Castles built on sand or predictive limnology in action? Part A: Evaluation of an integrated modelling framework to guide adaptive management implementation in Lake Erie

George B. Arhonditsis⁎, Alex Neumann, Yuko Shimoda, Dong-Kyun Kim, Feifei Dong, Gabriela Onandia, Cindy Yang, Aisha Javed, Meghan Brady, Ariola Visha, Felicity Ni, Vincent Cheng

Ecological Modelling Laboratory, Department of Physical & Environmental Sciences, University of Toronto, Toronto, Ontario, M1C1A4, Canada

ARTICLE INFO

Keywords:
Lake Erie
Best management practices
Uncertainty analysis
Adaptive management implementation
Integrated modelling framework

ABSTRACT

We present a technical analysis of all the recent modelling work that has been conducted to support the adaptive management process in Lake Erie; the most biologically productive system of the Great Lakes. With a wealth of models developed, Lake Erie represents a unique case study where an impressive variety of data-driven and process-based models have been developed to elucidate the major watershed and aquatic processes underlying the local water quality problems. In the Maumee River watershed, the primary contributor of total phosphorus loading (~30%) into Lake Erie, the modelling work is based on five independent applications of the same process-based model, i.e., the Soil and Water Assessment Tool (SWAT). The five SWAT models showed nearly excellent goodness-of-fit against monthly flow rates and phosphorus loading empirical estimates based on a single downstream station, but little emphasis was placed on evaluating the robustness of the hydrological or nutrient loading predictions with a finer (daily) temporal resolution, and even less so in capturing the impact of episodic/extreme precipitation events. The multi-model ensemble for the Lake Erie itself has been based on a wide range of data-driven and process-based models that span the entire complexity spectrum. Consistent with the general trend in the international modelling literature, the performance of the aquatic ecological models in Lake Erie declined from physical, chemical to biological variables. Temperature and dissolved oxygen variability were successfully reproduced, but less so the ambient nutrient levels. Model performance for cyanobacteria was inferior relative to chlorophyll a concentrations and zooplankton abundance. With respect to the projected responses of Lake Erie to nutrient loading reduction, we express our skepticism with the optimistic predictions of the extent and duration of hypoxia, given our limited knowledge of the sediment diagenesis processes in the central basin and the lack of data related to the vertical profiles of organic matter and phosphorus fractionation or sedimentation/burial rates. Our study also questions the adequacy of the coarse spatio-temporal (seasonal/annual, basin- or lake-wide) scales characterizing the philosophy of both the modelling enterprise and water quality management objectives in Lake Erie. We conclude by arguing that one of the priorities of the local research agenda must be to consolidate the ensemble character of the modelling work in Lake Erie. The wide variety of models that have been developed to understand the major causal linkages/ecosystem processes underlying the local water quality problems are a unique feature that should be cherished and further augmented.

1. Introduction

In the context of natural resource management, the central goals of policy analysis aim to identify the important drivers of environmental degradation, to elucidate the sources of controversy, and to put the necessary tools in place in order to anticipate the unexpected. Environmental problems have a way of resurfacing themselves and are rarely (if ever) solved completely. However, even if certain facets of a management problem change over time, the core issues remain the...
same, and thus it is critical to establish a framework that ensures both continuity in the decision-making process but also iterative adjustments to accommodate the extrinsic non-stationarity or intrinsic stochasticity (Allen et al., 2011; Williams et al., 2011). Viewed from this context, adaptive implementation is a pragmatic strategy that not only acts as a hedge against the ubiquitous uncertainty surrounding the study of environmental systems, but also paves the way for the dual pursuit of management and learning (Walters and Holling, 1990). Adaptive management offers flexibility in making decisions in the face of uncertainty, as scientific learning progressively advances from research, monitoring, and impartial evaluation of the outcomes of past and ongoing management actions (Fig. 1). Even though its core principles are often misconstrued as a “trial-and-error” process, adaptive management promotes a learning-while-doing mindset whereby policies or operations can be regularly updated (Conroy et al., 2011; Lyons et al., 2010; Williams et al., 2009). Thus, adaptive management should not be perceived as an end in itself, but rather as a means to galvanize our efforts to advance scientific understanding of the attributes and functioning of an “impaired” environmental system, as well as to crystallize the decision-making process and facilitate the integration of environmental concerns with socio-economic values (Arhonditsis et al., 2016; Williams and Brown, 2014).

Environmental modelling is one of the pillars of the adaptive management process, serving as an “information integrator” that brings together scientists, managers, and other stakeholders in a joint assessment of the degree of our understanding of the system being managed along with the compelling knowledge gaps we seek to answer through monitoring and research. Models can be qualitative and conceptual, quantitative and detailed, or anywhere in between. As simplistic representations of natural ecosystems, their application in an adaptive management context inevitably introduces the so-called approximation uncertainty (Arhonditsis et al., 2018). This uncertainty stems from the imperfect knowledge used and different assumptions made to determine model structure and inputs. Model input error mainly stems from the uncertainty related to the values of model parameters, forcing functions, and initial conditions, as well as from the fact that all models are simplifications of the natural system that approximate the actual processes, i.e., all parameters are effective (e.g., spatially and temporally averaged) values unlikely to be represented by fixed constants (Arhonditsis et al., 2006). Model structure error arises from (i) the selection of the appropriate state variables (model endpoints) to reproduce the key physical, chemical, and biological components given the management problem at hand; (ii) the selection of the more suitable mathematical expressions among a variety of equations used to describe the same natural process, e.g., linear, quadratic, sigmoidal, and hyperbolic functional forms to reproduce fish predation on zooplankton or Monod vis-à-vis Variable-Internal Stores equations to simulate the phytoplankton uptake of nutrients from the water column and their conversion into biomass; and (iii) the fact that our models are based on relationships which are derived individually in controlled laboratory environments but may not collectively yield an accurate picture of the entire environmental system (Arhonditsis et al., 2007, 2008a, 2008b). Coupled with other major sources of uncertainty (environmental variation, structural uncertainty, partial controllability, and partial observability), skeptical views question the ability of models to meaningfully assist the decision making process. The counter-argument to this criticism is that the problem of uncertainty is precisely where adaptive management is most valuable (Williams et al., 2009). This paradigm offers a coherent framework to quantify the uncertainty of initial forecasts of management actions and sequentially update them with post-implementation monitoring data; hence, this iterative synthesis of monitoring and modelling provides the basis to reduce the
uncertainty and revise management actions (Fig. 2). Even more so, there are voices arguing that adaptive management is better served by introducing multiple competing models that embed alternative hypotheses about the system functioning, as the iterative updating cycles will lend support to the optimal subset of models that more closely reflects the evolution of the environmental system and its potential responses to management actions (Ramin et al., 2012a; Williams et al., 2011).

It is thus axiomatic that the intertwined relationship between monitoring and modelling is key to the successful implementation of adaptive management. The commitment among managers, scientists, and other stakeholders to establish a sustainable monitoring and assessment program is imperative in this process. Both monitoring and assessment of environmental conditions should be designed to ensure that data are collected for all the relevant resource attributes and within the timeframe required for adaptive-decision making (Williams and Brown, 2014). Another critical aspect of the design of monitoring programs is the appropriate resolution in time and space in order to constrain the existing model(s) and effectively track our progress toward accomplishing management objectives. The information from monitoring is then used to evaluate management, improve understanding, and guide decision making, and it is thus critical to have the appropriate metrics to evaluate the knowledge (or value of information) acquired (Williams and Johnson, 2015; Yokota and Thompson, 2004). The latter endeavor represents our technical learning of the response of the environmental system to management interventions, which is conceptually (and operationally) nested within the so-called institutional learning or the potential to learn about the decision making process itself (Williams and Brown, 2016). Empirical evidence suggests that it may take multiple iterations of the technical learning cycle during which the institutional framework remains practically unaltered, and only a breakthrough in technical learning can prime the pump for revisiting and restructuring the basic institutional elements (Williams and Brown, 2014). More recently, this double-loop learning process has been expanded to distinguish between the socio-political and governance issues typically included within the institutional learning cycle (Johnson et al., 2015; Pahl-Wostl, 2009).

In this study, we present a technical analysis of all the recent modelling work that has been conducted to support the adaptive management process in Lake Erie; the most biologically productive system of the Great Lakes. In response to a binational remedial effort in the late 1970s and early 1980s, west-central phytoplankton biomass and central-basin hypoxia displayed significant improvement in Lake Erie, followed by a general increase since the mid-1990s which was likely the cumulative effect of several stressors, including the invasion of exotic species and increased agricultural loading of bioavailable phosphorus (Watson et al., 2016). Following the evolution of watershed and eutrophication modelling in the literature, Lake Erie represents a unique case study where a wide variety of models have been developed to understand the major causal linkages/ecosystem processes underlying the local water quality problems (Scavia et al., 2016a, 2016b; Scavia et al., 2017). Whether statistical (data-driven) or mechanistic (process-based), the basic premise of these models has been to support the decision-making process by linking watershed management actions with the response of the receiving waterbody.

The present study provides an overview of all the models used in the area by shedding light on their fundamental assumptions, structural features, and general consistency against empirical knowledge from the

Fig. 2. Sequential model updating in the context of adaptive management. The interphase between model uncertainty analysis and data collection is highlighted as a key process to advance our understanding and management actions in Lake Erie.
system. We subsequently offer a technical analysis of the forecasts drawn regarding the likelihood to mitigate the eutrophication phenomena in Lake Erie, based on changes in the agricultural management practices followed in the watershed. In a companion paper, we pinpoint knowledge gaps and monitoring assessment objectives that should be addressed to ensure that resource parameters are adequately measured and appropriately focused on relevant performance indicators (Arhonditsis et al., 2019). Our intent is neither to vilify the modelling enterprise in Lake Erie nor to roundly criticize all the models developed. We recognize that a great deal of novel modelling work has been done to offer insights into the watershed and lake processes, and our aim here is to impartially identify their strengths and weaknesses. In fact, some of the issues identified in our study are being addressed through on-going modelling work. Because of the challenges and complexity of the adaptive framework, we believe that all the experiences gained and lessons learned through the iterative monitoring-modeling-assessment cycles will consolidate our know-how with the management of one of the most intensively studied eutrophic systems worldwide (Michalak et al., 2013; Scavia et al., 2017). Our thesis is that critical thinking and effective monitoring are two fundamental prerequisites for hypothesis testing and robust model foundation in order to achieve one of the key aspirations with adaptive management; the reduction of uncertainty.

2. Study area

Lake Erie is the smallest and shallowest system of the Great Lakes and therefore tends to be the most susceptible to nutrient-driven water quality problems (Fig. 3). The shallowest section of Lake Erie is the western basin where depths average between 7.6 and 9.1 m; the central basin is deep enough to stratify during the summer (mean depth of 18.3 m) typically developing a “thin” hypolimnion with a small volume relative to the epilimnion; the eastern basin has an average depth of 24.4 m and large hypolimnetic volume. Recent evidence suggests that rapid ecological changes have been occurring in Lake Erie that primarily involve the severity of eutrophication phenomena, such as an increase in the magnitude and duration of harmful algal blooms (HABs), prolonged manifestation of hypoxia in the central basin, and excessive Cladophora growth in the eastern basin (Depew et al., 2011; Higgins et al., 2008; Michalak et al., 2013; Scavia et al., 2014; Watson et al., 2016). The western part of Lake Erie basin, comprising both the western basin and the Huron-Erie corridor (Detroit River Basin) receives > 60% of the external annual total phosphorus (TP) loading and among all the tributaries across the entire Lake Erie basin, the Maumee River watershed is the primary contributor of TP loading (~30%) into Lake Erie (Maccoux et al., 2016). The catchment area of Maumee River is approximately 16,480 km² extending over southern Michigan, northwestern Ohio, and northeastern corner of Indiana (Keitzer et al., 2016). Most of the watershed is predominantly agricultural (73%) and is characterized by a flat landscape (average slope: < 2%) and poorly-drained soils. > 90% of the agricultural land is drained by ditches and subsurface drainage, and 85% of phosphorus exports are reported to originate from agricultural inputs (e.g., fertilizer and manure applications) (Cubertson et al., 2016).

Planktonic blooms in Lake Erie have been characterized by a major shift of the dominant species and particularly by the predominance of toxic cyanobacteria (cHABs), such as Microcystis and other potentially toxic (Dolichospermum, Planktothrix) species (Kane et al., 2015; Allinger and Reavie, 2013; Millie et al., 2009). The majority of these cHABs frequently occur in the western basin of Lake Erie, where they cover extensive areas and can persist throughout the summer until late fall (Steffen et al., 2014; Michalak et al., 2013). Interestingly, the increased frequency in cHABs has occurred without distinct trend in total annual TP loading, but may be related to the timing, sources, and increased bioavailability of TP inflows and/or the increased frequency of extreme flow events (Obenour et al., 2014; Scavia et al., 2014). The cHAB severity has been causally linked to the volume of spring runoff from the Maumee River watershed, as well as to the elevated dissolved phosphorus inputs from the western basin tributaries (Watson et al., 2016; Stumpf et al., 2012). Growing evidence also highlights the broader role of dreissenid mussels in the ambient conditions (enhanced water clarity, selective particle removal, and soluble nutrient recycling), the supply and chemical speciation of nitrogen (N), the increasingly calm summer conditions, and the increasing reservoir of Microcystis seed colonies within Lake Erie’s benthos (Davis et al., 2015; Horst et al., 2014; Rinta-Kanto et al., 2009).

The proliferation of Cladophora along the northern shores of the eastern basin since the mid-1990s, was primarily attributed to the increased water clarity and suitable substrate following the colonization of the area by dreissenid mussels (Higgins et al., 2005). Nutrient-rich hypolimnetic masses of water transported in nearshore through upwelling events, excreted metabolic wastes, and/or particulate matter available for re-mineralization through egestion of non-edible algae by dreissenids are the likely suppliers of nutrients in the benthic environment (Valipour et al., 2016; Depew et al., 2011; Wilson et al., 2006).

The magnitude, duration, and frequency of hypoxia is predicted to be exacerbated by climate change via a number of mechanisms such as increased water temperature, deeper and longer stratification, and increased nutrient runoff during winter and spring (Jiang et al., 2012; Fang and Stefan, 2009). Because of its bathymetry, the severity of hypoxia in central Lake Erie can display both intra- and inter-annual variability driven by the impact of local weather conditions (e.g., wind, temperature) on physical processes such as mixing, within- and between-basin circulation, and water-column stratification (strength, depth), which in turn influence the rate of dissolved oxygen (DO) depletion, sediment-oxygen demand, and DO transfer across the thermocline (Rucinski et al., 2014; Zhou et al., 2013). Consistent with these predictions, a record-breaking hypoxic event occurred in Lake Erie in 2012, following a period of prolonged drought and low tributary flows (Stow et al., 2015).

3. Model ensembles and adaptive management implementation

Recognizing the need to address these ominous water quality trends, the International Joint Commission (IJC) formulated the Lake Erie Ecosystem Priority taskforce in 2012 to develop a sustained restoration plan by identifying knowledge and monitoring gaps, evaluating current conditions, and providing guidance for management targets (Watson et al., 2016; IJC, 2014). This initial IJC (2014) review, and associated work by binational taskforces, led to commitments for remedial action in the 2012 renewed Canada-USA Great Lakes Water Quality Agreement (GLWQA, 2016). Nutrients, algal biomass, and hypoxia are addressed under Annex 4 of this Agreement, with special emphasis on setting interim TP load and basin-specific concentration targets for Lake Erie. As part of the GLWQA (2016) review, a committee of modellers evaluated a set of Great Lakes eutrophication models that were designed to establish target TP loads in order to mitigate the eutrophication symptoms in the early 1970s (Scavia et al., 2016a). While the post-audit performance of the original eutrophication models in Lake Erie was satisfactory (e.g., Lesht et al., 1991; Di Toro et al., 1987), their basic structure was not deemed adequate to reproduce the relative importance of nearshore processes, the wide array of factors triggering HAB formation or the proliferation of nuisance benthic algae, and the role of other external stressors, such as invasive species or climate change (Scavia et al., 2017; Scavia et al., 2016a, 2016b). To address the wide range of conceptual and operational uncertainties typically characterizing any modelling exercise, the local community opted for a novel multi-model strategy that aimed to capitalize upon the wide variety of both empirical and process-based models of variant complexity that have been developed for Lake Erie over the past decade (Scavia et al., 2017; Scavia et al., 2016a, 2016b).
Being primarily a reflection of our current level of understanding and existing measurement technologies, the multi-model strategy adopted for Lake Erie accommodates the fact that many different model structures and many different parameter sets within a chosen model structure can acceptably reproduce the observed behavior of a complex environmental system (Arhonditsis et al., 2007, 2011; Beven and Freer, 2001; Christakos, 2003). While this very important notion is still neglected in the modelling literature, there are viewpoints suggesting that environmental management decisions relying upon a single, partially adequate, model structure can introduce bias and uncertainty that is much larger than the error stemming from a single, partially defensible, selection of model parameter values (Neuman, 2003). Importantly, the practise of basing ecological predictions on one single mathematical model implies that a valid alternative model may be omitted from the decision making process (Type I model error), but also that our forecasts could be derived from an erroneous model that was not rejected in an earlier stage (Type II model error). Recognizing that there is no true model of an ecological system, but rather several adequate descriptions of different conceptual basis and structure, the development of model ensembles is a technique specifically designed to address the uncertainty inherent in the model selection process. Instead of picking the single “best-fit” model to draw ecological forecasts, we can use a multi-model ensemble to derive a weighted average of the predictions from different models (Ramin et al., 2012a).

Notwithstanding the voices in the literature asserting that we are still missing rigorous methodological frameworks to develop multi-model ensembles (Neuman, 2003), the basic framework comprises several steps related to the development of “truly” distinct, site-specific conceptual models, selection of the optimal subset of both data-driven and process-based models, effective combination of these models to synthesize predictions, and subsequent assessment of the underlying uncertainty (Fig. 4). This methodological procedure involves three critical decisions aiming: (i) to identify the conceptual or structural differences of the existing models (ensemble members), and thus determine the actual diversity collectively characterizing the model ensemble; (ii) to determine the most suitable calibration/validation domain for evaluating model performance in time and space; and (iii) to establish an optimal weighting scheme in order to assign weights to
each ensemble member, when integrating over the individual predictions, and determine the most likely outcome along with the associated uncertainty bounds (Ramin et al., 2012a). In this study, we dissect the two model ensembles developed for the Maumee River watershed and the Lake Erie itself and evaluate their compliance with the aforementioned framework (Fig. 3).

In terms of model diversity, the watershed modelling work has been based on five applications of the same complex mathematical model (i.e., SWAT), each application reflected an independent attempt to characterize the processes that modulate phosphorus export from Maumee River. If we also consider the multiple sub-routines that can be chosen within SWAT (Scavia et al., 2017; Neitsch et al., 2011), the five models collectively capture some of the uncertainty related to our understanding of watershed attributes and functioning, and thus this exercise qualifies to be treated as an ensemble strategy (Kim et al., 2019). Alongside the five SWAT applications, an empirical model (SPARROW) was also used to validate a post-hoc delineation exercise of locations with higher likelihood of phosphorus export in the Maumee River watershed. By contrast, the lake ensemble has been based on a wide range of data-driven and process-based models to examine the achievability of different environmental targets under various external nutrient loading regimes (Shimoda et al., 2019). Beyond the structural diversity of the models in place, our study also attempts to identify their actual variability in terms of the characterization of fundamental biogeochemical processes modelled, ecological insights gained, and the nature of predictions drawn. Our thesis is that the true diversity of a model ensemble is not solely determined from the complexity of the mathematics or the number of system components simulated, but also from the characterization of the ecosystem processes, as determined by the parameter values assigned during the calibration and validation phase of the individual models, which ultimately shape the forecasts derived to guide the policy-making process. Regarding the model assessment, the typical criteria used are the degree of prediction errors (skill; the difference between predictions and observations); the relative goodness-of-fit for the different model endpoints (bias; differences in the model fit among flow, total and dissolved reactive phosphorus loading along with the nature of the error, i.e., over- or under-estimation); and the degree of divergence of the individual model forecasts when forced with various management scenarios (forecasting spread) (Arhonditsis et al., 2011; Arhonditsis and Brett, 2004). In this study, we will examine if the three criteria have been consistently evaluated, to what extent the model assessment was based on suitable calibration/validation domains and spatiotemporal scales relevant to the environmental management problems in question, and whether or not the performance assessment or the model complexity were used to determine the relative weights that each model carried during the synthesis of the individual predictions.

3.1. SWAT-based ensemble strategy for the Maumee River watershed: description, performance, and outstanding questions

In the Great Lakes area, a wide range of process-based models have been used to characterize the watershed processes associated with the hydrological cycle and nutrient fate and transport, but SWAT alone represents more than > 70% of the watershed modelling studies published for Lake Erie in the peer-reviewed literature (Dong et al., 2019;
Kim et al., 2019). SWAT is a semi-distributed watershed model that can operate on various time (sub-daily to annual) steps and its structure can accommodate the weather generation, watershed topography, hydrological processes, and agricultural practices (Neitsch et al., 2011). The watershed is disaggregated into subbasins and these subbasins are subsequently subdivided into hydrological response units (HRUs), which represent distinct combinations of soil and land-use characteristics. The water balance of each HRU in the watershed is based on four storage volumes: snow, top soil (0–2 m), shallow aquifer (2–20 m), and deep aquifer (> 20 m). Surface runoff is commonly estimated using the Curve Number (CN) method, and sediment erosion is calculated with the Modified Universal Soil Loss Equation (Arnold et al., 1998). Nutrient (nitrogen and phosphorus) dynamics are reproduced by considering both mineral (stable, active, and soluble) and organic (stable, active, and fresh) forms. The preference for SWAT over other watershed models is likely due to its ability to carry continuous long-term simulations in predominantly agricultural watersheds, as well as to reproduce the impact of episodic rainfall events at finer (i.e., daily or sub-daily) resolutions (Wellen et al., 2014a, 2014b; Neitsch et al., 2011).

Like many watershed models, SWAT is basing its predictive capacity on the explicit consideration of a wide array of physical, chemical, and biological processes that can shape downstream flow and nutrient export conditions. This strategy -in principle- renders greater assurance that all the potentially important mechanisms operating in a watershed context are accounted for, and thus the application domain of the model can be effectively extended (Beven, 2006; Pappenberger and Beven, 2006). However, increasing model complexity inevitably magnifies the disparity between what we want to learn from a model and what we can realistically measure/monitor given the available technology or resources at hand (Arhonditsis et al., 2008a, 2008b; Rode et al., 2016). Opting for an increased model complexity without a commensurate increase in the available empirical knowledge exacerbates the problem of under-determination, and therefore we experience a situation whereby several distinct choices of model inputs lead to the same model output, or alternatively many sets of parameters fit the data about equally well. This non-uniqueness of model solutions, also known as equifinality, negates the main function of mathematical models as inverse analysis tools, i.e., the available data for the dependent variables (flow, suspended solids, nutrient concentrations) are used through the model training (or calibration) phase to advance our knowledge of independent variables (model parameters) typically representing fundamental processes/fluxes of the water budget and/or nutrient cycles. To overcome the latter problem, the modelling work in Maumee River watershed was built upon five SWAT model applications from groups affiliated with Heidelberg University (HU), LimnoTech (LT), Ohio State University (OSU), Texas A&M University (TAMU), and the University of Michigan (UM) (Calbertson et al., 2016; Gildow et al., 2016; Kalicec et al., 2016; Keitzer et al., 2016; Muench et al., 2016). Based on different assumptions and independent calibrations, each model reflects a distinct characterization of the watershed processes modulating flow dynamics and phosphorus loading export. Thus, the five SWAT applications collectively represent an attempt to capture our incomplete knowledge of watershed attributes and functioning, and create an uncertainty envelope that can be propagated when the models are used for forecasting purposes and land-use management scenarios (Dong et al., 2019; Kim et al., 2019). It is also important to note that the characterization of the processes associated with the nitrogen cycle has not been studied yet, even though the relative contribution of different N species from the Maumee River watershed is likely to be one of the regulatory factors of the downstream water quality conditions, e.g., the composition of the phytoplankton community (Arhonditsis et al., 2019).

The calibration of the five SWAT models covered different time spans, 4 years for HU (2009–2012), 13 years for LT (1998–2013), 10 years for OSU (2000–2009) and TAMU (1999–1990), and 5 years for UM (2001–2005) (Scavia et al., 2017). Consequently, each model used different landscape (e.g., land-use/land-cover or LULC data) and meteorological information (see Table 1 in Kim et al., 2019). Moreover, the five SWAT models differed with respect to the agricultural inputs used, such as fertilizer and manure application rates. In particular, each postulated different fertilizer application rates from different sources of information, such as tri-state standards for HU and OSU, USDA-ARS for LT, agriculture census data for TAMU, and county fertilizer sales data for UM. Regarding the manure application, only two models (LT and UM) explicitly considered the role of this potentially important source of organic nutrient release in the soils. Another critical source of uncertainty revolved around the tile drainage assumptions. Three of the SWAT applications (LT, OSU, and UM) postulated that tiles are deployed at poorly-drained agricultural lands, but each model was based on different definitions of the categories of poorly-drained conditions (see Table 1 in Kim et al., 2019). Two models, HU and TAMU, also assumed that the tile drainage of the Maumee River watershed is highly correlated with low-slope areas, i.e., the HU model used a 3% slope to distinguish between steep and flat areas, while the cutoff point for TAMU was < 1% slope. All five SWAT models were calibrated against a single downstream station (Waterville, OH), while the LT model used two additional upstream sites (Blanchard and Tiffin) for a subsequent validation exercise (Scavia et al., 2016c). Consistent with the methodological practice recommended in the literature (Fig. 4), the five SWAT applications were based on identical validation time periods (from 2005 to 2014), which in turn (partly) insured their comparability for the purposes of a post-hoc ensemble synthesis (Kim et al., 2019).

Prior to conducting the validation exercise, Scavia et al. (2016c) specified the criteria of excellent fit against the flow rates to be lower than ± 10%, higher than 0.5, and higher than > 0.6 for the percent bias (PBIAS), modelling efficiency (MEF), and coefficient of determination ($r^2$), respectively. Likewise, the excellent performance for phosphorus loading was determined to be PBIAS < ± 25%, MEF > 0.4 and $r^2$ > 0.5. Based on these pre-specified performance standards, the five SWAT models showed nearly excellent goodness-of-fit against measured monthly flow rates and phosphorus loading empirical estimates (See Table S1 in the Supporting Information section). The OSU model reported the best agreement against the measured flow rates in

| Spatial Segment | Models |
|-----------------|--------|
| North-west      | H      | H     | H     | I     | H     |
| North-center    | H      | I     | H     | I     | H     |
| North-east      | L      | L     | I     | L     | L     |
| Center-west     | H      | H     | I     | I     | H     |
| Center          | I      | I     | I     | I     | I     |
| Center-east     | H      | L     | I     | L     | I     |
| South-west      | H      | H     | I     | I     | I     |
| South-center    | L      | H     | I     | I     | I     |
| South-east      | H      | H     | L     | I     | I     |

Results based on Scavia et al. (2016c). Some of these spatial projections are revisited on-going modelling work (Dr Margaret Kalicec, Pers. Comm.).
terms of MEF (0.91), PBIAS (10%), and $r^2$ (0.93). By contrast, the HU model was the only SWAT model that underestimated the flow rates (PBIAS: −7%) and showed the lowest MEF (0.82) and $r^2$ (0.86) among the five models. Regarding the TP and dissolved reactive phosphorus (DRP) loading predictions, the individual SWAT applications presented lower goodness-of-fit for both estimates, but the performance for TP (MEF: 0.56–0.82 and $r^2$: 0.70–0.82) was higher than for DRP (MEF: −0.02–0.71 and $r^2$: 0.51–0.71). The LT model outperformed the other four SWAT applications in its ability to capture the monthly TP and DRP load variability, whereas the HU model significantly overstated the TP (PBIAS: 37%) and DRP (PBIAS: 81%) loads. Given that the same SWAT application underestimated the flow rates, the latter result suggests that the simulated concentrations were distinctly overestimated to the extent that not only compensated for the underpredicted flows but also led to excessively higher phosphorus loading predictions (Kim et al., 2019). By contrast, the TAMU model clearly underestimated TP and DRP loads (−22% and −13%, respectively) compared with the overestimated flow rates (11%), which again is indicative of an even greater bias (i.e., under-prediction) for the corresponding concentrations.

One of the critical skill assessment tests of a dynamic watershed model, like SWAT, is its ability to support predictions at a finer (daily or subdaily) time scale (Wellen et al., 2015). Surprisingly, the SWAT modelling work in the Maumee River watershed placed little emphasis on evaluating the robustness of the hydrological or nutrient loading predictions with such a temporal resolution. To shed light on this important issue, Kim et al. (2019) independently evaluated the error associated with the daily outputs from the UM model and showed that the simulated TP and DRP concentrations are characterized by a substantial error (see Figs. 4 and SI 3 in Kim et al., 2019). Importantly, the same study evaluated the ability to capture the impact of episodic precipitation events, by using a peak-flow threshold of 1000 m$^2$ s$^{-1}$ to separate extreme events from the rest of the flow conditions over the course of a 5-year period (2001–2005). The results showed that the UM SWAT consistently underestimated the flow rates in 20 out of 22 events (> 91%), reinforcing the point that an excellent goodness-of-fit with a coarser (seasonal or monthly) resolution may simply stem from multiple daily errors/biases that cancel each other out when seasonally or monthly averaged, and thus does not necessarily guarantee acceptable performance against finer time scales (Arhonditsis and Brett, 2004). Given that the UM displayed the lowest discrepancy against the monthly flows (PBIAS: 6%) and TP loading (7%) among the five SWAT applications (Table S1), it stands to reason that the rest of the models likely fare worse with their daily simulations. If we also consider the emerging evidence of a disproportionally high fraction of the total annual loading occurring during event-flow conditions, it is particularly critical to strengthen the ability of the SWAT-ensemble to assess the impact of episodic/extreme rainfall events and elucidate their role on the broader watershed dynamics.

The potential impacts of climate change on hydrological extremes have received considerable attention during the last decade. It is predicted that global warming will amplify the hydrological cycle, and thus less frequent but more intense precipitation events could increase the severity of within-season drought, alter evapotranspiration, and generate greater runoff (Godess, 2013; Huntington, 2006; Knapp et al., 2008). In the same context, recent empirical evidence is on par with these projections suggesting that (i) total phosphorus and phosphate loads can vary by three orders of magnitude between wet and dry conditions; (ii) storm events and spring freshets play a predominant role with the peak daily loads in urban and agricultural watersheds, respectively; (iii) a significant fraction (> 50%) of the annual phosphorus loads can be generated during a small number of brief but intense precipitation events; and (iv) the flow-concentration relationship can be significantly influenced by the watershed physiography, land-use patterns, and antecedent soil moisture conditions (Green et al., 2007; Long et al., 2014, 2015). By contrast, an overarching flow-concentration paradigm for N species is even less clear relative to that for phosphorus. Contrary to our understanding of TP fate and transport, a greater proportion of total nitrogen (TN) is found in the dissolved phase due to relatively high solubility of nitrogen species, such as nitrite and nitrate, and can be transported by both overland and subsurface flow paths (Long et al., 2014). Furthermore, subsurface leaching of nitrate, and hence transport to groundwater, is generally greater than phosphate due to immobilization of phosphate by clay and other soil chemical constituents. Thus, one of the working hypothesis is that phosphorus and certain N species (ammonia, total Kjeldahl nitrogen) will be distinctly higher during precipitation and snowmelt events, while nitrate will display chemostatic behavior (i.e., apparent stability of the concentrations relative to the flow variability) (Long et al., 2014; Godsey et al., 2009). Another testable hypothesis is that the nutrient loads per unit area will be significantly higher from urban relative to agricultural watersheds and may be further exacerbated with climate change, as hydrological behavior can change above certain discharge or precipitation thresholds (Kim et al., 2016, 2017; Wellen et al., 2014a, 2014b). Climate warming will increase the vulnerability of soils to erosion in winter (snowpack decrease, early onset of spring snowmelt, frequent rainfall events, and snowmelt episodes), and consequently the contemporaneous sediment and nutrient loadings relative to current levels. It is critical to understand the potential changes in the interplay among catchment state, land use, and nutrient-export patterns induced by a changing climate, and this is another area where more work can be done in the Maumee River watershed.

3.2. Multi-model ensemble strategy for Lake Erie: description and performance

Counter to the watershed modelling framework for the Maumee River Watershed, the multi-model approach for the lake itself comprised a range of data-oriented and process-based models to examine the impact of nutrient loads to ecosystem integrity in Lake Erie and to evaluate the ERI achievability under different watershed management scenarios (Fig. 3; see also Scavia et al., 2016a, 2016b). The primary advantage of the suite of models used was their ability to capitalize upon the advantageous features of both statistical and process-based approaches; namely, the former models are derived from empirical parameter estimation that allows for rigorous assessment of predictive uncertainty, while the latter ones have the mechanistic foundation that can conceivably enable to draw predictions outside the domain used during their calibration or even validation (Arhonditsis et al., 2007). Regarding the diversity of the process-based modelling work for Lake Erie, the models developed came from the entire complexity spectrum with different strengths and weaknesses (Scavia et al., 2016a, 2016b). There are simple models in place with fewer unconstrained parameters that can be more easily subjected to uncertainty analysis, but they are also criticized as being crude oversimplifications not capable of reproducing the wide range of ecosystem behaviors. By contrast, there are complex models with numerous ecological processes and more sophisticated parameterization that are presumably more suitable to capture any potential non-linear system responses to distinct changes in external or internal conditions (e.g., land-use management, climate change, invasive species), but the main criticisms are their inevitably poor identifiability that may lead to ecosystem misconceptualizations, as well as their high computational demands that can be prohibitive in comprehensively quantifying their structural or parametric uncertainty (Arhonditsis et al., 2007; Arhonditsis and Brett, 2004). Models included in the multi-model ensemble strategy were two empirical models, UM/GLERI. Western Basin HAB model (Bertani et al., 2016) and NOAA Western Basin HAB model (Stumpf et al., 2016), and six process-based models: Total Phosphorus Mass Balance Model (TPMB; Chapra et al., 2016), 1-Dimensional Central Basin Hypoxia Model (1D-CBH; Rucinski et al., 2016), Ecological model of Lake Erie (EcoLE; Zhang et al., 2016), Western Lake Erie Ecosystem Model (WLEEM; Verhamme et.,
In reviewing the pertinent literature, a first notable finding is that the local modelling work has closely followed the recommended methodological protocol when developing models intended to assist environmental management (Fig. 2). Following the evolution of each model over time, we can find detailed sensitivity analysis exercises and goodness-of-fit statistics against a wide range of multi-year conditions. Of equal importance is that Scavia et al. (2016a, 2016b) attempted to draw parallels among the different models, thereby offering a much-needed synthesis in order to establish the multi-model forecasting tool. Nonetheless, because of the complexity of the existing mechanistic models in Lake Erie, the rigorous quantification of their uncertainty can be particularly challenging and thus has not received significant attention. A plausible way to control model uncertainty is to capitalize upon the findings of the sensitivity analysis exercises presented in the local literature by obtaining empirical estimates for critical/influential parameters or other inputs, which can be directly measured in the field or experimentally quantified (see also discussion in Kim et al., 2019). There are also promising statistical ways (e.g., linear or non-linear emulators) to overcome the cumbersome structure of complex models (e.g., ELCOM-CAEDYM, WLEEM, EcoLE) and quantify their uncertainty that should be considered in the next iteration of the modelling framework (Kim et al., 2014). Consistent with the general trend in the international modelling literature (Arhonditsis and Brett, 2004), we found that the performance of the aquatic ecological models in Lake Erie declined from physical, chemical to biological components of Lake Erie. The relative error for each state variable is based on the median value of all the corresponding graphs published in the peer-reviewed literature, in which field data were compared against simulated values. Counter to the practices followed in Scavia et al. (2016a, 2016b), our goodness-of-fit assessment was based on point comparisons in time and space instead of aggregated (seasonal/annual, basin- or lake-wide) scales.

4. Lessons learned from the Lake Erie modelling framework

4.1. SWAT-based ensemble strategy: delineation of “hot-spots” in the Maumee River watershed and analysis of BMP scenarios

Following the development of a spatially distributed model, the identification of high-risk areas (or “hot-spots”) with greater propensity for nutrient export and downstream delivery rates is an important exercise (Scavia et al., 2016c). In the Maumee River watershed, given that the calibration of all five SWAT applications was based on a single downstream station without any spatial information used to constrain the underlying processes, the (dis)agreement among the corresponding delineations could be primarily determined by two factors: (i) the discrepancies among the assumptions made or input data used during the spatial configuration (e.g., tile drainage, fertilizer/manure application rates, LULC data) of the individual models; and (ii) the differences in the characterization of processes pertaining to the water and nutrient cycles. The accommodation of the latter source of uncertainty, i.e., the inability to unequivocally quantify the relative importance of major hydrological and nutrient transport/transformation mechanisms, is one of the main benefits of the SWAT-ensemble strategy in the Maumee River watershed. By contrast, the errors arising from the mis-characterization of boundary conditions and forcing functions (e.g., fertilizer application rates, agricultural tile drainage network) among the ensemble members represent “nuisance” factors that inflate the uncertainty of the predictions drawn without necessarily advancing our

---

**Fig. 5.** Model performance for different physical, chemical, and biological components of Lake Erie. The relative error for each state variable is based on the median value of all the corresponding graphs published in the peer-reviewed literature, in which field data were compared against simulated values. Counter to the practices followed in Scavia et al. (2016a, 2016b), our goodness-of-fit assessment was based on point comparisons in time and space instead of aggregated (seasonal/annual, basin- or lake-wide) scales.
understanding of the watershed functioning or shedding light on the critical processes for achieving our environmental goals (Kim et al., 2019).

Bearing these two major sources of uncertainty in mind, Scavia et al. (2016c) showed an agreement among the models in identifying higher TP loading rates from the northwestern and southern parts of the Maumee River watershed, whereas a tendency for higher DRP export rates was mainly projected on the predominantly agricultural central area (Table 1 and Table S2). Considering that a significant amount of DRP (45–60%) can be exported to rivers through agricultural tile drainage (Kovacic et al., 2000; Xue et al., 1998), this disconnect between TP and DRP loading is likely associated with the tile-drainage assumptions made during the spatial configuration of the five models. In particular, Kim et al. (2019) showed that considerable variability exists among the individual estimates for surface runoff and tile flow (subsurface runoff), i.e., 191–275 mm and 50–139 mm, even though the total runoff (calculated as the sum of surface and tile flow, excluding groundwater) was fairly similar (301–334 mm) across the five SWAT models (see Fig. 3 in Kim et al., 2019). In the same context though, the differences in the spatial distribution of the tile-drainage among the models in conjunction with the fairly similar “hot-spot” mapping for a given phosphorus form (Table 1) may be an evidence of a retrofitting of the phosphorus-related parameters that counterbalances the differences in the simulated hydrology (see Table 2 in Kim et al., 2019). For example, the OSU model, based on a parameter specification that postulated low tile flow and high mineralization rates, predicted similar DRP loading patterns in the south-central Maumee River watershed with the TAMU model, in which high tile flow was combined with low P mineralization rates (Kim et al., 2019). Given the latter evidence along with the fact that the calibration of all the models was based on a single downstream station, it can be inferred that the capacity of the current SWAT ensemble to pinpoint nutrient-export hot-spots is uncertain and may simply be the result of multiple errors that cancel each other out and ultimately lead to the same output. It is thus critical to mitigate the confounding effects of the nuisance factors of the current watershed modelling framework by improving and/or sharing identical input data (e.g., fertilizer application rates, LULC data, and agricultural tile drainage network). In doing so, we will establish a common denominator across all of the SWAT applications upon which the implications of other critical sources of uncertainty will be examined.

As previously mentioned, notwithstanding the fact that the uncertainty envelope derived from an ensemble strategy is potentially inflated by factors not directly related to the watershed response to alternative BMP scenarios, there are still compelling arguments in favor of their use for guiding environmental policy decisions (Scavia et al., 2017). In the Maumee River watershed, ten land-use management scenarios were designed after considering issues related to their practical implementation and policy feasibility, i.e., commonly applied (fertilizer reduction, tillage replacement) versus less frequent management practices (land-use conversions, wetland/buffer restoration); the ability of SWAT to examine certain agricultural activities; and extensive consultation with agricultural and conservation stakeholders (Scavia et al., 2017; see also details in Table S3). Overall, little evidence was provided regarding the likelihood to achieve the March–July phosphorus loading targets of 186 metric tonnes of DRP and 860 metric tonnes of TP (GLWQA, 2016), or 40% reduction from the 2008 loads, across the different BMP scenarios examined (Fig. 6a,b). Interestingly, the attainability of the TP loading threshold seems to be more likely relative to the one for DRP loading. It is also worth noting that the forecasts associated with commonly applied BMP scenarios (S1–S4) were somewhat more conservative, in comparison with scenarios that are less frequently applied (S5 and S9) (Fig. 6a,b; see also discussion in Kim et al., 2019). Moreover, the degree of divergence of the individual model forecasts for the various BMP scenarios examined (or the forecasting spread) offered insights that can meaningfully influence the environmental policy analysis process. In particular, the forecasting spread increases significantly with the degree of deviation of BMP scenarios from the present conditions (Fig. 6c,d). The existing SWAT applications suggest that for every 50 metric tonnes of reduction achieved the standardized forecasting spread, or the deviation of the five models divided by their corresponding averaged prediction for a given scenario, increases by 1.5% and 13% for TP and DRP, respectively (see linear regression equations in Fig. 6c,d). To put it another way, the same scatterplots suggest that the standardized forecasting spread with the existing loading targets of 860 t for TP and 186 t for DRP is 26% and 33%, respectively.

The likelihood of a moderate reduction of DRP loading with nearly all the different management practices examined is consistent with recent empirical and modelling evidence from the Lake Erie watershed (Baker et al., 2017; King et al., 2017; Daloğlu et al., 2012). In particular, the increasing DRP loading trend after the mid-1990s has been attributed to the increased frequency of storm events, suboptimal fertilizer application rates and timing, and management practices that appear to increase phosphorus accumulation at the soil surface (Daloğlu et al., 2012). Counter to the (nearly) monotonic decline of the amount of nitrate in soils, which exhibited significant vertical mobility and was distinctly flushed out of the watershed after rainfall events, DRP appears to display a remarkable persistence across a wide range of spatial scales (King et al., 2017). The latter pattern highlights the role of legacy P as an important regulatory factor of the DRP concentration in soils. In particular, historical P fertilizer application rates seem to have led to soil reserves well-above critical levels in the area (Powers et al., 2016), which allowed to maintain high crop yields even after P fertilizer implementation declined (McDowell et al., 2016; Liu et al., 2015; Dodd et al., 2013). P build-up is hypothesized to have triggered a rapid, microbiologically mediated transformation of labile to bioavailable pools that are more susceptible to losses (King et al., 2017). If the latter mechanism holds true, then surface fertilizer applications could exacerbate soil P stratification. While this possibility casts doubt on the management practices designed to reduce top-soil P levels, the periodic soil inversion tillage to thoroughly mix the soil in the plow layer is recommended as the best strategy to avoid stratification increments (Baker et al., 2017). Moreover, alongside the buildup of legacy labile P fractions at the soil surface, recent research suggests the establishment of macropore flow pathways and increased tile drainage facilitate the hydrological connectivity and ultimately transport of soluble P among soil surface, subsurface drainage, and stream network (Jarvie et al., 2017; Sharpley et al., 2011). Given this emerging evidence, one challenging aspect for the evaluation of scenarios with the SWAT-ensemble is the proper consideration of legacy P (e.g., initialization that accommodates the spatial soil P variability, sufficient model spin-up period, parameter specification that reproduces the gradual P accumulation in the soils) and its ability to reproduce the critical hydrological and transformation mechanisms modulating the DRP loading in the Lake Erie basin.

4.2. Lake Erie multi-model ensemble strategy: Distinguishing between predictable patterns and sources of uncertainty with the load-response curves

In this section, we critically evaluate the credibility of the Scavia et al. (2016a, 2016b) forecasts on the achievability of the four ERIs, after forcing the Lake Erie multi-model ensemble with a series of nutrient loading reduction scenarios. To impartially conduct this analysis, we believe that these predictive statements cannot be viewed independently from the technical features of each of the modelling tools used, the characterization of the fundamental ecological processes modelled, and the ecological insights gained. In the same context, it is important to note that there is no consistent information from the individual members of the model ensemble regarding the quantification of all the fluxes pertaining to modelled biogeochemical cycles (carbon, nitrogen, phosphorus, oxygen), and therefore it is difficult to evaluate
the relative significance of critical ecological pathways (e.g., internal loading, importance of bacterial-mediated nutrient regeneration or nutrient excreta/egesta from zooplankton/dreissenids, seasonality of sedimentation fluxes) that can conceivably modulate the projected response of Lake Erie to exogenous nutrient loading reduction. In the next iteration of the modelling framework, we thus highlight the need to report for each ensemble member the resulting quantitative description of the simulated biogeochemical cycles (see Fig. 7 in Kim et al., 2013 or Fig. 4 in Gudimov et al., 2015), which not only will provide an essential piece of information to put the derived ecological forecasts into perspective, but will also allow to understand how distinctly different are the characterizations of the ecosystem functioning and consequently the degree of “diversity” of the modelling tools at hand. We should always bear in mind that the actual benefits of using multiple models do not stem from the complexity of the mathematics per se, but rather the ability to test alternative ecosystem conceptualizations or competing hypotheses regarding the role of certain facets of the underlying biogeochemistry.

4.2.1. Basin-specific overall phytoplankton biomass represented by summer average chlorophyll-a concentrations

Existing evidence from the international peer-reviewed literature suggests that the current generation of process-based models has satisfactory ability to reproduce the observed patterns of total phytoplankton biomass in a wide variety of aquatic systems and trophic conditions (Shimoda and Arhonditsis, 2016; Arhonditsis and Brett, 2004). Chlorophyll a as a bulk estimate of the total phytoplankton biomass across all the functional groups is a commonly used criterion to assess the trophic status of lakes. Lake Erie historically exhibits the highest lake-wide chlorophyll a concentrations with a distinct longitudinal west-east gradient (Cai and Reavie, 2018). Oligotrophic conditions prevail in the offshore waters of the central and eastern basins, rarely exceeding an average summer chlorophyll a concentration of 2.5 μg Chl a L⁻¹ over the last three decades (Dove and Chapra, 2015). On the contrary, seasonal average chlorophyll a levels in western basin often exceed the mesotrophic levels with the highest concentrations typically experienced during early fall (Zolfaghari and Duguay, 2016). Hence, the focus of the analysis of nutrient loading reduction scenarios and the development of load-response curves was on the western basin (Scavia et al., 2016a, 2016b). The models used for the latter exercise...
Fig. 7. (a) Predicted average summer chlorophyll a concentrations in the western basin of Lake Erie as a function of the corresponding annual TP load. Each response curve has been scaled to 100% at its maximum chlorophyll value to facilitate comparisons. The dashed line represents a 40% reduction from the 2008 (In developing loading scenarios, 2008 was chosen as a baseline year because its load was closest to the original 1978 Annex 3 target of 11,000 MT. Several loading scenarios, defined as different percentage reductions of the 2008 P load, were used to derive the load-response curves (Scavia et al., 2016)) annual load in the western basin. (b-d) Cyanobacteria bloom size as a function of spring Maumee River TP load as predicted from the NOAA Western Lake Erie HAB model (b), the U-M/GLERL Western Lake Erie HAB model (Because the three models considered different loading periods, to facilitate comparisons, spring loads from Maumee River are expressed in the response curves as average monthly load (Scavia et al., 2016b)) (c), and WLEEM (d). Solid lines are mean model predictions, dashed lines and shaded areas represent 95% predictive intervals, while the 70% predictive intervals are also included in (b). The horizontal line indicates the threshold for “severe” blooms, which equals 9600 MT for the first two models and was adjusted to an equivalent of 7830 MT for WLEEM. (e) August–September average hypolimnetic DO concentrations in the central basin as a function of the annual TP loading collectively added into the western and central basin of Lake Erie. The horizontal line represents the average concentration (4 mg DO L⁻¹) corresponding to initiation of statistically significant hypoxic areas. (f) Predicted Cladophora biomass, and associated 5th and 95th percentiles, by the coupled ELCOM-CAEDYM-EBC in the northern shoreline of the eastern basin as a function of whole lake annual TP loads. (All panels are modified with the publisher’s permission from Scavia et al., 2016b).
were very diverse, including an empirical model coupled with TPBM and three complex process-based models (EcoLE, WLEEM, and ELCOM-CAEDYM) (Scavia et al., 2016a, 2016b). The load-response curves produced from the four models displayed distinct differences in the predicted trajectories of the system. For example, Scavia et al. (2016b) noted that the annual TP loads into the western basin needed to bring about a 50% decrease in the maximum Chl a concentration ranged between 1130 MT and 3010 MT (Fig. 7a). The substantial uncertainty band raises the question to what extent it stems from the diversity of the models used and their conceptual/structural differences in terms of the suite of ecological drivers considered (i.e., the “desirable” variability that a model ensemble aims to accommodate) or from mischaracterizations of the system ecology and their individual biases that unnecessarily inflate the uncertainty of the model ensemble.

One of the members of the Lake Erie ensemble was based on a Chl a vs TP exponential regression model derived by data pooled from all of the Great Lakes (Dove and Chapra, 2015). As discussed in Appendix A, this empirical equation apparently underestimates the summer phytoplankton levels in the western Lake Erie basin, which likely stems from several lower-end (Lakes Huron and Michigan), and mid-range (central and eastern Lake Erie) observations that shape its intercept and slope, respectively (see Fig. 11 in Dove and Chapra, 2015). Building regression models from cross-sectional data is not an unusual practice in limnology, and this bias can be easily rectified by formulating a hierarchical structure that relaxes the assumption of globally common parameters and allows estimating system-specific regression coefficients (Cheng et al., 2010; Shimoda and Arhonditis, 2015). It is also important to consider other predictors that could conceivably influence the levels of total phytoplankton biomass, given that the existing regression model explained < 60% of the observed variability. We emphasize that it is essential to include and suitably update (at least) one data-driven model per ERI considered, as their empirical foundation offers a distinct alternative to the “growth-minus-loss” mass balance strategy characterizing the phytoplankton governing equation of process-based models.

The phytoplankton growth terms of EcoLE and ELCOM-CAEDYM bear conceptual resemblance in that they both explicitly consider the limitations posed from temperature, whereas all the other potentially limiting factors (e.g., light, nutrients) are included within the minimum formula of the Liebig’s Law and thus only one of those actively limits algal growth in a given time step. Regarding the plausibility of this strategy, Kim et al. (2014) argued that likely overstates the role of light availability as the predominant factor of the bottom-up forcing in the system, given that the low values assigned to half-saturation constants for phosphate uptake could be reducing the severity of phosphorus limitation in the summer. Counter to this assertion though, Zhang et al. (2016) showed that EcoLE simulations are suggestive of years (1997) when phosphorus is the sole limiting factor, even under the present (baseline) conditions, and other years (1998) when light and phosphorus interchangeably determine the degree of algal growth limitation during the summer period (see Fig. 11 in Zhang et al., 2016). According to forecasts drawn from EcoLE, it is also important to note that the initial benefits (i.e., decrease of total phytoplankton biomass) from the implemented nutrient loading reduction will be realized from the residents of the summer algal assemblage that possess inferior kinetics for P uptake (i.e., their non-diatom inedible algal group), whereas other species/superior P competitors (diatom-like or non-diatom edible algae) will display minimal decline (if not increase) because the severity of P limitation will be offset by the resulting improvement in the light environment (Zhang et al., 2016). The same pattern of a moderate response at the initial stages of loading reduction becomes more evident with the WLEEM load-response curve (see pages 14–15 in Scavia et al., 2016b), as the likelihood of a counterbalancing effect between P and light limitation carries more weight with the corresponding simulations, given that light availability is treated as an independent limiting factor (Verhamme et al., 2016). By contrast, the ELCOM-CAEDYM load-response curve is less steep than the rest of the models considered, even though the mathematics of the phytoplankton growth limitation term are fairly similar to EcoLE. One plausible explanation could be that P limitation with this model is based on a two-pronged process that first considers the nutrient uptake rate in relation to the ambient nutrients and subsequently the growth rate as a function of their internal nutrient storage. The latter strategy postulates that the intracellular P pool buffers the phytoplankton response to nutrient variability. Although we cannot unequivocally pinpoint which of the three representations of the same process (phytoplankton growth) is the “correct” one for Lake Erie, we believe that this is an excellent example of how multiple competing models can capture an influential source of uncertainty.

The characterization of the rest of the algal processes (respiration, excretion, senescence mortality, sedimentation rate) was fairly similar, and therefore their influence on the uncertainty of the load-response curves must have been negligible. One distinct structural difference among the mechanistic models was that only EcoLE explicitly considered the importance of zooplankton dynamics in Lake Erie. The limited consideration of the role of zooplankton is somewhat surprising, given the general attention to the mechanisms underlying the strength of trophic coupling in planktonic food webs and the local work on its diurnal and seasonal adaptive behavior to overcome the hypolimnetic hypoxia (Vanderploeg et al., 2009). Instead, the existing modelling work focused on the role of dreissenids as the primary factor modulating phytoplankton abundance and composition, which likely holds true in the nearshore zone (Boegman et al., 2008; Zhang et al., 2008). Thus, the improvement of zooplankton representation in the next iteration of Lake Erie modelling is important not only because it will allow to more effectively account for the likelihood of top-down control in offshore waters, but will also shed light on the broader implications of empirical evidence that certain residents (microzooplankton) of the zooplankton assemblage are more resilient than others (mesozooplankton) to toxic cyanobacterial blooms (Davis et al., 2012). In particular, an emerging hypothesis is that microzooplankton communities present a major nutrient source to fuel phytoplankton production during the period of summer stratification and low ambient nutrient compounds by heterotrophic zooplankton and microbes, which represent the major nutrient source to fuel phytoplankton production during the period of summer stratification and low ambient nutrient availability in the pelagic zone (Ramin et al., 2012b; Kamarainen et al., 2009; Teubner et al., 2003; Vanni et al., 2002). Earlier work from Goldman (1984) has described the intense microbially mediated regeneration as a rapidly turning “spinning wheel” by which nutrients are returned into the water column in short time scales (< 1 day) with minimal losses. There is abundant evidence from a wide range of morphologically and geographically diverse lakes that the excretion of inorganic phosphorus by zooplankton can potentially account for a significant fraction of the phytoplankton demands (Kowalewskiza-Madura et al., 2007; Conroy et al., 2005; Arhonditis et al., 2004; Gulati et al., 1995). EcoLE was the only member of the model ensemble that
explicitly considered the importance of this mechanism, suggesting that crustacean excretion may provide between 20 and 25% of the algal uptake (Zhang et al., 2016; Boegman et al., 2008). In the same context, dreissenid excreted/egested material could represent another major internal source of \( P \), and existing empirical and modelling evidence suggests that its contribution could be responsible for a significant fraction (15–30%) of the phytoplankton uptake in the western Lake Erie, especially when external loads are low (Zhang et al., 2016; Conroy et al., 2005; Arnott and Vanni, 1996; Mellina et al., 1995). Considering that the importance of this source will likely increase with the proliferation of dreissenid mussels in the western basin (Karateyev et al., 2014), increasing temperature and/or prolonged stratification (Johengen et al., 2013; Ramin et al., 2012b), it is critical to quantify the actual role of the corresponding internal nutrient fluxes to the overall \( P \) budget and revisit the existing models accordingly. Both ELCOM-CAEDYM and WLEEM have submodels that are specifically designed to accommodate the role of dreissenids.

Another important factor that could determine the response of Lake Erie to nutrient mitigation strategies is the likelihood of internal loading from the sediments. A recent study by Matisoff et al. (2016) attempted to shed light on this issue by obtaining estimates of the \( P \) diffusive fluxes from bottom sediments throughout the western basin of Lake Erie. A first striking finding was that the annual aerobic \( P \) flux for the entire a western basin is 378 MT P yr\(^{-1}\) with a 95% confidence interval of 359 and 665 MT P yr\(^{-1}\), which in turn could correspond to anything between 11 and 60% of the recommended target loads of 1130–3010 MT P yr\(^{-1}\) in order to achieve a 50% reduction in maximum chlorophyll \( a \) concentration reported by each model used (see Table 12 in Scavia et al., 2016b). In a similar manner, depending on the temperature correction factor used to calculate the aerobic \( P \) fluxes and under the assumption of complete mixing with an average water residence time of 50.7 days, Matisoff et al. (2016) estimated that the sediment contributes between 3.0 and 6.3 \( \mu \)gL\(^{-1}\) of dissolved \( P \) to the water column, which represents 20–42% of the \( UIC \) Target Concentration of 15 \( \mu \)gP L\(^{-1}\) for the western basin. It is also interesting to note that a recent study from the central basin reported \( P \) release from the sediments that could be up to 20% of the total external input of \( P \) to Lake Erie (Paytan et al., 2017). Moreover, Matisoff et al. (2016) found that the \( P \) fluxes under anaerobic conditions are (on average) 4 to 13 times larger than those under aerobic conditions. The greater \( P \) release is likely related to microbial reduction of \( Fe \) under anaerobic conditions, desorption of \( P \) into porewater, and subsequent diffusion into the overlying waters (Dittrich et al., 2013). While aerobic conditions at the sediment-water interface probably dominate redox chemistry in the western basin because of the shallow and actively mixed environment, the anaerobic fluxes may still be relevant since the shallow waters of the western basin can occasionally stratify for short time periods of up to 4–5 days and the bottom waters can become anaerobic.

It is thus critical to understand the intensive microbiological, geochemical, and physical processes occurring within the top few centimetres of the sediment and determine the fraction of organic matter and nutrients released into the overlying water. Diagenetic modelling is an indispensable tool to investigate the interplay among the sediment processes, to verify concepts, and to potentially predict system behaviors (Doan et al., 2018; Gudimov et al., 2016). This kind of diagenetic modelling as well as the data that necessitate to ground-truth those models (e.g., porewater analysis, phosphorus fractionation, organic matter profiles) are still missing in Lake Erie; especially from the central basin. Field, experimental, and modelling work should be designed to shed light on the mechanisms of phosphorus mobilization in the sediments and to identify process controls under a variety of conditions. Interestingly, although it cannot offer the mechanistic insights (e.g., primary and secondary redox reactions, mineral precipitation dissolution reactions, acid dissociation reactions, and \( P \) binding form reactions) that the new generation of sediment diagenetic models can, WLEEM has a potentially useful sediment mass-balance submodel that distinguishes between aerobic and anaerobic layers and uses linear partitioning to assign the total amount of a nutrient between dissolved and particulate fractions (Verhamme et al., 2016). A similar engineering-type approach is also incorporated to ELCOM-CAEDYM (Hipsey, 2008). Both submodels can provide insightful tools to dynamically track the response of the sediments under different nutrient loading regimes as long as several influential parameters (e.g., burial rates, particle mixing velocity among layers, resuspension rates, and aqueous mass transfer coefficients) are realistically constrained by empirical information from Lake Erie.

### 4.2.2. Cyanobacteria blooms in the western Lake Erie represented by the maximum 30-day average cyanobacteria biomass

For the purpose of establishing nutrient loading targets for this ERI, a 30-day average cyanobacteria biomass metric was selected as a proxy of the severity of harmful algal blooms in the western basin. In particular, a threshold of 9600 MT dry weight biomass has been selected to distinguish between “severe” and “mild” blooms (Stumpfl et al., 2012), based on existing records of satellite-estimated peak 30-day blooms since the early 2000s. Three models have been used to draw ecological forecasts and create load-response curves, i.e., \( UM/GLERL \) Western Basin \( HAB \) model, NOAA Western Basin \( HAB \) model, and WLEEM (Fig. 7b-d), but it should be noted that both ELCOM-CAEDYM and EcoLE consider multiple phytoplankton functional groups and could potentially be included in a similar exercise in the future. The synthesis of the predictions from two empirical equations and one mechanistic model has been somewhat problematic due to the different methods used to determine peak 30-day average cyanobacteria biomass. The former models are based on the satellite-derived estimates of maximum 30-day average bloom size calculated from consecutive 10-day composite images, which are in turn obtained by summing the highest biomass values observed at each pixel over each 10-day period (Stumpfl et al., 2012), whereas WLEEM derives a maximum 30-day moving average from basin-wide daily simulations of cyanobacteria biomass (Verhamme et al., 2016). To reconcile this mismatch, a WLEEM prediction of 7830 MT has been used as the equivalent to the satellite-derived bloom size of 9600 MT.

Notwithstanding their simplicity and small sample size used for their development, the consideration of two empirical models for the assessment of this particular ERI is critical, as evidence from the modeling literature provides little support to the ability of the current generation of mechanistic models to forecast structural shifts in the composition of phytoplankton assemblages and patterns of cyanobacteria dominance (Shimoda and Arhonditis, 2016; Anderson, 2005). In particular, a recent meta-analysis of 124 aquatic biogeochemical models found moderate fit statistics against empirical cyanobacteria biomass estimates, i.e., \( n = 70 \), median \( r^2 = 0.36 \), \( RE = 65\% \), \( MEF = 0.06 \) (Shimoda and Arhonditis, 2016). Of equal importance is the fact that there is considerable uncertainty with respect to the characterization of the ecophysiological traits of a “cyanobacteria-like” group, although there is a tendency to be defined as \( K \) strategists, displaying slower growth and metabolic rates that offer superior competitive skills in environments reaching their carrying capacity, with inferior \( P \) and superior \( N \) kinetics, adaptive capacity to tolerate turbid waters with low light availability, aptitude to sink slowly or even to regulate their vertical position within the water column in order to exploit favorable micro-environments, and limited preference (or even selective rejection) from zooplankton and dreissenids (Shimoda and Arhonditis, 2016). The mechanistic modelling work that is in place in Lake Erie does not deviate from this general delineation of cyanobacteria, but there is considerable variability with the actual values assigned to the individual ecophysiological parameters (Bocaniov et al., 2016; Verhamme et al., 2016; Zhang et al., 2016). The only notable discrepancy from these practices was WLEEM’s characterization of cyanobacteria as the fastest growing group of the simulated phytoplankton assemblage, which in turn may explain the steep slope (8.37
MT of cyanobacteria biomass per MT of spring TP load from the Maumee River) of the corresponding load-response curve (see Fig. 12 in Verhamme et al., 2016). Overall, the load-response curves of the three members of the model ensemble coalesced with respect to their predictions, suggesting that bloom sizes below the selected threshold can be achieved with cumulative Maumee March–July loads of 890–1150 MT or annual loads of 1679–2170 MT (see Table 2 in Scavia et al., 2016a). The questions arising though is do the models converge for the right reasons or better yet what is the degree of our confidence on these predictions?

If the skill assessment is one of the criteria to infer about the credibility of the existing HAB process-based modelling work, then our view differs from the practices followed in Lake Erie with respect to what constitutes an acceptable model fit or even an appropriate temporal scale to evaluate model performance. Namely, Scavia et al. (2016) reported a ~ 14% percent bias and 38% mean absolute relative error for WLEEM, when seasonally averaged cyanobacteria biovolume data were compared against the corresponding model outputs across four stations. Although this spatially and temporally aggregated assessment was intended to evaluate performance against the proposed metric for this ERI, abrupt and non-linear compositional shifts in phytoplankton assemblages are the very essence of HABs, and as such their study should be focused on finer time scales in order to understand the underlying mechanisms as well as their potential timing or actual magnitude. In particular, a careful inspection of the graphs included in Verhamme et al. (2016) lends little support to the ability of the model to predict cyanobacteria blooms (see their Fig. 8) and the reported goodness-of-fit statistics were clearly inflated from sites that did not exhibit significant cyanobacteria biovolume increase (e.g., GRI) during the study period. In fact, our independent assessment of the model fit with daily resolution was suggestive of a RE~100% and MEF < 0. Taken together with the substantial residual variability of the two empirical models, it can be inferred that the load-response curves for this ERI are surrounded by considerable predictive error and the relatively narrow uncertainty bounds are likely misleading.

Viewed from a heuristic perspective though, the existing cyanobacteria modelling work has collectively advanced our understanding of the factors triggering HAB formation. Considering that cyanobacteria dominance is largely the outcome of the resource competition among multiple phytoplankton species, one important lesson learned from both mechanistic and data-driven models was that both the dissolved reactive and particulate fractions of TP load must be taken into account when setting HAB-related load targets (Berti et al., 2016; Verhamme et al., 2016). Existing empirical estimates show significant variability of the bioavailable fraction of particulate phosphorus (20–45%) in the Maumee River, and several mechanisms (e.g., microbial mineralization, anoxic release from the sediments) could potentially determine the bioavailability of the inflowing material from the time of entry in early spring until the mid-summer initiation of *Microcystis* blooms (Stow et al., 2015; Ho and Michalak, 2015; Loewen et al., 2007). Another interesting finding from the modelling work in Lake Erie is that its susceptibility to HAB occurrence could be increasing (see also Appendix A), and this trend could be attributed to changing meteorological conditions, such as warmer temperatures and calmer summer conditions (Michalak et al., 2013). Nonetheless, the signature of climatic forcing is not always evident on the timing and magnitude of HAB occurrences (Michalak et al., 2013). Another fact that has received little attention is the importance of the inter-specific competition for various nitrogen forms; in particular, urea and ammonium are considered energetically favorable forms for protein synthesis and therefore predominant stimulants of *Microcystis* blooms (Chaffin and Bridgeman, 2014; Finlay et al., 2010). There is emerging evidence from other locations around the Great Lakes of a positive relationship between nitrogen concentration and toxin-producing *Microcystis* strains (Murphy et al., 2003) or microcystin production (Shimoda et al., 2016a; Kelly et al., 2019; Orr and Jones, 1998). Research also pinpoints to iron availability as another potential factor in triggering HAB events (Molot et al., 2014), but practically no work has been done to evaluate this hypothesis in Lake Erie.

Another major unknown is the potential role of intense precipitation events in initiating HAB events in Lake Erie. Our evolving understanding of their role in lakes identifies two classes of effects of weather events on abiotic conditions: short-lived effects of storms on lake thermal structure, and more prolonged effects of precipitation events on nutrient levels and water clarity (Drösch et al., 2009). The abrupt abiotic changes associated with extreme events can subsequently trigger changes in biotic ecosystem components (e.g., primary productivity, composition of plankton assemblages), but the magnitude of these shifts is predominantly modulated by the lake trophic status as well as fundamental system attributes (e.g., fetch, depth, water levels, wind regimes) (Shimoda et al., 2011). Consistent with these working hypotheses, Zhang et al. (2016) argued that even though the reduction in external phosphorus would result in distinct decline in algal biomass and *Microcystis* blooms in the western basin, several phosphorus input pulses stemming from storm events could induce favorable conditions for net *Microcystis* growth and ultimately dominance. Specifically, storm runoff events can bring into the system a substantial amount of nutrients but also elevate the turbidity, either through riverine input or sediment resuspension, which creates an environment that renders competitive advantage to *Microcystis* over the rest residents of the summer phytoplankton assemblage (Harke et al., 2016; Below and Giles, 1997). Along the same line of evidence, Stumpf et al. (2016) asserted that the inflows from Maumee River not only profoundly influence the abiotic environment, but also combined with the prevailing circulation patterns can determine the exact area in the western basin where HABs may initiate (Schwab et al., 2009). Thus, one of the important augmentations of the existing modelling framework is the establishment of direct linkages between watershed processes (as characterized by the SWAT-ensemble) and the receiving waterbody, in order to advance our understanding on the capacity of the perturbations induced by extreme runoff events to shape the complex interplay among physical, chemical, and biological components in western Lake Erie.

### 4.2.3. Central basin hypoxia represented by number of hypoxic days; average extent of hypoxic area during summer; and average hypolimnion DO concentration during August and September

Being different manifestations of hypoxia in the central basin of Lake Erie, the three metrics were intended to offer a comprehensive assessment of the magnitude of the problem in order to capture its broader ramifications on ecosystem integrity. Specifically, the threshold for the average August–September hypolimnetic minimum DO concentration was originally set at 2.0 mg L⁻¹, but a level of 4.0 mg L⁻¹ was ultimately selected, as it offers a more sensitive proxy to discern the onset of hypoxic areas (Zhou et al., 2013). The targeted levels for the average extent of hypoxic area during summer and number of hypoxic days were subsequently set to 2000 km² and 30 days, respectively. Counter to the practices followed with the previous two ERIs, the annual total phosphorus loads from tributaries to both western and central basins were identified as a better predictor for the hypoxia problem. Three models were used to generate the load-response curves with differences in the spatial resolution of the predictive statements drawn, i.e., 1D-CBH model provided horizontally-averaged DO values, whereas *ELCOM-CAEDYM* and *EcoLE* horizontally-resolved DO concentrations in the bottom layer (0.5–1.0 m for *ELCOM-CAEDYM*; 1.0 and 1–3 m for *EcoLE*). In addition, 1D-CBH model was forced with two distinct specifications of the load transport from the western to the central basin using: (i) the default net apparent attenuation loss rate due to settling;
and (ii) the daily outputs from WLEEM to account for the mass fluxes crossing the western-central basin boundary. Regarding the achievability of the 4.0 mg DO L⁻¹ threshold, the load-response curves provided a fairly wide uncertainty range, 2600–5100 MT, within which this target can be realized (Fig. 7e). Similarly, a load reduction anywhere between 3415 and 5955 MT was projected to reduce the average hypoxic extent to 2000 km² and the number of hypoxic days between 9 and 42 days (Scavia et al., 2016a, 2016b).

The response curves for August–September average hypolimnetic DO concentration displayed similar increasing trends with decreasing loads, while some discrepancies arose at lower nutrient loading conditions; especially when the threshold DO level of 4 mg L⁻¹ was reached (Fig. 7e). Although the latter uncertainty was partly attributed to the differences of the SOD formulations (Scavia et al., 2016a), the reality is that there were no significant conceptual differences among the three models used. The discrepancies were mainly introduced by the specific assumptions made to parameterize simple approximations of the potential response of the sediments to the reduced nutrient loading. In particular, 1D-CBH used an empirical equation to connect TP loading directly to SOD based on a cross-sectional dataset of 34 estuarine and coastal systems (see Appendix A), which was then used to predict the response of Lake Erie to loading reductions under the assumption that the large-scale (cross-sectional) patterns described in the model are also representative of the dynamics of individual systems (Shimoda and Arhonditsis, 2015). In doing so, we essentially assume that all the systems in the dataset have identical behavior and therefore the empirical relationship is the same among and within systems. Consequently, after fitted to the cross-sectional dataset, Rucinski et al.’s (2014) equation provides a SOD value equal to 0.50 g O₂ m⁻² d⁻¹ when the annual TP loads are 4000 MT year⁻¹, which in turn is predicted to bring ambient DO close to 4 mg L⁻¹. EcoLE expressed SOD as a function of temperature and oxygen concentration (the latter relationship was mathematically described by a rectangular hyperbola), with an oxygen half-saturation constant for SOD set equal to 1.4 mg O₂ L⁻¹ and a maximum SOD rate at 20 °C (0.22 g O₂ m⁻² d⁻¹) that was consistent with the range reported by Smith and Matisoff (2008; see their Table 1). Because of the values assigned to the two parameters, the oxygen losses to the sediments were predicted to be much lower (=0.16 g O₂ m⁻² d⁻¹) than 1D-CBH when DO exceeds the level of 4 mg O₂ L⁻¹, which led to a disproportional DO increase when the external loads are low, i.e., <2000 MT year⁻¹ (see Fig. 7e). Simply put, 1D-CBH and EcoLE estimate distinctly different SOD rates (0.50 vs 0.16 g O₂ m⁻² d⁻¹) under the same level of nutrient loading. Similar to the chlorophyll a predictions, ELCOM-CAEDYM resulted in a load-response curve that is less steep than the rest of the members of the model ensemble, which likely stems from the fact that an even higher maximum SOD rate at 20 °C (1.2 g O₂ m⁻² d⁻¹) was assumed for this exercise (Bocaniov et al., 2016). While the same model postulates that the maximum SOD rate decreases with decreasing TP loads, the estimated SOD fluxes are much higher than any other of the models used (see Table S4 in Bocaniov et al., 2016). Based on the SOD rates reported by Smith and Matisoff (2008), the values used by EcoLE appear to be closer to the recent trends in Lake Erie, but the likelihood of higher oxygen demand rates cannot be ruled out given the considerable spatial and temporal variability characterizing the system (Paytan et al., 2017; Matisoff and Neeson, 2005; Schloesser et al., 2005).

As previously mentioned, to effectively augment the diversity of models included in the Lake Erie ensemble, there is a need for a sediment diagenesis model that quantifies SOD as the sum of the organic matter mineralization processes and the oxidation of reduced substances within different sediment layers (Gudimov et al., 2016). Only the characterization of the redox-controlled processes and the determination of the vertical profiles of biodegradable organic matter can allow a reliable estimation of the mineralization half-life period and subsequently the response rates of the sediments following different management actions. It is important to recognize that none of the existing members of the model ensemble has the mechanistic foundation to account for the likelihood of a lagged sediment response and unexpected feedback loops (Smith and Matisoff, 2008). Along the same line of evidence, the meteorological forcing and its effects on the thermal structure (e.g., timing and duration of stratification, thickness of hypolimnion) along with the water level fluctuations in Lake Erie are two important confounding factors that could further broaden the uncertainty band around the load-response curves predicted by the multi-model ensemble (Watson et al., 2016). Both theoretical and empirical evidence suggests that the year-to-year variability related to local weather conditions and water levels can significantly impact the severity of hypoxia and degree of recovery (Rucinski et al., 2016; Watson et al., 2016; Liu et al., 2014). Even an increased frequency of multi-day storm events during winter and spring could result in increased runoff from agricultural watersheds, such as the Maumee River Basin, and ultimately exacerbate hypoxia (Cousino et al., 2015). Another emerging issue involves the likelihood of significant primary productivity during the ice cover period with possible strong linkages to DO availability in the system. Relatively high under-ice phytoplankton biomass has been recorded in the central and eastern basins, comparable to or higher than values typically observed in Lake Erie in the summer (Reavie et al., 2016; Oveisy et al., 2014). This elevated winter-spring production in the central basin of Lake Erie may become available for export to other trophic levels, including bacterial decomposition at the lake bottom later in the year, when the increase in water temperature could trigger the development of hypoxia (Oveisy et al., 2014; Twiss et al., 2012). Thus, there are views in the literature asserting that the easily biodegradable- biogenic material produced from winter-early spring algal blooms represents a considerable quantity that needs to be an integral part of any future modelling exercise that will revise the load-response curves.

4.2.4. Eastern basin Cladophora represented by dry weight biomass and stored P content

*Cladophora glomerata* is a filamentous green alga that has proliferated in the rocky nearshore zone of eastern Lake Erie since the mid-1990s, and has been responsible for the extensive fouling of the local beaches by decaying organic material (Higgins et al., 2008). Because of the meso-oligotrophic status of the offshore waters in the eastern basin, the presence of widespread *Cladophora* blooms in the nearshore area was attributed to the major reengineering of the biophysical littoral environment brought about by the invasion of dreissenids, including the profound alterations on the retention and recycling of nutrients (Tomlinson et al., 2010). Given the way this water quality problem manifests itself, the extent of beach fouling by sloughed material would have been the most suitable metric for this ERI. Nonetheless, the absence of a model (or an acceptable monitoring program) to evaluate the potential improvement against external nutrient loading reductions with such a metric, led to the pragmatic selection of algal dry weight biomass and stored phosphorus content as two proxies to track progress in the northeastern shoreline of Lake Erie (Scavia et al., 2016b). The achievability of a threshold value of 30 g dry weight biomass m⁻² was assessed by Valipour et al. (2016) using the high-resolution, three-dimensional ELCOM-CAEDYM coupled with the EBC model. Load-response curves were generated in an area centered on the Grand River and covering 40 km of the northern shoreline area out to the 15 m depth contour, and the predictions drawn suggested that P load reductions will bring about minor decline in the *Cladophora* biomass in the eastern basin (Fig. 7f). The predicted responses were influenced by the light availability, and the predicted changes were thus more noticeable at the non-light-limited shallow depths and less so at deeper depths. In fact, a small *Cladophora* biomass increase was predicted at deeper depths as a result of the improved water transparency induced by the most extreme loading reduction scenarios (Valipour et al., 2016).

Consistent with Scavia et al.’s (2016a) interpretation, we believe that there are three compelling reasons why the predicted load-
response curves should be viewed with extra caution. First, the domain within which *Cladophora* growth could be regulated by SRP concentrations is extremely low (0.2–1.0 μg P L$^{-1}$; see Tomlinson et al., 2010), while year-to-year variability even on the order of 1 μg P L$^{-1}$ could result in variations of depth-integrated biomass by a factor of 3.5 (Higgins et al., 2005). Second, except from the supply by dreissenid excreta (Ozersky et al., 2009), the SRP nearshore concentrations are also modulated by the inflows from the Grand River, as well as the nearshore-offshore exchanges. In particular, frequent upwelling events driven by favorable winds of 5–10 days period can easily increase P supply above saturation levels (Valipour et al., 2016). Third, although plausible explanations on the factors that accelerate the sloughing rates and their development within the *Cladophora* mats do exist, the mathematical representation of the associated processes is far from adequate. Sloughing rate is treated as a first-order loss process varying as a function of water temperature (physiological effect) and the depth of colonization (wind energy, benthic shear stress effect), but several important conceptual advancements are still overlooked (Higgins et al., 2008). For example, because of the absence of intercellular organelles or plasmodesmata, *Cladophora* filaments display limited transport of nutrient and other metabolic compounds from the surface of the algal mats, and thus cellular deterioration at the base of the *Cladophora* mats can occur while cells toward the surface of the mat still actively grow (Higgins et al., 2006). Recognizing that nuisance *Cladophora* growth are controlled by both local- and/or basin-level factors, our ability to examine the potential impact of large-scale watershed management actions and draw inference with the granularity required to understand nearshore, small-scale processes is confounded by considerable uncertainty. It is unrealistic to expect mechanistic models that have been calibrated to match average conditions in the offshore waters (along with all their inherent structural and parametric uncertainties) to support predictions in the nearshore zone with accuracy < 1 μg P L$^{-1}$. There is no evidence in the international literature that the current generation of models can go anywhere close to the required accuracy and spatiotemporal resolution required to address the *Cladophora* problem.

We thus emphasize that the local modelling efforts in Lake Erie will greatly benefit from a high-resolution monitoring of the nearshore zone to provide critical information for the existing model ensemble (see Table 1 in Arhonditsis et al., 2019). Rather than increasing the complexity of (already) over-parameterized models, the management efforts will be better supported by the development of two empirical models offering causal linkages between the abiotic conditions (e.g., SRP, light, temperature) in the surrounding environment and the internal P content and sloughing rates in *Cladophora* mats. Similar to Tomlinson et al. (2010)’s polynomial functions used to incorporate the role of temperature and light into EBC model, after digitizing the data originally reported by Graham et al. (1982), our proposition is to build an empirical modelling framework that will connect individual snapshots of the ambient conditions in the nearshore zone with the contemporaneous internal P levels at different depths during the growing season as well as the net production rates at the base of *Cladophora* mats. These two models could be used independently or in conjunction with the existing coupled ELCOM-CAEDYM-EBC, thereby introducing an extra layer of causality that connects the empirical characterization of microscale processes with coarser scale predictions of mathematical models (Shimoda et al., 2016b).

4.2.5. Take-home messages regarding our capacity to forecast the achievability of the four ERIs

In our attempt to distinguish between predictable patterns and sources of uncertainty with the load-response curves developed for the four ERIs, there are several lessons learned and important issues for future consideration: (i) the diversity of the existing modelling work in Lake Erie as well as the general evidence from the international literature suggest that the forecasting exercise related to the overall summer phytoplankton biomass in the western basin has a lot of potential to meaningfully assist the local management efforts. Our analysis highlighted the need to establish a more reliable empirical model (Chl a versus TP and other potential explanatory variables) and also improve our understanding (and subsequently their representation with the existing mechanistic models) of certain facets of phytoplankton ecology, such as the postulated degree of reliance of phytoplankton growth upon internal nutrient sources (e.g., microbially mediated regeneration, dreissenid or zooplankton excreted material in nearshore and offshore waters, respectively), the internal P loading from the sediments, and phytoplankton-zooplankton interactions. (ii) Considering the challenges with the modelling of individual phytoplankton functional groups, the forecasting exercise regarding the chAB likelihood of occurrence under different loading regimes is as robust as it can be realistically expected. The coupling of empirical and process-based models offers a healthy foundation to evaluate competing hypotheses and advance our knowledge on the suite of factors that may trigger cyanobacteria dominance in Lake Erie. We just caution though that the reported range of cumulative Maumee March–July loads of 890–1150 MT for achieving the chAB target is likely narrow and does not fully reflect the actual uncertainty with this ERI. (iii) Our study casts doubt on the ability of the existing models to support reliable predictions regarding the likelihood to alleviate the hypoxia in the central basin of Lake Erie, given that our mechanistic understanding of sediment diagenesis, i.e., the characterization of organic matter mineralization and redox-controlled processes within different sediment layers, is still inadequate. There is a rich research agenda that should be in place with the next iteration of the adaptive management cycle, before we are in a position to predict the degree and timing of the sediment response or the likelihood of unexpected feedback loops that could delay the realization of the anticipated outcomes. (iv) The modelling of *Cladophora* in the eastern basin has been insightful but carries little predictive value. The proposed high-resolution monitoring of the nearshore zone and subsequent establishment of causal linkages between abiotic conditions (e.g., SRP, light, temperature) in the surrounding environment and the internal P content and sloughing rates in *Cladophora* mats are two essential steps to further augment the existing modelling work.

5. Conclusions

In Lake Erie, a wide variety of both statistical and mechanistic models have been developed to evaluate the relationships among watershed physiography, land-use patterns, and phosphorus loading, and to predict the response of the lake to different management actions. Recognizing that these models have different conceptual and methodological strengths, one of the priorities of the local research efforts must be to maintain this diversity and further consolidate the ensemble character of the local forecasting tools. The key findings of our analysis are as follows:

- The watershed modelling work has been based on five independent applications of SWAT; a complex process-based model. Each application reflects different assumptions and process characterizations in order to quantify the relative importance of the mechanisms that determine phosphorus loading export from the Maumee River. Thus, this strategy—in principle—captures some of the uncertainties in our understanding of watershed attributes and functioning, and can be classified as a model ensemble.
- The five SWAT models showed nearly excellent goodness-of-fit against monthly flow rates and phosphorus loading empirical estimates based on a single downstream station, but little evidence was provided with respect to their ability to reproduce the measured phosphorus concentrations. The SWAT modelling work in the Maumee River watershed placed little emphasis on evaluating the robustness of the hydrological or nutrient loading predictions with a finer (daily) temporal resolution, and even less so in capturing the
impact of episodic/extreme precipitation events.

- In an attempt to delineate high-risk areas (or “hot-spots”) with greater propensity for nutrient export and downstream delivery rates, the SWAT applications coalesced in their projections and identified higher TP loading rates from the southwestern and southern parts of the Maumee River watershed, as well as a tendency for higher DRP export rates at the predominantly agricultural central areas. However, we caution that these projections require further ground-truthing by considering multiple sites across the Maumee River watershed to recalibrate the models, and by addressing some of their fundamental discrepancies regarding the fertilizer/manure application rates in the croplands or the spatial drainage of soils.

- Analysis of scenarios with commonly applied (fertilizer reduction, tillage replacement) and less frequent (wetland/buffer restoration) BMPs did not provide strong evidence regarding the likelihood to achieve the March–July phosphorus loading targets of 186 metric tonnes of DRP and 860 metric tonnes of TP, or 40% reduction from the 2008 loads. The forecasting spread also increases significantly with the degree of deviation of BMP scenarios from the present conditions. The existing SWAT applications suggest that the standardized forecasting spread with the existing loading targets of 860 t for TP and 186 t for DRP is 26% and 33%, respectively.

- The multi-model ensemble for the Lake Erie itself has been based on a wide range of data-driven and process-based models that span the entire complexity spectrum. Specifically, the models included in the multi-model ensemble strategy were two empirical models and six process-based models. Consistent with the general trend in the international modelling literature, the performance of the aquatic ecological models in Lake Erie declined from physical, chemical to biological variables. Temperature and dissolved oxygen variability were successfully reproduced, but less so the ambient nutrient levels. Model performance for cyanobacteria was inferior relative to chlorophyll a concentrations and zooplankton abundance.

- Our skill assessment analysis is based on a finer resolution in time and space, as opposed to the aggregated spatiotemporal (basin- or lake-wide, seasonal/annual time) scales adopted from the local modelling community, and therefore our goodness-of-fit statistics are less favorable. In principle, the selected coarse scales for evaluating model performance are defensible, as they are consistent with those used for the established nutrient loading targets and water quality indicators in Lake Erie. However, there are compelling technical and management reasons why this practice is problematic and it is recommended to be revisited during the next iteration of the modelling framework.

- General evidence from the international literature suggests that the prediction of the annual TP load reduction needed (1130–3010 MT) to decrease in half the summer average chlorophyll a concentration in the western basin is certainly in the right direction. Likewise, the coupling of empirical and process-based models offers a healthy foundation to evaluate competing hypotheses and support forecasts regarding the achievability of the target related to the maximum 30-day average cyanobacteria biomass in the same basin. To further reduce the predictive uncertainty, it is important to improve our understanding (and subsequently the representation with the existing process-based models) of several facets of phytoplankton ecology, such as the postulated degree of reliance of phytoplankton growth upon internal nutrient sources (e.g., microbially mediated regeneration, dreissenid or zooplankton excreted matter in nearshore and offshore waters, respectively), the internal P loading from the sediments, the role of nitrogen, and the trophic interactions with zooplankton.

- On a final note, our technical analysis casts doubt on the optimistic predictions of the extent and duration of hypoxia, given our limited knowledge of the sediment diagenesis processes in the central basin and the lack of data related to the vertical profiles of organic matter and phosphorus fractionation or sedimentation/burial rates. There are several research priorities that should be considered with the next iteration of the adaptive management cycle, before we are in a position to predict the degree and timing of the sediment response or the likelihood of unexpected feedback loops that could delay the realization of the anticipated outcomes. The same is true about our ability to forecast the control of Cladophora growth in the eastern basin by P load reductions. The management efforts will greatly benefit from a high-resolution monitoring of the nearshore zone to provide critical information regarding the causal linkages between the abiotic conditions (e.g., SRP, light, temperature) in the surrounding environment and the internal P content and sloughing rates in Cladophora mats. In our companion paper, Arhonditsis et al. (2019) identify knowledge gaps and monitoring assessment priorities to ensure that resource parameters are adequately measured and suitably focused on relevant performance indicators.

**Acknowledgements**

This project has received funding support from the International Joint Commission (Requisition Number 1042700743). We are grateful to Debbie Lee, Gavin Christie, Wang Lizhu, Craig Stow, Yingming Zhao, Erich B. Emery, Joe DePinto, and Michael Twiss for their constructive feedback on an earlier version of the manuscript.

**Appendix A. Description of the Multi-model Ensemble Strategy for Lake Erie**

**U-M/GLERL Western Lake Erie HAB model:** This is a regression model originally developed by Obenour et al. (2014) to predict peak HAB size in western Lake Erie as a function of spring TP loading from the Maumee River. An interesting feature of this empirical construct is the consideration of a temporal trend component aiming to capture the variability over time in the lake’s susceptibility to HAB formation. The optimal TP loading period (i.e., January–June) for predicting bloom size was assessed probabilistically using a weighting parameter that represents the temporal threshold, prior to which loading does not contribute to the late-summer algal bloom. Bertani et al. (2016) further refined the model by introducing an expression to estimate probabilistically the fraction of the particulate P load that becomes bioavailable. The model was jointly calibrated against three sets of bloom observations: (i) MERIS satellite remote sensing imagery (Stumpf et al., 2012); (ii) in situ measurements from a plankton sampling program (Bridgeman et al., 2013); and (iii) SeaWIFS satellite remote sensing imagery (Shuchman et al., 2006). A Bayesian hierarchical configuration enabled the characterization of the observation error within each dataset, as well as the realistic quantification of total predictive uncertainty of the derived annual bloom size. In its most recent calibration, the model accommodated the impact of dreissenids on P recycling by increasing the bioavailability of particulate phosphorus (PP) to algae (i.e., a prior probability was assigned to the relevant parameter postulating that the readily bioavailable PP discharged from tributaries should be at least 20%). The model explained over 91% of the year-to-year variability in bloom observations over the course of seventeen years (1998–2014), and the cross-validation performance remained relatively high, 80.6% (Scavia et al., 2016b).

While these results are certainly promising, it is important to note that the sample size (or time span covered) for the model training phase is still small, and thus each annual observation (i.e., pair of spring loading and bloom size values) carries significant weight on the form of the resulting equation as well as on the predictions drawn. In particular, a leave-one-year-out cross-validation exercise (i.e., removal of observations from one year, re-calibration of the model against the remaining data, and then forecasting of the excluded observations) revealed wide ranges for all the posterior estimates of the model.
parameters (see Table B2–4 in Scavia et al., 2016b). Along the same line of evidence, Bertani et al. (2016) noted that the additional years used for the calibration contain the markedly higher bloom sizes in 2013 and 2014 than in previous years, suggesting that the lake may becoming more susceptible to cyanobacteria blooms over time, with smaller loads apparently triggering larger blooms in most recent years. The latter finding was on par with Obenour et al.'s (2014) positive posterior estimate of the temporal trend term, which reinforces the point of an increased propensity for bloom formation in Lake Erie over time. A characteristic example of the latter finding is that the model predicts a bloom size of 15,700 MT (95% predictive interval: 5000–23,700) associated with a spring weighted TP load of 204 MT month$^{-1}$ for the 2013 conditions, while the same bloom size prediction is provided by a twice as high weighted TP load in 2008 (Scavia et al., 2016b).

**NOAA Western Lake Erie HAB model:** This is another regression model that attempts to elucidate the primary factors modulating the severity of summer cyanobacterial blooms in western Lake Erie, which in turn was first quantified by the Medium Resolution Imaging Spectrometer (MERIS) satellite imagery data. Using as a surrogate of total cyanobacteria biomass, the “cyanobacteria index” (CI) (Wynne et al., 2008, 2010), the original models presented by Stumpf et al. (2012) were based on a 10-year dataset (2002 – 2011). With the loss of MERIS data in April 2012, the Moderate Resolution Imaging Spectroradiometer (MODIS) data were employed for the subsequent four years (Stumpf et al., 2016), but it was noted that MERIS produces a more sensitive proxy (nominal uncertainty 10%) with less noise than MODIS (nominal uncertainty 25%). The CI is determined in 10-day composites by taking the highest cyanobacterial chlorophyll-related index at each pixel available from any of the daily images within a given 10-day time period, and thus remove interference and biasing from clouds. The annual CI used to define the bloom severity is the average of the three 10-day periods around the maximum severity of the bloom, so it is effectively a 30-day average which is then converted to biomass using a ratio of 4800 MT per CI (Scavia et al., 2016b).

According to Stumpf et al. (2016), the Maumee River discharge and total bioavailable phosphorus (TBP) loading from March through July, with July excluded only when June water temperatures were below the optimal temperature (20 °C), were the best predictors for Microcystis growth. Drawing similarities with Bertani et al. (2016), TBP was specified as the sum of the dissolved reactive phosphorus and the proportion of particulate phosphorus that is bioavailable, corrected for loss due to settling in the river. For the purposes of the multi-model ensemble forecast exercise, Scavia et al. (2016b) developed a variation of the previous models using as predictor the TP loading from March to July for all years, except 2003 and 2008, when the July temperature average was lower than 18 °C. The latter model explained 62% on the observed variability of the annual cyanobacteria bloom CI data, but the standard error of the slope was 24% of the mean slope estimate, and the uncertainty factor was approximately equal to 1.95 (or simply put, the CI predictions derived by the model are correct within a factor of 2). Similar to the U/M/GLERL Western Lake Erie HAB model, the “leave-one-out” test was indicative of a substantial sensitivity of the model intercept and slope to the individual years included in the calibration dataset; especially when the CI-TP loading pair for 2012 is considered (see Table B1–1 in Scavia et al., 2016b). Considering that this is a simple exponential model (e.g., without the ability to accommodate the long-term trends in the susceptibility of Lake Erie), the analysis of TP loading scenarios provided very optimistic forecasts, suggesting that a 40% load reduction will significantly reduce the likelihood of severe cyanobacteria blooms, even for years when river discharge will be excessively high (Scavia et al., 2016b). To put these forecasts into perspective, Stumpf et al. (2016) cautioned that the model in its current exponential form may not be able to capture the biomass-loading relationship during extreme years, like the 2015 bloom, and instead a logistic (sigmoidal-type) function could be a more suitable model. Notwithstanding the plausible ecological arguments to adopt an equation that postulates a “saturation-type” pattern (or a smaller increase rate of the bloom size) when excessively TP loading are experienced in Lake Erie (Stumpf et al., 2016), an alternative explanation could be that the capacity of satellite images to characterize the algal bloom severity has an upper bound, as the satellite data cannot quantify the additional biomass production once scum completely covers the entire area of water observed in each pixel.

**Total Phosphorus Mass Balance Model:** This is a simple phosphorus budget model originally used to establish phosphorus loading targets for the 1978 Great Lakes Water Quality Agreement (Chapra and Sonzogni, 1979; Chapra, 1977; Chapra and Robertson, 1977). The governing model equation predicts offshore TP concentrations, as a function of external loading, inter-segment hydrodynamic exchanges, and net sedimentation losses. Evolved from the original spatial segmentation, the model currently divides Lake Erie into three completely mixed compartments (western, central, and eastern basins), while the within-lake intersegment flow exchanges were derived by annual water balance estimates (tributary flows, lake-level variability, and over-lake precipitation minus evaporation). Alongside the advective transport, the model considers a bulk mixing coefficient which is a phenomenological parameter aiming to capture large-scale diffusive exchanges across open boundaries, such as large-scale eddy diffusion, and dispersion due to shear flow and spatial non-uniformities. The latter feature was parameterized with a chloride budget model (see Table 3 in Chapra et al., 2016), and the TPMB model was subsequently calibrated against median offshore TP concentrations by simply adjusting the segment-specific, apparent settling velocities. Thus, the model offers a parsimonious construct to obtain first-order approximations of in-lake processes that collectively modulate the ambient TP variability. The goodness-of-fit statistics were suggestive of satisfactory TPMB performance with a root mean square error between 3.6 and 5.4 μg L$^{-1}$ and a percent relative error (RE) between 26 and 29% (study period 1970–2010). The discrepancy between modelled and measured TP concentrations was attributed to the predominant role of the basin-specific tributary loadings, but other factors, such as errors stemming from the intersegment advective/diffusive mass exchanges or the segment-specific apparent settling velocities, were not ruled out (Chapra et al., 2016).

According to Occam’s razor (or principle of parsimony), models should be as simple as possible, but not simpler to the extent that we fail to consider important facets of the system modelled (Paudel and JAWITZ, 2012). The TPMB application in Lake Erie is no exception, as the fitting exercise pinpointed the need to assign higher apparent settling velocities in order to reproduce the higher P retention during the post-1990 period. A suite of mechanisms were suggested to provide the ecological underpinning of such structural augmentation, such as the dreissenid filtration that enhanced transport to the sediments; the proliferation of soft sediment settlers, i.e., quagga mussels (*Dreissena bugensis*), that may contribute to permanent P trapping in the sediments, structural shifts in algal community toward fast sinking diatoms due to low P availability, nearshore organic matter that could be getting sequestered in offshore bottom sediments, and gradual increase of bioavailable phosphorus in the exogenous loading (Chapra et al., 2016). Following the TPMB evolution over the course of the past thirty years, there are several illustrations on how additional complexity can be accommodated by the present structure, while maintaining its parsimonious character (Gudimov et al., 2012; Katsev, 2017; LeSht et al., 1991; Shimoda and Arhonditis, 2015; Yaksich et al., 1985). On a final note, the predicted TP concentrations for each lake segment were causally linked with two trophic indicator variables, summer chlorophyll $a$ and Secchi disk depth, based on the empirical relationships described in Dove and Chapra (2015); see their Figs. 11 and 12. It is important to recognize though that (i) the two equations were derived by pooled data from all of the Great Lakes, (ii) the corresponding regression models explained between 55 and 60% of the observed variability, and (iii) the reported offshore summer chlorophyll $a$ versus spring TP relationship distinctly
underestimates the summer phytoplankton levels in the western Lake Erie basin.

**1-Dimensional Central Basin Hypoxia Model:** The 1D-CBH model comprises a simple eutrophication model coupled with a 1-D thermal model that provides temperature and associated vertical mixing profiles in the offshore waters of Lake Erie central basin (Rucinski et al., 2010, 2014). The 1-D thermal model is based on the Princeton Ocean Model, which uses a Mellor-Yamada turbulence closure scheme to parameterize vertical mixing (Mellor and Yamada, 1982) and has been modified with an overland-overlake correction (Beletsky and Schwab, 2001). The eutrophication model explicitly simulates organic carbon, phytoplankton, zooplankton, and dissolved oxygen, while phosphorus is divided into two pools, i.e., available and unavailable (or particulate) P. The model is forced with carbon and phosphorus loading from both western and central basins, with the former loads routed to the central basin after accounting for a constant net apparent settling loss of 10 m·yr⁻¹ (Lesht et al., 1991). Both hydrodynamic and eutrophication models operate with the same vertical resolution, i.e., 48 vertical layers (each 0.5 m thick). The model has been tested against 19 years (1987–2005) of observed loading rates and meteorological conditions, with the focal question being the relative contribution of thermal stratification conditions versus P loading magnitude and timing on the severity of hypoxia in the central Lake Erie Basin. Because the model does not include ice-formation processes, the hydrodynamic and eutrophication models were initialized each year, using the earliest cruise sample concentration, and thus each year is simulated separately, as opposed to a continuous 19-year simulation.

The 1D-CBH model displayed excellent ability to capture the temporal evolution of the observed vertical temperature profiles (Rucinski et al., 2010), including the onset of stratification and thermocline development in summer. Maximum model error varied with depth, 1.9°C and 3.4°C for 1994 and 2005, while the overall RE was less than 5%. Nonetheless, the 1D-CBH tended to underestimate the sudden mixed layer depth increase in late August and early September caused by storms, as well as the effects of horizontal advection and internal wave propagation that could not be reproduced by a 1-D model (Rucinski et al., 2010). Dissolved oxygen was also reproduced fairly well, i.e., RE = 25%, MEF = 0.46, but the capacity to reproduce biological (phyto- or zooplankton) components, and other closely-related variables (e.g., DRP) is limited, i.e., RE > 50%, MEF < 0 (Rucinski et al., 2016; Rucinski et al., 2014). Overall, the model was suggestive that within the range of load seasonality observed from 1987 to 2005, the hypoxic response is far more dependent on stratification structure (i.e., early or prolonged stratification, deep thermocline) than on the timing of nutrient delivery into the system. The model consistently predicted that sediment oxygen demand (SOD) represents a substantial fraction of the overall oxygen demand. It is important to understand that the SOD-versus-TP loading (rectangular hyperbola or simply a Michaelis-Menten-type) equation used represents a convenient means to run loading scenarios and directly evaluate the changes in water quality variables, but does not offer any mechanistic insights about the actual response of the sediment diagenesis processes nor does it have solid empirical foundation. After all, the precursor of that equation from Borsuk et al. (2001) was based on cross-sectional data from 34 estuarine and coastal systems worldwide, with the majority of those bearing little resemblance to Lake Erie (see Table 1 in Borsuk et al., 2001).

**Ecological Model of Lake Erie:** EcoLE is based on the two-dimensional hydrodynamic and water quality model CE-QUAL-W2 (version 2.0), which was originally developed to simulate long and narrow waterbodies displaying longitudinal and vertical water quality gradients and lateral homogeneity (Zhang et al., 2008; Cole and Wells, 2006). EcoLE divided Lake Erie into 1 m vertical layers (maximum number 65 vertical layers) and 222 longitudinal segments from west to east. The model has six variables for hydrodynamic simulations: horizontal velocity, vertical velocity, free water surface elevation, pressure, density, and constituent concentration. The vertical eddy coefficient algorithm has also been modified to adjust the strength of seiches and improve the simulations of longitudinal currents in a large, wind-driven Lake Erie (Boegman et al., 2001). The ecological model includes 28 state variables, with flexibility to include more water quality variables (Cole and Wells, 2006). Phytoplankton was divided into non-diatom edible algae, non-diatom inedible algae, and diatoms, while zooplankton comprised cladocerans and four copepod variables (copepod eggs, nauplii, copepodes and copepods); the latter component was based on the stage-structured population model originally presented by Fennel and Neumann (2003). The ecological implications of dreissenids were considered by two processes: grazing on phytoplankton and nutrient recycling via excretion (Zhang et al., 2008). Both their clearance rate and excretion rates (N and P) were linked to the dreissenid population size per segment, which in turn was described as a function of depth-dependent density and sediment area (Jarvis et al., 2000), and basin-specific length-mass regressions (Zhang et al., 2008).

The hydrodynamics of the model were calibrated against water levels, currents, and temperature data during the growing season (May–September) from 1994 (Boegman et al., 2001). The water quality module was subsequently calibrated using phytoplankton and nutrient concentration data from 1994 and 1997 (Boegman et al., 2008; Zhang et al., 2008), and subsequently validated against data from 1998 and 1999 (Zhang et al., 2008). The latter skill assessment exercise was mainly intended to evaluate the ability to capture the longitudinal patterns in Lake Erie, while the temporal variability was phased out by using seasonally averaged field measurements. The goodness-of-fit statistics showed that the model could not accurately reproduce the longitudinal variability in the system with RE varying from 30% to 120% and MEF values that were almost consistently negative. That is, EcoLE displayed inferior performance relative to the simple use of a lake-wide average value for each variable included in the calibration dataset (Zhang et al., 2008). A recent model assessment against a snapshot of the vertical profiles for water temperature, DO, and chlorophyll a from the summer of 2008 provided more encouraging results, although some parameters (maximum sediment DO demand along with the associated half saturation, a coefficient in the cloud cover function) had to be readjusted in order to improve the representation of the vertical distribution of the water temperature and DO in the central basin (Scavia et al., 2016b). A key feature of the process characterization from EcoLE is that algal growth relies on P recycling within the upper water column, including crustacean excretion, organic matter decay, dreissenid mussel population excretion rates, while phosphorus release from the sediments occurred mainly in the central basin, when hypoxic conditions prevail. Similar to 1D-CBH, the latter prediction is based on a simple engineering approach (user-specified sediment release rates) that does little in shedding light on the diagenesis process, nutrient retention time in the sediments, and their potential response to reduced sedimentation fluxes of autochthonous material (Scavia et al., 2016b).

**Western Lake Erie Ecosystem Model:** WLEEM was introduced by Verhamme et al. (2016) as a three-dimensional (3-D), fine-scale modeling framework, developed to simulate water quality responses to changes in the meteorological conditions and discharges of nutrients and sediments from tributaries into the western basin Lake Erie. WLEEM comprises four process-based models: (i) EFDC (Environmental Fluid Dynamics Code), which is a 3-D Finite difference hydrodynamic thermodynamic model, origniary developed by Hamrick (1992), providing water level and velocity as forcing functions for SWAN, and current direction and velocity as forcing functions for SEDTRAN. (ii) SWAN (Simulating WAVes Nearshore), which simulates wave conditions based on wind, depth, frictions, and velocity. This submodel linked to the EFCD (input) and SEDTRAN (output) to provide wave forcing for circulation (i.e., wave height, period, and direction) and support the simulation of wind-driven resuspension. (iii) SEDTRSN (Sediment Transport Model), which is a 3-D sediment transport model developed based on SEDZLM model by Jones and Lick (2001). This
submodel offers flexible simulation options that provide sediment settling/deposition, resuspension/erosion, bed armoring, and 3-D transport of multiple cohesive and non-cohesive sediment classes. (iv) A2EM (Advanced Aquatic Ecosystem Model). The original modelling framework was developed for the Upper Mississippi River system by HydroQual, Inc. in the late 1990s (LimnoTech, 2009) and the same model was revised to add a finer representation of the plankton community to simulate their seasonal successional patterns. The water quality model A2EM was developed on a publicly available version of Row-Column AESOP (RCA) model code (LimnoTech, 2010, 2013) that simulates carbon, nitrogen, phosphorus, and oxygen cycles. The model has the capacity to simulate five phytoplankton functional groups, three zooplankton functional groups, two dreissenid mussel size-based classes, and benthic algae (i.e., Cladophora) (Verhamme et al., 2016). The seasonal dynamics of only three major phytoplankton functional groups (i.e., diatoms or winter algal assemblage, green algae or summer algal assemblage, and cyanobacteria) were considered by Verhamme et al. (2016). WLEEM was calibrated using monitoring data from March to November 2011–2013, which represent a wide range of environmental conditions in Lake Erie with both high and low flow, nutrient loads, wind patterns, and significant variability of nutrient concentration and algal biomass (Verhamme et al., 2016). Model calibration first focused on physical (hydrodynamics, temperature, suspended solids), followed by chemical (TP and soluble reactive phosphorus or SRP, nitrogen species) components. The calibration of phytoplankton state variables involved around the timing and peak magnitude of the simulated functional groups; namely, diatoms, green, and cyanobacteria during spring, early summer, and middle to later summer, respectively. WLEEM successfully reproduced the temperature variability (RE < 5% and MEF > 0.830), and less so the ambient TP levels (43% < RE < 53% and = 0.949 < MEF < 0.067) in response to high-flow, spring-loading events, with concentrations being attenuated (diluted) with distance from the Maumee River mouth (Shimoda et al., 2019). SRP (53% < RE < 75% and = 0.054 < MEF < 0.448) and NO$_3$ + NO$_2$ (32% < RE < 41% and 0.351 < MEF < 0.725) concentrations were also reproduced reasonably well. Because of the limited data availability, the processes underlying the onset of spring (March–May) bloom were not adequately characterized. The model also tends to initiate HAB development earlier than the monitoring data suggest by approximately four weeks (see Fig. 8 in Verhamme et al., 2016). Model performance against cyanobacteria (96% < RE < 106% and = 0.302 < MEF < 0.097) was distinctly inferior relative to chlorophyll a concentrations (58% < RE < 71% and = 0.315 < MEF < 0.025).

Estuary and Lake Computer Model-Computational Aquatic Ecosystem Dynamics Model: ELCOM-CAEDYM is a coupled 3-D hydrodynamic (ELCOM) and biological (CAEDYM) model, and is considered one of the most commonly used models worldwide. ELCOM on its own has been used to examine the response of the thermal structure (Liu et al., 2014) and circulation patterns (León et al., 2012) in Lake Erie to changes in meteorological conditions (e.g., air temperature, wind speed). The same model has also been used to simulate 3-D transport of organic materials/organisms, such as walleye larvae floating near the lake surface (Zhao et al., 2009). During its initial application, ELCOM has been calibrated to vertical temperature profile and data from 1994 for all basins, and subsequently against data from 2001 to 2003 in the eastern basin (León et al., 2005). Most recent simulations of the water temperature patterns displayed very satisfactory performance, i.e., RE < 10% and MEF > 0.780 (Bocaniov et al., 2016; Bocaniov and Scavia, 2016; Oveisy et al., 2014; Bocaniov et al., 2014).

The existing ELCOM-CAEDYM applications in Lake Erie opted for a simpler ecological structure than its built-in capacity. Specifically, CAEDYM has a total of 112 state variables that consider water, sediment, chemical, biological processes, and can simulate up to 7 phytoplankton functional groups (PFGs), 5 zooplankton functional groups (ZFGs), 6 fish groups, 4 macroalgal groups, 3 invertebrate groups, 3 mussel size classes, macrophytes, seagrass, jellyfish (Hipsey and Hamilton, 2008). The application of ELCOM-CAEDYM to Lake Erie is limited to five phytoplankton groups (i.e., early bloom diatoms, lake bloom diatoms, cyanophytes, cooler and deeper water flagellate and warmer and brighter water flagellates) (Bocaniov et al., 2014, 2016; León et al., 2011). State variables that represent zooplankton have never been activated with ELCOM-CAEDYM applications in Lake Erie, and thus the role of zooplankton has been accommodated solely by the increased grazing-induced losses of phytoplankton (Bocaniov et al., 2014, 2016; León et al., 2011). A novel feature of ELCOM-CAEDYM in Lake Erie is the explicit representation of dreissenid dynamics. The built-in dreissenid sub-model was developed based on a clam model for a temperate estuarine lagoon (Spillman et al., 2008). Dreissenid biomass density was modulated by grazing (on phytoplankton and detrital particle) and respiration, excretion, egestion, and mortality losses (Bocaniov et al., 2014). Nutrient release from the sediments and ice formation modules are also incorporated to ELCOM-CAEDYM. Both modules have been used to consider the role of sediments on the elevated TP concentrations in the northeastern nearshore zone (León et al., 2011), as well as to evaluate the effects of ice cover on winter productivity and subsequent hypoxia development (Oveisy et al., 2012).

ELCOM-CAEDYM was calibrated against DO observations (León et al., 2006). Another version with additional submodules (i.e., PFGs, dreissenids, and ice-cover) has been also recalibrated against temperature, chlorophyll a, light attenuation, and nutrient concentration data from 2002 (Bocaniov et al., 2014; León et al., 2011) and temperature, DO, ice-cover and thickness from the 2004–2005 winter period (October–April) data (Oveisy et al., 2014). The most recent studies (i.e., Karatayev et al., 2017; Bocaniov et al., 2016; Bocaniov and Scavia, 2016) improved the external forcing functions with finer resolution meteorological data and more comprehensive discharges from tributaries. The simulated thermal structure was validated against USEPA cruise data from 2008, satellite-derived, lake-wide surface observations (Bocaniov and Scavia, 2016), and DO, chlorophyll a, nutrient concentrations from two basin-wide cruises occurred in 2008 (Bocaniov et al., 2016). With the most recent studies, the reproduction of chlorophyll a concentrations was generally satisfactory with the skill assessment by Bocaniov et al. (2014), 18% < RE < 42% and = 0.467 < MEF < 0.868, and less so with the one presented by Valipour et al. (2016), 64% < RE < 88% and = 3.081 < MEF < 0.250.

Eastern Basin Cladophora Model: The EBC model is a simple mechanistic model aiming to predict Cladophora standing biomass and P stored in plant tissues (Auer et al., 2010). The original framework to simulate Cladophora growth/biomass developed by Canale and Auer (1982) has been modified by several authors (Auer et al., 2010; Tomlinson et al., 2010; Higgins et al., 2005, 2006), which generally focused on refinements of three terms: growth, loss by respiration and sloughing. The growth rate of Cladophora is expressed as a function of the maximum gross specific growth rate and limitation multipliers that account for the role of light, temperature, internal P, and maximum biomass density (carrying capacity). The respiration rate considers both a dark/basal rate that varies only with temperature as well as a light-enhanced rate determined by temperature and light intensity (Tomlinson et al., 2010). In its updated version, sloughing is modelled as a first-order loss process with the rate coefficient varying as a function of water temperature and the depth of colonization, which in turn reflects the effect of wind energy (i.e., momentum leading to detachment) and benthic shear stress effect. The physiological effects of temperature initiate with a minimum temperature until an optimum temperature level is reached at and above which the sloughing rate is at its maximum value (Tomlinson et al., 2010). In a recent study, Valipour et al. (2016) incorporated EBC with ELCOM-CAEDYM in order to investigate the interplay among external phosphorus loading, nearshore phosphorus, and Cladophora growth. The introduction of EBC into a 3-D hydrodynamic environment intended to shed light on both nearshore and basin-scale physical factors, such as wind, surface runoff, degree of
stratification, upwelling events (period of 5–10 days), seiches (~14 h), and near-inertial waves (~17 h) that shape offshore-nearshore exchanges. Upwelling events during the months of May–early July appear to be the predominant drivers of offshore-nearshore mass exchanges, whereby SRP is injected from the hypolimnion into the nearshore northeastern Lake Erie; especially since these events usually coincide with optimal light and water temperature conditions (Valipour et al., 2019), while the median RE of all the “observed-versus-predicted” comparisons presented in the literature was larger than 60% (Fig. 5). In a similar manner, the corresponding values for the P tissue content varied from −2.795 to −0.016 (Shimoda et al., 2019). Except from our limited understanding of the drivers underlying the sloughing processes, some of the identified weaknesses include the lack of a comprehensive SRP database for nearshore habitat colonized by Cladophora, which inevitably leads to a reliance upon measurements from offshore sites or at water intakes located at depths often beyond the domain of colonization (Auer et al., 2010). Moreover, there is no empirical evidence of Cladophora biomass levels immediately prior to the dreissenid invasion, and thus there is no direct way to reconcile their role on Cladophora “resurgence” in Lake Erie (Higgins et al., 2005). In addition, the internal phosphorus is rarely measured in natural populations, and the values used typically span a wide range (Tomlinson et al., 2010; Malkin et al., 2008; Higgins et al., 2005; Auer and Canale, 1982; Jackson and Hamdy, 1982). There also seems to be some inconsistency with respect to the sampling/analytical protocols followed during the determination of Cladophora biomass used to constrain the existing models.

Appendix B. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ecoinf.2019.05.014.

References

Allen, C.R., Fontaine, J.J., Pope, K.I., Garmestani, A.S., 2011. Adaptive management for a turbulent future. J. Environ. Manag. 92, 1339–1345. https://doi.org/10.1016/j.jenvman.2010.11.019.

Alling, L.E., Reavie, E.D. 2013. The ecological history of Lake Erie as recorded by the phytoplankton community. J. Great Lakes Res. 39 (3), 365–382. https://doi.org/10.1016/j.jglr.2013.06.014.

Anderson, T.R., 2005. Plankton functional type modelling: Running before we can walk? J. Plankton Res. 27 (11), 1078–1081. https://doi.org/10.1093/plankt/boh076.

Arhonditsis, G.B., Brett, M.T., 2004. Evaluation of the current state of mechanistic aquatic biogeochemical modelling: Mar. Ecol. Prog. Ser. 271, 13–26. https://doi.org/10.1016/j.marpolprog.2004.06.030.

Arhonditsis, G.B., Brett, M.T., Windm., Brett, M.T., Schindler, D.E., 2004. Patterns and mechanisms of phytoplankton variability in Lake Washington (USA). Water Res. 38, 4013–4027. https://doi.org/10.1016/j.watres.2004.06.030.

Arhonditsis, G.B., Adams-VanHarms, B.A., Nielsen, L., Stow, C.A., Reckhow, K.H., 2006. Evaluation of the current state of mechanistic aquatic biogeochemical modelling: citation analysis and future perspectives. Environ. Sci. Technol. 40 (21), 6547–6554. https://doi.org/10.1021/es060103q.

Arhonditsis, G.B., Qian, S.S., Stow, C.A., Lamon, E.C., Reckhow, K.H., 2007. Evaluating biogeochemical modeling of phosphorus in western Lake Erie during bloom conditions. J. Appl. Phycol. 19, 291–304. https://doi.org/10.1007/s10811-006-9059-0.

Arhonditsis, G.B., Papantou, D., Zhang, W., Perhar, G., Massos, E., Shi, M., 2008a. Bayesian calibration of mechanistic aquatic biogeochemical models and benefits for environmental management. J. Marine Syst. 73, 83–100. https://doi.org/10.1016/j.jmarsys.2007.07.004.

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

Arhonditsis, G.B., Brett, M.T., 2004. Evaluation of the current state of mechanistic aquatic biogeochemical modelling: Mar. Ecol. Prog. Ser. 271, 13–26. https://doi.org/10.1016/j.marpolprog.2004.06.030.

Arhonditsis, G.B., Papantou, D., Zhang, W., Perhar, G., Massos, E., Shi, M., 2008a. Bayesian calibration of mechanistic aquatic biogeochemical models and benefits for environmental management. J. Marine Syst. 73, 83–100. https://doi.org/10.1016/j.jmarsys.2007.07.004.

Arhonditsis, G.B., Perhar, G., Zhang, W., Massos, E., Shi, M., Das, A., 2008b. Addressing equifinality and uncertainty in eutrophication models. Water Resour. Res. 44 https://doi.org/10.1029/2007WR005862. (W01420).
cyano bacteri al bloom formation: the critical role of anoxia and ferrous iron. Freshw. Biol. 59, 1323–1340. https://doi.org/10.1111/fwb.12334.

Muenich, R.L., Kalic, M., Scavia, D., 2016. Evaluating the impact of legacy P and agricultural conservation practices on nutrient loads from the Maumee River watershed. Environ. Sci. Technol. 50, 8416–8418. https://doi.org/10.1021/acs.est.0c01427.

Murphy, D.V., Recous, S., Stockdale, E.A., Fillery, I.R.P., Jensen, L.S., Hatch, D.J., Ort, D.R., Boysen, K., Caves, B., Confesor, R., DePinto, J., Gildow, M., Martin, J., 2016. Informing Lake Erie Agriculture Nutrient Management Via Scenario Evaluation. University of Michigan: Ann Arbor, MI, USA.

Scavia, D., DePinto, J.V., Auer, M., Bertani, I., Bocanov, S., Chapra, S., Leon, M., McCrimmon, C., Obenour, D., Peterson, G., Rucinski, D., Schles, D., Steger, L., Stumpf, R., Yerubandi, R., Zhang, H., 2016b. Great Lakes Water Quality Agreement Nutrient Annex Task Team Emerging Science Synthesis. Great Lakes National Program Office. USEPA, Chicago Available at: https://www.epa.gov/sites/production/files/2016-09/documents/nutrientannexmultimodelingreportfinal3aug2016.pdf.

Steffen, M.M., Belisle, B.S., Watson, S.B., Boyer, G.L., Wilhelm, S.W., 2014. Status, causes and controls of cyanobacterial blooms in Lake Erie. J. Great Lakes Res. 40 (2), 215–225. https://doi.org/10.1016/j.jglr.2013.12.012.

Stowe, A., Oya, R., Johnson, T., Richardson, R.P., 2015. Long-term and seasonal trend decomposition of Maumee River nutrient inputs to western Lake Erie. Environ. Sci. Technol. 49, 3392–3400. https://doi.org/10.1021/acs.est.5b02648.

Stumpf, R.P., Wynne, T.T., Baker, D.B., Fahnengstelt, G.L., 2012. Interannual variability of cyanobacterial blooms in Lake Erie. PloS One 7 (8), e42444. https://doi.org/10.1371/journal.pone.0042444.

Stumpf, R.P., Johnson, L.T., Wynne, T.T., Baker, D.B., 2016. Forecasting annual cyanobacterial bloom biomass to inform management decisions in Lake Erie. J. Great Lakes Res. 42 (6), 1174–1183. https://doi.org/10.1016/j.jglr.2016.08.006.

Teunher, C., Krobie, N.D., Donauban, K., Kabas, W., Kirschner, A.K.T., Pister, G., Salbrechter, M., Dokulil, M.T., 2020. Enhanced phosphorus accumulation efficiency by the pelagic community at reduced phosphorus supply: a lake experiment from bacteria to metazoan zooplankton. Limnol. Oceanogr. 65, 1141–1149. https://doi.org/10.1002/lno.10933.
Wellen, C., Arhonditsis, G.B., Long, T., Boyd, D., 2014b. Accommodating environmental uncertainty: the value of information. Ecol. Model. 222 (18), 3429–3436. https://doi.org/10.1016/j.ecolmodel.2011.07.003.

Wilson, K.A., Howell, E.T., Jackson, D.A., 2006. Replacement of zebra mussels by quagga mussels in the Canadian nearshore of Lake Ontario: the importance of substrate, round goby abundance, and upwelling frequency. J. Great Lakes Res. 32, 11–28. https://doi.org/10.3394/0380-1330(2006)32[11:ROZMBQ]2.0.CO;2.

Wynne, T.T., Stumpf, R.P., Tomlinson, M.C., Warner, R.A., Tester, P.A., Dyble, J., Fahnentiel, G.L., 2008. Relating spectral shape to cyanobacterial blooms in the Laurentian Great Lakes. Int. J. Remote Sens. 29, 3665–3672. https://doi.org/10.1080/01431160802007640.

Wynne, T.T., Stumpf, R.P., Tomlinson, M.C., Dyble, J., 2010. Characterizing a cyanobacterial bloom in western Lake Erie using satellite imagery and meteorological data. Limnol. Oceanogr. 55, 2025–2036. https://doi.org/10.4319/lo.2010.55.5.2025.

Xue, Y., David, M.B., Gentry, L.E., Kovacic, D.A., 1998. Kinetics and modelling of dissolved phosphorus export from a tile-drained agricultural watershed. J. Environ. Qual. 27 (4), 917–922. https://doi.org/10.2134/jeq1998.00472425002700040028x.

Yaksich, S.M., Mell, D.A., Baker, D.B., Kramer, J.W., 1985. Lake Erie nutrient loads, 1970–1980. J. Great Lakes Res. 11, 117–131. https://doi.org/10.1016/S0380-1330(85)71751-0.

Yokota, F., Thompson, K.M., 2004. Value of information analysis in environmental health risk management decisions: past, present, and future. Risk Anal. 24, 635–650. https://doi.org/10.1111/j.0272-4330.2004.00464.x.

Zhang, H., Culver, D., Boegman, L., 2008. A two-dimensional ecological model of Lake Erie: application to estimate dreissenid impacts on large lake plankton populations. Ecol. Model. 214, 219–241. https://doi.org/10.1016/j.ecolmodel.2008.02.005.

Zhao, Y., Jones, M.L., Shuter, B.J., Roseman, E.F., 2009. A biophysical model of Lake Erie walleye (Sander vitreus) explains interannual variations in recruitment. Can. J. Fish. Aquat. Sci. 66 (1), 114–125. https://doi.org/10.3394/0338-1340(2009)08<0089:TQAFBS>2.0.CO;2.

Zolfaghari, K., Duguay, C., 2016. Estimation of Water Quality Parameters in Lake Erie from MERIS Using Linear Mixed Effect Models. Remote Sens. 8 (6), 473. https://doi.org/10.3390/rs8060473.
CASTLES BUILT ON SAN D OR PREDICTIVE LIMNOLOGY IN ACTION?

PART A: EVALUATION OF AN INTEGRATED MODELLING FRAMEWORK TO GUIDE ADAPTIVE MANAGEMENT IMPLEMENTATION IN LAKE ERIE

[SUPPORTING INFORMATION]

George B. Arhonditsis*, Alex Neumann, Yuko Shimoda, Dong-Kyun Kim, Feifei Dong, Gabriela Onandia, Cindy Yang, Aisha Javed, Meghan Brady, Ariola Visha, Felicity Ni, Vincent Cheng

Ecological Modelling Laboratory
Department of Physical & Environmental Sciences, University of Toronto,
Toronto, Ontario, Canada, M1C1A4

* Corresponding author

Email: georgea@utsc.utoronto.ca, Tel.: +1 416 208 4858; Fax: +1 416 287 7279
Table S1: Predictive validation of the five SWAT models against monthly flow and phosphorus loading at Waterville, Ohio, from 2005-2014. Values close to 0 for percent bias (PBIAS) and 1 for the modelling efficiency (MEF) and the coefficient of determination ($r^2$) indicate higher model performance.

| Measures of Model Fit* | Models |  |
|------------------------|--------|--------|--------|--------|--------|--------|
|                        | HU     | LT     | OSU    | TAMU   | UM     | Model Average |
| PBIAS                  | -7%    | 10%    | 10%    | 11%    | 6%     | 6%     |
| MEF                    | 0.82   | 0.90   | 0.91   | 0.86   | 0.89   | 0.88   |
| $r^2$                  | 0.86   | 0.91   | 0.93   | 0.88   | 0.91   | 0.90   |
| PBIAS                  | 37%    | -6%    | -7%    | -22%   | 7%     | 2%     |
| MEF                    | 0.64   | 0.82   | 0.73   | 0.56   | 0.70   | 0.69   |
| $r^2$                  | 0.74   | 0.82   | 0.75   | 0.71   | 0.70   | 0.75   |
| PBIAS                  | 81%    | 1%     | 16%    | -13%   | -13%   | 14%    |
| MEF                    | -0.02  | 0.71   | 0.51   | 0.52   | 0.46   | 0.44   |
| $r^2$                  | 0.55   | 0.71   | 0.54   | 0.70   | 0.51   | 0.60   |
Table S2: Projections of DRP export from nine spatial segments of the Maumee River watershed (Scavia et al., 2016)*. The categories low (L), intermediate (I), and high (H) of our classification scheme refer to the projected range of nutrient export from each of the five SWAT models, and therefore they do not represent numerically comparable levels of areal loading.

| Spatial Segment | HU | LT | OSU | TAMU | UM |
|-----------------|----|----|-----|------|----|
| North-west      | I  | H  | H   | L    | I  |
| North-center    | I  | I  | I   | I    | I  |
| North-east      | L  | L  | L   | L    | L  |
| Center-west     | I  | I  | H   | I    | I  |
| Center          | H  | I  | L   | H    | I  |
| Center-east     | I  | I  | I   | L    | I  |
| South-west      | I  | I  | H   | I    | I  |
| South-center    | I  | H  | I   | I    | I  |
| South-east      | I  | I  | I   | I    | I  |

*Scavia, D., Kalcic, M., Muenich, R.L., Aloysius, N., Arnold, J., Boles, C., Confesor, R., DePinto, J., Gildow, M., Martin, J., Read, J., Redder, T., Robertson, D.M., Sowa, S.P., Wang, Y.-C., White, M., Yen, H., 2016. Informing Lake Erie agriculture nutrient management via scenario evaluation. Ann Arbor, Michigan.
Table S3: Description of the different SWAT-ensemble model scenarios, categorized based on the policy questions they were intended to address. Modified from Scavia et al. (2016).

| Number of Scenarios | Name                                                                 | Description                                                                                                                                                                                                 | Policy question                                                                 |
|---------------------|----------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------|
| S1                  | Nutrient management at 25% random adoption                           | 25% of row crop acreage was randomly chosen and the following practices were applied: a 50% reduction in P fertilizer application, fall timing of P applications, and subsurface placement of P into the soil. | What level of nutrient management will be sufficient to reach phosphorus targets? |
| S2                  | Nutrient management at 100% adoption                                  | The following practices were applied to 100% of row crop fields throughout the watershed: a 50% reduction in P fertilizer application, fall timing of P applications, and subsurface placement of P into the soil. | Can nutrient management alone achieve targets?                                   |
| S3                  | Commonly recommended practices at 100% random adoption               | Four practices were applied to a separate 25% of the crop fields: a 50% reduction in P fertilizer application, subsurface application of P fertilizers, continuous no tillage, and medium-quality buffer strips. | What extent of adoption of commonly recommended practices will be needed to achieve the targets? |
| S4                  | Continuous no-tillage and subsurface placement of P fertilizer at 50% random adoption | 50% of row crop acres were randomly chosen to apply a combination of continuous no-tillage and subsurface application of P fertilizers.                                                                         | Is no-tillage effective provided P is applied below the soil surfaces?           |
| S5                  | Cropland conversion to grassland at 10% (5a), 25% (5b), and 50% (5c) targeted adoption | In this set of three scenarios, 10%, 25%, and 50% of the row croplands with the lowest crop yields and greatest TP losses were converted to switchgrass and managed for wildlife habitat with limited harvesting for forage and no P fertilization. | How much row cropland would be needed to be converted to grassland to achieve the targets? |
| S6                  | Series of practices at 50% targeted adoption                          | A series of practices were targeted to the 50% of row cropland with the highest TP loss in the watershed. Practices: subsurface application of P fertilizers, cereal rye cover crop in the winters without wheat, and application of medium-quality buffer strips. | What extent of targeted infield and edge-of-field practices reach the targets?    |
Table S1 (Continued): Description of the different model scenarios, categorized based on the policy questions they were intended to address.

| Number of Scenarios | Name                                | Description                                                                 | Policy question                                      |
|---------------------|-------------------------------------|-----------------------------------------------------------------------------|------------------------------------------------------|
| S7                  | **Series of practices** at 50% random adoption | A series of practices were applied at random to 50% of row cropland. Practices: subsurface application of \( P \) fertilizers, cereal rye cover crop in the winters without wheat, and application of medium-quality buffer strips. | What if in-field and edge-of-field practices were applied at random? |
| S8                  | **Diversified rotation** at 50% random adoption | An alternative corn-soybean-wheat rotation with a cereal rye cover crops all winters without wheat was applied over a randomly chosen 50% of row cropland. | What is the impact of returning to winter wheat and winter cover crops? |
| S9                  | **Wetlands and buffer strips** at 25% targeted adoption | Wetlands were targeted to the 25% of sub-watersheds with the greatest \( P \) loading and assumed to intercept half of overland and tile \( \text{fL} \), and medium-quality buffer strips were targeted to the 25% of row crop acreage responsible for the greatest \( TP \) loss. | How much \( P \) reduction can be achieved through structural practices? |
| S10                 | **In-field practices** at 25% random adoption | A group of four in-field practices was applied to a random 25% of row cropland. Practices included a 50% reduction in \( P \) fertilizer application, fall timing of \( P \) applications, subsurface placement of \( P \) fertilizers, and a cereal rye cover crop. | What can be achieved at 25% application of in-field practices? |