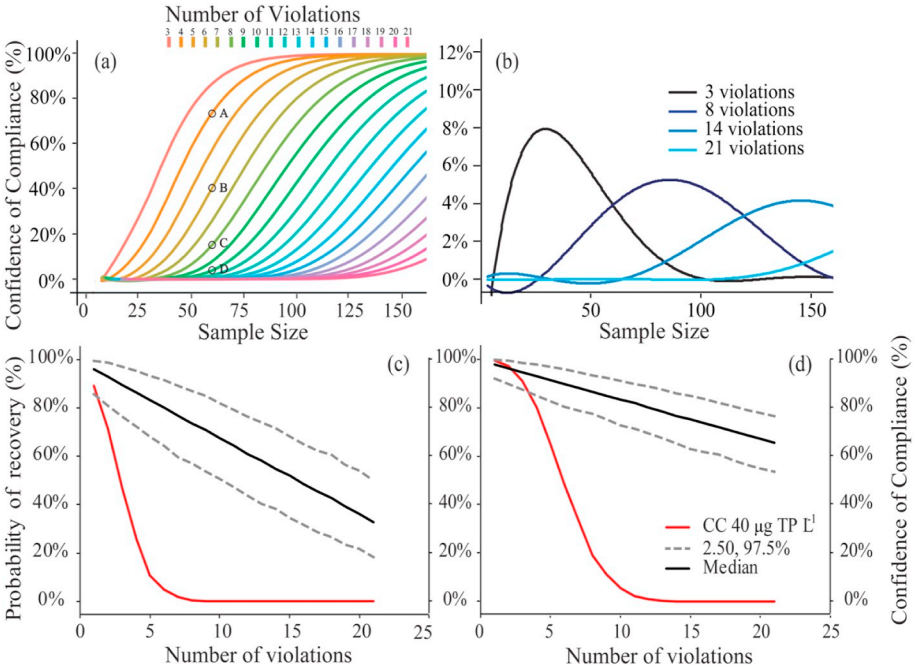

Fig. 5. (a) Degree of confidence that the 40 µg TP L^{-1} criterion will not be exceeded by > 10% in time and space when different sample sizes are collected from the Bay of Quinte and no prior knowledge -uniform reference prior, *Beta*(1, 1)- is assumed about the prevailing conditions. (b) Variations in our confidence-of-compliance assessment when the predictions are drawn from prior beliefs that are informed by empirical evidence from the system during the 1996–2017 period. The left Y axes of the *bottom panels* express the probability of recovery (or 100%-expected exceedance frequency of the water quality criterion) as a function of the number of samples with measured TP concentration > 40 µg L^{-1} (X axes), after monitoring three sites on a biweekly basis during (c) one growing season (n = 30) or (d) two growing seasons (n = 60). The solid black and gray dashed lines provide the mean (and associated 95% credible intervals) probability of the system to be restored, even though violations of the TP criterion still occur with different frequency. Red lines in the same bottom panels represent our confidence that the 40 µg TP L^{-1} criterion will not be exceeded by > 10% in time and space for a given number of violations recorded (right Y axes). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

It is within the context of the inherent uncertainty and/or our imperfect knowledge of natural systems that our statistical framework aims to assist with the development of realistic water quality criteria. In particular, the Bayesian nature of our work gives direct answers to questions of confidence of compliance with percentile standards; and as such can effectively guide management decisions and monitoring practices that consider both risks of reaching false conclusions, i.e., falsely inferring a breach of standard or falsely inferring compliance (McBride and Ellis, 2001). In particular, our modelling framework introduces a new continuous variable, the *degree of recovery*, which quantifies the likelihood of prevalence of the desirable/reference conditions in the system (100%-expected exceedance frequency of the water quality criterion), as a function of the frequency of standard violations (solid black and gray dashed lines in Fig. 5c and d). In the next step, instead of using this probability as the basis for a binary water quality assessment (impaired/non-impaired system), the estimated probabilities are translated into confidence statements to express the degree of our certainty that the targeted threshold values will not be exceeded by > 10% within the compliance assessment domain (red lines in Fig. 5c and d). In other words, we introduce an extra dimension of uncertainty by targeting the probability of an acceptable exceedance risk (or margin of safety) across all water samples taken (Wild et al., 1996; Zhang and Arhonditsis, 2008; Mahmood et al., 2014).

The question arising though is whether the precautionary spirit of our approach with the introduction of two layers of uncertainty is necessary, or if it is overly conservative and runs the danger to perpetrate the delisting process of impaired waterbodies with an overly alarmist mindset. After all, someone may question if it really matters that the TP or chlorophyll *a* threshold levels are exceeded more frequently than what is stipulated by the corresponding water quality criteria? With respect to the latter question, the perspective of the public in the Bay of Quinte area has been positive that it does matter. Based on the responses of > 1500 local residents and tourists, Ramin et al. (2018) indicated that there is a dramatic change in the public sentiment between the beginning and end of summer season, following the gradual deterioration of the water quality conditions. Moreover, in an attempt to associate the degree of satisfaction of the public with the contemporaneous water quality conditions, the same study found that the majority of the public is satisfied when chlorophyll *a* concentrations remain below the 10 µg L^{-1} threshold or ambient TP levels are lower than 20–25 µg L^{-1}, while the appreciation level increases significantly for every incremental decrease of the two water quality variables below their corresponding cutoff levels (see Figs. 8 and SI 6 in Ramin et al., 2018). The fact that the subjective judgments and sentiment of the public are tightly connected with the prevailing environmental conditions offers ammunition to develop and implement an ambitious long-term management plan that protects the Bay of Quinte from excess nutrients associated with urban runoff, sewage treatment plants, and agricultural land uses. Bearing in mind that the two critical conditions to embrace the precautionary principle are (*i*) the existence of a threat of an undesirable ecosystem shift, and (*ii*) the scientific uncertainty as to the extent of possible damage (Arrow and Fischer, 1974; Foster et al., 2000), we believe that the Bay of Quinte is an excellent example of a system where we have to be extra cautious in declaring success with respect to the degree of restoration and future resilience of the new ecosystem state.

4. Conclusions

We presented a statistical framework that aims to provide support for the technical challenges arising from the binary comparison between the conditions typically prevailing in impaired systems and the "desired" ones of reference sites. The main thrust of our framework is the adoption of percentile standards, whereby the water quality goals revolve around the extremes (and not the average conditions) and the line between impairment/non-impairment is drawn by explicitly acknowledging an inevitable risk of threshold crossings, the level of which could be subjected to decisions that reflect the different priorities of the local stakeholders and public regarding the integrity of various ecosystem services (Ramin et al., 2018; Kim et al., 2018). We contend that a single-value averaged standard, based on monitoring of offshore waters, is neither reflective of the range of spatiotemporal dynamics typically experienced in a natural system nor does it allow to evaluate our progress with ecosystem services at the degree of granularity required to influence positively the public sentiment. The degree of public satisfaction is primarily determined by the prevailing conditions at a particular recreational site in given date, and not by the average water quality over the entire system and growing season. Although our study

did not explicitly address the "offshore-versus-inshore" issue, we were able to show that even if we base the delisting decisions in multiple offshore sites of the studied system, the underlying variability and range of water quality conditions warrants the consideration of probabilistic criteria.

Our statistical methodology can be easily customized to examine exceedance probabilities and confidence compliance with environmental targets in any other impaired system around the Great Lakes or anywhere else in the world. Our case study, the Bay of Quinte, Lake Ontario, was an ideal system to showcase the merits of the proposed framework for four basic reasons: (*i*) the system is on the verge of being delisted as an AOC, after a long history of eutrophication problems and successful implementation of remedial measures that resulted in tangible water quality improvements; (*ii*) the prevalence of desirable conditions is occasionally interrupted by the occurrence of water quality extremes (e.g., high ambient nutrient levels, harmful algal blooms) that make the decision-making process less straightforward; (*iii*) the environmental policy-making process is founded upon a broader public engagement and the decisions are shaped by a multitude of factors (scientific understanding, public knowledge, and stakeholder perspectives) that do not always coalesce in terms of their perspectives; and (*iv*) there is a wealth of data to depict the water quality history of the system and sequentially update our beliefs about its resilience in response to ever-changing external stressors (i.e., urbanization, agriculture intensification, climate change), as the on-going monitoring will be supplying more empirical evidence. Notwithstanding the significant progress made thus far, our analysis suggests that there is still space for improvement until the prevailing conditions in the Bay of Quinte more reliably resemble those stipulated by the two water quality criteria. In the meantime, because of the presence of active feedback loops (e.g., internal loading) and its susceptibility to extreme events (e.g., harmful algal blooms), we believe that the establishment of rigorous compliance rules, such as $< 10\%$ exceedance frequency of the 40 μg TP L^{-1} and 10 μg chl*a* L^{-1} with $> 75\%$ confidence of compliance over the course of a 2- or 3-yr period, are essential steps until we can unequivocally declare the successful restoration of the system.

Acknowledgment

Funding for this study was provided by the Natural Sciences and Engineering Research Council of Canada (NSERC) through a Discovery Grant (George Arhonditsis). The authors are grateful to all the scientists and technical personnel, who have been involved with the development of the water quality dataset in the Bay of Quinte.

References

Arhonditsis, G.B., Stow, C.A., Steinberg, L.J., Kenney, M.A., Lathrop, R.C., McBride, S.J., Reckhow, K.H., 2006. Exploring ecological patterns with structural equation modelling and Bayesian analysis. Ecol. Model. 192, 385–409.

Arhonditsis, G.B., Papantou, D., Zhang, W., Perhar, G., Massos, E., Shi, M., 2008. Bayesian calibration of mechanistic aquatic biogeochemical models and benefits for environmental management. J. Mar. Syst. 73, 8–30.

Arhonditsis, G.B., Kim, D.-K., Shimoda, Y., Zhang, W., Watson, S., Mugalingam, S., Dittrich, M., Geater, K., McClure, C., Keene, B., Morley, A., 2016. Integration of best management practices in the Bay of Quinte watershed with the phosphorus dynamics in the receiving waterbody: what do the models predict? Aquat. Ecosyst. Health. 19 (1), 1–18.

Arhonditsis, G.B., Kim, D.-K., Kelly, N., Neumann, A., Javed, A., 2018. Uncertainty analysis by Bayesian inference. In: Recknagel, F., Michener, W. (Eds.), Ecological Informatics, 3rd edition. Springer, pp. 215–249.

Arrow, K.J., Fischer, A.C., 1974. Environmental preservation, uncertainty and irreversibility. Q. J. Econ. 88, 312–319.

Borsuk, M.E., Stow, C.A., Reckhow, K.H., 2002. Predicting the frequency of water quality standard violations: a probabilistic approach for TMDL development. Environ. Sci. Technol. 36, 2109–2115.

Dennis, B., 1996. Discussion: should ecologists become Bayesians? Ecol. Appl. 6, 1095–1103.

Doan, P.T.K., Watson, S.B., Markovic, S., Liang, A., Guo, J., Mugalingam, S., Stokes, J., Morley, A., Zhang, W., Arhonditsis, G.B., Dittrich, M., 2018. Phosphorus retention and internal loading in the Bay of Quinte, Lake Ontario, using diagenetic modelling. Sci. Total Environ. 636, 39–51.

Environment Canada (EC), United States Environment Protection Agency (USEPA), 2013. Great Lakes Water Quality Agreement. pp. 56.

Foster, K.R., Vecchia, P., Repacholi, M., 2000. Science and the precautionary principle. Science 288, 979–981.

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., Rubin, D.B., 2013. Bayesian data analysis. Chapman and Hall/CRC (675 pp).

George, T.K., Boyd, D., 2007. Limitations on the development of quantitative monitoring plans to rack the progress of beneficial use impairment restoration at Great Lakes Areas of Concern. J. Great Lakes Res. 33, 686–692.

Gudimov, A., Ramin, M., Labencki, T., Wellen, C., Shelar, M., Shimoda, Y., Boyd, D., Arhonditsis, G.B., 2011. Predicting the response of Hamilton Harbour to the nutrient loading reductions: a modeling analysis of the "ecological unknowns". J. Great Lakes Res. 37, 494–506.

Howard-Williams, C., Allanson, B.R., 1981. Phosphorus cycling in a dense *Potamogeton pectinatus* L. Bed. Oecologia 49, 56–66.

International Joint Commission (IJC), 2003. Status of restoration activities in Great Lakes areas of concern: a special report. April 2003. Washington DC and Ottawa. https://www.ijc.org/sites/default/files/aoc_report-e.pdf.

Janse, J.H., Scheffer, M., Lijklema, L., Van Liere, L., Sloot, J.S., Mooij, W.M., 2010. Estimating the critical phosphorus loading of shallow lakes with the ecosystem model PCLake: sensitivity, calibration and uncertainty. Ecol. Model. 221, 654–665.

Kim, D.-K., Zhang, W., Rao, Y.R., Watson, S., Mugalingam, S., Labencki, T., Dittrich, M., Morley, A., Arhonditsis, G.B., 2013. Improving the representation of internal nutrient recycling with phosphorus mass balance models: a case study in the Bay of Quinte, Ontario, Canada. Ecol. Model. 256, 53–68.

Kim, D.-K., Ramin, M., Cheng, Y.S., Javed, A., Kaluskar, S., Kelly, N., Kobiliris, D., Neumann, A., Ni, F., Peller, T., Perhar, G., Shimoda, Y., Visha, A., Wellen, C., Yang, C., Mugalingam, S., Arhonditsis, G.B., 2018. An integrative methodological framework for setting environmental criteria: evaluation of stakeholder perceptions. Ecol. Inform. 48, 147–157.

Mahmood, M., Blukacz-Richards, E.A., Baumann, P.C., McMaster, M., Hossain, M., Arhonditsis, G.B., 2014. A Bayesian methodological framework for setting fish tumor occurrence delisting criteria: a case study in St. Marys River area of concern. J. Great Lakes Res. 40, 88–101.

McBride, G.B., Ellis, J.C., 2001. Confidence of compliance: a Bayesian approach for percentile standards. Water Res. 35, 1117–1124.

Minns, C.K., Moore, J.E., Doka, S.E., St. John, M.A., 2011. Temporal trends and spatial patterns in the temperature and oxygen regimes in the Bay of Quinte, Lake Ontario, 1972–2008. Aquat. Ecosyst. Health. 14, 9–20.

Munawar, M., Fitzpatrick, M., Munawar, I.F., Niblock, H., Kane, D., 2012. Assessing ecosystem health impairments using a battery of ecological indicators: Bay of Quinte, Lake Ontario example. Aquat. Ecosyst. Health. 15, 430–441.

Nicholls, K.H., 1999. Effects of temperature and other factors on summer phosphorus in the inner Bay of Quinte, Lake Ontario: implications for climate warming. J. Great Lakes Res. 25, 250–262.

Nicholls, K.H., Heintsch, L., Carney, E., 2002. Univariate step-trend and multivariate assessments of the apparent effects of P loading reductions and zebra mussels on the phytoplankton of the Bay of Quinte, Lake Ontario. J. Great Lakes Res. 28, 15–31.

Ramin, M., Cheng, V.Y.S., Kim, D.-K., Ni, F.J., Javed, A., Kelly, N.E., Yang, C., Midlane-Jones, S., Mugalingam, S., Arhonditsis, G.B., 2018. A Bayesian methodological framework for coupling public perception with the water quality criteria setting process. Ecol. Econ. 147, 298–311.

Reckhow, K.H., Arhonditsis, G.B., Kenney, M.A., Hauser, L., Tribo, J., Wu, C., Elcock, K.J., Steinberg, L.J., Stow, C.A., McBride, S.J., 2005. A predictive approach to nutrient criteria. Environ. Sci. Technol. 39, 2913–2919.

Shabman, L., Smith, E., 2003. Implications of applying statistically based procedures for water quality assessment. J. Water Res. Plan. Man. 129, 330–336.

Shimoda, Y., Watson, S.B., Palmer, M.E., Koops, M.A., Mugalingam, S., Morley, A., Arhonditsis, G.B., 2016. Delineation of the role of nutrient variability and dreissenids (Mollusca, Bivalvia) on phytoplankton dynamics in the Bay of Quinte, Ontario, Canada. Harmful Algae 55, 121–136.

Smith, E.P., Canale, R.P., 2015. An analysis of sampling programs to evaluate compliance with numerical standards: total phosphorus in Platte Lake, MI. Lake Reserv. Manage. 31, 190–201.

Wild, P., Hordan, R., LePlay, A., Vincent, R., 1996. Confidence intervals for probabilities of exceeding threshold limits with censored log-normal data. Environmetrics 7, 247–259.

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

Zhang, W., Kim, D.-K., Rao, Y.R., Watson, S., Mugalingam, S., Labencki, T., Dittrich, M., Morley, A., Arhonditsis, G.B., 2013. Can simple phosphorus mass balance models guide management decision?: a case study in the Bay of Quinte, Ontario, Canada. Ecol. Model. 257, 66–79.