https://doi.org/10.5194/hess-2021-296
Preprint. Discussion started: 2 June 2021
© Author(s) 2021. CC BY 4.0 License.

Spatially Referenced Bayesian State-Space Model of Total Phosphorus in western Lake Erie

Timothy J. Maguire[1], Craig A. Stow[2], Casey M. Godwin[1]

[1]Cooperative Institute for Great Lakes Research, School for Environment and Sustainability, University of Michigan, Ann Arbor, Michigan 48109, United States
[2]NOAA Great Lakes Environmental Research Laboratory, Ann Arbor, Michigan, 48108, United States

Correspondence to: Timothy J. Maguire (maguiret@umich.edu)

Abstract. Collecting water quality data across large lakes is often done under regulatory mandate, however it is difficult to connect nutrient concentration observations to sources of those nutrients and to quantify this relationship. This difficulty arises from the spatial and temporal separation between observations, the impact of hydrodynamic forces, and the cost involved in discrete samples collected aboard vessels. These challenges are typified in Lake Erie, where binational agreements regulate riverine loads of total phosphorus (TP) to address the impacts from annual harmful algal blooms (HABs). While it is known that the Maumee River supplies 50% of the nutrient load to Lake Erie, the details of how the Maumee River TP load changes Lake Erie TP concentration have not been demonstrated. We developed a hierarchical spatially referenced Bayesian state-space model with an adjacency matrix defined by surface currents. This was applied to a 2km-by-2km grid of nodes, to which observed lake and river TP concentrations were joined. The model generated posterior samples describing the unobserved nodes and observed nodes on unobserved days. We quantified the impact plume of the Maumee River by experimentally changing concentration data and tracking the change of in-lake predictions. Our impact plume represents the spatial and temporal variation of how river concentrations correlate with lake concentrations. We used the impact plume to scale the Maumee River spring TP load to an effective Maumee River TP spring load for each node in the lake. By assigning an effective load to each node the relationship between load and concentration is consistent throughout our sampling locations. A linear model of annual lake node mean TP concentration and effective Maumee River load estimated that in the absence of the Maumee River load lake concentrations at the sampled nodes would be 23.1 µg l^{-1} (± 1.75, 95% credible interval, CI) and that for each 100 tons of spring TP effective load delivered to Lake Erie, mean TP concentrations increase by 11 µg l^{-1} (± 1, 95% CI). Our proposed modelling technique allowed us to establish these quantitative connections between Maumee TP load and Lake Erie TP concentrations which otherwise would be masked by the movement of water through space and time.

1 Introduction

In a collective response to the economic, human health, and environmental damage caused by pollution, assessing water quality is a regulatory mandate across many waterbodies. In many aquatic ecosystems nutrient concentrations are a primary water quality analyte collected. Observed concentrations are driven by both point and non-point sources. Excessive nutrient export primarily from agricultural watersheds leads to eutrophication, harmful algae blooms (HABs), and threatens drinking water contamination (Brooks et al., 2016; Schneider and Bláha, 2020). Individual water bodies present data collection challenges, particularly large lakes. Even for those locations with well-funded sample collection schema, trying to describe the spatiotemporal heterogeneity in nutrients is difficult. Discernible trends are difficult to assess as samples represent discrete spatial data within a system of constantly moving water and asymmetrical spatial extents of riverine nutrient plumes.

Lake Erie is an example of a waterbody that is challenging to model (Ho and Michalak, 2017; Steffen et al., 2017; Stow et al., 2015). While Lake Erie is large (25,700 km2), its western basin is relatively shallow (mean depth 6 m) (Bolsenga and Herdendorf, 1993) and intense nutrient export from the agriculturally dominated Maumee River watershed leads to episodic HABs (Watson et al., 2016). Commercial fisheries, drinking water, and human health within Lake Erie are all impacted because of the combination of nutrient addition, HABs, and physical lake properties (Wituszynski et al., 2017). Because of these intersecting concerns, a binational effort to regulate phosphorus entering Lake Erie has been active since 1978 (GLWQA, 2012). Nutrient concentration and physical lake data are pivotal in understanding the causes of western basin Lake Erie water quality issues and have been collected by a broad range of federal, state/province, and local agencies throughout western Lake Erie (Fig. 1). The goal has been to collect these data at a spatial and temporal scale which should lead to a defined relationship of how river nutrient load effects lake nutrient concentration; understanding of the influence of riverine load through time and space; and ultimately the ability to predict how river load reductions would manifest as altered lake concentrations.

Yet, while western Lake Erie is routinely monitored and the nutrient concentration and flow are estimated daily in rivers, a generalizable connection between Maumee River phosphorus load and Lake Erie phosphorus concentrations remains undefined (i.e., if phosphorus load increases by 100 tons what is the response in lake concentration?) (Rowland et al., 2019). Spring Maumee River soluble reactive phosphorus (SRP) export correlates with western Lake Erie HABs extent; this pattern has been observed since the SRP loads started to increase in the 1990s (Ho and Michalak, 2017; Michalak et al., 2013; Stow et al., 2015). The challenge is that nutrients in the lake move with the water currents, resulting in a complex relationship of upstream and downstream current dependence. Moreover, within-lake phosphorus cycling is dynamic and impacted by biological and physical processes (Li et al., 2021; Matisoff et al., 2016). Additionally, the time between sampling events within this time-series and the size of the lake-river system where models need to be applied inherently adds uncertainty and reduces the predictive efficacy of transport models linked with hydrodynamic models (Schwab et al., 2009).

Bayesian frameworks offer a methodology that can quantify uncertainty in the effect of nutrient load on nutrient distribution within a dynamic system such as Lake Erie. The goal of this study was to quantify the impact of river nutrient delivery across western Lake Erie through time and space. While Spatial models have been used in the Great Lakes for predicting HABs biomass, HABs extent, and nutrient transport (Fang et al., 2019; Schwab et al., 2009), we proposed a Bayesian framework for similar spatial data. We showed that phosphorus concentration, along with an informed uncertainty, can be estimated within a state-space model that incorporates concentration data from within the lake, the rivers, and lake surface currents. Although this approach is informed by the currents, it is does not include all the explicit biogeochemical and physical processes that are part of mechanistic models (Rowe et al., 2019). Here, we quantified how well our model fits the data, generated predictions of TP concentrations across western Lake Erie, experimentally manipulated the available concentration observations in order to estimate the spatial and temporal impact from the Maumee River plume and tested the hypothesis that when water movement is incorporated, there is a linear relationship between river load and western Lake Erie water TP concentrations.

2 Methods

2.1 Study Area and Data Curation

We limited the model spatial window to western Lake Erie (bounded by the portion of the lake west of -83.1° W, Fig. 1) which left ~600 km2 to be defined. We gathered surface concentrations of TP (μg/l) from publicly available databases through Environment Climate Change Canada's Offshore Water Quality Survey, the U.S. Environmental Protection Agency's Great Lakes National Program Office, the Canadian Ministry of the Environment, Conservation and Parks Great Lakes Intake Program, the U.S. National Oceanographic and Atmospheric Administration (NOAA) Great Lakes Environmental Research Laboratory (GLERL) Ecosystem Dynamics Long-Term Research program, and NOAA GLERL Western Lake Erie (WLE) Sampling (Table A1). The data used here ranged from 2008 to 2018. For riverine TP concentrations from the Maumee River and River Raisin across the 2008 to 2018 interval we downloaded data from the National Center for Water Quality Research (NCWQR) at Heidelberg University (Table A1). When multiple samples were collected from a station on a single day the sample average was used.

2.2 Model Description

We created a model where day *t* TP concentrations are predicted based on the concentrations "upstream" at day *t-1*. The spatial adjacency of "upstream" relationships was defined by the direction and magnitude of surface currents.

To build our adjacency matrix we first defined a hypothetical distance and direction that surface water moved based on surface currents. We used surface current data retrieved from the NOAA Great Lakes Coastal Forecasting System (GLCFS, Table A1) database. These data are defined by hourly eastward and northward water velocity (m s^{-1}) predicted across Lake Erie on a 2 km-by-2 km grid (Fig. 1). These data were converted to mean daily northward and eastward velocity (m day^{-1}) for each node for years 2008 to 2018. The surface current direction in radians (dLat and dLon) was calculated for each node using the node latitude (Lat_0) and longitude (Lon_0), the Earth's radius (R, 6378137 m), the northward velocity offset in meters (dN), and eastward offset in meters (dE) (Eqs 1 and 2). The direction water travelled in radians was used to determine the latitude (Lat_1) and longitude (Lon_1) which represents the mean daily position of the surface water movement (Eqs 3 and 4).

$$dLat = \frac{dN}{R} \tag{1}$$

$$dLon = \frac{dE}{R*Cos\left(\pi*\frac{Lat_0}{180}\right)} \tag{2}$$

$$Lat_1 = Lat_0 + dLat * \left(\frac{180}{\pi}\right) \tag{3}$$

$$Lon_1 = Lon_0 + dLon * \left(\frac{180}{\pi}\right) \tag{4}$$

The limited model spatial window of western Lake Erie was represented by 254 nodes (Fig. 1). The Lake Erie surface water TP concentrations were associated with their closest nodes of the same 2 km-by-2 km grid nodes used in the surface current datasets. NCWQR concentration data were collected for the Maumee River (41.5° N, -83.712778° W) and the River Raisin (41.960556° N, -83.531111° W) locations ~30 and ~18 km, respectively, inland from Lake Erie. River concentrations were assigned to the node closest to the river mouth. The assumption that these concentrations represent the conditions at the terminus of the rivers adds uncertainty to our modelling, however the spatial extent of this extra uncertainty should end where the Lake Erie TP concentration data begins to inform the model posterior samples.

2.2.1 State-Space Models

We constructed hierarchical, spatially referenced Bayesian state-space models for each year to estimate TP concentrations for each node on each day. The temporal range annually was May 20 to October 2, to coincide with the majority of the WLE sampling. The distance between each the daily offset surface current location (Lat_1, Lon_1) and each 2 km-by-2 km concentration node was measured and the node *n* with the shortest distance defined the adjacency matrix to associate each node *n* on day *t* with the node *k* on day *t-1*. For nodes subject to the river models, the latent state ($x_{n,t,y}$) was defined by previous river time-step while the lake models use the adjacency matrix to identify which latent state should be used.

$$y_{n,t,y} \sim N\left(x_{n,t,y}\, , \sigma^2\right) \tag{5}$$

$$x_{n,t,y} \sim N\left(xp_{n,t,y}\, , \tau^2\right) \quad Trunc(a \leq x_{n,t,y} \leq b) \tag{6}$$

$$xp_{n,t,y} = \begin{cases} x_{k,t-1,y} * \beta_{mau} & if\ k = Maumee\ River\ Node \\ x_{k,t-1,y} * \beta_{rai} & if\ k = River\ Raisin\ Node \\ x_{k,t-1,y} * \beta_{self} & if\ n = k \\ x_{k,t-1,y} * \beta_{lake} & if\ n \neq k \end{cases} \tag{7}$$

Log-transformed TP concentration observations (*y*) at the n^{th} node on the t^{th} day of the y^{th} year was estimated with a normal data model sampled from the state variable (*x*) at the n^{th} node on the t^{th} day of the y^{th} year with standard deviation σ (Eq 5). The latent state ($x_{n,t,y}$; Eq 6) is sampled from a normal distribution of a predicted latent state ($xp_{n,t,y}$, Eq 7) and standard deviation τ. $x_{n,t,y}$ was truncated by the detection limit of TP laboratory analysis (5 µg l^{-1}, *a*, Eq 6) and the maximum value observed in each year (*y*) within the Maumee River (*b*, Eq 6), $xp_{n,t,y}$ was defined depended on the node *n* as being a river or lake node (Eq

7). River nodes were described by previous state variable of that river ($x_{n,t-1,y}$). River models were fit using β_{mau} and β_{rai} for the Maumee River and River Raisin, respectively. These βs are fit in a hierarchical framework to a hyperparameter β with uninformed normal (N(0,0.01)) and gamma hyperpriors (*gamma*(0.001,0.001)). River model coefficients were fit hierarchically because the ecological and anthropogenic processes enacted on these watersheds are similar, if at different scales. The two lake models were fit with two independent β coefficients depending on if the nearest adjacent node k is the same as the estimated node n (β_{self}) or if a different node k is the nearest (β_{lake}), each with non-informative normal priors. Separate independent in-lake models were used to capture different potential drivers of TP concentration through time depending on whether each node was subject to little surface water movement (β_{self}) or active surface water movement (β_{lake}). In 2012 there were no River Raisin observations and so the model in 2012 treats the River Raisin node as a lake node. The model was run in R (version 4.0.2) and JAGS (version 4.3.0) (Eddelbuettel, 2017; Microsoft Corporation and Weston, 2020; Plummer, 2019). Each year's model iteration count was 50,000 with a thin of 10, representing 5,000 effective samples along three independent Markov chains. Initial conditions for the latent state $x_{n,t=1,y}$ were defined as the mean and variance of the river or lake data within each year's model window. The efficacy of the models was described via posterior predictive p-values (Gelman, 2013) and Bayesian R^2 (Vehtari et al., 2017), while the performance between years and across nodes was assessed with K-Fold cross validation (CV) utility (Geisser and Eddy, 1979; Piironen and Vehtari, 2017) (Eq 8).

2.2.2 State-Space Model Fit

Posterior predictive p-values were calculated by model year with test statistic mean TP concentration, to compare means of observations to the means of the model outputs. For each year, a posterior p-value distribution was described by 15,000 bootstraps of 100 resamples from the observed node posteriors. Bayesian R^2 defined as the resolved variance (var_{fit}) divided by the sum of var_{fit} and the residual variance (var_{res}) was calculated for each model year. Model var_{fit} was the variance of the modelled predictive mean, while var_{res} is the variance of model predictive mean subtracted from the observations (Gelman et al., 2019). K-Fold coefficient of variation (CV) utility compared model predictive performance across years and across nodes. The cross validated utility was applied by removing all d observations from a randomly selected node k with at least 10 observations collected during the model year. K-Fold CV was calculated 3 times per year, for years that had less than 3 nodes with at least 10 observations, all nodes that satisfied the 10-observation cut-off were used. K-Fold CV is the mean leave-one-node-out log predictive density from posterior samples of the omitted d observation $\hat{y}_{n,t,y}$ at node n, day t, year y, were compared to observed concentrations at $y_{n,t,y}$ (Piironen and Vehtari, 2017) (Eq 8).

$$\text{K} - \text{Fold CV}_{n,y} = \frac{1}{d}\sum_{d=1}^{d} \log p\left(y_{n,t,y} \middle| \hat{y}_{n,t,y}\right) \tag{8}$$

Using a model that describes posterior predictive distributions of mean K-Fold CV across years and nodes we examined if our state-space approach preferentially generated predictions that contain the observed values (Eq 9 and 10). 95% credible differences between group means (for nodes μ_n or for years μ_y) that do not contain 0 were used to determine if groups were different (e.g., if the 95% credible difference from μ_{2018}- μ_{2017} contains 0 these means are not considered different).

$$\mathrm{K-Fold\ CV}\ \sim N(\mu_{on} + \textstyle\sum_n^1 \mu_n, \sigma_n{}^2) \tag{9}$$

$$\mathrm{K-Fold\ CV}\ \sim N\big(\mu_{oy} + \textstyle\sum_y^1 \mu_y, \sigma_y{}^2\big) \tag{10}$$

μ_{on} and μ_{oy} were fit with normal priors $\left(N\left(\overline{K - Fold\ CV}, 5 * \sigma_{KFCV}{}^2\right)\right)$, σ_{KFCV} was defined as the standard deviation of the K-Fold CVs. The μ_n and μ_y's were given normal priors $\left(N\left(0.1,\ \sigma_{\mu}{}^2\right)\right)$, and σ_μ which functions as the within-group variance has a gamma prior with rate and shape estimated from the mode and standard deviation of the K-Fold CVs (Kruschke, 2014). Finally, σ_n and σ_y which represent the between-group variance were fit with a uniform prior $\left(uniform(100^{-1} * \sigma_{KFCV}, 10 * \sigma_{KFCV})\right)$. $\sum \mu_n$ and $\sum \mu_y$ were constrained to 0 when fitting μ_{on} and μ_{oy}.

2.3 Model Experimentation

Our state-space models were used to test the hypothesis that western Lake Erie TP concentrations are a linear function of Maumee River TP load when surface water movement is incorporated. We incorporate water movement into our linear model by first estimating the spatial impact of the Maumee River. The Maumee River impact plume was estimated by artificially reducing the Maumee River TP concentrations by 50% $(\grave{y}_{Maumee,t,y})$, the model was run again (Eq 5-7). The model output for each node was examined and the position of each node's concentration $(\grave{y}_{n,t,y})$ 95% predictive intervals (PI) was compared to the original model ($y_{n,t,y}$). The change from the original 95% PI of $y_{n,t,y}$, which we call the deflection, was interpreted as evidence that the Maumee River node was, at some time-step, influencing node *n*. The annual mean root squared sum of the $\grave{y}_{n,t,y}$ 95% PI change compared to the $y_{n,t,y}$ 95% PI was then normalized by the largest value for that model year (*y*), this normalized estimate of PI change ($d_{n,y}$) across the 254 nodes within our spatial window was used to define effective Maumee River spring TP impact within Lake Erie. We estimated Maumee River spring load estimates (l_y, tons TP) by multiplying NCWQR daily flow and TP concentration data (Table A1) from March 1 to July 31 annually. Finally, we multiplied $d_{n,y}$ and l_y to represent a spatially explicit effective Maumee River TP spring load at each node $(\grave{l}_{n,y})$.

A linear model of mean TP concentration ($\bar{y}_{n,y}$) per year per node (*n*, where node *n* had at least one observation) as a function of effective spring Maumee River TP load $(\grave{l}_{n,y})$ was used to test for a linear relationship between Maumee River load and Lake Erie surface water TP concentrations. The model was fit in a Bayesian framework which allowed us to fit the heteroskedastic relationship of concentration and effective load by fitting a positive linear relationship to model variance and effective load (Eq 11). $\beta_{1,2}$ were given uninformative normal priors (N(0,0.001)) while $\alpha_{1,2}$ were given uninformative log normal priors ($logN$(0,0.001)) because they must be positive random variables.

$$\bar{y}_{n,y} \sim N\left(\beta_1 + \beta_2 * \grave{l}_{n,y}, \left(\alpha_1 + \alpha_2 * \grave{l}_{n,y}\right)^2\right) \tag{11}$$

3 Results

The annual data sets defined by TP concentration observations and riverine TP data on our 2-km by 2-km grid in western Lake Erie contained an average of 99.1 % missing vales. The number of nodes that contained observations ranged from 14 to 40 among years. The mean number of samples available at each observed lake node during the model year ranged from 2 to 9. Within the 252 Lake Erie nodes across the available 11 years, a total of 1,218 observations were collected, our hierarchical spatially referenced Bayesian state-space model was then able to provide estimates for the 375,774 unobserved TP concentrations. Between the Maumee River and River Raisin, a total of 2,258 observations were available in the dataset and the missing 734 values were also described by posterior distributions.

3.1 State-Space Model Fit

To assess the efficiency of the model fit we determined annual Bayesian R^2, posterior predictive p-values, and k-fold cross validated utility. The 11 years of models had mean Bayesian R^2 values from 0.86 to ~1 and mean posterior predictive p-values from 0.42 to 0.59 (Table 2). Posterior predictive p-values of 0.5 indicate a good fit between model output and observations and the 95% CI of all our yearly posterior p-value distributions contain 0.5. Finally, the results of the k-fold cross validation utility 95% credible difference showed no difference across all pairwise comparisons of mean K-fold CV by year or node.

3.2 State-Space Model Outputs

Posterior distributions for each node on each day provide estimates for TP concentrations where observations are present and in the absence of observations (Fig. 2, Appendix B). Mean and 95% PI model posterior samples of each node at every day defined our predicted concentration. By example, the Maumee River node in 2018 shows the model following the data and widening PIs where observations are missing (Fig. 2a). For Lake Erie nodes that contained observations the posterior samples follow the broad trend in the observed data (Fig. 2b). Nodes without any observations also follow the trend in downstream observed nodes, and while the uncertainty is larger at unobserved nodes the PIs stay within expected values (Fig. 4c).

3.3 Model Experimentation

After artificially reducing the Maumee River concentrations by 50%, the nodes where TP concentration PIs were altered were defined as being within the Maumee River area of impact. The mean square root of each node's summed squared deflection annually normalized by the largest mean value ($d_{n,y}$) in general was highest near the mouth of the Maumee. The impacted area spread south and east along the State of Ohio coast most years, but some years were subject to larger plumes distributed further north (Fig. 3, Video Supplement 1).

The normalized annual mean Maumee impact estimates ($d_{n,y}$) generated per node were used to adjust spring load to an effective spring load $\left(\grave{l}_{n,y}\right)$ at each node where samples were collected. Lake Erie TP concentration was linearly correlated to the effective Maumee River TP spring load (Mean node concentration = 23.1 (± 1.75, 95% CI) + 0.11 (± 0.01,95% CI) * Effective Spring Load (tons TP); Fig. 4). The heteroskedastic error in the mean concentration ($\bar{y}_{n,y}$) and effective load $\left(\grave{l}_{n,y}\right)$ relationship was defined by a linear function $\left(\alpha_1 + \alpha_2 * \grave{l}_{n,y}\right)$. α_1 was estimated to be 2.9 (± 1.4,95% CI) and α_2 was 0.04 (± 0.008,95% CI).

4 Discussion

4.1 State-Space Model Fit

By amending the western Lake Erie TP observations with riverine data and surface currents within a Bayesian model framework we were able to generate estimates of TP across time and space. The models consistently generated plausible posterior samples for mean TP concentration as each 95% CI of annual posterior predictive p-values included 0.5 and annual Bayesian R^2 95% CI values ranged from 0.83 to 0.99 (Table 1). These indicators of model fit support the use of our framework in predicting water quality within large water bodies even with sparse observations within the data. The k-fold CV results generated by removing all the observations of a randomly selected node with at least 10 observations showed that model predictions were equally accurate across years and by node. Predicting equally well across the nodes and within any year additionally supports this framework as being a useful application of Bayesian methods in water quality modelling.

4.2 State-Space Model Output

An important property of this modelling approach is that the surface current derived adjacency matrix (k) we used to define our predicted spatially explicit latent state concentration ($x_{n,t,y}$) also produced estimates of TP concentration at nodes where no observations were available. This approach takes discrete measurements in western Lake Erie and establishes connections across the lake surface and through the model year. The model does this across 136 days and 254 2km by 2km nodes, yet model uncertainties are within the range of TP concentrations expected for western Lake Erie. The model framework allows information from discrete grab samples to be shared across any waterbody where the movement pattern of water is available. This modelling technique could be applied at lower temporal resolution (weeks or months) with broadly defined patterns of water movement if daily fine scale surface current data are unavailable. Our model also can generate estimates at unobserved node or at unobserved time-steps of observed nodes without requiring defined biogeochemical processes of a mechanistic model.

Within Lake Erie, having estimates for unobserved nodes and nodes that are infrequently sampled allows a connection between discrete point data collected by boat and data layers which cover large sections of lake surface. The spatial distance and temporal disconnect between the data generated by multiple actors on the same waterbody often precludes the combined use of data from multiple projects. However, state-space models which explicitly incorporate time as well as space via water movement could harness more of the available data to make predictions beyond the spatial bounds of the original projects. For example, remote sensed data layers would be especially useful in this model structure, both as predictor variables and as the response variables. Connecting estimates of TP concentrations to chlorophyll-a concentrations would enhance existing predictive models of cyanobacteria distribution and biomass (Fang et al., 2019).

The agencies and organizations collecting grab samples within Lake Erie would also be able to use the state-space model output to select sample locations specific to hypotheses. The model can be used to predict the movement of high nutrient water masses which investigators could target. Additionally, projects examining the impact of the Maumee River could sample in and out of the Maumee River impact plume. Beyond the impact to field work, this modelling approach can also be used in model selection. Since this modelling approach is based entirely on observations in the absence of independent explanatory variables, it should be used as a benchmark model for future mechanistic or more complex models to be tested against.

4.3 Model Experimentation

Having demonstrated the functional capacity of our hierarchical spatially referenced Bayesian state-space model predicting TP concentrations, our hypothesis was that a linear relationship exists between spring Maumee River load and observed Lake Erie concentrations. By experimentally reducing the concentrations for the Maumee River and rerunning the model we were able to track the "downstream" repercussions to the lake node predicted values and infer the Maumee River impact plume. Tracking a plume of TP impact using the grab samples was not previously possible because of the distances between sampling locations and the fact that the number of unobserved days outnumber the observed days. The plume extent in general follows the southern coast (Fig. 3), which would be expected because of the movement associated with the Coriolis effect. Importantly, this is not a plume that displays high concentration of TP, rather this is the impact plume of the Maumee River. Concentrations outside the impact plume are not influenced, or weakly influenced, by the Maumee and thus load reductions within the Maumee River would not impact lake concentrations in those areas. Our linear model (Eq 11) estimated that when the effective load of the Maumee was 0, the mean annual concentration in the area where samples were collected would be 23.1 µg l^{-1} ± 1.75, 95% CI.

Each year the Maumee River TP impact plume dimension and intensity changed. Rowland et al. (2019) demonstrated how a linear model of Lake Erie TP observations as a function of Maumee River spring loads defined positive relationships at the closest stations. Here, we were able to fit parameters that define the load to concentration relationship across all western Lake Erie. Much of the regulatory attention in addressing Lake Erie HABs has focused on Maumee River spring export and providing this quantitative connection is important in furthering watershed TP reduction efforts. Our model estimated that for each 100 tons of spring TP effective load delivered to Lake Erie, TP concentrations in the lake increase by 11 µgl l^{-1} (± 1, 95% CI). We could use our defined linear relationship for hindcasting expected concentration reductions in western Lake Erie based on Maumee spring TP loads which were reduced by 40% for all our model years. Additionally, given a mean concentration maximum, we could predict the load reductions required in previous years to meet that target. Using our linear relationship between lake concentration and spring load to make forecasts for future years is harder. The size and shape of the Maumee River impact ($d_{n,y}$) changes each year (Video supplement 1) and our method defines the river impact from observations. Without being provided an estimate of $d_{n,y}$, a forecast of mean western Lake Erie TP concentrations based on a proposed spring TP load is not achievable. An achievable next step for this modelling framework could be linking the size of the Maumee River plume and Lake Erie TP concentrations to HABs biomass and toxin production, the spatial aspect of such a model could explain why the relationship between bloom biomass and Maumee TP export is not linear (Obenour et al., 2014).

5 Conclusions

Our state-space model framework was shown to fit the data well, generated reasonable estimates of concentration at observed and unobserved locations, was modified experimentally to estimate a river impact plume, and used the experimentally derived plume to test a regulatory relevant hypothesis. Adequately characterizing water quality in a large waterbody is difficult. Sampling and laboratory analysis is expensive and too time consuming to feasibly cover even a portion of Lake Erie with high temporal and spatial resolution. However, we demonstrate that a Bayesian state-space framework informed by an adjacency matrix defined by surface currents can generate daily TP concentration which are constrained by uncertainties appropriate for lake conditions. By amending the data from two rivers entering the lake our model (Eq 5-7) enables the rivers to inform the observed and unobserved lake nodes, the observed lake nodes inform the unobserved lake nodes, and unobserved lake nodes also inform unobserved and observed nodes. This information sharing across time and space empowers this model to connect sparse data across large distances. By experimenting with the model, we were able to estimate a plume of impact from the Maumee River and apply the experimental results to hypothesis testing. The model is amenable to using remote sensing data and can effectively connect lake wide datasets with discrete grab samples. The application here used TP, but any analyte could be modelled in this same structure to generate estimates through time and space, hypothesis test, or to build baseline models to test process-based models against.

Code and Data availability

Template code for reproducing our model is available publicly on Zenodo; at doi: 10.5281/zenodo.4884997. All the data used here were from publicly available sources which we provide in Appendix A. On the Zenodo site we made our curated data for 2018 available.

Video supplement

A supplemental video of Maumee River impact plume through time is available through Copernicus.

Author contribution

TJM, CAS, and CMG designed the models, the model fitting quality control and the model experiments. TJM developed the model code and performed the simulations. TJM prepared the manuscript with contributions from CAS, and CMG.

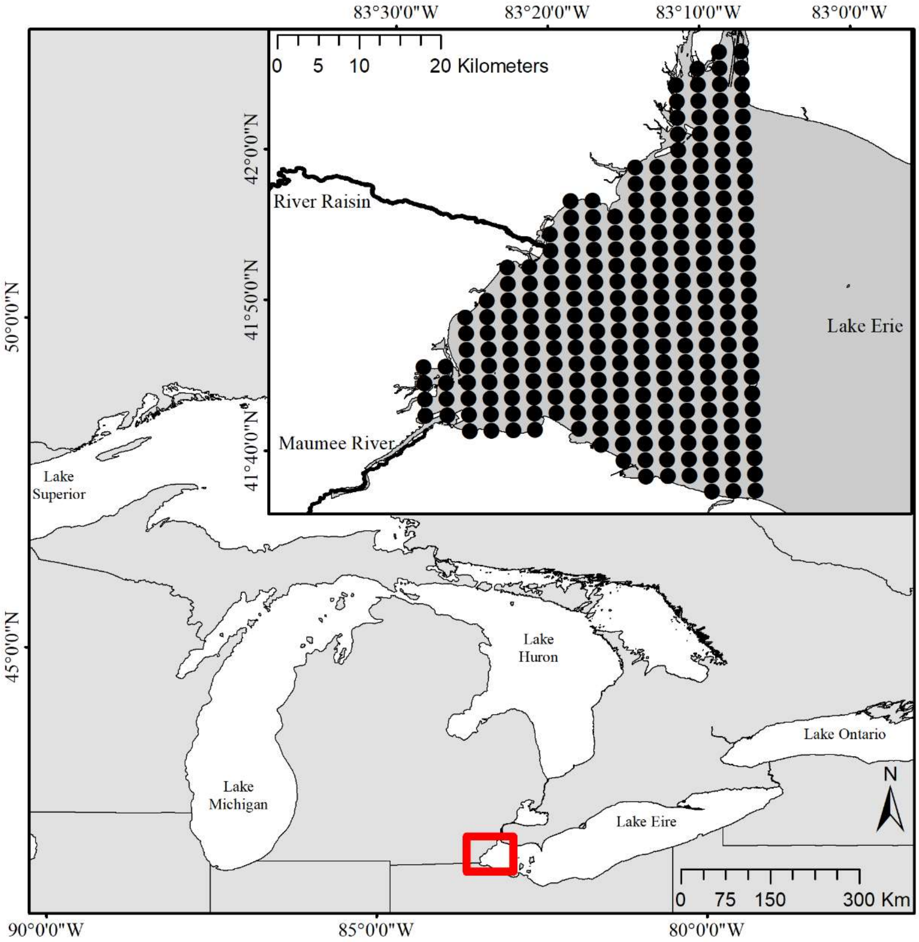
Competing interests

The authors declare that they have no conflict of interest.

Acknowledgements

The authors would like to acknowledge Peter Alsip of University of Michigan Cooperative Institute for Great Lakes Research (CIGLR) for collating the data from sources listed in Appendix A. The authors would also like to acknowledge the field personnel of the CIGLR/NOAA GLERL WLE sampling team whose field work inspired this manuscript. Funding was awarded to the CIGLR through the NOAA Cooperative Agreement with the University of Michigan (NA17OAR4320152).

https://doi.org/10.5194/hess-2021-296
Preprint. Discussion started: 2 June 2021
© Author(s) 2021. CC BY 4.0 License.

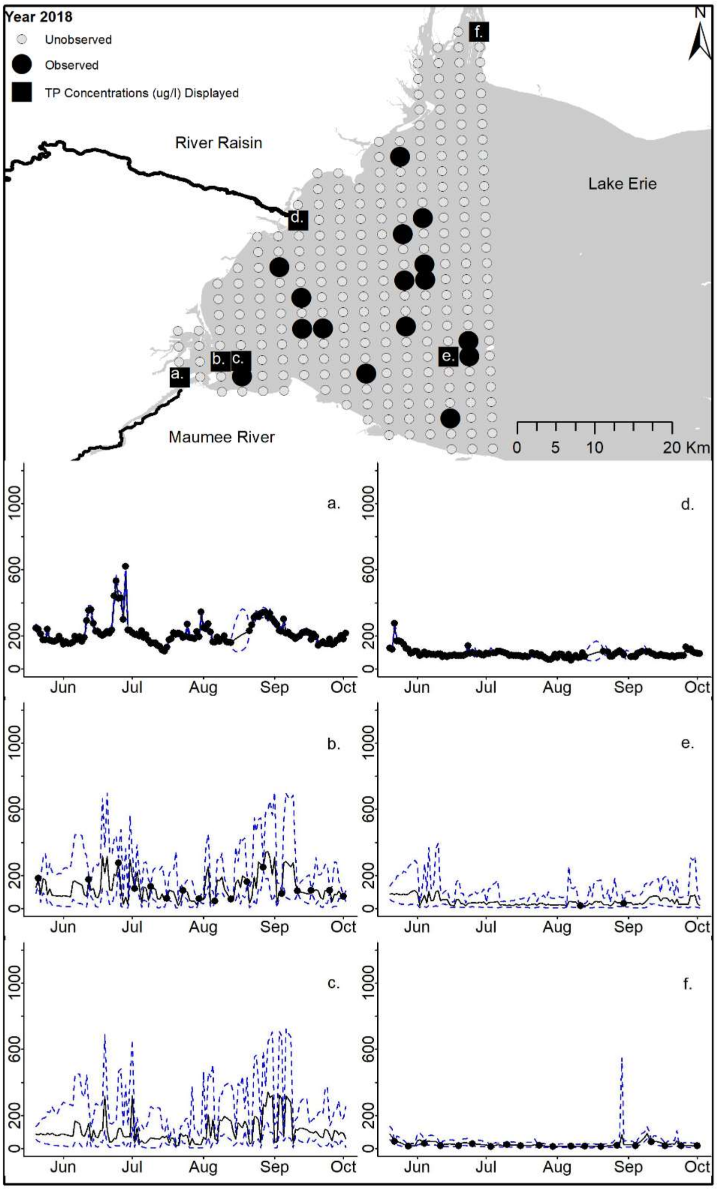


325 **Figure 1. Map showing the location of study region in western Lake Erie. The inset map shows the tributaries and loke nodes that were included in the model. Our site boundary was defined by the western portions of Lake Erie. A grid of 2 km-by-2 km nodes was used to snap existing concentration data and define an adjacency matrix based on surface currents.**

330 **Table 1. Bayesian model assessment via p-value (posterior predictive p-values of 0.5 are indicative of a good fit and 95% credible intervals (CI) of our yearly results each containing 0.5) and R^2 (each year > 0.8) showed the model generated posterior samples similar in structure to the observations.**

Year	Posterior Predictive p-values		R^2	
	95% CI		95% CI	
2008	0.4	0.6	0.99	0.999
2009	0.49	0.68	0.915	0.961
2010	0.4	0.59	0.835	0.882
2011	0.37	0.56	0.993	1
2012	0.4	0.6	0.994	1
2013	0.32	0.51	0.984	0.996
2014	0.4	0.59	0.988	0.999
2015	0.37	0.57	0.95	0.984
2016	0.41	0.6	0.933	0.979
2017	0.41	0.61	0.994	1
2018	0.39	0.59	0.973	0.999

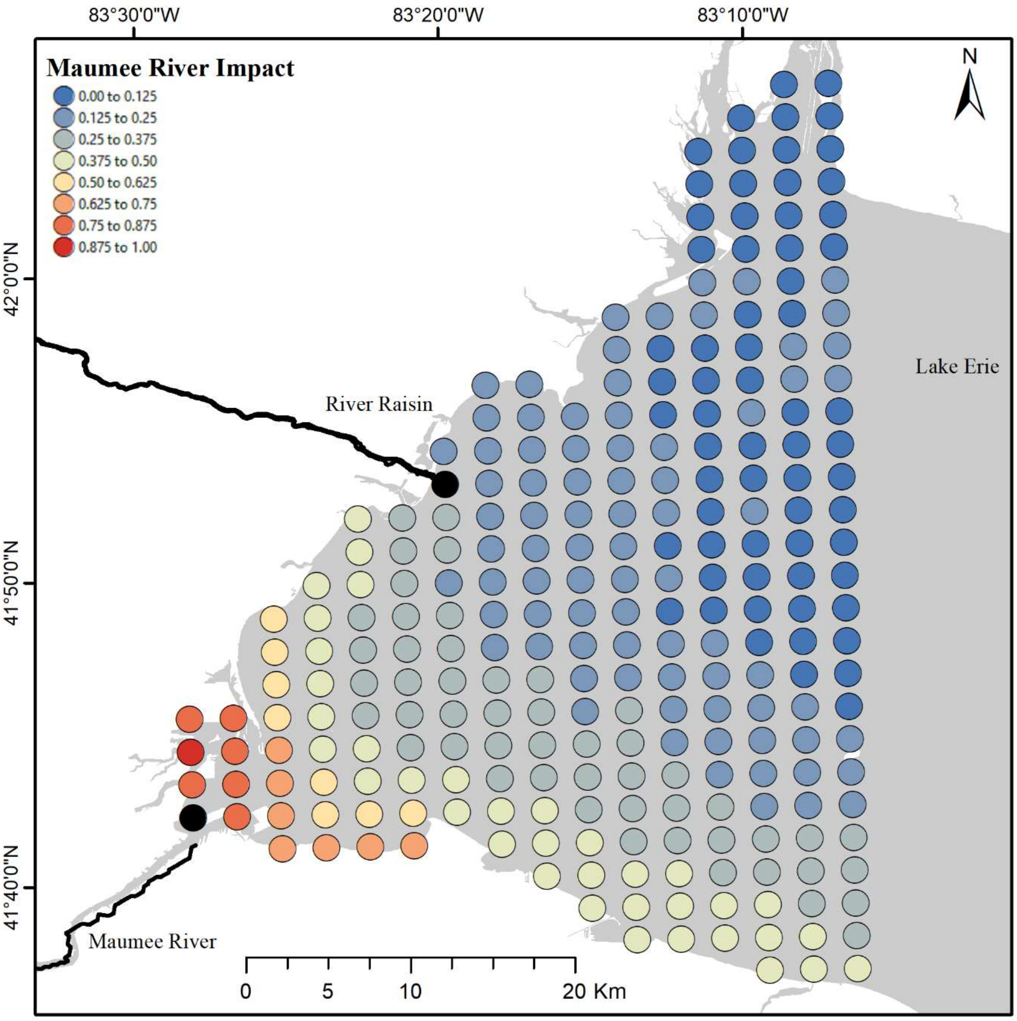
© Author(s) 2021. CC BY 4.0 License.



335

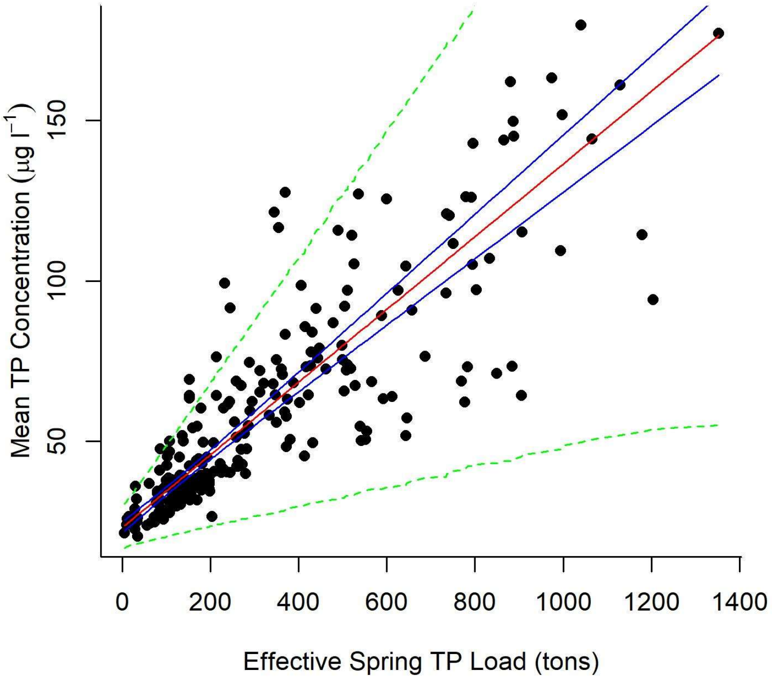
Figure 2. For 2018 the total phosphorus concentration (µg l^{-1}) at observed and unobserved nodes were estimated from the model posterior samples. Mean (solid black line) and 95% PI (dashed blue line) for the model posterior samples of each node at every day for (a) the Maumee River, (b,c,e,f) western Lake Erie nodes, and (d) the River Raisin.

https://doi.org/10.5194/hess-2021-296
Preprint. Discussion started: 2 June 2021
© Author(s) 2021. CC BY 4.0 License.



340

Figure 3. Heatmap of the mean Maumee River impact plume from 2008 to 2018.



345 **Figure 4. Mean concentration at the observed nodes for each year was modelled as a function of the relative Maumee River spring TP load (Mean concentration = 23.1 (± 1.75, 95% CI) + 0.11 (± 0.01,95% CI) * Effective Load), where variance in concentration increased linearly with effective load. Effective load was defined by multiplying the normalized river impact generated by experimentally tracing the Maumee River's impact on Lake Erie nodes annually. 95% predictive intervals of the data (green dotted lines) and 95% credible intervals of the linear relationship (blue solid lines) were generated from the model output.**

350 **Appendices**

Appendix A

Table A1. The data sources for total phosphorus concentrations and surface currents were all retrieved from publicly available online repositories.

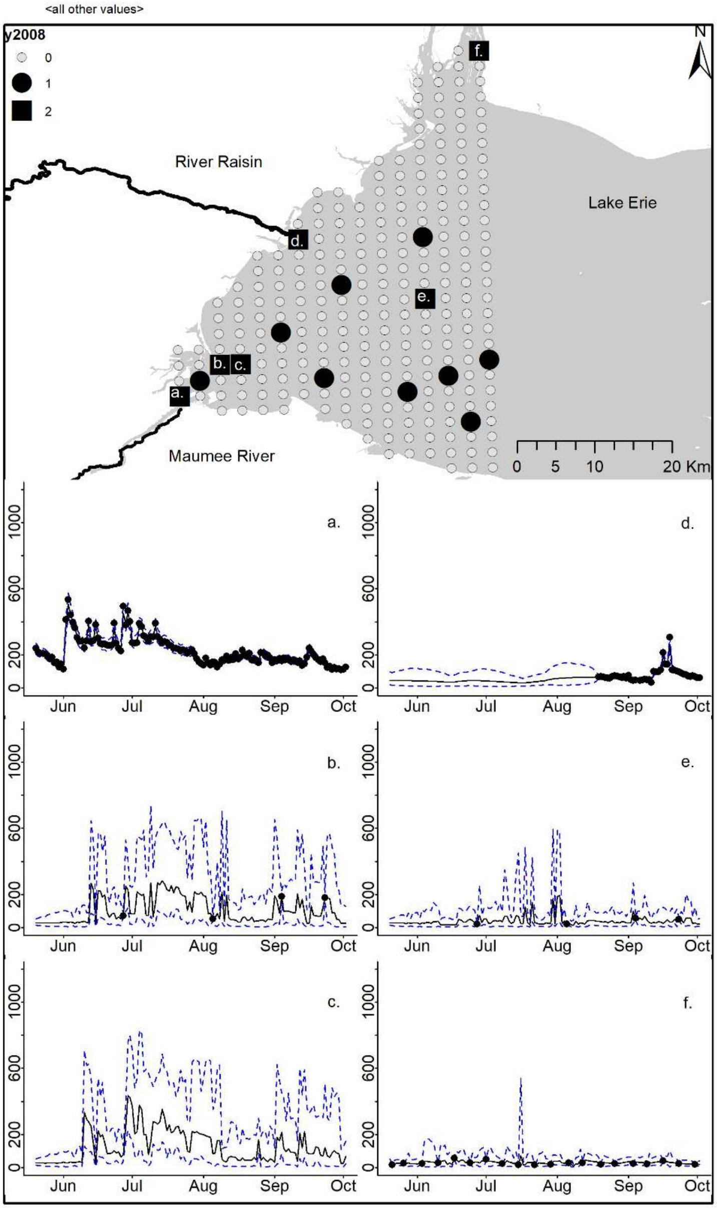
Agency	Link	Data Type	n (2008 to 2018)
Environment Climate Change Canada's Offshore Water Quality Survey	Digital Object Identifier: 10.18164/495eb10d-d423-432a-980f-264ef287d45b	Total Phosphorus Concentration (µg l^{-1})	121
U.S. Environmental Protection Agency's Great Lakes National Program Office	https://cdx.epa.gov/	Total Phosphorus Concentration (µg l^{-1})	149
Ministry of the Environment, Conservation and Parks Great Lakes Intake Program	http://files.ontario.ca/moe_mapping/ downloads/2Water/GLIP/All_Lakes_GLIP.csv	Total Phosphorus Concentration (µg l^{-1})	637
National Oceanographic and Atmospheric Administration (NOAA) Great Lakes Environmental Research Laboratory (GLERL) Ecosystem Dynamics Long-Term Research program	Digital Object Identifier: doi.org/10.25921/11da-3x54	Total Phosphorus Concentration (µg l^{-1})	111
NOAA GLERL Western Lake Erie Sampling	Digital Object Identifier: doi.org/10.25921/11da-3x54	Total Phosphorus Concentration (µg l^{-1})	1145

National Center for Water Quality Research at Heidelberg University	https://ncwqr-data.org/	Total Phosphorus Concentration (μg l^{-1})	2258
NOAA Great Lakes Coastal Forecasting System	https://www.glerl.noaa.gov/res/glcfs/	Surface Currents (m North, m East)	1020318

https://doi.org/10.5194/hess-2021-296
Preprint. Discussion started: 2 June 2021
© Author(s) 2021. CC BY 4.0 License.

Appendix B

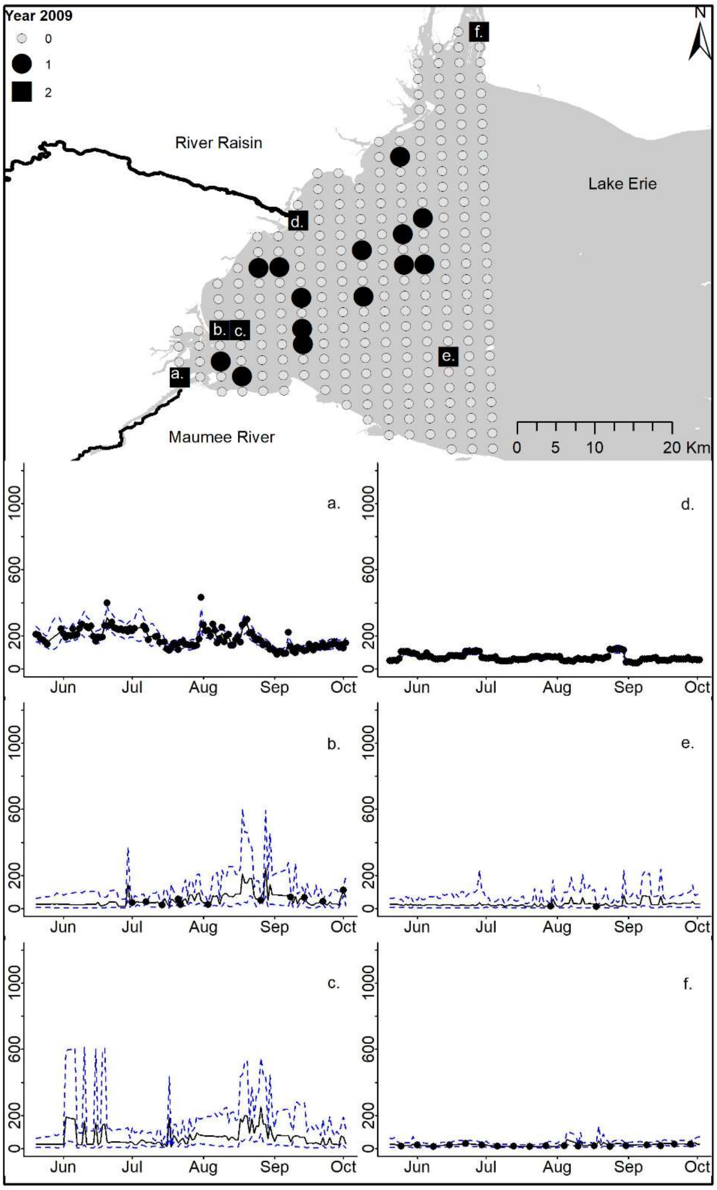
https://doi.org/10.5194/hess-2021-296
Preprint. Discussion started: 2 June 2021
© Author(s) 2021. CC BY 4.0 License.



© Author(s) 2021. CC BY 4.0 License.

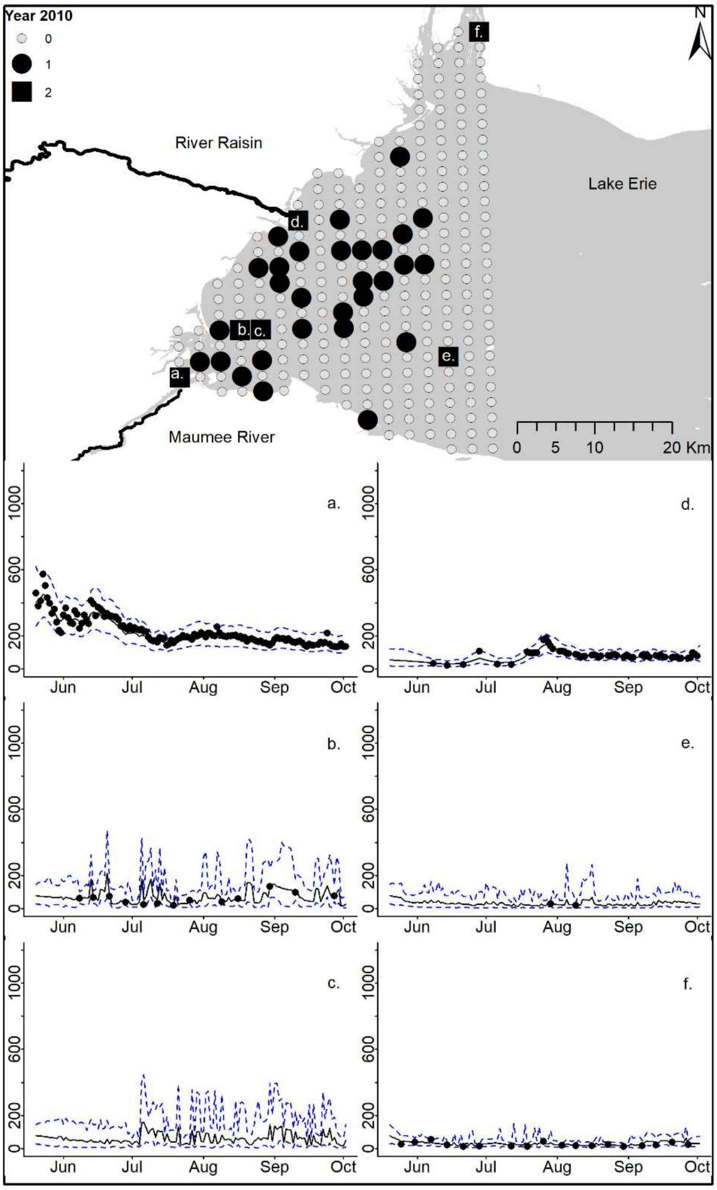
Figure B1. For 2008 the total phosphorus concentration (µg l^{-1}) at observed and unobserved nodes were estimated from the model posterior samples. Mean (solid black line) and 95% PI (dashed blue line) for the model posterior samples of each node at every day for (a) the Maumee River, (b,c,e,f) western Lake Erie nodes, and (d) the River Raisin.

https://doi.org/10.5194/hess-2021-296
Preprint. Discussion started: 2 June 2021
© Author(s) 2021. CC BY 4.0 License.



360

Figure B2. For 2009 the total phosphorus concentration (µg l^{-1}) at observed and unobserved nodes were estimated from the model posterior samples. Mean (solid black line) and 95% PI (dashed blue line) for the model posterior samples of each node at every day for (a) the Maumee River, (b,c,e,f) western Lake Erie nodes, and (d) the River Raisin.


Year 2010
0
1
2
N
River Raisin
Lake Erie
Maumee River
a.
b.
c.
d.
e.
f.
0 5 10 20 Km
0
200
600
1000
Jun
Jul
Aug
Sep
Oct

365 **Figure B3. For 2010 the total phosphorus concentration (µg l^{-1}) at observed and unobserved nodes were estimated from the model posterior samples. Mean (solid black line) and 95% PI (dashed blue line) for the model posterior samples of each node at every day for (a) the Maumee River, (b,c,e,f) western Lake Erie nodes, and (d) the River Raisin.**

https://doi.org/10.5194/hess-2021-296
Preprint. Discussion started: 2 June 2021
© Author(s) 2021. CC BY 4.0 License.

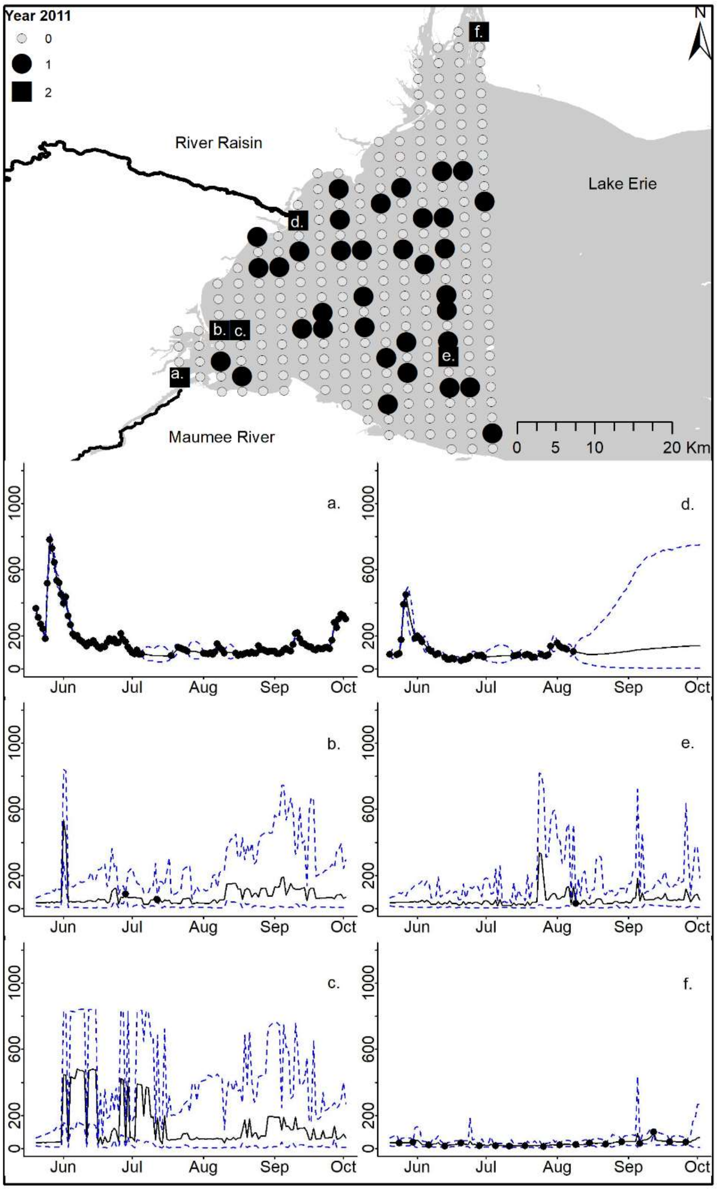


Figure B4. For 2011 the total phosphorus concentration (µg l^{-1}) at observed and unobserved nodes were estimated from the model 370 **posterior samples. Mean (solid black line) and 95% PI (dashed blue line) for the model posterior samples of each node at every day for (a) the Maumee River, (b,c,e,f) western Lake Erie nodes, and (d) the River Raisin.**

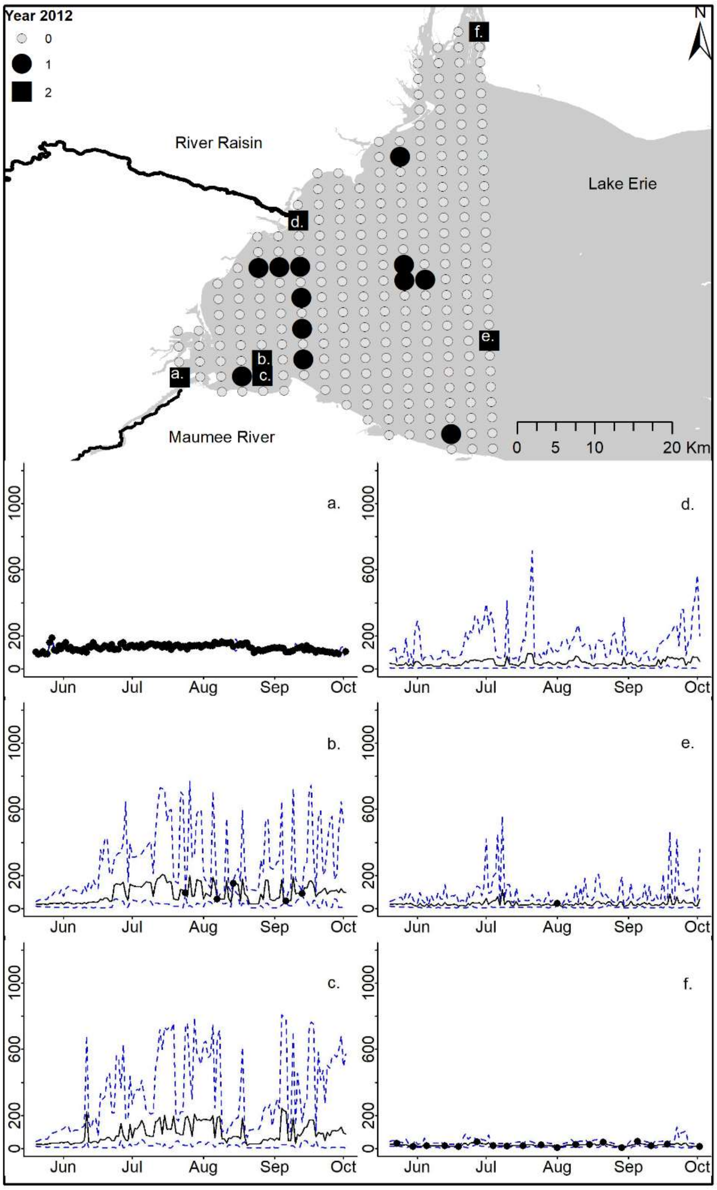

Year 2012
0
1
2
N
River Raisin
Lake Erie
Maumee River
a.
b.
c.
d.
e.
f.
0 5 10 20 Km
0
200
600
1000
Jun
Jul
Aug
Sep
Oct

Figure B5. For 2012 the total phosphorus concentration (µg l^{-1}) at observed and unobserved nodes were estimated from the model posterior samples. Mean (solid black line) and 95% PI (dashed blue line) for the model posterior samples of each node at every day for (a) the Maumee River, (b,c,e,f) western Lake Erie nodes, and (d) the River Raisin.

https://doi.org/10.5194/hess-2021-296
Preprint. Discussion started: 2 June 2021
© Author(s) 2021. CC BY 4.0 License.

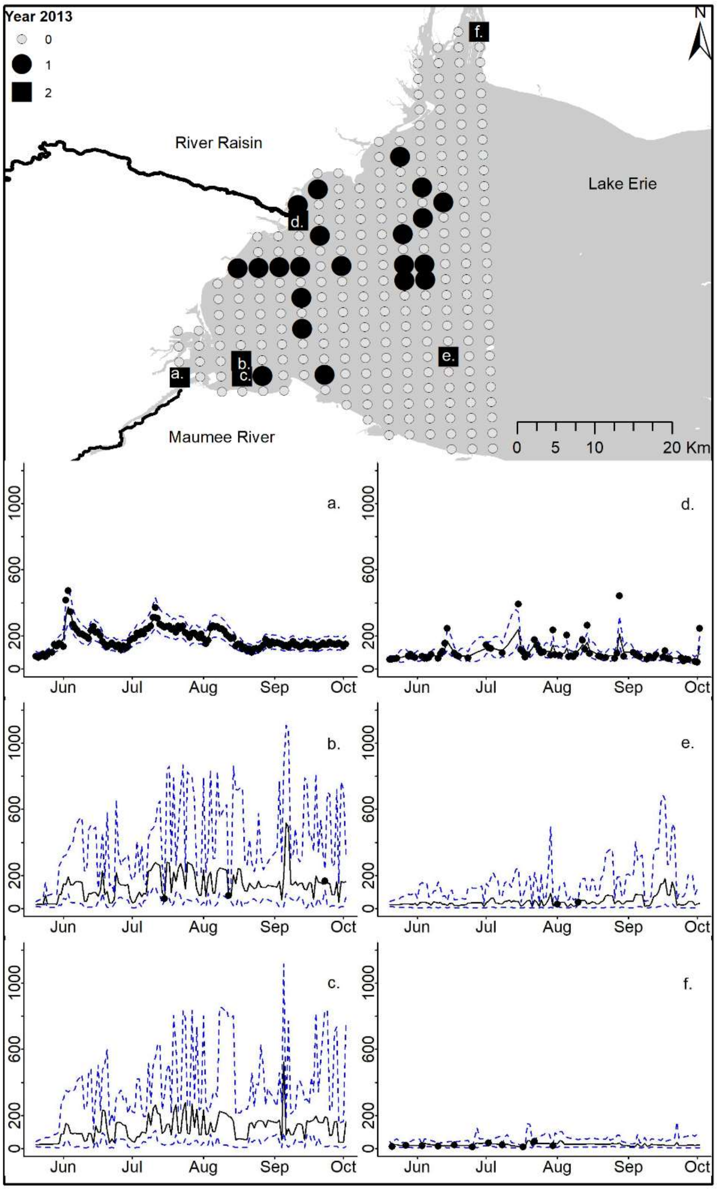
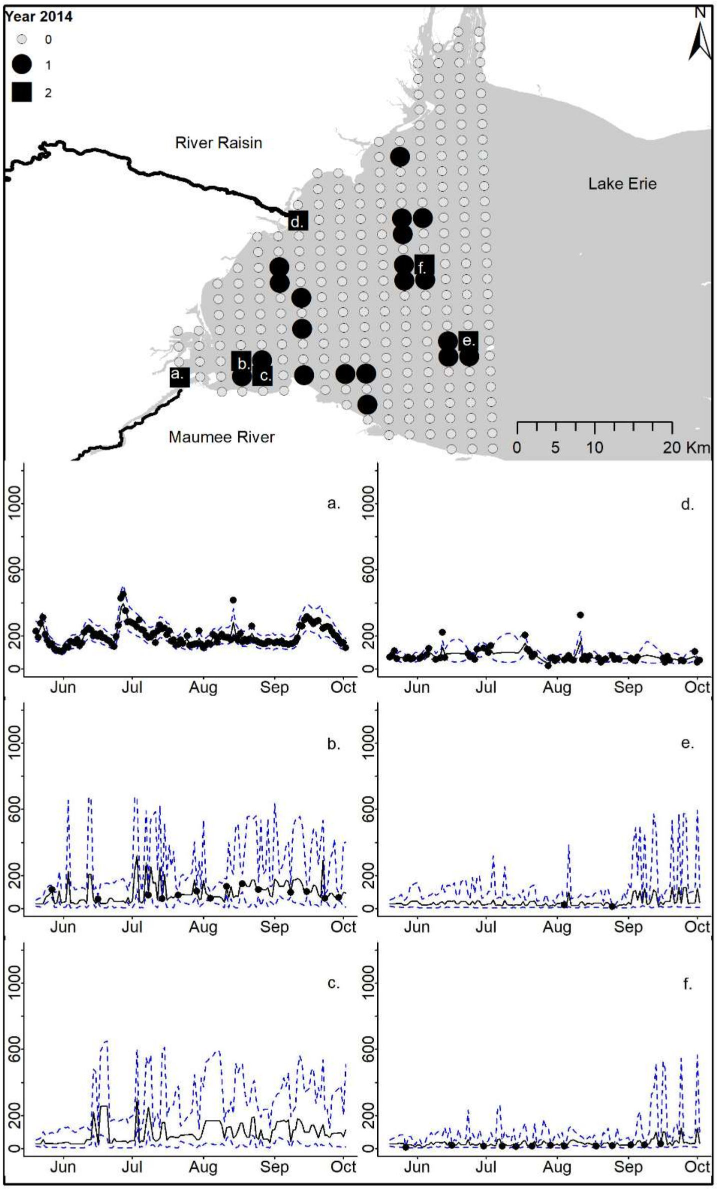
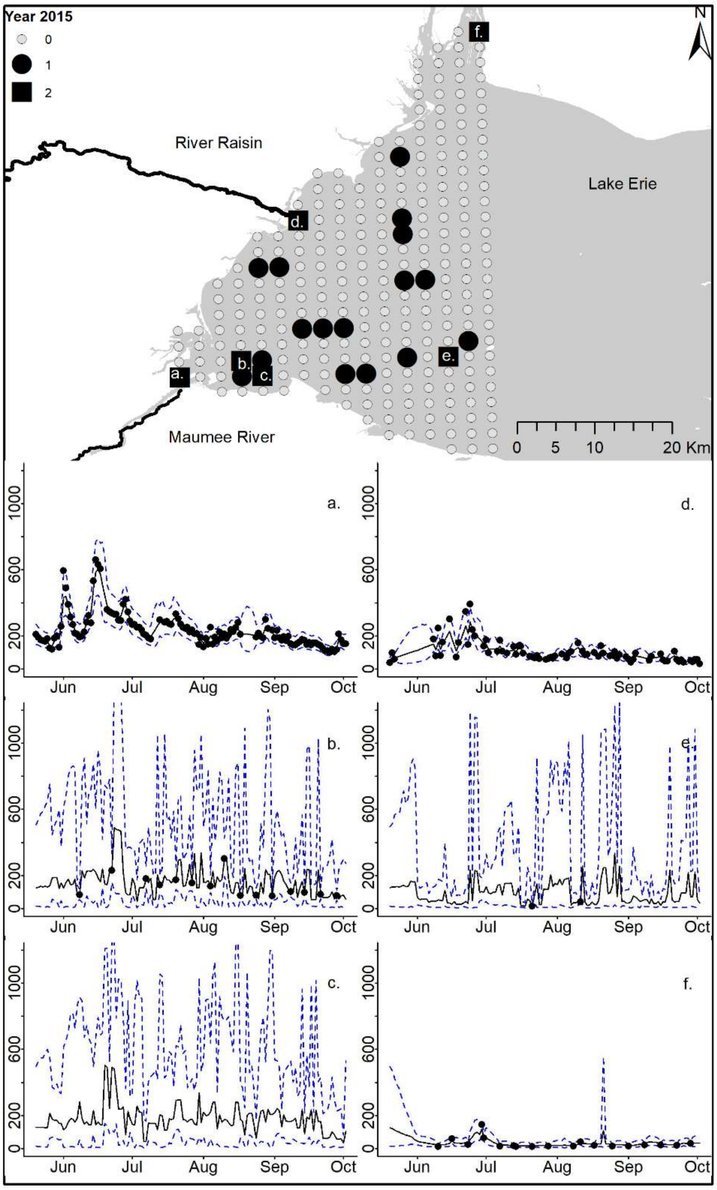


Figure B6. For 2013 the total phosphorus concentration (µg l^{-1}) at observed and unobserved nodes were estimated from the model posterior samples. Mean (solid black line) and 95% PI (dashed blue line) for the model posterior samples of each node at every day for (a) the Maumee River, (b,c,e,f) western Lake Erie nodes, and (d) the River Raisin.


Year 2014
0
1
2
N
River Raisin
Lake Erie
Maumee River
a.
b.
c.
d.
e.
f.
0 5 10 20 Km
0
200
600
1000
Jun
Jul
Aug
Sep
Oct

380

Figure B7. For 2014 the total phosphorus concentration (µg l^{-1}) at observed and unobserved nodes were estimated from the model posterior samples. Mean (solid black line) and 95% PI (dashed blue line) for the model posterior samples of each node at every day for (a) the Maumee River, (b,c,e,f) western Lake Erie nodes, and (d) the River Raisin.


Year 2015
0
1
2
N
River Raisin
Lake Erie
Maumee River
a.
b.
c.
d.
e.
f.
0 5 10 20 Km
0
200
600
1000
Jun
Jul
Aug
Sep
Oct

385 **Figure B8. For 2015 the total phosphorus concentration (µg l^{-1}) at observed and unobserved nodes were estimated from the model posterior samples. Mean (solid black line) and 95% PI (dashed blue line) for the model posterior samples of each node at every day for (a) the Maumee River, (b,c,e,f) western Lake Erie nodes, and (d) the River Raisin.**

https://doi.org/10.5194/hess-2021-296
Preprint. Discussion started: 2 June 2021
© Author(s) 2021. CC BY 4.0 License.

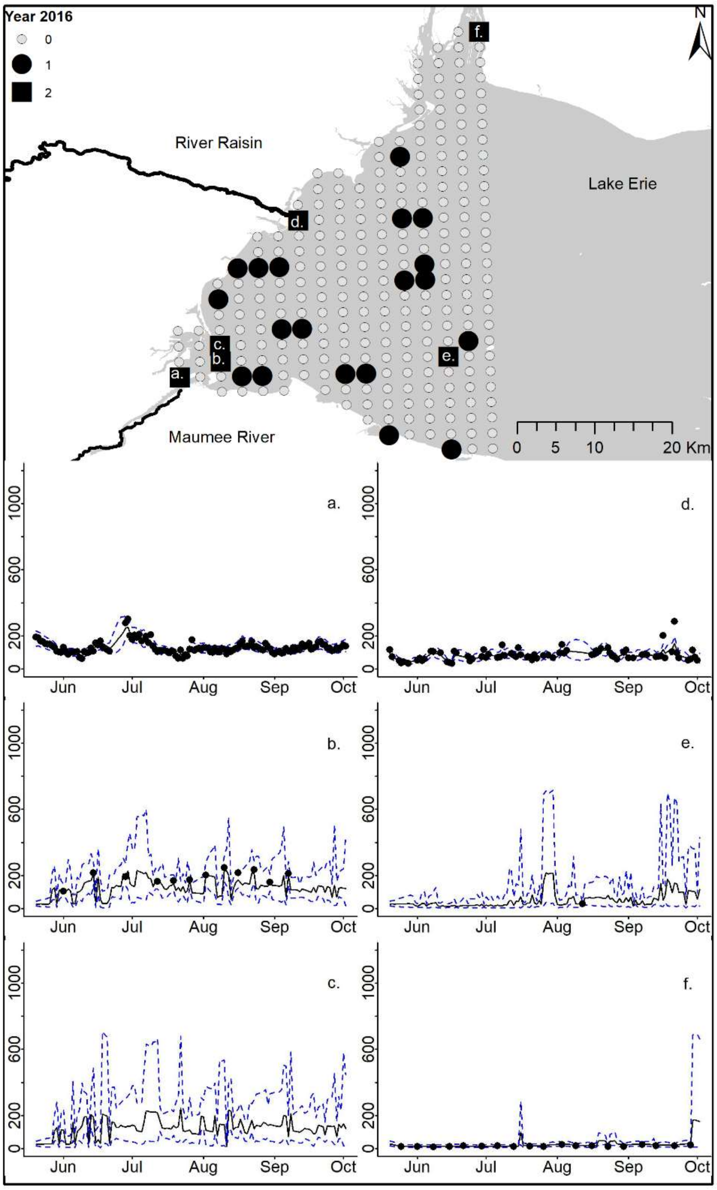


Figure B9. For 2016 the total phosphorus concentration (µg l^{-1}) at observed and unobserved nodes were estimated from the model 390 posterior samples. Mean (solid black line) and 95% PI (dashed blue line) for the model posterior samples of each node at every day for (a) the Maumee River, (b,c,e,f) western Lake Erie nodes, and (d) the River Raisin.

https://doi.org/10.5194/hess-2021-296
Preprint. Discussion started: 2 June 2021
© Author(s) 2021. CC BY 4.0 License.

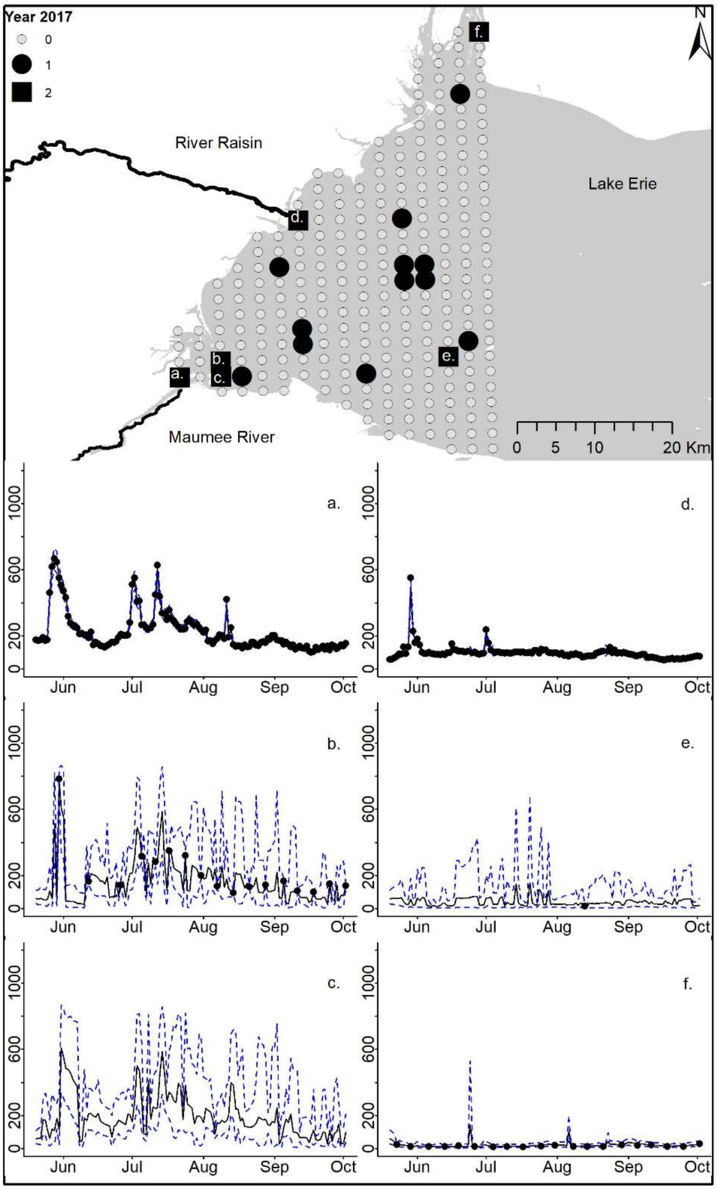


Figure B10. For 2017 the total phosphorus concentration (µg l^{-1}) at observed and unobserved nodes were estimated from the model posterior samples. Mean (solid black line) and 95% PI (dashed blue line) for the model posterior samples of each node at every day 395 **for (a) the Maumee River, (b,c,e,f) western Lake Erie nodes, and (d) the River Raisin.**

References

Bolsenga, S. J. and Herdendorf, C. E.: Lake Erie and Lake St. Clair Handbook, Wayne State University Press., 1993.

Brooks, B. W., Lazorchak, J. M., Howard, M. D. A., Johnson, M. V, Morton, S. L., Perkins, D. A. K., Reavie, E. D., Scott, G. I., Smith, S. A. and Steevens, J. A.: Are harmful algal blooms becoming the greatest inland water quality threat to public health and aquatic ecosystems?, Environ. Toxicol. Chem., 35(1), 6–13, 2016.

Eddelbuettel, D.: random: True Random Numbers using RANDOM.ORG, [online] Available from: https://cran.r-project.org/package=random, 2017.

Fang, S., Del Giudice, D., Scavia, D., Binding, C. E., Bridgeman, T. B., Chaffin, J. D., Evans, M. A., Guinness, J., Johengen, T. H. and Obenour, D. R.: A space-time geostatistical model for probabilistic estimation of harmful algal bloom biomass and areal extent, Sci. Total Environ., 695, 133776, 2019.

Geisser, S. and Eddy, W. F.: A predictive approach to model selection, J. Am. Stat. Assoc., 74(365), 153–160, 1979.

Gelman, A.: Two simple examples for understanding posterior p-values whose distributions are far from uniform, Electron. J. Stat., 7, 2595–2602, 2013.

Gelman, A., Goodrich, B., Gabry, J. and Vehtari, A.: R-squared for Bayesian regression models, Am. Stat., 2019.

GLWQA: Great Lakes Water Quality Agreement; Protocol Amending the Agreement Between Canada and the United States of America on Great Lakes Water Quality, 1978, as Amended on October 16, 1983 and on November 18, 1987, 2012.

Ho, J. C. and Michalak, A. M.: Phytoplankton blooms in Lake Erie impacted by both long-term and springtime phosphorus loading, J. Great Lakes Res., 43(3), 221–228, 2017.

Kruschke, J.: Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan, 2014.

Li, J., Ianaiev, V., Huff, A., Zalusky, J., Ozersky, T. and Katsev, S.: Benthic invaders control the phosphorus cycle in the world's largest freshwater ecosystem, Proc. Natl. Acad. Sci., 118(6), 2021.

Matisoff, G., Kaltenberg, E. M., Steely, R. L., Hummel, S. K., Seo, J., Gibbons, K. J., Bridgeman, T. B., Seo, Y., Behbahani, M. and James, W. F.: Internal loading of phosphorus in western Lake Erie, J. Great Lakes Res., 42(4), 775–788, 2016.

Michalak, A. M., Anderson, E. J., Beletsky, D., Boland, S., Bosch, N. S., Bridgeman, T. B., Chaffin, J. D., Cho, K., Confesor, R. and Daloğlu, I.: Record-setting algal bloom in Lake Erie caused by agricultural and meteorological trends consistent with expected future conditions, Proc. Natl. Acad. Sci., 110(16), 6448–6452, 2013.

Microsoft Corporation and Weston, S.: doParallel: Foreach Parallel Adaptor for the "parallel" Package, [online] Available from: https://cran.r-project.org/package=doParallel, 2020.

Obenour, D. R., Gronewold, A. D., Stow, C. A. and Scavia, D.: Using a B ayesian hierarchical model to improve L ake E rie cyanobacteria bloom forecasts, Water Resour. Res., 50(10), 7847–7860, 2014.

Piironen, J. and Vehtari, A.: Comparison of Bayesian predictive methods for model selection, Stat. Comput., 27(3), 711–735, 2017.

Plummer, M.: rjags: Bayesian Graphical Models using MCMC, [online] Available from: https://cran.r-project.org/package=rjags, 2019.

Rowe, M. D., Anderson, E. J., Beletsky, D., Stow, C. A., Moegling, S. D., Chaffin, J. D., May, J. C., Collingsworth, P. D., Jabbari, A. and Ackerman, J. D.: Coastal upwelling influences hypoxia spatial patterns and nearshore dynamics in Lake Erie, J. Geophys. Res. Ocean., 124(8), 6154–6175, 2019.

Rowland, F. E., Stow, C. A., Johengen, T. H., Burtner, A. M., Palladino, D., Gossiaux, D. C., Davis, T. W., Johnson, L. T. and Ruberg, S.: Recent patterns in Lake Erie phosphorus and chlorophyll a concentrations in response to changing loads, Environ. Sci. Technol., 54(2), 835–841, 2019.

Schneider, M. and Bláha, L.: Advanced oxidation processes for the removal of cyanobacterial toxins from drinking water, Environ. Sci. Eur., 32(1), 1–24, 2020.

Schwab, D. J., Beletsky, D., DePinto, J. and Dolan, D. M.: A hydrodynamic approach to modeling phosphorus distribution in Lake Erie, J. Great Lakes Res., 35(1), 50–60, 2009.

Steffen, M. M., Davis, T. W., McKay, R. M. L., Bullerjahn, G. S., Krausfeldt, L. E., Stough, J. M. A., Neitzey, M. L., Gilbert, N. E., Boyer, G. L. and Johengen, T. H.: Ecophysiological Examination of the Lake Erie Microcystis Bloom in 2014: Linkages between Biology and the Water Supply Shutdown of Toledo, OH, Environ. Sci. Technol., 51(12), 6745–6755, 2017.

Stow, C. A., Cha, Y., Johnson, L. T., Confesor, R. and Richards, R. P.: Long-term and seasonal trend decomposition of Maumee River nutrient inputs to western Lake Erie, Environ. Sci. Technol., 49(6), 3392–3400, 2015.

Vehtari, A., Gelman, A. and Gabry, J.: Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC, Stat. Comput., 27(5), 1413–1432, 2017.

Watson, S. B., Miller, C., Arhonditsis, G., Boyer, G. L., Carmichael, W., Charlton, M. N., Confesor, R., Depew, D. C., Höök, T. O. and Ludsin, S. A.: The re-eutrophication of Lake Erie: Harmful algal blooms and hypoxia, Harmful Algae, 56, 44–66, 2016.

Wituszynski, D. M., Hu, C., Zhang, F., Chaffin, J. D., Lee, J., Ludsin, S. A. and Martin, J. F.: Microcystin in Lake Erie fish: risk to human health and relationship to cyanobacterial blooms, J. Great Lakes Res., 43(6), 1084–1090, 2017.