Ecological Modelling 430 (2020) 109136

Contents lists available at ScienceDirect

Ecological Modelling

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

Original research paper

Predicting lake surface water phosphorus dynamics using process-guided machine learning

Paul C. Hanson[1,*], Aviah B. Stillman[1], Xiaowei Jia[2], Anuj Karpatne[3], Hilary A. Dugan[1], Cayelan C. Carey[4], Joseph Stachelek[5], Nicole K. Ward[4], Yu Zhang[6,7], Jordan S. Read[8], Vipin Kumar[2]

[1] University of Wisconsin-Madison, Center for Limnology
[2] University of Minnesota, Department of Computer Science and Engineering
[3] Virginia Tech, Department of Computer Science
[4] Virginia Tech, Department of Biological Sciences
[5] Department of Fisheries and Wildlife, Michigan State University, East Lansing, Michigan
[6] Department of Civil and Environmental Engineering, Penn State University, University Park, PA, USA, 16802
[7] Earth and Environmental Sciences Division, Los Alamos National Laboratory, Los Alamos, NM, USA, 87545
[8] U.S. Geological Survey

ARTICLE INFO

Keywords:
Phosphorus
Lake Mendota
Model
Machine learning
Lake
Long-term

ABSTRACT

Phosphorus (P) loading to lakes is degrading the quality and usability of water globally. Accurate predictions of lake P dynamics are needed to understand whole-ecosystem P budgets, as well as the consequences of changing lake P concentrations for water quality. However, complex biophysical processes within lakes, along with limited observational data, challenge our capacity to reproduce short-term lake dynamics needed for water quality predictions, as well as long-term dynamics needed to understand broad scale controls over lake P. Here we use an emerging paradigm in modeling, process-guided machine learning (PGML), to produce a phosphorus budget for Lake Mendota (Wisconsin, USA) and to accurately predict epilimnetic phosphorus over a time range of days to decades. In our implementation of PGML, which we term a Process-Guided Recurrent Neural Network (PGRNN), we combine a process-based model for lake P with a recurrent neural network, and then constrain the predictions with ecological principles. We test independently the process-based model, the recurrent neural network, and the PGRNN to evaluate the overall approach. The process-based model accounted for most of the observed pattern in lake P; however it missed the long-term trend in lake P and had the worst performance in predicting winter and summer P in surface waters. The root mean square error (RMSE) for the process-based model, the recurrent neural network, and the PGRNN was 33.0 μg P L^{-1}, 22.7 μg P L^{-1}, and 20.7 μg P L^{-1}, respectively. All models performed better during summer, with RMSE values for the three models (same order) equal to 14.3 μg P L^{-1}, 10.9 μg P L^{-1}, and 10.7 μg P L^{-1}. Although the PGRNN had only marginally better RMSE during summer, it had lower bias and reproduced long-term decreases in lake P missed by the other two models. For all seasons and all years, the recurrent neural network had better predictions than process alone, with root mean square error (RMSE) of 23.8 μg P L^{-1} and 28.0 μg P L^{-1}, respectively. The output of PGRNN indicated that new processes related to water temperature, thermal stratification, and long term changes in external loads are needed to improve the process model. By using ecological knowledge, as well as the information content of complex data, PGML shows promise as a technique for accurate prediction in messy, real-world ecological dynamics, while providing valuable information that can improve our understanding of process.

AUTHOR CONTRIBUTION STATEMENT

PCH and ABS co-led the entire manuscript effort. PCH, ABS, XJ, AK, and VK formulated the research objectives and experimental design. PCH, XJ, ABS, and HAD designed and wrote the model code, and HAD, JS, YZ, and JSR provided the data and data transformations. CCC, JS, and NKW provided interpretation of results and relevance to water quality. All authors wrote and edited the paper.

DATA AVAILABILITY STATEMENT

Data and metadata will be made available in the Environmental Data Initiative (EDI) data repository.

* Corresponding author.
E-mail address: pchanson@wisc.edu (P.C. Hanson).

https://doi.org/10.1016/j.ecolmodel.2020.109136
Received 1 October 2019; Received in revised form 14 May 2020; Accepted 18 May 2020
Available online 30 May 2020

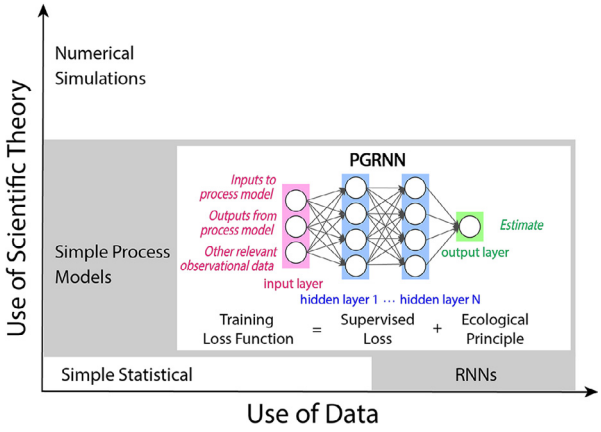

Figure 1. Adapted from Karpatne et al. (2017; Fig. 1b), this graph represents the intersection of empirical and theory driven approaches to science. Process-guided machine learning (PGML) exploits the strengths of each approach. The implementation of PGML in this study, which is process-guided recurrent neural networks (PGRNN), happens at the intersection of recurrent neural networks (RNNs) and simple process models.

1. Introduction

Phosphorus (P) is a critical limiting nutrient for phytoplankton and microbes (Schindler et al., 2016) and thus is an important driver of water quality in many lakes and reservoirs (Wetzel 2001). Where lakes and reservoir water quality is driven by phosphorus (P) dynamics, long-term P budgets can be used to understand lake responses to land use and climate change (Motew et al., 2017). However, predicting lake P dynamics is challenging because of the variety of factors that control P cycling and the time scales at which they occur (Chapra, 2008). For example, rapid P uptake by organisms with high turnover rates, sedimentation of those organisms, release of P from anoxic nutrient-rich sediments, and hydrologic load and export all are important processes that operate on minute to annual time scales (Cooke et al., 2005).

In response to the challenges of predicting aquatic ecosystem processes, such as P cycling, new modeling approaches are being developed that leverage recent increases in the availability of observational data (Porter et al., 2012). However, taking advantage of the information content of complex data remains difficult, in part due to the lack of integration between differing approaches to modeling ecosystems (e.g., Karpatne et al., 2017)(Fig. 1).

Traditionally, ecologists have focused on empirical and mechanistic approaches, which range in complexity from relatively simple statistical and process models to highly complex numerical simulations (Fig. 1, y-axis). Process-based models are commonly built to capture physical, chemical, and biological interactions governing natural systems (Starfield et al., 1994). Despite extensive use, the majority of process-based models implement approximate forms of relationships, either due to incomplete knowledge of certain processes or for practical computing purposes (Clark et al., 2016; Robson, 2014). This commonly introduces biases in predictions and significantly impacts the accuracy and the generalizability of the models (Allen & Hoekstra, 2015; Arhonditsis & Brett, 2004).

With advances in computing infrastructure and data availability, there is opportunity to incorporate machine learning into the ecological modeler's toolbox (Fig. 1, x-axis). Machine learning has shown tremendous success in many commercial applications, such as natural language processing, due to the capacity of computer algorithms to detect complex patterns in large volumes of data (Wang & Deng, 2019). In ecology, machine learning approaches are sometimes regarded with skepticism, because as a black-box approach they lack the ability to interpret the relationships between input drivers and the targeted responses, and therefore cannot resolve underlying mechanism (Jia et al., 2018, 2019; Lazer et al., 2014). Moreover, since the algorithms' training processes are ignorant of any theoretical principles, there is no guarantee that the predictions are consistent with physical laws or ecological theory (Peters et al., 2014).

An emerging modeling approach, process-guided machine learning (PGML; Karpatne et al., 2017), shows promise for improving model skill and ecological inference by exploiting the strengths of both mechanistic models and machine learning (Fig. 1). By combining ecological process with machine learning, PGML improves prediction accuracy while ensuring that predictions do not violate relevant physical laws or ecological principles. For example, Jia et al. (2018) has demonstrated that combining machine learning with a hydrodynamic lake model leads to predictions that are more accurate than either of the two approaches used independently, while remaining true to lake thermodynamics. PGML has also been used effectively to reduce the amount of environmental observations necessary for using powerful machine learning tools (Jia et al., 2019; Read et al., 2019), which is an important consideration, given that most ecosystems have a sparsity of observational data (Read et al., 2017).

In this work, we advance both prediction accuracy and scientific discovery of P cycling in eutrophic Lake Mendota (Wisconsin, USA) by comparing three different modeling approaches: a process-based model, machine learning, and a third approach that implements PGML as a hybrid of process-based and machine learning models and that are constrained by ecological principles. By comparing the three modeling approaches, we address three objectives: 1) develop a lake P budget that includes P inputs and outputs and P exchange between lake sediments and the water column. An accurate P budget will help us better understand the controls of water column P, assess the water quality implications of factors affecting external loading, and weigh those against factors controlling internal P loading; 2) compare the ability of the three models to produce accurate and unbiased predictions for epilimnetic P concentration over scales ranging from days to decades, because P is a limiting nutrient for primary production and the formation of harmful algal blooms in Lake Mendota; and 3) use the PGML to identify what processes might be missing from what we assume to be the dominant controls over lake P cycling, as instantiated in the process model. By addressing these objectives, we aim to gain a better understanding of whether PGML may be generalizable to other ecosystem applications, as well as compare its ability to predict P versus other methods.

2. Methods

2.1. Study site and model objectives

We used Lake Mendota, Wisconsin, a 40 km^2, 25 m deep eutrophic lake (Fig. 2), as our study system because of the extensive long-term data available through the North Temperate Lakes Long Term Ecological Research program (NTL LTER) (Magnuson et al., 2006) and past work focused on understanding catchment P dynamics and long-term water quality (Kara, Heimerl, et al., 2012; Lathrop & Carpenter, 2013). Lake Mendota exhibits high P concentrations as a result of agricultural and urban runoff in its catchment. High P leads to summer cyanobacterial blooms and degraded water quality, which negatively impact recreational use and produce adverse effects on wildlife and human health (Lathrop, 2007). Morphometric, hydrologic, and biogeochemical information, including data sets and their sources, is available in Tables 1, 2.

Given our objectives of producing a whole-lake P budget and surface water P predictions, we devised a time dynamic approach for predicting lake surface P concentrations over scales ranging from days to decades. Short time scales are necessary to address the summer months of July and August when water quality is at its worst and to capture critical transitions that occur during spring and fall turnover (i.e., when the lake fully mixes). Longer time scales are necessary for

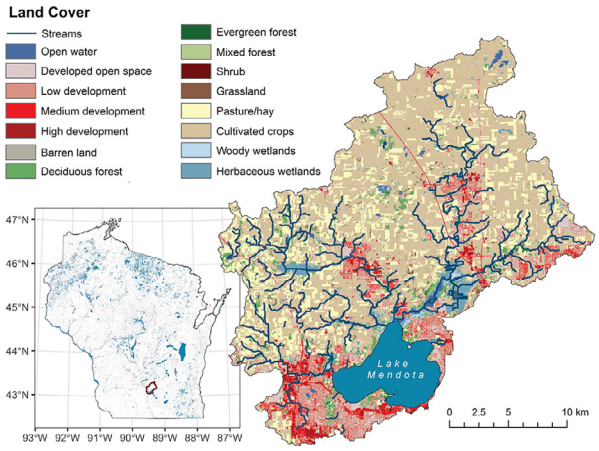

Figure 2. The study site, Lake Mendota, Wisconsin, is located in the Yahara River watershed of southern Wisconsin (see inset map). Land use in the watershed, as defined by the US National Land Cover Database, is predominantly cultivated crops, urban, and pasture/hay. A long history of agriculture in the watershed has contributed to the eutrophic state of the lake.

addressing slow changes due to interannual and long-term variation in precipitation, P loads, and temperature, as well as possible influences of changing P mass storage in lake sediments.

Our analytical approach blends an ecosystem process model (hereafter referred to as PROCESS) with the machine learning technology, recurrent neural networks (hereafter referred to as RNN). When combined, they are an implementation of PGML that we term a Process-Guided Recurrent Neural Network (PGRNN; after Jia et al. 2018). The purposes of the PGRNN are to improve overall prediction accuracy, reduce bias of PROCESS, and reduce over-fitting of RNN, while providing insights into the design of PROCESS.

2.2. PROCESS model description

We developed a simple process-based model to predict water column total phosphorus (P) on a daily time step over a period of decades. We prioritized summer epilimnetic P concentration, as well as reproducing the annual signal. There were few examples in the literature of ecosystem scale P models applied over decades and that used relatively simple approaches. Jensen et al. (2006) formulated a model with most of the features that interested us, was relatively simple, and had been used on multi-year data. We devised PROCESS as a time dynamic model (after Jensen et al., 2006) with the processes we assumed to represent the dominant fluxes of water column P: external load, sedimentation, recycling, and export (Hakanson & Bryhn, 2008). PROCESS (Fig. 3) is a mass-balance model that tracks three different pools of P within the lake system: epilimnetic P (P_{Epi}; Eq. 1), hypolimnetic P (P_{Hypo}; Eq. 2), and sediment P (P_{Sed}; Eq. 3). We used a mass balance approach so that we could estimate P retention, the relative contributions of internal versus external loads of P to the water column, and so that we could potentially run scenarios of changing lake condition or changing P load. We also felt a simple model could be applied to other lakes. Data requirements and parameter values are in Tables 1 and 2.

P load in the epilimnion is described by the mass balance differential equations as follows:

$$\frac{dP_{Epi}}{dt} = Load_{DP} + Entrainment - Sedimentation - Export \quad (1)$$

Where $Load_{DP}$, the proportion of the load that is dissolved P, is calculated using Equation 1a:

$$Load_{DP} = Load_{Total}{}^{*}\left(1 - C_{Load_{PP}}\right) \quad (1a)$$

And $Load_{Total}$ is the total daily P load (g d^{-1}), and $C_{Load_{PP}}$ is a constant, unitless parameter with a value of 0.5 that is the fraction of the input load that is particulate P. Although we do not have direct evidence to select that value with precision, it is based on expert advice from scientists at the Wisconsin Department of Natural Resources. We note that this value is simply a scalar that does not affect the time dynamics of lake P. Hydrological inflow was obtained from the USGS gage on the Yahara River at Windsor, WI (USGS-05427718), and these data were used to calibrate the catchment hydrology model, PIHM (Penn State Integrated Hydrologic Model) (Qu & Duffy, 2007). PIHM provided the lake water budget, and phosphorus loads were estimated from the water budget and the USGS loading functions (Appling, Leon, & McDowell, 2015).

Entrainment is calculated differently for two different scenarios. Scenario one is when the epilimnion is increasing in depth (growing). The entrainment is calculated by multiplying the change in epilimnion volume by the phosphorus concentration of the hypolimnion as shown in Equation 1b$_1$:

Table 1
Parameters for the model, PROCESS. Parameters estimated during model optimization (i.e., free parameters) are in bold. NTL LTER denotes the North Temperate Lakes Long-Term Ecological Research program.

(Abbreviation, if any) Description	Value and units	Source
Lake area	39,393,719 m^2	NTL LTER
Mean depth	12.8 m	NTL LTER
Maximum depth	25 m	NTL LTER
Hypsometry	m^2 per 1 m increment in lake depth	NTL LTER
Mean hydrologic residence time (not used in model, but for information only)	4.5 y	Approximated from hydrology data. Not used directly in any of the calculations.
(*SedDepth*) Lake sediment available to P recycling	0.1 m	(Nurnberg, 1988)
Sediment bulk density	1.33 × 10^6 g m^{-3}	Standard bulk density for clay/silt USDA
Sediment available P	0.806 mgP g^{-1} (sediment dry weight)	(Holdren & Armstrong, 1980)
(C_{Burial})Permanent sediment burial rate	1.2 mm y^{-1}	Manually fit in the process model to ensure approximately constant sediment P over the simulation period
(C_{Sed}) Phosphorus sedimentation rate	**0.0137 d^{-1}**	**Free parameter fit by optimization**
(a) First constant in the recycling equation	-4.30 mg P m^{-2} d^{-1}	(Nurnberg, 1988)
(b) Second constant in the recycling equation	**22.86 g dry weight m^{-2} d^{-1}**	**Free parameter fit by optimization**
($\theta_{Sedimentation}$) Temperature scaling factor for P sedimentation	**1.065 (unitless)**	**Free parameter fit by optimization**
($\theta_{Recycling}$) Temperature scaling factor for P recycling	**1.172 (unitless)**	**Free parameter fit by optimization**
($TBase_{Sed}$) Base temperature for sedimentation temperature scaling	10°C	Manually set to improve optimization of temperature scaling
($TBase_{Rec}$) Base temperature for recycling temperature scaling	10°C	Manually set to improve optimization of temperature scaling

Table 2
Time series driver datasets used in the three models, including abbreviation used for the driver, the units, the source, and the model for which the data were used, as indicated by an 'X'.

(Abbreviation) Description	Units	Source	Used in PROCESS	Used in RNN	Used in PGRNN
(T_{Epi}) Epilimnetic water temperature	°C; mean daily water temperature at 1 m and 20 m	NTL LTER	X	X	X
(T_{Hypo}) Hypolimnetic water temperature	°C; mean daily water temperature at 1 m and 20 m	NTL LTER	X	X	X
(*Z*) Thermocline depth	m	GLM	X	X	X
(Q_{In}) Discharge, overland flow	$m^3 d^{-1}$	PIHM	X	X	X
($Load_{Total}$) Lake P load	$g\ d^{-1}$	USGS	X	X	X
(T_{Air}) Air temperature	°C	NLDAS	X	X	X
(*EpiVolume*) Volume of the epilimnion	m^3	GLM		X	X
(*HypoVolume*) Volume of the hypolimnion	m^3	GLM		X	X
(*Strat*) Stratification	boolean	GLM		X	X
(*Precip*) Precipitation	m	NLDAS-2		X	X
(*SW*) Shortwave radiation	$W\ m^{-2}$	NLDAS-2		X	X
(*U*) Wind speed	ms^{-1}	NLDAS-2		X	X
(*LL*) Lake water level	m	NTL LTER		X	X
($P\text{-}PROCESS_{Epi}$) Epilimnetic P concentration	$\mu g\ P\ L^{-1}$	PROCESS predictions			X
($P\text{-}PROCESS_{Hypo}$) Hypolimnetic P concentration	$\mu g\ P\ L^{-1}$	PROCESS predictions			X

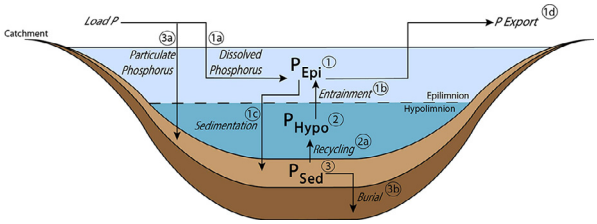

Figure 3. Conceptual diagram illustrating the three state variables (P_{Epi}, P_{Hypo}, P_{Sed}) and the fluxes (→) in the simple process model, PROCESS, of a temperate dimictic lake. The numbers correspond to the equations that describe the state variable or fluxes.

$$Entrainment = \frac{dEpiVolume}{EpiVolume} * P_{Hypo}\ (Growing\ epilimnion) \quad (1b1)$$

Scenario two is when the epilimnion is decreasing in depth (shrinking). The entrainment is calculated by multiplying the change in epilimnion volume by the phosphorus concentration of the epilimnion as shown in Equation 1b$_2$:

$$Entrainment = \frac{dEpiVolume}{EpiVolume} * P_{Epi}\ (Shrinking\ epilimnion) \quad (1b2)$$

Sedimentation is the sum of the fraction of the total load inputs that are particulate P and the amount of P within the epilimnion that contributes to sedimentation as shown in Equation 1c:

$$Sedimentation = \left(Load_{Total} * C_{Load_{PP}}\right) + \left(P_{Epi} * C_{sed} * \theta_{Sedimentation}^{(T_{Epi} - TBase_{Sed})}\right) \quad (1c)$$

Where C_{sed} (d^{-1}) is a free parameter and a first order decay rate for the epilimnetic P pool, $\theta_{Sedimentation}$ is a unitless free parameter (i.e., estimated in model optimization) and an Arrhenius coefficient for temperature scaling of sedimentation, T_{Epi}, is the observed epilimnetic temperature, and $TBase_{Sed}$ is the base temperature for sedimentation temperature scaling.

Export is calculated by multiplying the observed volumetric daily outflow, Q_{Out}($m^3\ d^{-1}$), by the P concentration in the epilimnion as shown in Equation 1d:

$$Export = Q_{Out} * P_{Epi} \quad (1d)$$

The second P pool, P_{Hypo}, is described by the mass balance differential equation shown below:

$$\frac{dP_{Hypo}}{dt} = Recycing - Entrainment \quad (2)$$

Recycling is the flux of P from the sediment P pool (P_{Sed}) back into the water column and is modeled as the product of three different terms described in Equation 2a. Although multiple mechanisms have been proposed for efflux of P from sediments, our goal is to provide a simple, empirically-based approach that subsumes more than one possible process within the parameters. This approach provides flexibility and generalizability.

$$Recycling = LakeArea_{Hypo} * C_{FSedP} * \theta_{Recycling}^{(T_{Hypo} - TBase_{Rec})} \quad (2a)$$

$LakeArea_{Hypo}$ is the calculated area based on lake hypsometry at the thermocline depth, which was estimated from General Lake Model (GLM) that used observed data as inputs (Hipsey, Salmon, & Mosley, 2014). $\theta_{Recycling}$ is a unitless free parameter and an Arrhenius coefficient for temperature scaling of recycling, T_{Hypo} is the observed hypolimnetic temperature, and $TBase_{Rec}$ is the base temperature for recycling temperature scaling. C_{FSedP} is a recycling term calculated using the regression developed by Nurnberg (Nurnberg, 1988):

$$C_{FSedP} = a + b * P_{Sed} \quad (2b)$$

In Equation 2b, a is a constant value of -4.3 mg P m^{-2} d^{-1}, b is a free parameter that is the slope of the recycling equation, and P_{Sed} is the phosphorus mass in the lake sediments.

The third P pool, P_{sed}, is described by the mass balance differential equation:

$$\frac{dP_{sed}}{dt} = Load_{PP} + Sedimentation - Recycling - Burial \quad (3)$$

Where $Load_{PP}$ is the P load that is particulate phosphorus as shown in Equation 3a:

$$Load_{PP} = Load_{Total} * C_{Load_{PP}} \quad (3a)$$

Burial is the P_{sed} that becomes no longer available to recycling, as in Equation 3b:

$$Burial = C_{Burial} * 365^{-1} * P_{Sed} * SedDepth^{-1} \quad (3b)$$

Where C_{Burial} is a constant value of 1.2 mm y^{-1} that determines the rate at which P is no longer available for recycling and is permanently sequestered. *SedDepth* is a constant value of 0.1 m (Hoffman, Armstrong, & Lathrop, 2013) that quantifies the sediment available for recycling.

Four parameters for PROCESS were estimated as part of an optimization routine. Optimization used the predictions of P_{Epi}, converted to concentration units, and the observations of P concentration that were taken from surface waters of Lake Mendota over the study period. Hypolimnetic P observations were not used during optimization, because P_{Hypo} from PROCESS was sensitive to our estimates of the

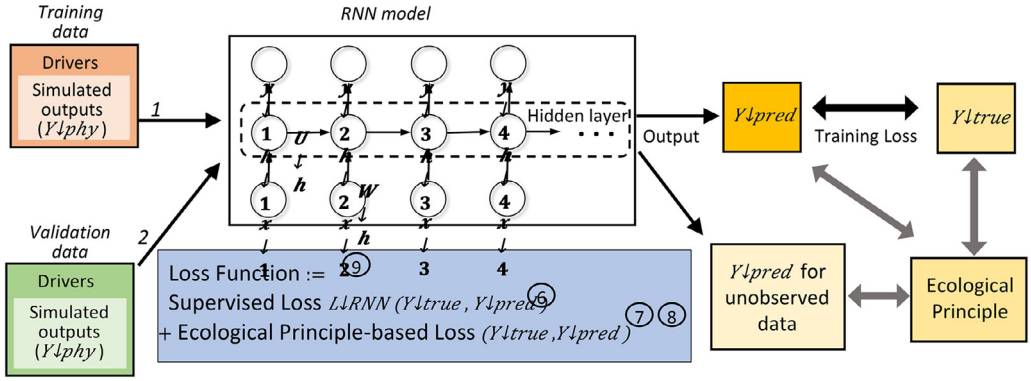

Figure 4. The architecture of the PGRNN model. We use RNN as the machine learning model to make predictions of surface water P (Y_{pred}). The training data include observed driver data ("Drivers") and the output of PROCESS ("Simulated outputs") from 1999-2015. After training, the model is validated on data from 1995-1998. The loss function includes the supervised loss on the training data and the process-based loss, which represents constraints imposed on predictions by ecological principles ("Physical laws").

hypolimnetic volume, especially when hypolimnetic volumes were small. Moreover, our primary concern was accurate estimates of P_{Epi} because of its importance to water quality. PROCESS was optimized using the Levenberg-Marquardt algorithm, which is a least-squares method, in the FME package (Soetaert & Petzoldt, 2010). Parameters estimated in model fitting are based on values referenced, with generous ranges to accommodate broad searching space. The parameters are phosphorus sedimentation rate (C_{Sed}), the second constant in the recycling equation (b), temperature scaling factor for P sedimentation ($\theta_{Sedimentation}$), and temperature scaling factor for P recycling ($\theta_{Recycling}$), which are constrained between 0.0010 – 0.10 d^{-1} (Soranno, Carpenter, & Lathrop, 1997), 0-40 g dry weight m^{-2} d^{-1} (Nurnberg, 1988), 1.0-1.4 (unitless), and 1.0-1.4 (unitless), respectively. While this relative simplicity helps ensure generalizability by reducing the number of free parameters to four (Table 1), it also subsumes potentially important processes unaccounted in the model that may explain important variance in the observed P, which we explored below with the machine learning approaches.

2.3. RNN model descriptions

A standard RNN model aims to learn a mapping relation from a series of input drivers $X = \{x_1, x_2, ..., x_T\}$ on a total of T consecutive dates to a series of target variables $Y = \{y_1, y_2, ..., y_T\}$ (i.e., phosphorus concentration) on corresponding dates through a black-box function f_{RNN}. The power of RNN lies in its ability to capture the data dependencies across time. Specifically, the RNN model introduces hidden representation h_t at each time step t, which summarizes the information at both current time step and previous time steps. The hidden representation h_t can be computed using a transition function, as follows:

$$h_t = \sigma(W_h x_t + U_h h_{t-1} + b_h), \tag{4}$$

where $\sigma(\cdot)$ denotes the sigmoid function, W_h, U_h, and b_h are model parameters that need to be tuned using true observations. It can be seen from this equation that the hidden representation h_t at time t not only depends on input drivers at the same step (i.e., x_t), but also depends on the hidden representation from the previous time step (i.e., h_t).

Once obtaining the hidden representation, we can compute the final model outputs as:

$$y_t = \sigma(W_y h_t + b_y), \tag{5}$$

where W_y and b_y are another set of model parameters.

While standard RNN can model the dependency across time (Eq. 4), it can hardly memorize long-term patterns as the model will gradually lose its memory towards long history. To address this problem, various black-box transition functions have been used to replace the transition function used in the standard RNN (Eq. 4). Gated Recurrent Unit (GRU) (Chung, Gulcehre, Cho, & Bengio, 2014) is one of the most widely structures to define transition functions, which has shown promising results to memorize both short-term and long-term patterns in a variety of real-world applications. In our problem, successfully modeling surface P is contingent on capturing dynamics at both short time scale and long time scale, and thus we adopt the GRU to define transition functions in this work. However, due to the sole dependence on the training data, this model is likely to produce results that are inconsistent with our knowledge of ecological process.

The tuning of model parameters within each machine learning model requires a training process that optimizes toward a specific loss function. In the RNN model, we use a supervised loss function that minimizes the RMSE between the predicted values $\hat{y}_t$ and the observed values y_t at each time step. This supervised loss function can be defined as follows:

$$L_{RNN} = \sqrt{\sum_t (\hat{y}_t - y_t)^2 / T} \tag{6}$$

where y_t and $\hat{y}_t$ denote the observed target value and the predicted target value at time t, respectively. The $\sum_t \cdot$ operation takes the summation over all the time steps when the observed data are available.

2.4. PGRNN model description

The Process Guided Recurrent Neural Network (PGRNN) model makes two important advancements beyond PROCESS and RNN. The first is to combine PROCESS and RNN in a hybridized model and the second is to use ecological principles to constrain predictions from the hybrid model. Together, these techniques are the PGRNN.

Hybrid-process-data (HPD) model: We can use the process-based model $f_{PROCESS}$ (described in 2.2) to simulate the value of the target variable $Y_{PROCESS}$. $Y_{PROCESS}$ will provide an incomplete representation of the target variable due to simplified or missing processes in $f_{PROCESS}$, thus resulting in model discrepancies with respect to observations. To overcome the model deficiencies, we combine f_{RNN} and $f_{PROCESS}$ to leverage information in both ecological process and data (Fig. 4). The PGRNN architecture makes it possible to leverage the knowledge embedded in traditional process-based models by using the process-based model output ($Y_{PROCESS}$) as additional inputs to the machine learning for training. This results in the hybrid-process-data (HPD) model: f_{HPD}: $[X, Y_{PROCESS}] \rightarrow Y$.

The HPD model goes far beyond that of many traditional approaches (e.g., Eq. 4) to fitting observational data. For example, a common approach to leveraging the simulated outputs from process-based models is through residual modeling, which is widely used in ecological research (Hamilton, Lloyd, & Flores, 2017; Xu & Valocchi, 2015). In residual modeling, an additional function $g(\cdot)$ is fitted to close the gap between the simulated output $Y_{PROCESS}$ and the true observations Y. The residual modeling can be summarized as:

$$f_{Residual}(x) = f_{PROCESS}(x) + g(x) \tag{7}$$

HPD improves upon Eq. 7. Since the deep neural networks can serve as a universal learner with sufficient layers (Hornik, Stinchcombe, &

White, 1989), the HPD model has the potential to capture complex relationships between x, $Y_{PROCESS}$, and Y. In other words, HPD model has the capacity to learn arbitrary functions to map the input [x, $Y_{PROCESS}$] to the true observations Y. Here we consider a case when the learned $f_{HPD}(x, Y_{PROCESS})$ function can be decomposed into two separate parts $g(x) + Y_{PROCESS}$. This is equivalent to the residual modeling method. However, the hybrid-process-data model can also learn coupled relation between x and $Y_{PROCESS}$, which cannot be easily decomposed into separate modeling components on x and on $Y_{PROCESS}$. Hence, the HPD model goes far beyond the capacity of many traditional approaches such as the residual modeling method.

Ecological constraints: Another major innovation of our PGRNN modeling approach is to incorporate constraining principles into the traditional loss function that guide the learning of weights of the RNN. For instance, in the implementation of PGRNN to hydrodynamic modeling, the laws of energy conservation were used to constrain thermal profile predictions (Jia et al., 2018). However, for P cycling the constraining physical principle is mass balance, which is already incorporated in $f_{PROCESS}$. As an alternative to physical constraint for fitting the PGRNN, we use an ecological principle common to data from natural systems. We consider the power scaling laws, which dictate that power in lower frequencies exceeds power in higher frequencies (e.g., Kara et al., 2012; Fig. 5, as an application to lake data). In our case, this means that changes in P will tend to be smaller at short time scales, such as daily, than at longer time scales, such as monthly. This rule is useful in regularizing the daily change of model predictions and eliminating spuriously high or low values that in the real world would require an unrealistic P flux. Specifically, the PGRNN model makes predictions at daily scale, which is at a higher frequency than the observed data. Therefore, we include an additional term in the loss function of the PGRNN to penalize any daily prediction changes that are larger than the 80 percentile of the changes of the first differences of the observed values, as follows:

$$L_{deltas} = \sum_{t} ReLU(|\hat{y}_{t+1} - \hat{y}_t| - p_{0.8}(\Delta y))/T, \tag{8}$$

Where $p_{0.8}(\Delta y)$ denotes the 80 percentile of the changes of the first differences of the observed values. The first difference was used to eliminate long-term signals in the observational data. The function ReLU(·) is the rectified linear unit, which is expressed as $\mathrm{ReLU}(x) = \max(0, x)$. The 80 percentile was chosen through manual tuning to ensure that daily predictions had variability that conformed to reasonable expectations of daily mass change in epilimnetic P.

An additional constraint to PGRNN is applied and that also relates to ecological scaling. Even if the predicted data and the observed data are sampled at different frequency, they should maintain similar statistical property over the spectrogram. Therefore, we apply the Short Time Fourier Transform, with a window length 256 on predicted outputs and observed data to generate their spectrogram over time, represented as $\hat{s}$ and s, respectively. Then we penalize large power discrepancy between them, as follows:

$$L_{stft} = \frac{\sum_t \sum_f |\hat{s}_{t,f} - s_{t,f}|}{TF}, \tag{9}$$

where TF is the total number of frequency levels, $\hat{s}$ and s represent the obtained spectrogram from $\hat{y}$ and y, respectively, $s_{t,f}$ and $\hat{s}_{t,f}$ represent the power of the spectrum at time t and frequency f. Combining the Eqs. 5–7, we have the final training loss function, as follows:

$$L_{training} = L_{RNN} + \lambda L_{deltas} + \mu L_{stft}, \tag{10}$$

Where λ and μ are hyper-parameters to control the weight of each component in the training objective function. In our test, we set λ as 400 and μ as 0.05 based on a cross-validation test. The learning model is trained using the standard back-propagation algorithm with the Adam optimizer (Kingma & Ba, 2015).

In summary, the PGRNN incorporates ecological knowledge from two aspects. First, it indirectly incorporates process knowledge hidden in the existing process-based model through HPD model. Second it directly includes known ecological principles (e.g., power scaling) by generalizing the loss function of standard RNN to include ecological constraints.

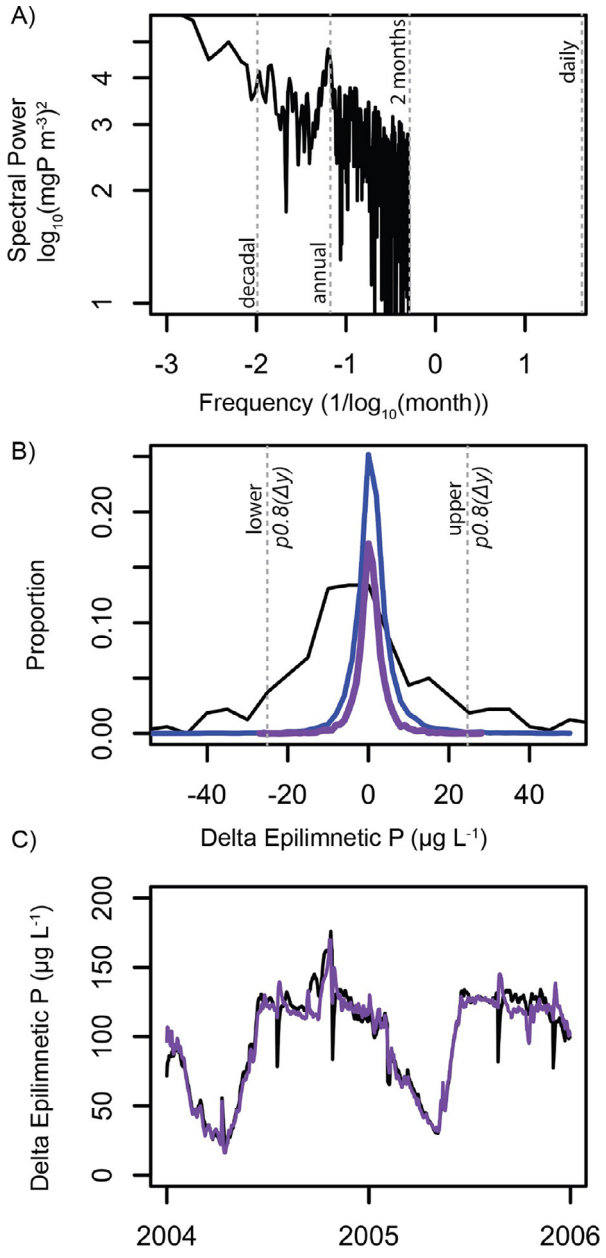

Figure 5. Data informing the ecological scaling principal. (A) Power spectrum of observed epilimnetic P shows higher power in lower frequencies, with the highest frequency being bi-monthly (ca. twice the sampling frequency). The modeled data, which are daily, are assumed to have lower power at the daily scale than the observed data have at the bi-monthly scale. The daily frequency indicated for reference. (B) A density plot of the first difference of observed epilimnetic P (black line), modeled P from RNN (blue line) and from PGRNN (purple line). Vertical dashed lines represent the constraints in Eq. 8. Note that observed data are bi-weekly, and modeled data are daily. (C) The predictions from PGRNN without the ecological principal in the loss function (black line) and with the ecological principle in the loss function (purple line).

We fit all models to epilimnetic P concentration, which was sampled approximately monthly for 20 years. Hypolimnetic P observations were not used for model fitting, because our primary concern was accurate estimates of P_{Epi} because of its importance to water quality. Moreover,

P_{Hypo} from PROCESS was sensitive to our estimates of the hypolimnetic volume, especially when hypolimnetic volumes were small.

In this work, we implement a one-layer RNN model with the GRU structure. The model has 10 units in the hidden layer (i.e., the dimension of $h_t$is 10). During the training process, we create data chunks with 100 time steps each (i.e., $T = 100$) and 15 input drivers (Table 2) for each time step. Hence, the input data for each chunk is a 100-by-15 matrix. We preprocess the input data by standardizing the data along each of 15 feature dimensions. The output is a 100-dimension vector for the predicts outputs at every time step.

We conducted a 6-fold cross validation test by dividing our study period for a single data set equally into six parts. Each time we trained the PGRNN and RNN using the data from five selected sections and then tested the model on the remaining section. In particular, we divide 7,200 samples into six parts. In each test, we train the model using 6,000 samples and test on the remaining 1,200 samples. We repeat this process for six time periods so as to eliminate the possibility that the model is biased towards specific parts. We chose to use section 1 as the validation section for model comparison because it provided overall best model performance.

Direct interpretation of neural network parameters (i.e., nodes in the network) is not informative, given their large numbers and the abstract nature of layering in RNNs. Thus, we determined the relative importance of 15 drivers (Table 2) in PGRNN by comparing the average absolute weights of network parameters that connect drivers to hidden network layers (Cho et al. 2014; Fig. 6). For the most important drivers, we determined the effect on epilimnetic P by setting all but one driver to their mean values, and plotting epilimnetic P as a function of the remaining unaltered driver. Comparison of drivers in this way, including PROCESS as a potential driver, provides insights into the design of PROCESS and our understanding of how P cycling works in Lake Mendota.

We evaluated all three models (PROCESS, RNN, and PGRNN) for goodness-of-fit using root mean squared error (RMSE) and visual inspection of predictions and observations over annual and decadal time scales. For the entire time series, as well as for summer-only (mean of July and August) and winter-only (mean of January and February) periods, we evaluated long-term trends using Mann-Kendall (MK, A.I. McLeod, 2011) statistics for observed and modeled data. Residuals were analyzed for long-term trends using MK statistics, and for auto-correlation using Auto Correlation Function (ACF, R Core Team, 2015).

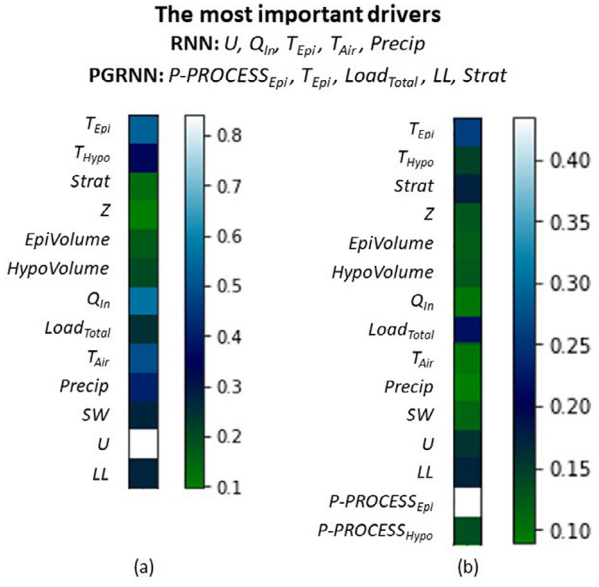

Figure 6. Criteria for selection of the most important drivers for the (a) RNN and (b) PGRNN. Each row depicts the level of importance of the corresponding driver in the respective models. The level of importance is calculated as the average absolute values of the weight parameters that connect each driver to different hidden units in the RNN model (Cho et al., 2014).

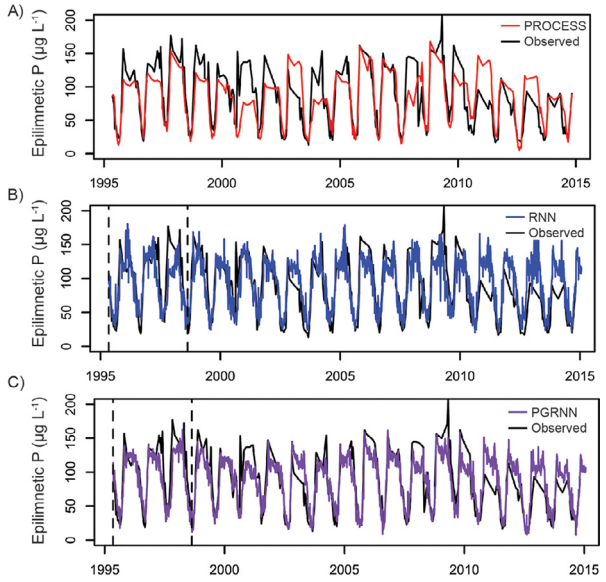

Figure 7. Observed epilimnetic P and model predictions from (A) PROCESS, (B) RNN, and (C) PGRNN. The vertical dashed lines bound the time range of the validation data.

3. Results

3.1. Lake phosphorus budget

PROCESS produced an ecosystem-scale balanced P budget, with predictions that reproduced the annual surface water P dynamics reasonably well (Fig. 7A). Each year, P_{Epi} rises rapidly during fall turnover (i.e., when the water column mixes) and remains high during winter. In spring, there is gradual decrease in P_{Epi}, followed by a precipitous decrease in summer, due to settling of P, presumably in the form of particulate organic phosphorus (POM) (mostly phytoplankton) from the epilimnion through the hypolimnion to the sediments. During the summer stratified period, P_{Epi} remains low, while P builds rapidly in the hypolimnion (not shown) due to efflux of P from the sediments and capture of fallout from the epilimnion. We note that in our model, P sedimenting from the epilimnion is added to the sediment P pool for simplicity. Mean annual load was 1.80 g m^{-2} lake area y^{-1}. The lake retained 72.3% of the load (burial ~1.29 g m^{-2} y^{-1}) and exported the remainder (~0.51 g m^{-2} y^{-1}). On average, P in the epilimnion due to external and internal loads was 22.6% and 77.4%, respectively, indicating the importance of the large lake sediment P pool to surface water concentrations. The magnitudes of these fluxes should be treated with caution, given the issues with model fit identified immediately below.

3.2. Model evaluation

Overall, the PGRNN performed better than PROCESS and RNN in reproducing both long-term (Fig. 7) and seasonal P cycling (Fig. 8). In winter, epilimnetic P is under-predicted by PROCESS early in the time series and over predicted later in the time series (Fig. 8A). RNN nearly matches the winter slope of the observed data, but is biased high (Fig. 8B), and PGRNN better reproduces the trend and has less overall bias (Fig. 8C). In summer, PROCESS has no long-term trend (Fig. 8D), and RNN has a modest decreasing trend (Fig. 8E), while PGRNN reproduces the long-term trend and captures some of the large

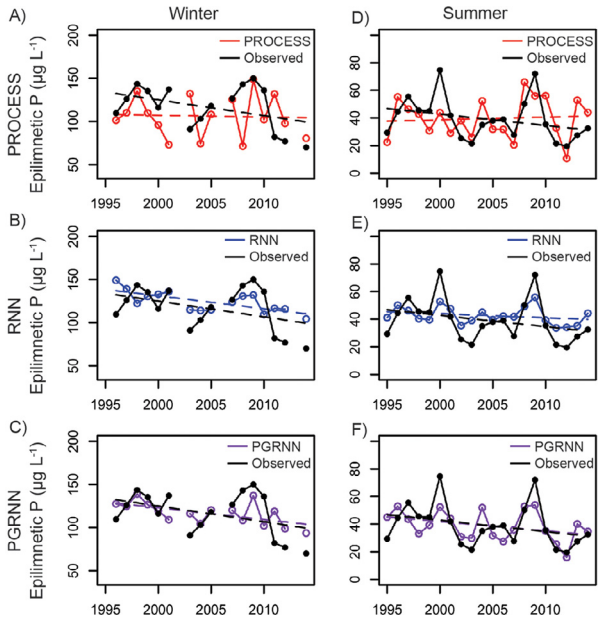

Figure 8. Winter (mean of January, February) predictions for (A) PROCESS, (B) RNN, and (C) PGRNN. Summer (mean of July, August) predictions for (D) PROCESS, (E) RNN, and (F) PGRNN. The dashed black lines are long-term trends of observed data and the dashed colored lines are long-term trends of the model predictions.

Table 3
Analysis of trends in observational data and model predictions for the complete time series (All data), winter (mean of December, January), and summer (mean of July, August). Mann-Kendall (MK) statistics were used to determine trend, with "Score" representing either downward (-) or upward (+) trend and its significance (*p* value). Separate tests were run for summer (July and August) and winter (December and January).

Test	Observational data	PROCESS	RNN	PGRNN
All data				
MK Score	-8689	-1459	-4387	-5689
MK tau	-0.186	-0.0311	-0.0934	-0.121
MK *p* value	<0.001	0.417	0.0147	0.00155
Winter				
MK Score	-22	-12	-56	-32
MK tau	-0.183	-0.100	-0.467	-0.533
MK *p* value	0.344	0.620	0.0133	0.00456
Summer				
MK Score	-55	12	-42	-48
MK tau	-0.290	0.0632	-0.221	-0.253
MK *p* value	0.0796	0.721	0.183	0.127

interannual changes missed by the other two models (Fig. 8F). Trends for observations, model predictions, and residuals are quantified in Tables 3 and 4.

Residual error for PROCESS was lowest of the three models for the validation segment, highest for the overall time series, and had the greatest downward trend (Fig. 9A, Table 3). RNN improved fits over PROCESS for the whole time series, with lower and less downward trend in the residuals (Fig. 9B). Although fit improved, RNN introduced high variance at the daily scale and occasionally spuriously high or low predictions (Fig. 7B). The PGRNN provided better fit to the data over RNN, and like the RNN captured additional scales of variability missed by PROCESS (Fig. 9C, Tables 3, 4). The PGRNN did not produce spuriously high or low predictions for P. For both winter and summer, PGRNN (Fig. 10C, F, respectively) had best overall model performance (Tables 3, 4) with the least biased residuals.

The PGRNN improved prediction by using additional features (i.e., explainable patterns) in the driving data, as well as adjusting

Table 4
Analysis of model residuals for the complete time series (All data), winter (mean of December, January), and summer (mean of July, August). Mann-Kendall (MK) statistics were used to determine trend, with "Score" representing either downward (-) or upward (+) trend and its significance (*p* value). Values for auto-correlation function (ACF) are the lags that are significant. Root mean squared error (RMSE) is for the full time series, with values in parentheses referring to the validation segment.

Test	PROCESS	RNN	PGRNN
All data			
MK Score	-11715	-8715	-8166
MK tau	-0.249	-0.186	-0.174
MK *p* value	<0.001	<0.001	<0.001
ACF	1-7	1-10	1-10
RMSE (µg L^{-1})	28.0 (19.8)	23.8 (25.4)	21.1 (22.7)
Winter			
MK Score	-20	4	-2
MK tau	-0.167	0.0333	-0.0167
MK *p* value	0.392	0.893	0.964
ACF	None	None	None
RMSE (µg L^{-1})	33.0	22.7	20.6
Summer			
MK Score	-46	-44	-6
MK tau	-0.242	-0.232	-0.0316
MK *p* value	0.144	0.163	0.871
ACF	None	None	None
RMSE (µg L^{-1})	14.3	10.9	10.7

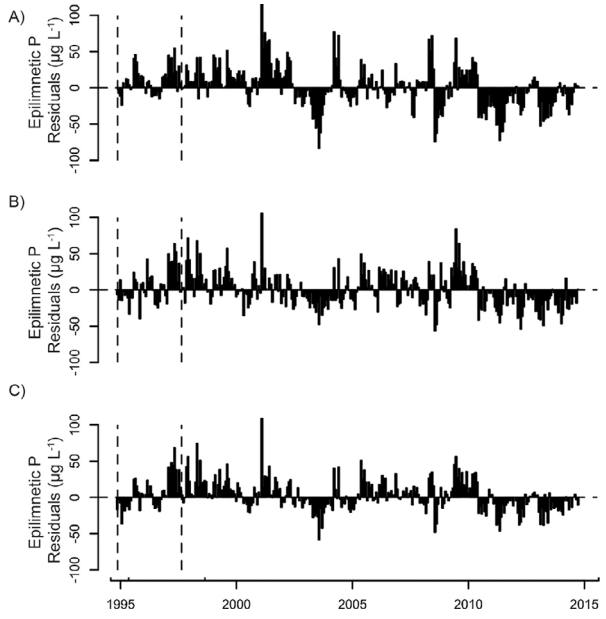

Figure 9. Residuals of model predictions for full time series of (A) PROCESS, (B) RNN, and (C) PGRNN.

predictions from PROCESS. The five most significant drivers for PGRNN model prediction were determined to be PROCESS, epilimnetic temperature, P load, lake level, and thermocline depth (Fig. 6). Mann-Kendall statistics for trends in the five predictors are as follows: PROCESS (MK τ = -0.0122, $p > 0.1$); T (MK τ = -0.0118, $p > 0.1$); P load (MK τ = -0.069, $p < 0.001$); lake level (MK τ = -0.093, $p < 0.001$); and Z (MK τ = -0.059, $p < 0.001$).

Because parameters in neural networks have no directly interpretable values, we show the importance of predictors through their influence on the predictions. In Figure 11A, high frequency variability shows the recurrent nature of the PGRNN. The influence of PROCESS on the predictions (Fig. 11B, red line) is approximately double that of temperature (Fig. 11C), which is approximately double that of the remaining predictors (Fig. 11D). The difference between the pink and red

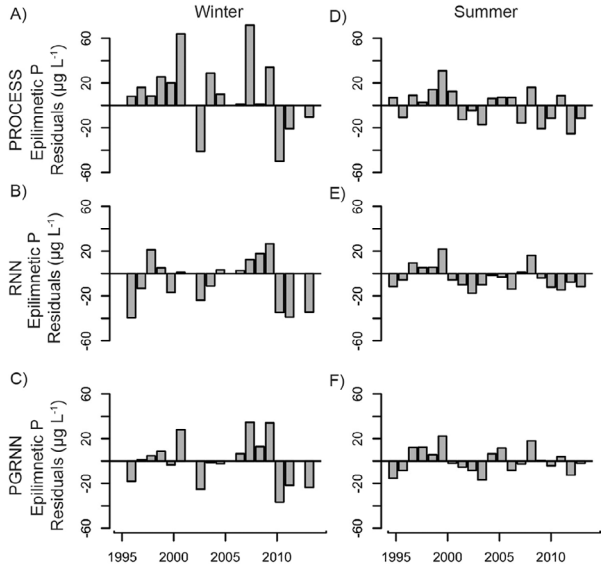

Figure 10. Residuals of winter (mean of January, February) model predictions for (A) PROCESS, (B) RNN, and (C) PGRNN. Residuals of summer (mean of July, August) model predictions for (D) PROCESS, (E) RNN, and (F) PGRNN.

lines in Figure 11B shows that PGRNN reassigns some of the variance from PROCESS (pink line) to the other predictors. Figure 11E shows for a few years how the PGRNN predictions (purple line) are the result of the additive effects shown in panels A-D, with the final predictions shown in purple to be consistent in color with other figures. The long-term downward trend in the observational data, and missed by PROCESS, was partially described by PGRNN, which assigned significant negative trends for P load, lake level, and thermocline depth. Notably, PGRNN indicates that neither temperature nor PROCESS contribute to long-term downward trends.

4. Discussion

Process-guided machine learning (PGML) shows promise in accurately predicting lake surface water P dynamics. Our implementation of PGML, the PGRNN model, melds PROCESS and RNN and outperforms both of those models in a number of ways. PGRNN provides better accuracy, and it rarely predicts outside the range of magnitudes of realistic biogeochemical fluxes (Fig. 5). The improvement in mean error for summer predictions from PROCESS to PGRNN (Table 4, Fig. 10) is about 4 µg L^{-1}, which is approximately 10% of the mean long-term summer P concentration in the lake (39.8 µg L^{-1}). Model performance in previous lake P modeling work rarely is documented (e.g., Bennett et al., 2013), and there are few examples of models applied for more than about a decade, making it challenging to compare our performance with the literature. However, our RMSE values were about half those of the model on which PROCESS was based (Jensen et al., 2006).

Our application of a common ecological principle, spectral power scaling, improved the quality of the predictions for the PGRNN, in part by eliminating outlier predictions. While a similar 'smoothing' of the predictions could have been accomplished by post processing them, it would not have fed back to improve the estimation of the PGRNN parameters. Thus, we argue that building such principles into the model's loss function (Eq. 10) provides for a better model calibration. Previous work on PGML used physical laws of water density to constrain lake temperature predictions for the machine learning components (Jia et al., 2018), and the PGML approach outperformed the original physical model in prediction accuracy, while maintaining fidelity to physical laws. It could be argued for lake P, however, that occasional spurious predictions could better represent the observational

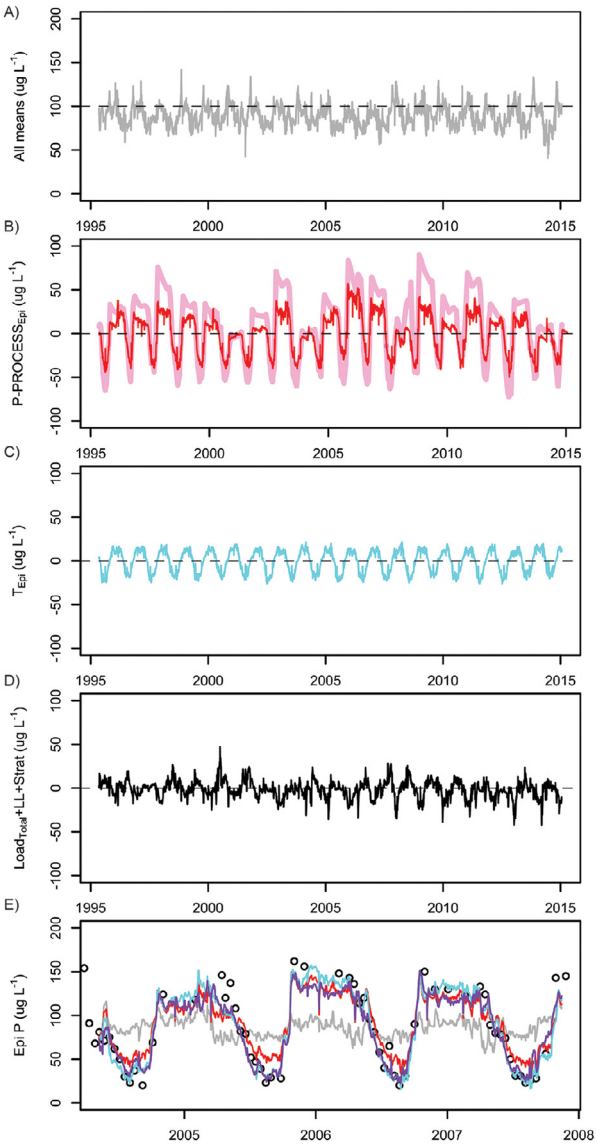

Figure 11. The influence of multiple drivers, including the output from PROCESS, on epilimnetic P predictions from the PGRNN: (A) all drivers set to mean values shows the mean of the predictions and the dynamics inherent in the neural network coefficients; (B) the effect of the PROCESS output on PGRNN predictions (red), and for comparison the original predictions from PROCESS (pink); (C) the effect of epilimnetic water temperature; (D) the combined effects of lake P load ($Load_{Total}$), lake level (LL), and thermocline depth (Strat). (E) The additive effects (purple line, as in Fig. 7C) of predictions in panels A-D, as well as the observed surface water P concentration (black circles) for three annual cycles.

data features. In this case, a different principle could be applied that would force predictions to conform to statistical characteristics of the observed distributions, which might work particularly well for extreme value distributions (Carpenter, Booth, Kucharik, & Lathrop, 2015).

The use of ecological knowledge as part of PGML helped us learn about PROCESS and guides future work on that model. Although we were encouraged that the predictions from PROCESS were the most important input to PGRNN for surface water P predictions (Fig. 11B,E), the PGRNN suggests that PROCESS could be improved in predicting annual cycles by including an additional temperature component (Fig. 11C). PGRNN also selected P load, lake level, and thermocline depth (Fig. 11D) to improve the annual cycle. Interestingly, PGRNN uses these last three terms to create a downward trend missed by PROCESS over the time series. While our confidence in lake level and

thermocline depth is high because they are directly measured with high-frequency sensors, our confidence in P load is lower, suggesting the downward trend lies in that variable. There are many reasons to suspect the P load. There are challenges in developing the rating curve for P load via surface inflows, because inflow from big storms is difficult to characterize, even though a few storms might account for half of the annual P load to Lake Mendota (Carpenter et al., 2015). Inter-annual variability of loads is high as well, ranging from about 0.5-2.0 g P m^{-2} y^{-1} (Carpenter et al., 2015), and our long-term estimate of 1.8 g P m^{-2} y^{-1} is at the high end of that range. Furthermore, P measurements in the inflows are sparse through time, and not all inflows have observed P data. The results of PGRNN suggest we investigate further a possible downward trend in P load to the system over the past few decades. The inclusion of these factors could improve both the predictions of the annual cycle and the long-term downward trend.

There is some evidence that a downward step-change in the observational data for epilimnetic P around 2010 may be real, and that the inaccuracies from model predictions appear to be due to over-prediction in winter (Fig. 10). All models over-predict P concentrations from 2010 onward (Fig. 9), suggesting that neither the driving data nor PROCESS can account for the change. Food web changes may be the answer. Introduction of spiny water fleas to Lake Mendota (high concentrations beginning ~2010) has perturbed the zooplankton community in ways that alter nutrient dynamics (Walsh et al., 2016). By preying on daphnia that consume phytoplankton, spiny water fleas potentially support higher phytoplankton biomass and therefore additional sedimentation. Recent invasion by zebra mussels (discovered in year 2015) has likely altered the phytoplankton community in ways that affect P cycling (Mellina et al., 1995); however, given their low densities and the narrow littoral zone, the effects on water clarity are likely too small to detect. Other catchment changes, such as increased plant biomass capturing P in river outlets, may be decreasing P loading into Lake Mendota, but these have yet to be well-studied and documented and are important topics for future work.

Recommended modifications to PROCESS thus far would be relevant to modeling Lake Mendota, but what about application of PROCESS to other lakes? For dimictic temperate lakes, it may be important to discriminate between processes related to iron in lake sediments, because iron can have strong control over P exchange between lake sediments and the water column. Alternative mechanisms between high iron and low iron P cycling were described by Hoffman et al. (2013). In high iron lakes, hypolimnetic buildup of P results primarily from the dissociation of P and Fe under anoxic conditions, but also from mineralization of POP from the epilimnion to the hypolimnion. During fall turnover, the water column becomes oxygenated, Fe binds P, and the precipitate settles back to the sediments (Campbell & Torgersen, 1980). However, in low iron lakes, such as Lake Mendota, efflux of P from the sediments is dominated by mineralization of POP sedimented from the epilimnion, as well as legacy P. For Lake Mendota specifically, an empirical study showed that approximately 2/3 of the hypolimnetic P buildup can be attributed to epilimnetic POP sources (Hoffman et al., 2013). This corresponds closely to our modeling results, in which annual sedimentation is roughly 65% of annual recycling. In our model, we do not discriminate alternative sediment-water column biogeochemical processes. We lump them all into recycling, under the assumption that particulate P settles to the sediments, whether bound to Fe or organic matter. While this simplification conceptually blurs the separation of water column and sediments, we argue that, in practice, that line is blurred in Mendota by a substantial layer of loose, flocculent matter that rests between the sediment and the hypolimnion. If we were to apply PROCESS to a high iron lake, we might consider adding a P scavenging process that would occur during mixing or oxygenation events.

While our application of PGML shows promise, many avenues of PGML evaluation have yet to be explored. PGRNN has provided useful information about PROCESS and its model structure, which can be the greatest source of error in predictions (Arhonditsis & Brett, 2004). However, we have not evaluated the parameter uncertainties for PROCESS nor have we assessed the importance of observational uncertainty. In addition, the relative contributions of two defining features of PGRNN – PROCESS as an input to RNN and the ecological principle in constraining predictions – have not been fully assessed. It may be the case that an RNN with ecological principles built into the loss function (i.e., PGRNN without PROCESS as an input) predicts nearly as accurately as PGRNN. However, in cases where observational data are sparse, training the RNN with process (i.e., putting PROCESS back into PGRNN) may reduce the amount of data needed for accurate predictions (Jia et al., 2018), and for most lakes and reservoirs, we have sparse data for important characteristics, such as water quality (Read et al., 2017).

PGML offers a framework to both improve prediction accuracy and update understanding, both of which are at the heart of the scientific endeavor. Often in ecology, we have insufficient data to exploit the most powerful machine learning techniques. We can relieve machine learning of the burden of capturing ecological signals by essentially training them with known process (Lee, Ros, Li, & Gaidon, 2018). This allows the machine learning to attribute remaining ecosystem variance to a set of drivers. We find it heartening that in an era of “big data” and “deep learning” ecological knowledge retains an important role. At the same time, we must be cautious in our predictions, because the largest source of error tends to be in our assumptions of which processes are most important, and therefore included in our models (Arhonditsis & Brett, 2004). Predicting P dynamics is a critical aspect of understanding lake water quality, and as we apply our process models to predicting P dynamics in a much broader suite of lakes, we will need techniques that leverage both our knowledge and the growing data from Earth observing systems.

4. Conclusions

Process guided machine learning, implemented as a process guided recurrent neural network (PGRNN) in this study, is a particularly valuable approach to modeling the lake phosphorus problem because it helped in overcoming challenges in both understanding and predicting surface water phosphorus accurately. PGRNN made the most accurate predictions, outperforming the process based model and the recurrent neural network in prediction accuracy, reducing prediction bias, and reproducing both short term and long term variability in observed lake phosphorus. As importantly, the PGRNN provided additional valuable information relevant to ecosystem process. The process model was designed to be parsimonious and to reproduce ecological processes known to be relevant to the annual dynamics of surface water phosphorus concentration. The PGRNN identified scales of variability missed by the process-based model. There was a long-term downward trend in lake phosphorus that occurred over decades, which the process model did not predict but that PGRNN attributed to a long and slow change in the phosphorus loads to the lake. There was a small but significant annual cycle related to water temperature that was identified by the machine learning but not predicted by the process model. These insights resulted from the hybridization of ecological knowledge and machine learning, which shows tremendous potential for advancing our prediction capabilities and understanding of dynamical natural systems.

Declaration of Competing Interests

None

ACKNOWLEDGEMENTS

We thank our CNH colleagues and our GLEON colleagues for valuable discussions of the ideas herein and Samantha Oliver for reviewing

the manuscript. We are grateful for Y. Gil, who catalyzed the collaboration between ecologists and computer scientists. Two anonymous reviewers provided helpful criticisms. The NTL LTER (DEB-1440297) provided context and data for the study. Funding: The U.S. National Science Foundation provided funding through the CNH-Lakes project (ICER-1517823), DEB-1753639, OAC-1934633 and DEB-1753657.

References

McLeod, A.I., 2011. Kendall: Kendall rank correlation and Mann-Kendall trend test. R package version 2.2. Retrieved from. http://cran.r-project.org/package=Kendall.

Allen, T.F.H., Hoekstra, T.W., 2015. Toward a unified ecology, 2nd ed. Columbia University Press.

Appling, A.P., Leon, M.P., McDowell, W.H., 2015. Reducing bias and quantifying uncertainty in watershed flux estimates: The R package loadflex. Ecosphere 6 (12), 1–25. https://doi.org/10.1890/ES14-00517.1.sm.

Arhonditsis, G.B., Brett, M.T., 2004. Evaluation of the current state of mechanistic aquatic biogeochemical modeling. Marine Ecology Progress Series 271 (April), 13–26. https://doi.org/10.3354/meps271013.

Bennett, N.D., Croke, B.F.W., Guariso, G., Guillaume, J.H.A., Hamilton, S.H., Jakeman, A.J., …, Andreassian, V., 2013. Characterising performance of environmental models. Environmental Modelling and Software 40, 1–20. https://doi.org/10.1016/j.envsoft.2012.09.011.

Campbell, P., Torgersen, T., 1980. Maintenance of iron meromixis by iron redeposition in a rapidly flushed monimolimnion. Canadian Journal of Fisheries and Aquatic Sciences 37 (8), 1303–1313.

Carpenter, S.R., Booth, E.G., Kucharik, C.J., Lathrop, R.C., 2015. Extreme daily loads: role in annual phosphorus input to a north temperate lake. Aquatic Sciences 77 (1), 71–79. https://doi.org/10.1007/s00027-014-0364-5.

Chapra, S.C., 2008. Surface Water-Quality Modeling. Waveland Press.

Cho, K., Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. https://doi.org/10.1074/jbc.M608066200.

Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling, 9. https://doi.org/10.1109/ICORR.2015.7281186.

Clark, M.P., Schaefli, B., Schymanski, S.J., Samaniego, L., Luce, C.H., Jackson, B.M., …, Ceola, S., 2016. Improving the theoretical underpinnings of process-based hydrologic models. Water Resources Research 2350–2365. https://doi.org/10.1002/2015WR017910.Received.

Cooke, D.G., Welch, E.B., Peterson, S., Nichols, S.A., 2005. Restoration and Management of Lakes and Reservoirs. CRC Press.

Hakanson, L., Bryhn, A.C., 2008. A Dynamic Mass-balance Model for Phosphorus in Lakes with a Focus on Criteria for Applicability and Boundary Conditions. Water Air Soil Pollution 187, 119–147.

Hamilton, F., Lloyd, A.L., Flores, K.B., 2017. Hybrid modeling and prediction of dynamical systems. PLoS Computational Biology 13 (7), 1–20. https://doi.org/10.1371/journal.pcbi.1005655.

Hipsey, M.R., Salmon, S.U., Mosley, L.M., 2014. A three-dimensional hydro-geochemical model to assess lake acidification risk. Environmental Modelling and Software 61, 433–457. https://doi.org/10.1016/j.envsoft.2014.02.007.

Hoffman, A.R., Armstrong, D.E., Lathrop, R.C., 2013. Influence of phosphorus scavenging by iron in contrasting dimictic lakes. Canadian Journal of Fisheries and Aquatic Sciences 70 (7), 941–952. https://doi.org/10.1139/cjfas-2012-0391.

Holdren, G.C.J., Armstrong, D.E., 1980. Factors Affecting Phosphorus Release from Intact Lake Sediment Cores. Environmental Science & Technology 14 (1), 79–87. https://doi.org/10.1021/es60161a014.

Hornik, K., Stinchcombe, M., White, H., 1989. Multilayer Feedforward Networks are Universal Approximators. Neural Networks 2, 359–366. https://doi.org/10.1016/0893-6080(89)90020-8.

Jensen, J.P., Pedersen, A.R., Jeppesen, E., Søndergaard, M., 2006. An empirical model describing the seasonal dynamics of phosphorus in 16 shallow eutrophic lakes after external loading reduction. Limnology and Oceanography 51 (1part2), 791–800. https://doi.org/10.4319/lo.2006.51.1_part_2.0791.

Jia, X., Karpatne, A., Willard, J., Steinbach, M., Read, J., Hanson, P.C., …, Kumar, V., 2018. Physics Guided Recurrent Neural Networks for Modeling Dynamical Systems: Application to Monitoring Water Temperature and Quality in Lakes. Computing Research Repository 3 Retrieved from. http://arxiv.org/abs/1810.02880.

Jia, X., Willard, J., Karpatne, A., Read, J., Zward, J., Steinbach, M., Kumar, V., 2019. Physics Guided RNNs for Modeling Dynamical Systems: A Case Study in Simulating Lake Temperature Profiles. ArXiv Eprints, arXiv 1810 https://doi.org/arXiv:1810.13075v1.

Kara, E.L., Hanson, P., Hamilton, D., Hipsey, M.R., McMahon, K.D., Read, J.S., …, Kratz, T., 2012. Time-scale dependence in numerical simulations: Assessment of physical, chemical, and biological predictions in a stratified lake at temporal scales of hours to months. Environmental Modelling and Software 35, 104–121. https://doi.org/10.1016/j.envsoft.2012.02.014.

Kara, E.L., Heimerl, C., Killpack, T., Van de Bogert, M.C., Yoshida, H., Carpenter, S.R., 2012. Assessing a decade of phosphorus management in the Lake Mendota, Wisconsin watershed and scenarios for enhanced phosphorus management. Aquatic Sciences 74 (2), 241–253.

Karpatne, A., Atluri, G., Faghmous, J., Steinbach, M., Banerjee, A., Ganguly, A., …, Kumar, V., 2017. Theory-guided Data Science: A New Paradigm for Scientific Discovery from Data. TKDE 29 (10), 2318–2331. https://doi.org/10.1109/TKDE.2017.2720168.

Kingma, D.P., & Ba, J.L. (2015). Adam: A method for Stochastic Optimization. In *ICLR* (p. 15). https://doi.org/10.1063/1.4902458.

Lathrop, R.C., 2007. Perspectives on the eutrophication of the Yahara lakes. Lake and Reservoir Management 23 (4), 345–365. https://doi.org/10.1080/07438140709354023.

Lathrop, R.C., Carpenter, S.R., 2013. Water quality implications from three decades of phosphorus loads and trophic dynamics in the Yahara chain of lakes. Inland Waters 4, 1–14. https://doi.org/10.5268/IW-4.1.680.

Lazer, D., Kennedy, R., King, G., Vespignani, A., 2014. Supplementary Materials for The Parable of Google Flu: Traps in Big Data Analysis. Science 343 (March), 1203–1206. https://doi.org/10.1126/science.1248506.

Lee, K.-H., Ros, G., Li, J., & Gaidon, A. (2018). Spigan: Privileged adversarial learning from simulation. *ArXiv Preprint ArXiv:1810.03756*.

Magnuson, J.J., Kratz, T.K., & Benson, B.J. (2006). *Long-Term Dynamics of Lakes in the Landscape: Long-Term Ecological Research on North Temperate Lakes*.

Mellina, E., Rasmussen, J.B., Mills, E.L., 1995. Impact of zebra mussel (*Dreissena polymorpha*) on phosphorus cycling and chlorophyll in lakes. Canadian Journal of Fisheries and Aquatic Sciences 52 (12), 2553–2573. https://doi.org/10.1139/f95-246.

Motew, M., Chen, X., Booth, E.G., Carpenter, S.R., Pinkas, P., Zipper, S.C., …, Kucharik, C.J., 2017. The Influence of Legacy P on Lake Water Quality in a Midwestern Agricultural Watershed. Ecosystems 20 (8), 1468–1482. https://doi.org/10.1007/s10021-017-0125-0.

Nurnberg, G.K., 1988. Prediction Release Rates from Total and Reductant-Soluble Phosphorus in Anoxic Lake Sediments. Can.J.Fish.Aquat.Sci 45, 453–462.

Peters, D.P.C., Havstad, K.M., Cushing, J., Tweedie, C., Fuentes, O., Villanueva-Rosales, N., 2014. Harnessing the power of big data: infusing the scientific method with machine learning to transform ecology. Ecosphere 5 (6), 1–15. https://doi.org/10.1890/ES13-00359.1.

Porter, J.H., Hanson, P.C., Lin, C.C., 2012. Staying afloat in the sensor data deluge. Trends in Ecology and Evolution 27 (2), 121–129. https://doi.org/10.1016/j.tree.2011.11.009.

Qu, Y., Duffy, C.J., 2007. A semidiscrete finite volume formulation for multiprocess watershed simulation. Water Resources Research 43 (8), 1–18. https://doi.org/10.1029/2006WR005752.

R Core Team, 2015. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria Retrieved from. https://www.r-project.org/.

Read, E.K., Carr, L., De Cicco, L., Dugan, H.A., Hanson, P.C., Hart, J.A., …, Winslow, L.A., 2017. Water quality data for national-scale aquatic research: The Water Quality Portal. Water Resources Research 53 (2), 1735–1745. https://doi.org/10.1002/2016WR019993.

Read, J.S., Jia, X., Willard, J., Appling, A.P., Zwart, J.A., Oliver, S.K., …, Steinbach, M., 2019. Process-guided deep learning predictions of lake water temperature. Water Resources Research 55 (11), 9173–9190. https://doi.org/10.1029/2019WR024922.

Robson, B.J., 2014. State of the art in modelling of phosphorus in aquatic systems: Review, criticisms and commentary. Environmental Modelling and Software 61, 339–359. https://doi.org/10.1016/j.envsoft.2014.01.012.

Schindler, D.W., Carpenter, S.R., Chapra, S.C., Hecky, R.E., Orihel, D.M., 2016. Reducing phosphorus to curb lake eutrophication is a success. Environmental Science and Technology 50 (17), 8923–8929. https://doi.org/10.1021/acs.est.6b02204.

Soetaert, K., Petzoldt, T., 2010. Inverse Modelling, Sensitivity and Monte Carlo Analysis in R Using Package FME. Journal of Statistical Software 33 (3), 28. https://doi.org/10.18637/jss.v033.i03.

Soranno, P.A., Carpenter, S.R., Lathrop, R.C., 1997. Internal phosphorus loading in Lake Mendota : response to external loads and weather. Can. J. Fish. Aquat. Sci. 54, 1883–1893.

Starfield, A.M., Smith, K.A., Bleloch, A.L., 1994. How to model it: Problem solving for the computer age, 1st ed. McGraw-Hill, New York.

Walsh, J.R., Carpenter, S.R., Vander Zanden, M.J., 2016. Invasive species triggers a massive loss of ecosystem services through a trophic cascade. Proceedings of the National Academy of Sciences 113 (15), 4081–4085. https://doi.org/10.1073/pnas.1600366113.

Wang, M., Deng, W., 2019. Deep Face Recognition: A Survey. Computing Research Repository 1–26 https://doi.org/arXiv:1804.06655v3.

Xu, T., Valocchi, A.J., 2015. Data-driven methods to improve baseflow prediction of a regional groundwater model. Computers and Geosciences 85, 124–136. https://doi.org/10.1016/j.cageo.2015.05.016.