Water Resources Research

RESEARCH ARTICLE
10.1029/2019WR024922

Special Section:
Big Data & Machine Learning in Water Sciences: Recent Progress and Their Use in Advancing Science

Process-Guided Deep Learning Predictions of Lake Water Temperature

Jordan S. Read[1], Xiaowei Jia[2], Jared Willard[2], Alison P. Appling[1], Jacob A. Zwart[1], Samantha K. Oliver[1], Anuj Karpatne[3], Gretchen J. A. Hansen[4], Paul C. Hanson[5], William Watkins[1], Michael Steinbach[2], and Vipin Kumar[2]

[1]U.S. Geological Survey, Reston, VA, USA, [2]Department of Computer Science and Engineering, University of Minnesota, Minneapolis, MN, USA, [3]Department of Computer Science, Virginia Tech, Blacksburg, VA, USA, [4]Department of Fisheries, Wildlife, and Conservation Biology, University of Minnesota, Minneapolis, MN, USA, [5]Center for Limnology, University of Wisconsin-Madison, Madison, WI, USA

Key Points:
- Process-Guided Deep Learning (PGDL) models integrate advanced empirical techniques with process knowledge
- We used PGDL to accurately predict lake water temperatures for various conditions
- PGDL performance improved significantly when pretraining data included diverse conditions generated by an existing process-based model

Supporting Information:
- Supporting Information S1

Correspondence to:
J. S. Read,
jread@usgs.gov

Citation:
Read, J. S., Jia, X., Willard, J., Appling, A. P., Zwart, J. A., Oliver, S. K., et al. (2019). Process-guided deep learning predictions of lake water temperature. *Water Resources Research*, *55*. https://doi.org/10.1029/2019WR024922

Received 3 FEB 2019
Accepted 23 OCT 2019
Accepted article online 8 NOV 2019

Abstract The rapid growth of data in water resources has created new opportunities to accelerate knowledge discovery with the use of advanced deep learning tools. Hybrid models that integrate theory with state-of-the art empirical techniques have the potential to improve predictions while remaining true to physical laws. This paper evaluates the Process-Guided Deep Learning (PGDL) hybrid modeling framework with a use-case of predicting depth-specific lake water temperatures. The PGDL model has three primary components: a deep learning model with temporal awareness (long short-term memory recurrence), theory-based feedback (model penalties for violating conversation of energy), and model pretraining to initialize the network with synthetic data (water temperature predictions from a process-based model). In situ water temperatures were used to train the PGDL model, a deep learning (DL) model, and a process-based (PB) model. Model performance was evaluated in various conditions, including when training data were sparse and when predictions were made outside of the range in the training data set. The PGDL model performance (as measured by root-mean-square error (RMSE)) was superior to DL and PB for two detailed study lakes, but only when pretraining data included greater variability than the training period. The PGDL model also performed well when extended to 68 lakes, with a median RMSE of 1.65 °C during the test period (DL: 1.78 °C, PB: 2.03 °C; in a small number of lakes PB or DL models were more accurate). This case-study demonstrates that integrating scientific knowledge into deep learning tools shows promise for improving predictions of many important environmental variables.

1. Introduction

Scientific knowledge advances through progress in empiricism and theory. Empirical observations of environmental dynamics give researchers hints about how systems are structured and how they may function, while theory synthesizes information into conceptual frameworks where data are used to test or refine scientific understanding. The rapid growth of data along with advances in computation have led to powerful empirical tools such as deep learning (DL; see LeCun et al., 2015; or Shen, 2018 for DL background tailored for water scientists) that can make accurate predictions based on data alone, which has sparked a discussion regarding the reduced need for theory in science (Mazzocchi, 2015). Indeed, DL tools have been applied to water resource challenges ranging from model parameterization (e.g., Gentine et al., 2018) to image processing (e.g., Islam et al., 2018; Karpatne et al., 2016; Rezaee et al., 2018) with promising results. In contrast, process-based models encode our understanding of the world (theory) developed from decades of observations and experiments. Process-based models are also successful in tackling water resource challenges (Kollet & Maxwell, 2008), predicting outside the range of data sets on which they were developed (Me et al., 2018; Scibek & Allen, 2006; Winslow et al., 2017), and providing the basis for exploring scenarios of environmental change (Cobourn et al., 2018). Process-based models remain the preferred choice for environmental modeling (Fatichi et al., 2016; Hipsey et al., 2015).

Deep learning and process-based modeling approaches each have drawbacks that can reduce trust in outcomes and limit their application. DL models often require extensive training data sets in order to learn the dynamics of complex systems (e.g., H. Chen et al., 2018). Since these models have no

assumptions of the processes underlying the data, known laws or theory are ignored (such as conservation of energy) and this omission can lead to spurious and inaccurate predictions (e.g., Lazer et al., 2014; Nayak et al., 2013), particularly when predictions are made outside the range of data used to train the DL models. In contrast, while process-based models are strongly rooted in scientific theory, their implementation represents a subset of the real processes controlling ecosystems, leading to a number of constraints (Hilborn & Mangel, 1997). When confronted with data from environmental systems, the calibration of these models may be strongly influenced by real-world processes not included in the model (Arhonditsis & Brett, 2004) and process-based models often diverge from theory (Clark et al., 2016). Moreover, these models are typically designed with rigid relationships between numerical code and data; additional data beyond what is used to configure and drive process-based models cannot be integrated without major effort (e.g., adding new predictors), increasing the lag between data growth and modeling improvements.

A new modeling paradigm—"Theory-Guided Data Science" (TGDS; Karpatne et al., 2017)—is designed to combine the strengths of empiricism and theory. TGDS models use advanced empirical methods to extract pattern from data while also imposing structure or rules based on scientific theory. Because these hybrid models can be designed to remain true to accepted theory or physical laws while also learning very complex relationships when data are abundant, their predictions tend to be physically and biologically realistic, and more accurate than process-based models (see Fang et al., 2017; Humphrey et al., 2016; Hunter et al., 2018; Jia et al., 2019). Under the broad umbrella of TGDS models, Process-Guided Deep Learning (PGDL) pairs Earth systems process understanding with the most promising class of predictive tools. Deep learning currently represents the state-of-the-art model architectures for prediction, as evidenced by their success in a number of challenging tasks, including computer vision, natural language processing, and drug design (H. Chen et al., 2018; Hof, 2013; Howard, 2013; LeCun et al., 2015). Novel examples of PGDL relevant to environmental predictions include the use of domain knowledge to reduce training data needs for computer vision tasks (Stewart & Ermon 2014), the use of a physical constraint as a loss term in an artificial neural network (Karpatne et al., 2018), and the recent advancements by Jia et al. (2019) to include penalties for energy conservation violations and initialize recurrent neural network weights by "pretraining" on the predictions from existing process-based models. Despite the compelling results from early PGDL efforts, applications to Earth-science modeling will likely be limited until PGDL can be shown to provide reliable generalizability to multiple systems and extrapolation beyond the original training data set.

Societally relevant water resources information and new research pathways will emerge from improved predictive accuracy for foundational environmental variables. Water temperature is one such variable, considered an ecosystem "master factor" because it controls metabolism, influences water chemistry, and is directly linked to growth, survival, and reproduction of fish (Brett, 1971; Magnuson et al., 1979). Further, the thermal regimes of aquatic ecosystems are changing (O'Reilly et al., 2015), and understanding the physical, biological, and economic consequences of these changes is a primary challenge for water resource scientists and managers. Unfortunately, water temperature data are lacking at the relevant spatial and temporal scales needed for decision-making, including in waterbodies with sparse observations or during previously unobserved time periods. Existing predictions of lake temperature at scale are limited to surface waters (e.g., Bachmann et al., 2019; Toffolon et al., 2014) or, when extended to deeper waters, have limited accuracy (J. S. Read et al., 2014; Winslow et al., 2017). Process-Guided Deep Learning models have shown great promise to improve prediction accuracy of water column lake temperatures; however, implementations are currently limited to one or two well-observed lakes where models were trained and evaluated under similar conditions (Jia et al., 2019; Karpatne et al., 2018). To fill this information gap, additional efforts are needed to explore the use of PGDL models at increased spatial scale and in conditions where representative training data do not exist.

Here we evaluate Process-Guided Deep Learning for lake temperature prediction. Our objectives are to (1) evaluate the effect of incomplete data (sparse data and out-of-bound data) on predictions from process-based, deep-learning-only, and PGDL models; (2) Improve the understanding of how synthetic (model-generated) data can augment limited environmental observations for PGDL training; and (3) test whether we can scale predictive modeling beyond highly studied systems by applying the PGDL modeling approach to dozens of lakes in the Midwestern United States.

2. Methods

2.1. Overview

We describe a method for integrating process knowledge (i.e., theory), advanced empirical techniques, and observational data sets to predict environmental variables, specifically lake water temperature at multiple depths. We evaluated three methods for predicting water temperature, one method was based on theory, another on empiricism, and the third method was a hybrid of the first two. First, we used an existing process-based model to generate predictions that are consistent with the known effects of meteorological drivers on lake temperature. Next, we developed deep learning tools to learn empirical relationships between meteorological drivers and patterns in lake water temperatures to make temperature predictions. Lastly, by combining these two approaches, we created a Process-Guided Deep Learning (PGDL) model by integrating process knowledge of energy conservation into the DL modeling framework (this formulation was first implemented and described by Jia et al., 2019). The PGDL model learned patterns in lake temperature data from predictions generated by a default configuration of the process-based model, and then refined those predictions based on training from lake temperature observations and a loss term that penalized predictions that did not conserve energy. We built on the PGDL framework of Jia et al. (2019), with additional exploration of the limits of PGDL and background written specifically for a water resources audience. Here data sparsity experiments originally presented by Jia et al. (2019) were refined and used to reevaluate our understanding of the impact of sparsity on modeling approaches, and new experiments were designed to examine model performance during out-of-bound conditions, contrast the effects of using different amounts of pretraining data, and apply PGDL at scale for a diverse collection of temperate lakes in the Midwest United States. Descriptions of the data, model components, PGDL integration, and model evaluations are found below, and more detailed information and links to code and data sets can be found in Texts S1–S5 in the supporting information.

The objectives of this study were to build and evaluate models for predicting lake temperature dynamics at multiple depths for temperate lakes. As such, a brief background on lake thermal regimes is relevant to understanding this challenge (see Wetzel & Likens, 2000 for a more detailed overview). Lake temperatures in temperate regions respond to seasonal changes in air temperature and solar radiation, and often form a surface layer of ice cover during the winter. The minimum temperatures for these lakes are typically bounded by the freezing point of water (at or near 0 °C), and maximum temperatures are reached in middle to late summer, with values in our study region between 25 and 35 °C. Most sources of incoming energy are absorbed by surface waters, warming the water and decreasing water density. If mixing energy (e.g., from wind) is insufficient to redistribute surface heating, a vertical density gradient forms in the lake and temperature dynamics of deeper waters begin to diverge from surface waters. This phenomenon is called thermal stratification. In the temperate region, thermal stratification can generate differences between surface and bottom waters of up to 30 °C. Accurate prediction of lake water temperature requires that models incorporate temperature changes from prevailing weather conditions while also reproducing features resulting from the presence, absence, strength, and duration of thermal stratification, including differing dynamics of surface and bottom waters. Additionally, while the thermal dynamics above were described in a one-dimensional (vertical) context, there are numerous factors that can generate horizontal heterogeneity in lake temperatures (e.g., differences in spatial patterns of heating, cooling, and/or mixing energy sources). These three-dimensional patterns in drivers and water temperatures were ignored for this study as all three modeling approaches focus on predicting one-dimensional lake temperatures through time.

2.2. Data Sources

Lake temperature in situ measurements were used to train and test all models. Sources of lake temperature observations included the Water Quality Portal (E. K. Read et al., 2017), North Temperate Lakes Long-Term Ecological Research program, and databases from the Minnesota and Wisconsin Department of Natural Resources. The search was limited to observations from lakes in the U.S. states of Minnesota and Wisconsin in the years 1980–2018. Most data were discrete water temperature profile measurements. A smaller number of lakes were instrumented with buoys that measured temperature continuously for certain periods of the year. All data were reduced to a single vertical temperature profile per lake-day. In the case of continuous measurements or multiple discrete measurements per lake-day, observations closest to noon local time were used to represent the daily profile. In the case of multiple sampling locations per lake-day,

the mean temperature across locations was used. Lakes were chosen for this analysis (see Experiment 3 below) based on the availability of temperature observations; lakes were included if they had at least 200 sampling dates with at least five observation depths per date, and were stratified (temperature differential of more than 1 °C between the shallowest and deepest depths measured) for >70% of those profiles. We chose stratified lakes to model as they represent more challenging tests for all of the models used (versus predicting temperature in a single well-mixed water layer).

Meteorological data were gathered from gridded data sets and formatted as predictors for water temperature modeling. Gridded data are necessary when scaling modeling domains beyond well-monitored individual systems. Following Winslow et al. (2017), we downloaded North American Land Data Assimilation System (NLDAS-2; Mitchell, 2004; Xia et al., 2012) primary forcing data for each grid cell that contained the centroid of a lake in our study, and transformed the variables into process-based model inputs (see Hipsey et al., 2019). The sum of incoming and outgoing energy fluxes is the primary control on lake temperature change (Lenters et al., 2005; Wetzel & Likens, 2000), and these fluxes were computed from daily time series inputs that included air temperature, shortwave radiation, longwave radiation, windspeed, relative humidity, and precipitation. All inputs were normalized for use as features in the DL network, and a second nonnormalized copy of the energy fluxes was used to calculate a process constraint within the same model (only used for the PGDL formulation; see details below) and also as drivers for lake-specific process-based models.

Weather stations that measured local meteorological conditions provided an enhanced source of predictor data for two of the study lakes. There exists a trade-off between coverage and accuracy for meteorological data products, as national/global data sets have known biases and inaccuracies (Xia et al., 2012) not present in carefully maintained and appropriately instrumented local observatories. For Lake Mendota (43.1113°N, −89.4255°E) and Sparkling Lake (46.0091°N, −89.6995°E), we assembled daily time series of the same variables described above which were measured from fixed stations approximately 2 and 10 km away from the two lakes, respectively. Gridded NLDAS data sources for solar radiation and relative humidity were used in place of weather station data for Sparkling Lake because instrumentation issues were present during significant parts of the modeling period for these two variables.

Contextual data collated for each lake included water clarity, lake size, depth, shape, and surrounding landscape type (Table S1 in the supporting information). Earlier multilake modeling studies in this region assembled data for process-based modeling (J. S. Read et al., 2014; Winslow et al., 2017), and we drew from their data sets for these variables. See Winslow et al. (2017) for a detailed description of the sources and processing of these data, but in brief: bathymetric maps from a variety of sources were digitized to generate depth/area relationships for each lake, water clarity estimates from in situ observations (E. K. Read et al., 2017; Soranno et al., 2017) and remote sensing images (for methods, see Torbick et al., 2013) were averaged into a single water clarity estimate for each lake, the National Hydrography Data set medium resolution "Permanent identifiers" (version 2; Simley & Carswell, 2009) were used to connect data to each lake's surface geometry, and lastly, the dominant land-cover type within a 100-m shoreline buffer from the 2011 National Land-Cover Database (Homer et al., 2015) was used to estimate the degree of wind-sheltering (Markfort et al., 2010) for each lake. These data were used to set parameters in the process-based lake temperature model but were omitted from the DL modeling process (with the exception of the depth/area relationship, which was used in the PGDL formulation).

2.3. Model Components

To extract pattern from lake water temperature observations, we used a Long Short-Term Memory (LSTM) network as our primary deep learning model component (Gers et al., 1999; Hochreiter & Schmidhuber, 1997). An LSTM is a type of Recurrent Neural Network (RNN) that includes specialized memory cells that can capture multitime-step relationships, such as recognizing that water temperature change is slow and muted in the fall in contrast to the highly variable spring warming period. Temperature dynamics in lakes fluctuate with differences in heating and cooling. These patterns have basic temporal structure (e.g., seasonal differences between summer and winter) coupled with short-term responses to prevailing weather conditions. With appropriate training and predictor data, LSTMs have the potential to simulate these time series dynamics. We built DL models as LSTMs (these empirical-only models are referred to simply as "DL" below)

with 20 hidden units and the normalized meteorological inputs specified in section 2.2 for simulating daily lake water temperatures at multiple depths.

To make predictions of water temperature based on theory, we used the General Lake Model (GLM version 2; Hipsey et al., 2019) for process-based modeling. GLM is an open-source, one-dimensional lake hydrodynamic model that balances fluxes of mass and energy on a daily (or subdaily) time step, and tracks state variables (such as temperature) with Lagrangian layers resolved in the vertical dimension (see https://github.com/AquaticEcoDynamics/GLM). GLM contains complex vertical mixing routines that redistribute heat in response to prevailing conditions and external forcing. Processed-based modeling of lake temperatures has a long history of software development and application, and many other models exist that are similar to GLM both in performance and formulation of physical processes (SIMSTRAT: Goudsmit et al., 2002; DYRESM: Imberger, 1981; MINLAKE: Riley & Stefan, 1988); many similar models were reviewed and compared by Perroud et al. (2009). GLM was chosen for this study because of its proven ability to simulate thermal dynamics in lakes and reservoirs, and because the open-source codebase and supported integration with other modeling modules (via the Framework for Aquatic Biogeochemical Models; Bruggeman & Bolding, 2014) gives modelers additional control over simulation complexity.

To establish a theoretical underpinning for hybrid water temperature modeling, we used the law of conservation of energy. Conservation of energy is the primary law implemented in GLM and other similar process-based models, and it is a critical component for evaluating the physical reasonableness of water temperature predictions. A simplified numerical energy budget (see equation (1)) that could be used in a hybrid formulation was designed by evaluating the magnitude of terms (Lenters et al., 2005). We reduced complexity by including only primary contributions to energy change at a daily time step, ignoring what are typically relatively small energy fluxes in lakes, such as sediment heating and advection (precipitation, in/outflows, and other water budget components). Collectively, the fluxes that were not included in this simplified model typically account for less than 1% of the daily energy budget terms (Lenters et al., 2005). Additionally, the processes of ice formation, ice melt, time-varying albedo, and sublimation were considered to be beyond the scope of this simplified energy budget model (see Hamilton et al., 2018 for details on lake ice simulations) and were ignored. The resulting formulation was therefore not valid when the lakes were likely ice covered and the constraint was not enforced for predictions during those conditions. The details of this formulation are as follows:

$$\frac{dU}{dt} = \phi_{SW_{in}} - \phi_E - \phi_H + \phi_{LW_{in}} - \phi_{LW_{out}} \tag{1}$$

For process-based thermodynamic models (such as GLM), the conservation of energy requires the volume-averaged change in thermal energy (dU) to match the net energy flux over the same time period (dt). $\phi_{SW_{in}}$ is the incoming shortwave radiation minus reflected (7%), ϕ_E and ϕ_H are the latent and sensible heat fluxes (respectively), $\phi_{LW_{in}}$ is the incoming longwave radiation minus reflected (3%), and $\phi_{LW_{out}}$ is the longwave radiation emitted from the lake. Both values for reflected percentages are used here as constants, and these values (7% and 3%) are commonly used in energy budget studies (e.g., Lenters et al., 2005). All terms are in W/m^2. ϕ_E, ϕ_H, and $\phi_{LW_{out}}$ fluxes require an estimate of the temperature at the surface of the lake (an expansion of these terms and their relationship to other meteorological drivers can be found in Jia et al. (2019) or Hipsey et al. (2019)). The thermal energy at any point in time can be estimated based on water temperatures and depth-specific area of the lake:

$$U = \frac{c_w}{A_s} \sum_{i=1}^{n_{layers}} \rho_i T_i V_i \tag{2}$$

where c_w is the specific heat capacity of water (4,186 J kg^{-1} °C^{-1}), A_S is the surface area of the lake (in m^2), ρ_i is the water density (kg/m^3), T_i is the water temperature (C°), and V_i is the volume (m^3), all for each layer of the model i.

2.4. Integrating Process Into Deep Learning Models

Our method for guiding deep learning with existing theory involved identifying parsimonious constraints from theory and translating these constraints into a deep learning framework. As mentioned

above, conservation of energy is an obvious choice for a physical constraint in the use-case of temperature simulations. Our simplified energy budget formulation was created for this purpose, as each flux term can be calculated directly from DL inputs or from a combination of inputs and DL-predicted surface water temperatures (Jia et al., 2019). The expected balance of these terms—the net thermal energy change over the model time step—can be calculated by combining DL-predicted temperatures with lake geometry (see equation (2)). When predictions and inputs fail to close the energy budget as defined by equation (1), the model framework can penalize model performance at each time step that violates energy conservation.

The penalty for energy conservation violations can be implemented as an element of the DL training objective function. Training a DL model involves iteratively adjusting model parameters to minimize the objective function, which is most commonly accomplished using a process called backpropagation (Werbos, 1988). The objective function quantifies the loss at each training iteration and is simply a weighted sum of the loss from the accuracy, any standard regularization loss terms that penalize for increased model complexity, and any user-designed penalties. The weights for each term are hyperparameters defined by the network's engineer in order to balance the importance of each penalty for the modeling application. For this study, the loss terms included the sum of squared errors between predicted and observed temperature and the mean absolute value of energy conservation violations beyond an assumed error threshold, with zero loss assessed for lesser violations. A loss threshold was used because penalizing all minor energy conservation violations implies all energy conservation terms are present (see section 2.3 description of minor terms that were ignored) and that the measurement of incoming fluxes is without observation errors, neither of which are true in this case. For directly measured energy fluxes, the error threshold was set to 24 W/m^2 (experiments 1 and 2; this threshold choice was informed by the assumed error distribution of measured energy flux components from an earlier study by Lenters et al. (2005)) and set to 36 W/m^2 for the less accurate gridded flux data (Experiment 3 below; this threshold is based on a comparison between gridded data and station data for the location of Lake Mendota). The Adam stochastic gradient-based optimization algorithm (Kingma & Ba, 2014) was used to minimize the loss function with a learning rate of 0.005, and all DL and PGDL training was continued for 400 epochs. Additional details regarding the formulation of the network, the objective function, hyperparameter choices, example code, and links to reproducible PGDL examples can be found in the sections S4 and S5 in the supporting information.

Water temperature predictions from the uncalibrated process-based model were used to pretrain PGDL models to initialize the network structure in advance of training with true observations. Layer-wise pretraining (e.g., Erhan et al., 2010) and network-level pretraining (Jia et al., 2019; Lee et al., 2018) have been shown to significantly increase the prediction accuracy and generalizability of DL models. Following Jia et al. (2019), we performed network-level supervised pretraining (hereafter referred to as pretraining) on PGDL models by using depth-resolved GLM temperature predictions as labels. These process-based simulations were configured with observable lake-specific parameters (e.g., water clarity and lake depth) and driven with meteorological data from 1980 to 2018, and neither the model nor any lake-specific parameters were altered in response to performance relative to temperature observations (see J. S. Read et al., 2014; Winslow et al., 2017 for details; we refer to these models in the following text as "uncalibrated"). These pretraining models differ from the calibrated GLM models described in section 2.5, which instead used temperature observations to inform selection of model parameters that improved prediction in the training period. Pretraining data included all daily temperature outputs at all depths from GLM with the exception of test periods. After training PGDL on pretraining data, the final network parameters became the initial network parameters for the final training procedure which used true temperature observations. Pretraining was used for all PGDLs in this study and provides a potential predictive advantage to the PGDL framework that was not available to the DL models used for comparison. As such, we clarify our intents were to use DL models (and PB models) as a baseline for comparison as opposed to attempting to evaluate models that share the same complexity and initialization procedures (PGDL, PB, and DL differ in both). All three model formulations used the same daily time-varying inputs: air temperature, shortwave radiation, longwave radiation, wind speed, relative humidity, and precipitation (as rain or snow). The resulting Process-Guided Deep Learning models included meteorological data as inputs, an initial network structure established by pretraining, and a loss term that penalized conservation of energy violations (Figure 1 and Texts S4 and S5).

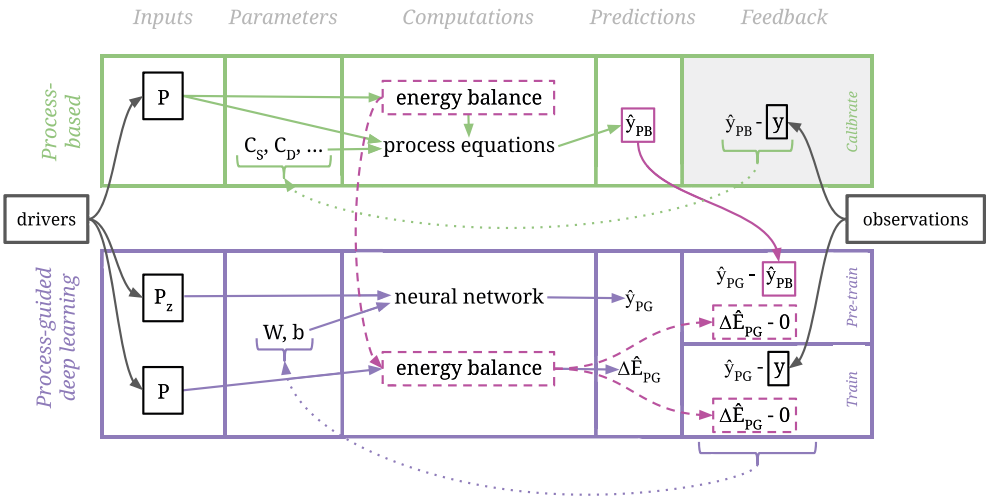

Figure 1. Comparison between process-based model (PB) and process-guided deep learning model (PGDL). Conceptual links between the two are in pink, showing the integration of the energy balance concept (dashed pink lines) and process-model-generated pretraining data (solid pink line) from PB into PGDL. Both models accept data in the form of drivers and observations (black lines; P = predictors, P_z = normalized predictors, y = temperature). Although the models differ greatly in their structures, they have in common that they accept the same raw inputs, use parameters and computations to generate predictions (green and purple solid lines), and revise the parameters based on feedback (called "calibration" for PB or "pretraining" and "training" for PGDL; green and purple dotted lines). PB calibration (gray box and green dotted line) is used for calibrated PB models but is omitted for the experiments in this manuscript when generating uncalibrated predictions ($\hat{y}_{PB}$) for use in PGDL pretraining.

2.5. Model Experiments

We designed experiments to evaluate the performance of different modeling approaches with data that mimic three real-world water temperature prediction challenges. Experiment 1 required models to make temperature predictions for a single lake with a range of monitoring data density, spanning extremely sparse to nearly comprehensive measurements. Experiment 2 tested model performance for two lakes when key time periods were removed from temperature monitoring data, artificially creating scenarios where lake monitoring would only cover certain seasons or years. Experiment 3 tested the ability of models to reconstruct the unobserved past in many lakes using broadly available weather data. Lastly, a fourth experiment evaluated the impact of using different pretraining data sets on the PGDL predictions in Experiments 1 and 2. These experiments did not evaluate the impact of data quality or errors in model inputs or observations on prediction quality, but such an effort could add insights beyond what is presented here. We estimated the model's overall predictive performance by calculating the root-mean-square error (RMSE) between predictions and observations. Although useful, other error metrics such as estimates of bias or variance were not included in the study.

The model training (or the analogous "calibration" step in process-based modeling) and testing procedures involved dividing observations into separate data sets based on sampling date. Calibrated process-based GLM models are hereafter referred to as "PB" to differentiate from the uncalibrated GLM models (these uncalibrated models are referred to below as "PB_0," signifying "0" dates were used for training) that are used to generate labeled data for PGDL pretraining. Test data were used only to independently evaluate the accuracy of temperature predictions after training was complete, and the test data set had no influence on the model training or calibration phase. Experiments 1 and 2 (described below) required five iterations of test and train data. The logic for how observation dates were assigned to train or test data sets varied based on the experiments and details regarding calibration and training can be found in Texts S2–S4 and Figures S4–S14 in the supporting information, in addition to descriptions and links to archives for reproducible code and data sets.

Experiment 1: Effects of Data Sparsity or Abundance on Model Performance

Recognizing that rich data sets are commonly assumed to be necessary for machine-learning-based approaches, we generated tests to evaluate the impact of data sparsity for the three model types. This

experiment used data from Lake Mendota, and divided temperature observations into data sets with 540 test dates and 980, 500, 100, 50, 10, or 2 training dates. Each "date" included observations at multiple depths, with most dates consisting of 23 observations between 0 and 20 m downward from the lake surface. For each training data set (with the exception of the two-profile data sets, which were only used to train the PGDL and PB), DL and PB models were trained or calibrated with the training data set, while PGDL models were first pretrained with output from an uncalibrated Lake Mendota PB_0 model and then trained on the temperature profiles. All models were run continuously through the entire multiyear simulation periods at a daily time step. This experiment is similar to Jia et al. (2019), including the use of Lake Mendota as the study lake. We consider our design to be a refined and improved version of the original experiment that supports more robust conclusions regarding differences between empirical and theory-based approaches. Modifications to the original design are as follows: higher quality locally measured meteorological data were used during test and train periods (compared to less accurate gridded data), more comprehensive PB calibrations were performed (see Text S3), shorter test and train time periods were used that included a greater number of observations from an automated measurement buoy (2009–2017 with 35,242 temperature observations versus 1980–2014 with 13,158 observations), the sparsity range was extended and spanned 46 to 22,776 observations (2 and 980 days, respectively, compared to a range of 161 to 8,037 observations), the method used for creating sparser training involved randomly removing full sampling dates (i.e., all observations on a given date) instead of randomly removing individual observations, and the test periods were set based on random continuous blocks of time instead of a single fixed range (to reduce the impact of differences in the predictability of water temperature across years).

Experiment 2: Assessing Transferability When Predicting Outside the Bounds of Training Data

In order to understand the transferability of process-guided deep learning models to time periods on which they were not trained, predictions for PB, PGDL and DL models were tested for conditions that were purposefully dissimilar to training data. Experiment 2 used the same temperature profiles from Lake Mendota as Experiment 1, as well as data from Sparkling Lake, which is also outfitted with an automated buoy. Sparkling Lake is smaller, clearer, and slightly shallower than Lake Mendota, and is located in northern Wisconsin. Details regarding Mendota and Sparkling can be found in Table S1. The effects of data sparsity were not considered for this experiment, and training data for each prediction challenge included 500 dates. We defined *in-bound* test and training data for comparison purposes using the same collections of 500 profiles from Experiment 1, and two out-of-bound prediction challenges that included (1) prediction of temperatures in warmer years when trained on data for cooler years (referred to as *years experiment* in this section) and (2) the prediction of summer temperatures when summer observations were withheld from the training data set (referred to as *seasons experiment* in this section). According to annual average air temperature, the warmest three years during the experiment period (2009–2017) were 2012, 2016, and 2017. Temperature profiles during these years were assigned to *years experiment* test data sets. For the *seasons experiment*, any profiles taken between Julian dates 173 and 264 (inclusive, representing the summer period from ~ 22 June to ~r21 September) were assigned to the *seasons experiment* test data set. Training data sets were comprised of 500 profiles from the observations remaining after removing test data. Following the same process as outlined in experiment 1, PGDL models were pretrained with PB_0 (pretraining data began in 1980 and were therefore more extensive than the training period), and PB, DL, and PGDL models were trained with the training data sets as described above.

Experiment 3: Scalability: Applying PGDL to Broad-Scale Modeling

One potential advantage of a PGDL approach is a reduced need for site-specific calibration of a process model, which requires a substantial amount of additional handling time and expert judgement to keep PB parameters within appropriate ranges and to avoid overfitting. Though such calibration can improve accuracy, it may be possible to apply PGDL at broad scales and achieve reasonably accurate predictions in many lakes even without PB model calibration. Sixty-eight lakes met the data requirements for model construction and testing that we established for this experiment, including having a least 200 unique observation dates where temperature profiles were taken at five or more depths (Table S2). Temperature data were split into training and test data sets by using the first two thirds of the observations for training (ordered by date), and the remainder for testing. All model formulations were also trained with three iterations of a random selection of 10 unique observation dates from the full training data set to evaluate the impact of sparse

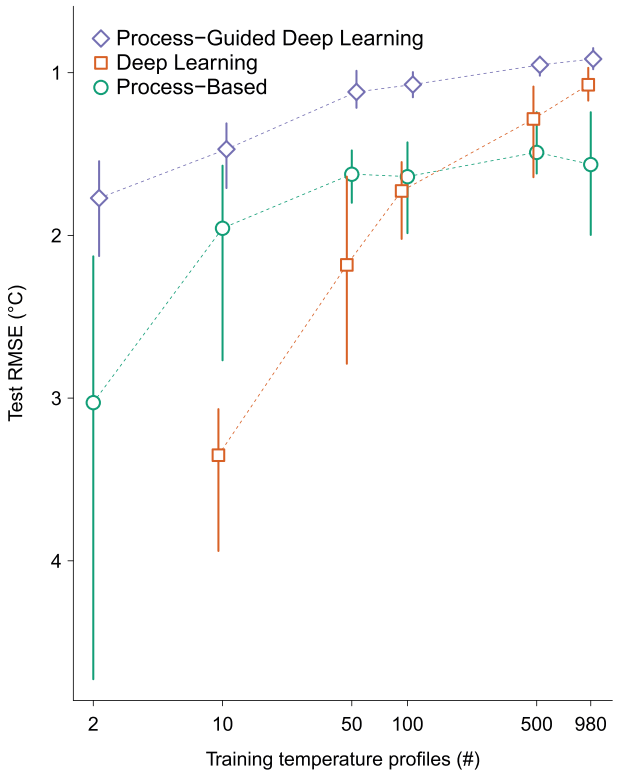

Figure 2. Water temperatures were estimated for Lake Mendota using three different model formulations. The *Process-Guided Deep Learning* model is a Long Short-Term Memory (LSTM) Recurrent Neural Network with energy conservation as a process constraint and temperature pretraining from a process-based model. The *Deep Learning* model is the same LSTM model used in *Process-Guided Deep Learning* (including the same time-varying predictors), but does not have process constraints or pretraining. The *Process-Based* model is a calibrated one-dimensional Lagrangian vertical layer hydrodynamic "General Lake Model" (GLM; Hipsey et al., 2019). Models were trained using subsets of temperature profiles from Lake Mendota, and predictive performance (as measured by root-mean-square error (RMSE)) was calculated for temperature observations from separate test periods. Vertical lines represent the range of RMSE from five iterations of the train/test experiment and markers are the mean.

data conditions across different lakes. As with the PGDL formulations mentioned above, uncalibrated process-based model temperature predictions from PB_0 were used to generate an informative pretraining data set for initializing the DL network, followed by using the site-specific observations from the training data sets to train each of the 68 PGDL models. The pretraining models do not represent the best-available GLM predictions for any one lake, but they do represent what is often done as a tractable scaling approach to predict temperatures in hundreds or thousands of lakes (e.g., Winslow et al., 2017). Pretraining data were generated from GLM simulations that began in 1980 but data were truncated to remove the entire test period (and any dates that preceded the test period). We believe that pretraining data can include the test period in this context without compromising the validity of predictions since the pretraining data consists of untrained model output that is not altered in any way by observations. However, these data were withheld here to make an extremely clear distinction between training and test periods in this manuscript. To allow comparison between PB, DL and PGDL models, PB and DL models were trained or calibrated on training data with the same approaches as above and as described in Texts S3 and S4. All models were then evaluated based on their ability to predict water temperature during the test period.

Experiment 4: Effects of Reduced Pretraining Data on PGDL Performance

In order to quantify the impact of exposing PGDL to a less comprehensive range of conditions during pretraining, the Lake Mendota sparsity and out-of-bound prediction experiments were repeated a second time using reduced pretraining data sets. We refer to these two contrasting configurations of pretraining data as "limited pretraining" and "extended pretraining" (for the reduced and original data sets, respectively). Pretraining data are products from an uncalibrated GLM, which produces continuous temperature estimates at the time step of the model (daily) that were sampled at 0.5-m depth intervals from the water surface to generate a uniform matrix of labeled data. The original pretraining data sets for Experiments 1 and 2 began in April of 1980 and extended into the training periods, with test periods masked from the pretraining process (see test period masks in Figures S3–S8 and S10–S13). The additional *limited pretraining* data sets were created by removing all data preceding the start of the training period from these matrices (1980 to April of 2009 data were removed), greatly reducing the data volume and variety as compared to Experiments 1 and 2. Each of the sparse and out-of-bound predictions for Lake Mendota were then completed in the same way as the originals, with the exception being that *limited* pre*training* data were used instead to pretrain PGDLs. To further illustrate the contrast in data sets, the *limited pretraining* data set in the out-of-bound season experiment (Experiment 2) was less than 20% of the *extended pretraining* data set size and contained no data from any summer period (compared to the 29 summers appearing in the *extended pretraining* data sets from 1980 to 2008). After PGDL pretraining, the models were refined with the same training observations from Experiments 1 and 2 and prediction accuracy was calculated based on the differences between predictions and observations in the test period.

3. Results

3.1. Effects of Data Sparsity or Abundance on Model Performance (Experiment 1)

Process-Guided Deep Learning predictions of Lake Mendota water temperatures were more accurate than predictions from empirical-only and process-based models for all training data conditions from Experiment 1 (Figure 2). The accuracy of all models decreased as fewer observations were used for

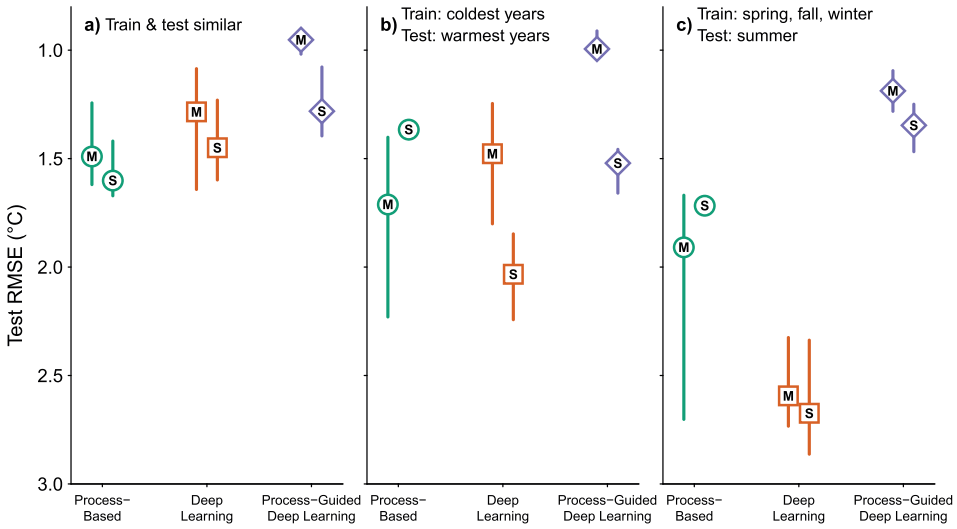

Figure 3. Water temperatures predicted using three model formulations for Lake Mendota ("M") and Sparkling Lake ("S") with contrasting periods for test and training data sets. The vertical lines and markers represent the range and the mean of five iterations, respectively. In (a), the test and training data sets were chosen randomly to represent similar periods following Figure 2. In (b), test data were from the warmest years (2012, 2016, and 2017), and training data were randomly sampled from the remaining years (2009–2011 and 2013–2015). In (c), test data were from the summer period (day of year 173 to 264) for all years (2009–2017) and training data were randomly sampled from the remaining dates.

training, although the impact on model accuracy was different for each of the three model types. Empirical-only (basic RNN as *Deep Learning*) performance suffered the most and the hybrid approach (*Process-Guided Deep Learning*) performance was impacted the least as the amount of training data was artificially reduced (Figure 2). Although the PGDL and DL models had similar performance when trained with a high volume of temperature observations, the performance of the two models quickly diverged when data were increasingly sparse.

When sufficient observed temperature data existed, the empirical-only and hybrid models (DL and PGDL) outperformed the calibrated process-based (PB) model (based on RMSE; Figure 2; see 980 profiles). Training (or calibrating) each model on 980 days of observations of Lake Mendota water temperatures and evaluating performance on 540 independent days (referred to above as "test") resulted in mean RMSEs (means calculated as average of RMSEs from the five data set iterations) of 0.92, 1.07, and 1.56 °C for the PGDL, DL, and PB models, respectively. The differences between training and test errors was greatest for the DL model (average train RMSE: 0.75 °C, average test RMSE: 1.07 °C), followed by PGDL (0.74 versus 0.92 °C), with the smallest differences occurring for PB simulations (1.5 versus 1.56 °C).

The relative performance of DL versus PB depended on the amount of training data. The accuracy of Lake Mendota temperature predictions from the DL was better than PB when trained on 500 profiles (1.28 and 1.49 °C, respectively) or more, but worse than PB when training was reduced to 100 profiles (1.73 and 1.64 °C, respectively) or fewer. The PGDL prediction accuracy was more robust compared to PB when only two profiles were provided for training (1.77 and 3.03 °C, respectively). As an indication of how common these different monitoring regimes are in practice, the multilake data set contained 2 lakes with at least 980 profiles (Mendota and Sparkling Lakes), 9 lakes with at least 500 profiles, 267 with 100, 558 with 50, 1,736 with 10, and 3,602 lakes with at least two temperature profiles (see Figure S2).

3.2. Assessing Transferability When Predicting Outside the Bounds of Training Data (Experiment 2)

The accuracy of predictions from the PGDL model was superior in most cases to the performance of the empirical-only (DL) and process-based (PB) models when applied to conditions outside the bounds of training data (Figures 3b and 3c). The exception to this result was that PB models outperformed PGDL models in Sparkling lake for the *years* experiment, when the three warmest years were withheld from model

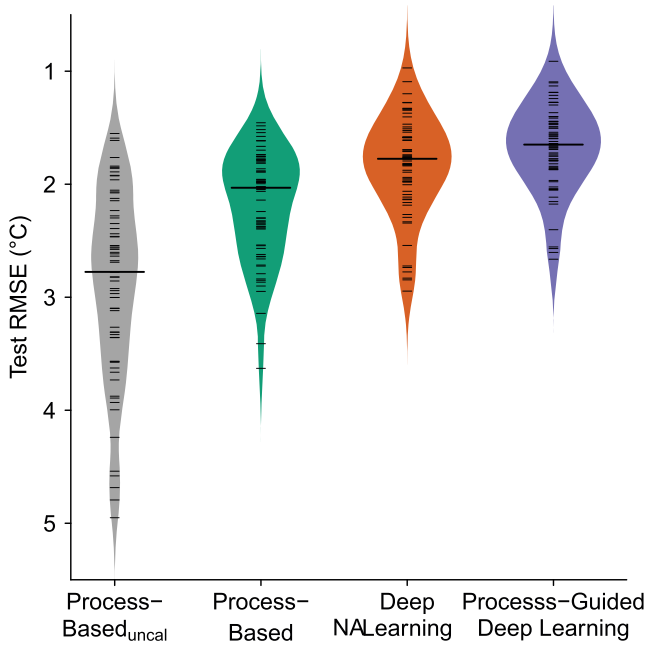

Figure 4. Kernel density plots for accuracy of historical water temperature predictions as measured by RMSE in 68 lakes with difference model formulations, including *Process-Based* (PB), *Deep Learning* (DL), and *Process-Guided Deep Learning* (PGDL). *Process-Based$_{uncal}$* models were not calibrated with any observations (these models are source of uncalibrated process-based model output used in pretraining PGDL models) and all other models were trained from the most recent two thirds of each lake's observation data set (varied from a minimum of 133 to a maximum of 1,788 unique training dates). The prediction accuracy of individual lakes is shown as short black dashes and the median of all lakes is shown as a longer and thicker horizontal black line. Accuracy was calculated as root-mean-square error between the model predictions and the observations in the test period (test period was the earliest one-third of each lake's observation data set).

construction (1.37 and 1.52 °C RMSE for PB and PDGL, respectively). *In-bound* predictions (Figure 3a) were generally more accurate for all three modeling approaches compared to out-of-bound predictions (Figures 3b and 3c), with both exceptions appearing in the *years* predictions. Lake Mendota's PGDL *years* predictions were approximately the same accuracy as the *in-bound* prediction mean (1.71 and 1.49 °C), while Sparkling Lake's PB *years* predictions were an improvement over *in-bounds* (1.37 and 1.6 °C). Process-based models were more accurate in their out-of-bound predictions than DL models except for Lake Mendota in the years experiment (1.71 °C for PB and 1.48 °C for DL; see also Figures S14–S17). Sparkling Lake predictions were less accurate than Lake Mendota predictions for six of the nine combinations of model type and prediction challenge evaluate, with the three exceptions being PB *years*, PB *seasons*, and DL *seasons*. Near-surface summertime water temperature predictions from DL *seasons* (Figure 3c) for both lakes were biased cold, while PGDL and PB predictions at the same depths were more accurate (Figures S16 and S17).

3.3. Scalability: Applying PGDL to Broad-Scale Modeling (Experiment 3)

Predictions from PGDL models applied to 68 lakes were more accurate or as accurate (within ±0.05 °C RMSE) as all but five of the calibrated PB models and five of the DL models (Figure 4 (see PGDL, PB, and DL) and Figure S18 for detailed lake-specific results; all RMSE values reported here correspond to model performance in the test period). The median RMSE (across all lakes) was 1.65 °C for PGDL, 1.78 °C for DL, and 2.03 °C for PB. The range of prediction accuracy for PGDL models was 0.91 to 2.66 °C, 0.97 to 2.95 °C for DL, and 1.46 to 3.63 °C for PB. The PB_0 predictions (which were used to pretrain PGDL) had a median RMSE of 2.78 °C, with a range of 1.55 to 4.95 °C (see Process-Based$_{uncal}$ in Figure 4). The improvements in accuracy of PGDL compared to the PB_0 predictions used for pretraining were variable across lakes, with 11 lakes improving RMSE by over 2° compared to pretrainer RMSEs, 55 lakes improving by smaller amounts, and 2 lakes with PGDL prediction accuracy that was approximately equal to the pretrainer accuracy in the test period (within ±0.05 °C RMSE). When comparing performance of predictions on individual lakes, the difference in RMSE between PB and PGDL ranged from −0.23 to 2.01 °C and −0.28 to 1.15 °C for DL to PGDL (positive values indicate better performance by PGDL; see also Figure S18). When observations were artificially removed to leave only 10 dates for training, predictions from PGDL models were more accurate or as accurate as 50 of the calibrated PB models (73.5% of total) and more accurate than all DL models (Figure S18; see $PGDL_{10}$, PB_{10}, and DL_{10}, respectively).

3.4. Effects of Reduced Pretraining Data on PGDL Performance (Experiment 4)

When observations were sparse or environmental conditions differed between the training and test periods, models pretrained with more comprehensive synthetic temperature data were substantially more accurate than models with smaller pretraining data sets (Figure 5; sim_2, sim_{10}, sim_{50}, sim_{100}, $year_{500}$, and $seas_{500}$). However, the impact of reduced pretraining data on PGDL accuracy was minimal when training data were plentiful and training periods were similar to the test periods (Figure 5; sim_{980} and sim_{500}). The two largest differences in prediction accuracy between the extended and limited pretraining data sets were when only two profiles were used for training (1.77 versus 2.67 °C RMSE; Figure 5; sim_2) and when models were trained using data from colder seasons and used to predict summer temperatures (1.19 versus 1.98 °C RMSE; Figure 5; $seas_{500}$). Excluding summer data from pretraining decreased PGDL performance and resulted in worse temperature estimates than the calibrated PB model (as measured by RMSE; 1.98 versus 1.91 °C, respectively; Figure 3c; *Process-Based* "M" versus gray filled marker in Figure 5; $seas_{500}$). When models were trained on colder years and used to predict warmer years, the additional complete years included in the

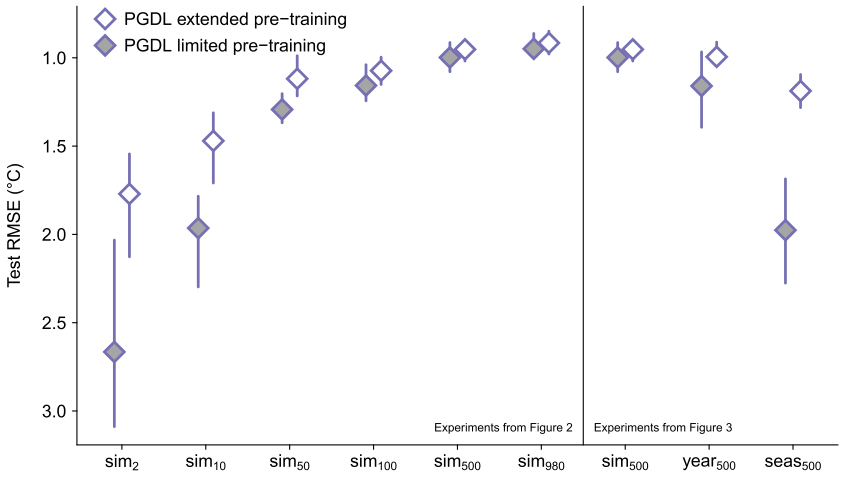

Figure 5. Effect of pretraining data volume and variety on accuracy of temperature predictions from *Process-Guided Deep Learning* (PGDL) models as measured by root-mean-square error (RMSE) for Lake Mendota. Different amounts and time periods of synthetic data were generated from a process-based model to pretrain PGDL models for the experiments presented in Figures 2 and 3. RMSE values for each prediction as aligned along the *x* axis and labeled with shorthand for the experiment and a subscript for the number of training profiles used. "sim" is shorthand for similar, where training and test periods in Figure 2 were designed to have similar conditions. "Year" and "seas" are shorthand for the warm year and warm season (summer) test periods shown in Figure 3. The "extended pretraining" (open diamonds) included pretraining data that began in April of 1980 and excluded test periods, and is identical to the Lake Mendota results presented in Figures 2 and 3. "Limited pretraining" (gray-filled diamonds) included pretraining data only from the training period (also excluding the test period). The vertical bars show the range of prediction accuracy for five iterations of training data splits and the markers are the mean.

extended pretraining data set were likely responsible for the 0.17 °C RMSE improvement over predictions from the limited case (0.99 versus 1.16 °C, respectively; Figure 5; $year_{500}$), and for reducing the difference in accuracy between *similar* and *years* prediction challenges for Lake Mendota from 0.16 °C to a negligible difference (difference between gray filled diamonds for $similar_{500}$ (0.95 °C) and $year_{500}$ (0.99 °C) versus differences between open diamonds for the same in Figure 5). The difference in prediction accuracy between limited and extended pretraining PGDL models monotonically decreased as data became more plentiful (Figure 5, left to right for $sim_{\{n\}}$).

4. Discussion

We show here that Process-Guided Deep Learning can be used to address important prediction problems in aquatic science and has the potential to accelerate knowledge discovery by explicitly combining empiricism and theory. Many in the environmental sciences have called on the community to embrace machine learning as a powerful predictive tool (Hampton et al., 2013; Mosavi et al., 2018; Olden et al., 2008), and build theoretical knowledge directly into these models (Karpatne et al., 2017; Shen et al., 2018). Indeed, Shen et al. (2018) suggest the first step in pursuit of deep learning-powered scientific advances for hydrology is to "integrate physical knowledge, process-based models, and DL models." The hybrid modeling approach highlighted in this study combines process-based and deep learning models and has great relevance to disciplines historically dominated by the development and application of process-based models, such as physical limnology and hydrology.

Several components were responsible for the PGDL's superior predictive accuracy, including temporal recurrence, process constraints, and pretraining (Figure 1). The choice of an LSTM to be the underlying predictive engine for PGDL helped the models recognize critical time series patterns and relationships that led to improved predictions, evidenced by the LSTM's ability to estimate temperatures when trained on adequate data (see *Deep Learning* in Figure 2 and Figure 4). Most process-based models have core formulae or encoded physical laws that ground simulations to reality, and energy conservation is the law at the heart of thermodynamic models. By adding energy conservation as a process-based loss term in the PGDL's objective

function, the model was able to learn physically valid responses to meteorological drivers, likely advantageous when asked to predict out of bounds from the conditions of training data (Figures 3b and 3c). Tracking energy flow required an augmentation to the standard LSTM architecture (see Jia et al., 2019 for details) in order to expose each flux component to the loss term as a calculation from model inputs or a combination of inputs and PGDL-predicted surface water temperatures. Pretraining DL models with synthetic data can be used to overcome conditions when environmental observations would otherwise be too sparse or not representative enough of test periods to support advanced machine learning techniques such as LSTMs. In Figure 4, we showed that PGDL improvements over the accuracy of the trainer can be substantial and that this framework can be scaled to many lakes, while the results from Figure 5 suggest that adequate volume and variety of pretraining data can help compensate for limited training data.

Despite the data deluge from new types of environmental data, many present-day modeling challenges will require accurate predictions where data are scarce. Even with the growth of new sensing technologies, the majority of lakes and streams are unmonitored or have only a few observations. Direct environmental measurements have continued to accumulate, but the exponential data growth that has made DL methods more tractable has happened elsewhere. Large-scale model outputs (e.g., multiple configurations of MODFLOW, Fienen et al., 2018; the National Water Model, Hooper et al., 2017; climate models, Scher, 2018), remote sensing data (e.g., Karpatne et al., 2016; Schaeffer et al., 2018), and hybrid modeled/observed gridded data sets (e.g., NLDAS, Mitchell, 2004) are the "big data" foundation for water resources (Y. Chen & Han, 2016), and the volume of direct in situ observations is small by comparison. Even the "richest" case of observations used in the study presented here included hundreds (not thousands or millions) of daily profiles for training (Figure 2). As such, approaches that combine existing models or theory with new ways to utilize data are needed (Fang et al., 2017; Karpatne et al., 2017). We have shown here that the greatest gains of PGDL temperature modeling (when compared to process-based models or more traditional formulations of DL) were when observations were relatively scarce (Figure 2). Even a small number of water temperature sampling dates (e.g., 2–100) can be used to train a PGDL model that results in substantial improvements over empirical-only and process-based approaches (Figure 2). At the scale of tens to thousands of lakes, the traditional option has been to accept reduced model performance of an uncalibrated process-based model or substantially increase the required handling time and expert knowledge to calibrate process-based models for individual lakes. PGDL therefore represents a promising method for scaling prediction to the scale of environmental problems, and these models performed well across a diverse set of lakes (Figure 4). This finding is important for future hybrid modeling efforts since temperature is one of the most widely measured variables in water resources (E. K. Read et al., 2017) and data scarcity challenges are likely greater for other applications in water quality modeling.

Pretraining was integrated successfully into a PGDL water temperature prediction framework by Jia et al. (2019) and builds on other modeling concepts that continue to be relevant for future exploration, such as statistical bias correction (e.g., quantile mapping; Panofsky & Brier, 1958) and model emulators. Complex process-based model outputs have been successfully approximated elsewhere with deep learning (e.g., Scher, 2018) and other machine learning techniques (e.g., Fienen et al., 2018; Yan & Minsker, 2006). Alternatively, the DL pretraining process (Coto-Jimenez, 2018; Erhan et al., 2010) consider DL models trained by synthetic data to be an intermediate (as opposed to final) product of the predictive framework. With the pretrained DL model as a starting point, final training is completed with actual observations and imperfections and/or biases in the source model (the "pretrainer") can be reduced or eliminated through the iterative training process. In this way, pretrained DL models have the opportunity to exceed the performance of their trainer (Jia et al., 2019; Lee et al., 2018) in contrast to earlier efforts designed to emulate physical models.

The quantity and information content of the synthetic data used to pretrain PGDL models was critical to prediction accuracy (Figure 5). The current PGDL formulation can be thought of as a DL emulator of a process-based model, which is then further refined (or debiased) with observation-based training and encouraged to remain physically realistic via an energy conservation loss term (Figure 1). Process-based model outputs are used to build the emulator, and therefore, the properties of these data are an important factor in determining how effective the pretraining process is at establishing a useful initial network state for PGDL. We contrasted the performance of models pretrained with extended versus limited synthetic data by shortening the data sets and removing key periods (Experiment 4; Figure 5). We found that it was critical to expose PGDL to

pretraining data that contained similar conditions to those found in the test period, most clearly illustrated by the difference in performance for predicting summertime temperatures when pretraining data included prior summers versus did not (Figure 5; $seas_{500}$ open versus gray markers, respectively). Greater quantities of pretraining data were likely responsible for improved predictions in data sparse conditions, but the differences between limited and extended pretraining data decreased as observation-based training data increased (Figure 5, left side). Although we did not separately test the contributions of energy conservation and pretraining to improvements in predictive accuracy, it is likely that pretraining is responsible for the majority of predictive gains. Jia et al. (2019) isolated the effects of the energy conservation loss term and found small but consistent improvements in accuracy when compared to an otherwise identical LSTM (although this finding was not tested for out-of-bound training data), and larger improvements when models added a pretraining step. As such, the energy conservation penalty likely serves as a regularization term that improves the generalizability of the model, while pretraining acts as a powerful surrogate for real observations when comprehensive training data are limited. The evaluation of pretraining data set design performed in our study was not comprehensive and presents new opportunities for deeper exploration, including assessing data accuracy impacts (e.g., using calibrated instead of uncalibrated models for pretraining), understanding pretraining volume and variety needs (e.g., pretraining with artificially elevated ranges for weather conditions or using ensembles of models with a variety of formulations), exploring improvements to pretraining routines by altering loss terms and stopping criteria (our implementation used identical parameters and procedures for both pretraining and training), and testing the utility of pretraining for predicting other environmental phenomena and other process-based model formulations.

The work presented here is a significant step forward in the pursuit of hybrid process-based deep learning predictions, but many opportunities exist to extend and improve the approach. Our study limited the predictive challenge to a collection of lakes that were treated as independent model systems (no relationships between lakes were used or explored to improve models) and relatively simple neural network structures were used for DL and PGDL. Future work could rely on the high degree of physical coherence across lakes (Benson et al., 2000; Palmer et al., 2014) to share information and improve models in data poor or unmonitored lakes. Given the importance of interconnections in water resources modeling (e.g., Wagener et al., 2010), this research avenue is an important next step and has parallels with predicting streamflow in unmonitored basins (e.g., Hrachowitz et al., 2013; Tongal & Booij, 2018; Worland et al., 2018). Also, although it was clear that the uncalibrated process models used in pretraining varied in how well they represented temperature dynamics (see Figures 4 and S18), evaluating the impact of pretrainer model quality on PGDL predictions was beyond the scope of this paper (see also Shen et al., 2018 for a similar charge: "The extent to which errors in [PB] model results affect DL outcomes remains to be explored"). Despite this, we expect that improved pretrainer models would translate into improved PGDL predictions. Additionally, while PGDL models clearly outperformed DL and PB predictions when the inputs and temperature observations were carefully curated (Lake Mendota and Sparkling Lake; Figures 2 and 3), the results were nuanced when multiple monitoring campaigns were included (most lakes other than Sparkling and Mendota had multiple spatial sampling locations that were treated as the same) and lower quality inputs were used (gridded weather data versus directly observed). Perhaps the substitution of less accurate gridded data for two metrological inputs used in Sparkling Lake prediction was a factor in why *years* predictions are worse for Sparkling than those for Lake Mendota (Figures 3b and S15 versus S14, respectively). For the 68 lakes simulated, PGDL predictions of temperature were generally an improvement over PB or DL models but were not more accurate in every case (Figures 4 and S18). Future exploration of model structures could improve predictions for PGDL and DL models, as the LSTM models used here were deep in time, but shallow in layers. We used this architecture because prior domain knowledge exists that connects the output (temperature) to the inputs (meteorology) through simple thermodynamic relationships governing temperature change. Additional DL layers provide additional data abstraction, and deeper-layered networks may be useful for more complex relationships (e.g., modeling water quality, Maier et al., 2010; or stream discharge, Shortridge et al., 2016), but are not always necessary (Ba & Caruana, 2014).

It is increasingly necessary for water scientists to make predictions for unseen time periods (including the uncertain future and the unobserved past) using a variety of model architectures and assumptions. Our evaluation of out-of-bound predictions from empirical-only, process-based, and hybrid process-guided deep learning models showed that in an isolated case, PGDL predictions were more accurate than the tested

alternative models (for five of six cases; Figure 3) and suffered a smaller drop in out-of-bound performance when compared to other models (Figure 3). While these experiments were designed to mimic the challenge of modeling an unseen but warmer future (Figure 3b) and a season-ahead forecast (Figure 3b), real forecasts are necessary to begin the cycle of learning (Dietze et al., 2018). As evidence from another domain that the pursuit into forecasting should consider hybrid model architectures, Dueben and Bauer (2018) evaluated the potential for making simplified forecasts of global weather with DL, concluding that while many benefits of DL are clear (especially for short-term forecasts), Earth systems domain knowledge is still a requirement for developing models capable of comprehensive forecasts.

The modeling subdiscipline of "Theory Guided Data Science" (TGDS; Karpatne et al., 2017) covers a spectrum of modeling approaches that blend theory and data-driven models, ranging from simple ML approaches that predict residuals from process-based models (e.g., Demissie et al., 2009) to proposed model structures with tight internal coupling between ML and PB components. Our PGDL model codified process components (theory) into LSTM loss terms, but we did not explore other hybrid architectures, such as the integration of DL components into a process model. Future work could consider using DL components in place of uncertain elements of PB models to allow more flexible data-driven learning. As knowledge, predictive tools, and data co-evolve, the collection of modeling approaches under the TGDS umbrella offers researchers the flexibility to meet science questions with models suited to purpose. Leveraging this new modeling paradigm for water resources requires both an embrace of the role DL can play in predictions, and deliberate efforts to design for flexibility and iteration in future model architectures.

We utilized the PGDL modeling framework to produce improved predictions of lake water temperature, a critically important variable for improving understanding of aquatic ecosystems. PGDL predictions achieved a 0.5 °C reduction in RMSE relative to a calibrated process-based model (Figure 2), and we showed that this modeling framework could be scaled up and used for predictions in many lakes while maintaining favorable accuracy compared to calibrated process-based and deep learning models (Figures 4 and S18). Because temperature is an ecosystem "master variable" (Magnuson et al., 1979), these improvements could translate directly into improved modeling of biota (e.g., Hansen et al., 2017; Mainali et al., 2015; Paerl & Huisman, 2008) and forecasting of ecosystem conditions and services (Dietze et al., 2018). For water temperature and many other variables, models that explicitly combine empiricism and theory can help accelerate our path to future knowledge discovery.

Acknowledgments
See supporting information for data access, extended methods details, and example code. See https://doi.org/10.5066/P9AQPIVD for this study's data release and https://doi.org/10.5281/zenodo.3497495 for the versioned code repository. This research was funded by the Department of the Interior Northeast and North Central Climate Adaptation Science Centers, a Midwest Glacial Lakes Fish Habitat Partnership grant through F&WS, NSF Expedition in Computing Grant 1029711 to the University of Minnesota, a postdoctoral fellowship awarded to J.A.Z. under NSF EAR-PF-1725386, and a seed grant from the Digital Technology Center at the University of Minnesota. Access to computing facilities was provided by the University of Minnesota Supercomputing Institute and USGS Advanced Research Computing, USGS Yeti Supercomputer (https://doi.org/10.5066/F7D798MJ). We thank North Temperate Lakes Long-Term Ecological Research (NSF DEB-1440297) and Global Lake Ecological Observatory Network (NSF 1702991) for modeling discussions and data sharing, and Arka Daw, Randy Hunt, Jeff Sadler, Emily Read, and Mike Fienen for the PGDL discussions and ideas. We thank Luke Winslow, Noah Lottig, Madeline Magee, and along with MN DNR and WI DNR for temperature and bathymetric data, with special thanks to Pete Jacobson, Katie Hein, and Madeline Humphrey for collating thousands of temperature records, and Dave Wolock, the editorial group at WRR, and three anonymous reviewers for input that was used to improve this paper.

References

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

Ba, J., & Caruana, R. (2014). Do Deep Nets Really Need to be Deep? In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems 27* (pp. 2654–2662). Curran Associates, Inc. Retrieved from. http://papers.nips.cc/paper/5484-do-deep-nets-really-need-to-be-deep

Bachmann, R. W., Sharma, S., Canfield, D. E., & Lecours, V. (2019). The Distribution and Prediction of Summer Near-Surface Water Temperatures in Lakes of the Coterminous United States and Southern Canada. *Geosciences*, *9*(7). https://doi.org/10.3390/geosciences9070296

Benson, B. J., Lenters, J. D., Magnuson, J. J., Stubbs, M., Dillon, P. J., Hecky, R. E., & Lathrop, R. C. (2000). Regional coherence of climatic and lake thermal variables of four lake districts in the Upper Great Lakes Region of North America. *Freshwater Biology*, *43*(3), 517–527. https://doi.org/10.1046/j.1365-2427.2000.00572.x

Brett, J. R. (1971). Energetic Responses of Salmon to Temperature. A Study of Some Thermal Relations in the Physiology and Freshwater Ecology of Sockeye Salmon (Oncorhynchus nerkd). *American Zoologist*, *11*(1), 99–113. https://doi.org/10.1093/icb/11.1.99

Bruggeman, J., & Bolding, K. (2014). A general framework for aquatic biogeochemical models. *Environmental Modelling & Software*, *61*, 249–265. https://doi.org/10.1016/J.ENVSOFT.2014.04.002

Chen, H., Engkvist, O., Wang, Y., Olivecrona, M., & Blaschke, T. (2018). The rise of deep learning in drug discovery. *Drug Discovery Today*, *23*(6), 1241–1250. https://doi.org/10.1016/J.DRUDIS.2018.01.039

Chen, Y., & Han, D. (2016). Big data and hydroinformatics. *Journal of Hydroinformatics*, *18*(4), 599–614. https://doi.org/10.2166/hydro.2016.180

Clark, M. P., Schaefli, B., Schymanski, S. J., Samaniego, L., Luce, C. H., Jackson, B. M., et al. (2016). Improving the theoretical underpinnings of process-based hydrologic models. *Water Resources Research*, *52*, 2350–2365. https://doi.org/10.1002/2015WR017910

Cobourn, K. M., Carey, C. C., Boyle, K. J., Duffy, C., Dugan, H. A., Farrell, K. J., et al. (2018). From concept to practice to policy: modeling coupled natural and human systems in lake catchments. *Ecosphere*, *9*(5), e02209. https://doi.org/10.1002/ecs2.2209

Coto-Jimenez, M. (2018). Pre-training Long Short-term Memory Neural Networks for Efficient Regression in Artificial Speech Postfiltering. In 2018 IEEE International Work Conference on Bioinspired Intelligence (IWOBI) (pp. 1–7). IEEE. https://doi.org/10.1109/IWOBI.2018.8464204

Demissie, Y. K., Valocchi, A. J., Minsker, B. S., & Bailey, B. A. (2009). Integrating a calibrated groundwater flow model with error-correcting data-driven models to improve predictions. *Journal of Hydrology*, *364*(3–4), 257–271. https://doi.org/10.1016/J.JHYDROL.2008.11.007

Dietze, M. C., Fox, A., Beck-Johnson, L. M., Betancourt, J. L., Hooten, M. B., Jarnevich, C. S., et al. (2018). Iterative near-term ecological forecasting: Needs, opportunities, and challenges. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(7), 1424–1432. https://doi.org/10.1073/pnas.1710231115

Dueben, P. D., & Bauer, P. (2018). Challenges and design choices for global weather and climate models based on machine learning. *Geoscientific Model Development*, *11*(10), 3999–4009. https://doi.org/10.5194/gmd-11-3999-2018

Erhan, D., Bengio, Y., Courville, A., Manzagol, P.-A., Vincent, P., & Bengio, S. (2010). Why Does Unsupervised Pre-training Help Deep Learning? Pierre-Antoine Manzagol Pascal Vincent Samy Bengio. *Journal of Machine Learning Research*, *11*, 625–660. Retrieved from. http://www.jmlr.org/papers/volume11/erhan10a/erhan10a.pdf

Fang, K., Shen, C., Kifer, D., & Yang, X. (2017). Prolongation of SMAP to Spatiotemporally Seamless Coverage of Continental U.S. Using a Deep Learning Neural Network. *Geophysical Research Letters*, *44*, 11,030–11,039. https://doi.org/10.1002/2017GL075619

Fatichi, S., Vivoni, E. R., Ogden, F. L., Ivanov, V. Y., Mirus, B., Gochis, D., et al. (2016). An overview of current applications, challenges, and future trends in distributed process-based models in hydrology. *Journal of Hydrology*, *537*, 45–60. https://doi.org/10.1016/J.JHYDROL.2016.03.026

Fienen, M. N., Nolan, B. T., Kauffman, L. J., & Feinstein, D. T. (2018). Metamodeling for Groundwater Age Forecasting in the Lake Michigan Basin. *Water Resources Research*, *54*, 4750–4766. https://doi.org/10.1029/2017WR022387

Gentine, P., Pritchard, M., Rasp, S., Reinaudi, G., & Yacalis, G. (2018). Could Machine Learning Break the Convection Parameterization Deadlock? *Geophysical Research Letters*, *45*, 5742–5751. https://doi.org/10.1029/2018GL078202

Gers, F. A., Schmidhuber, J., & Cummins, F. (1999). Learning to forget: continual prediction with LSTM. In 9th International Conference on Artificial Neural Networks: ICANN '99 (Vol. 9, pp. 850–855). IEE. https://doi.org/10.1049/cp:19991218

Goudsmit, G.-H., Burchard, H., Peeters, F., & Wüest, A. (2002). Application of k-ε turbulence models to enclosed basins: The role of internal seiches. *Journal of Geophysical Research*, *107*(C12), 3230. https://doi.org/10.1029/2001JC000954

Hamilton, D. P., Magee, M. R., Wu, C. H., & Kratz, T. K. (2018). Ice cover and thermal regime in a dimictic seepage lake under climate change. *Inland Waters*, *8*(3), 381–398. https://doi.org/10.1080/20442041.2018.1505372

Hampton, S. E., Strasser, C. A., Tewksbury, J. J., Gram, W. K., Budden, A. E., Batcheller, A. L., et al. (2013). Big data and the future of ecology. *Frontiers in Ecology and the Environment*, *11*(3), 156–162. https://doi.org/10.1890/120103

Hansen, G. J. A., Read, J. S., Hansen, J. F., & Winslow, L. A. (2017). Projected shifts in fish species dominance in Wisconsin lakes under climate change. *Global Change Biology*, *23*(4), 1463–1476. https://doi.org/10.1111/gcb.13462

Hilborn, R., & Mangel, M. (1997). *The Ecological Detective: Confronting Models with Data - Ray Hilborn, Marc Mangel - Google Books* (1st ed.). Princeton University Press.

Hipsey, M. R., Bruce, L. C., Boon, C., Busch, B., Carey, C. C., Hamilton, D. P., et al. (2019). A General Lake Model (GLM 3.0) for linking with high-frequency sensor data from the Global Lake Ecological Observatory Network (GLEON). *Geoscientific Model Development*, *12*(1), 473–523. https://doi.org/10.5194/gmd-12-473-2019

Hipsey, M. R., Hamilton, D. P., Hanson, P. C., Carey, C. C., Coletti, J. Z., Read, J. S., et al. (2015). Predicting the resilience and recovery of aquatic systems: A framework for model evolution within environmental observatories. *Water Resources Research*, *51*, 7023–7043. https://doi.org/10.1002/2015WR017175

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735–1780. Retrieved from. http://www7.informatik.tu-muenchen.de/~hochreithttp://www.idsia.ch/~juergen, https://doi.org/10.1162/neco.1997.9.8.1735

Hof, R. D. (2013). Deep Learning--With massive amounts of computational power, machines can now recognize objects and translate speech in real time. Artificial intelligence is finally getting smart. Retrieved January 29, 2019, from https://www.technologyreview.com/s/513696/deep-learning/

Homer, C., Dewitz, J., Yang, L., Jin, S., Danielson, P., Xian, G., et al. (2015). Completion of the 2011 National Land Cover Database for the Conterminous United States-Representing a Decade of Land Cover Change Information. *Photogrammetric Engineering & Remote Sensing*, *81*(5), 345–354. https://doi.org/10.14358/PERS.81.5.345

Hooper, R., Nearing, G., & Condon, L. (2017). Using the National Water Model as a Hypothesis-Testing Tool. *Open Water Journal*, *4*(2). Retrieved from). https://scholarsarchive.byu.edu/openwater/vol4/iss2/3

Howard, J. (2013). The business impact of deep learning. In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '13 (p. 1135). New York, New York, USA: ACM Press. https://doi.org/10.1145/2487575.2491127

Hrachowitz, M., Savenije, H. H. G., Blöschl, G., McDonnell, J. J., Sivapalan, M., Pomeroy, J. W., et al. (2013). A decade of Predictions in Ungauged Basins (PUB)—a review. *Hydrological Sciences Journal*, *58*(6), 1198–1255. https://doi.org/10.1080/02626667.2013.803183

Humphrey, G. B., Gibbs, M. S., Dandy, G. C., & Maier, H. R. (2016). A hybrid approach to monthly streamflow forecasting: Integrating hydrological model outputs into a Bayesian artificial neural network. *Journal of Hydrology*, *540*, 623–640. https://doi.org/10.1016/J.JHYDROL.2016.06.026

Hunter, J. M., Maier, H. R., Gibbs, M. S., Foale, E. R., Grosvenor, N. A., Harders, N. P., & Kikuchi-Miller, T. C. (2018). Framework for developing hybrid process-driven, artificial neural network and regression models for salinity prediction in river systems. *Hydrology and Earth System Sciences*, *22*(5), 2987–3006. https://doi.org/10.5194/hess-22-2987-2018

Imberger, J. (1981). *A dynamic reservoir simulation model-DYRESM. In Transport Models for Inland and Coastal Waters* (pp. 310–361). Academic Press. Retrieved from. https://ci.nii.ac.jp/naid/80001418574/

Islam, K. A., Pérez, D., Hill, V., Schaeffer, B., Zimmerman, R., & Li, J. (2018). Seagrass Detection in Coastal Water Through Deep Capsule Networks. In Chinese Conference on Pattern Recognition and Computer Vision (PRCV) (pp. 320–331). Springer. https://doi.org/10.1007/978-3-030-03335-4_28

Jia X., Willard J., Karpatne A., Read J.S., Zwart J., Steinbach M., Kumar V., (2019) Physics guided RNNs for modeling dynamical systems: A case study in simulating lake temperature profiles. InProceedings of the 2019 SIAM International Conference on Data Mining 2019 May 6 (pp. 558-566). Society for Industrial and Applied Mathematics. https://doi.org/10.1137/1.9781611975673.63

Karpatne, A., Atluri, G., Faghmous, J. H., Steinbach, M., Banerjee, A., Ganguly, A., et al. (2017). Theory-Guided Data Science: A New Paradigm for Scientific Discovery from Data. *IEEE Transactions on Knowledge and Data Engineering*, *29*(10), 2318–2331. https://doi.org/10.1109/TKDE.2017.2720168

Karpatne, A., Khandelwal, A., Chen, X., Mithal, V., Faghmous, J., & Kumar, V. (2016). *Global Monitoring of Inland Water Dynamics: State-of-the-Art, Challenges, and Opportunities* (pp. 121–147). Springer, Cham. https://doi.org/10.1007/978-3-319-31858-5_7

Karpatne, A., Watkins, W., Read, J.S., V Kumar. (2018). Physics-guided neural networks (PGNN): An application in lake temperature modeling. arXiv preprint https://arxiv.org/abs/1710.11431

Kingma, D. P., & Ba, J. (2014). Adam: A Method for Stochastic Optimization. Retrieved from http://arxiv.org/abs/1412.6980

Kollet, S. J., & Maxwell, R. M. (2008). Capturing the influence of groundwater dynamics on land surface processes using an integrated, distributed watershed model. *Water Resources Research*, *44*, W02402. https://doi.org/10.1029/2007WR006004

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The Parable of Google Flu: Traps in Big Data Analysis. *Science*, *343*(6176), 1203–1205. https://doi.org/10.1126/science.1248506

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Lee, K.-H., Ros, G., Li, J., & Gaidon, A. (2018). SPIGAN: Privileged Adversarial Learning from Simulation. Retrieved from http://arxiv.org/abs/1810.03756

Lenters, J. D., Kratz, T. K., & Bowser, C. J. (2005). Effects of climate variability on lake evaporation: Results from a long-term energy budget study of Sparkling Lake, northern Wisconsin (USA). *Journal of Hydrology*, *308*(1–4), 168–195. https://doi.org/10.1016/J.JHYDROL.2004.10.028

Magnuson, J. J., Crowder, L. B., & Medvick, P. A. (1979). Temperature as an Ecological Resource. *American Zoologist*, *19*(1), 331–343. https://doi.org/10.1093/icb/19.1.331

Maier, H. R., Jain, A., Dandy, G. C., & Sudheer, K. P. (2010). Methods used for the development of neural networks for the prediction of water resource variables in river systems: Current status and future directions. *Environmental Modelling & Software*, *25*(8), 891–909. https://doi.org/10.1016/J.ENVSOFT.2010.02.003

Mainali, K. P., Warren, D. L., Dhileepan, K., McConnachie, A., Strathie, L., Hassan, G., et al. (2015). Projecting future expansion of invasive species: comparing and improving methodologies for species distribution modeling. *Global Change Biology*, *21*(12), 4464–4480. https://doi.org/10.1111/gcb.13038

Markfort, C. D., Perez, A. L. S., Thill, J. W., Jaster, D. A., Porté-Agel, F., & Stefan, H. G. (2010). Wind sheltering of a lake by a tree canopy or bluff topography. *Water Resources Research*, *46*, W03530. https://doi.org/10.1029/2009WR007759

Mazzocchi, F. (2015). Could Big Data be the end of theory in science?: A few remarks on the epistemology of data-driven science. *EMBO Reports*, *16*(10), 1250–1255. https://doi.org/10.15252/embr.201541001

Me, W., Hamilton, D. P., McBride, C. G., Abell, J. M., & Hicks, B. J. (2018). Modelling hydrology and water quality in a mixed land use catchment and eutrophic lake: Effects of nutrient load reductions and climate change. *Environmental Modelling & Software*, *109*, 114–133. https://doi.org/10.1016/J.ENVSOFT.2018.08.001

Mitchell, K. E. (2004). The multi-institution North American Land Data Assimilation System (NLDAS): Utilizing multiple GCIP products and partners in a continental distributed hydrological modeling system. *Journal of Geophysical Research*, *109*, D07S90. https://doi.org/10.1029/2003JD003823

Mosavi, A., Ozturk, P., Chau, K., Mosavi, A., Ozturk, P., & Chau, K. (2018). Flood Prediction Using Machine Learning Models: Literature Review. *Water*, *10*(11), 1536. https://doi.org/10.3390/w10111536

Nayak, P. C., Venkatesh, B., Krishna, B., & Jain, S. K. (2013). Rainfall-runoff modeling using conceptual, data driven, and wavelet based computing approach. *Journal of Hydrology*, *493*, 57–67. https://doi.org/10.1016/J.JHYDROL.2013.04.016

Olden, J. D., Lawler, J. J., & Poff, N. L. (2008). Machine Learning Methods Without Tears: A Primer for Ecologists. *The Quarterly Review of Biology*, *83*(2), 171–193. https://doi.org/10.1086/587826

O'Reilly, C. M., Sharma, S., Gray, D. K., Hampton, S. E., Read, J. S., Rowley, R. J., et al. (2015). Rapid and highly variable warming of lake surface waters around the globe. *Geophysical Research Letters*, *42*, 10,773–10,781. https://doi.org/10.1002/2015GL066235

Paerl, H. W., & Huisman, J. (2008). Climate. Blooms like it hot. *Science*, *320*(5872), 57–58. https://doi.org/10.1126/science.1155398

Palmer, M. E., Yan, N. D., & Somers, K. M. (2014). Climate change drives coherent trends in physics and oxygen content in North American lakes. *Climatic Change*, *124*(1–2), 285–299. https://doi.org/10.1007/s10584-014-1085-4

Panofsky, H.A., & Brier, G.W., 1958. Some applications of statistics to meteorology. Mineral Industries Extension Services, College of Mineral Industries, Pennsylvania State University.

Perroud, M., Goyette, S., Martynov, A., Beniston, M., & Annevillec, O. (2009). Simulation of multiannual thermal profiles in deep Lake Geneva: A comparison of one-dimensional lake models. *Limnology and Oceanography*, *54*(5), 1574–1594. https://doi.org/10.4319/lo.2009.54.5.1574

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

Read, J. S., Winslow, L. A., Hansen, G. J. A., Van Den Hoek, J., Hanson, P. C., Bruce, L. C., & Markfort, C. D. (2014). Simulating 2368 temperate lakes reveals weak coherence in stratification phenology. *Ecological Modelling*, *291*, 142–150. https://doi.org/10.1016/J.ECOLMODEL.2014.07.029

Rezaee, M., Mahdianpari, M., Zhang, Y., & Salehi, B. (2018). Deep Convolutional Neural Network for Complex Wetland Classification Using Optical Remote Sensing Imagery. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *11*(9), 3030–3039. https://doi.org/10.1109/JSTARS.2018.2846178

Riley, M. J., & Stefan, H. G. (1988). Minlake: A dynamic lake water quality simulation model. *Ecological Modelling*, *43*(3–4), 155–182. https://doi.org/10.1016/0304-3800(88)90002-6

Salman, A., Jalal, A., Shafait, F., Mian, A., Shortis, M., Seager, J., & Harvey, E. (2016). Fish species classification in unconstrained underwater environments based on deep learning. *Limnology and Oceanography: Methods*, *14*(9), 570–585. https://doi.org/10.1002/lom3.10113

Schaeffer, B. A., Iiames, J., Dwyer, J., Urquhart, E., Salls, W., Rover, J., & Seegers, B. (2018). An initial validation of Landsat 5 and 7 derived surface water temperature for U.S. lakes, reservoirs, and estuaries. *International Journal of Remote Sensing*, *39*(22), 7789–7805. https://doi.org/10.1080/01431161.2018.1471545

Scher, S. (2018). Toward Data-Driven Weather and Climate Forecasting: Approximating a Simple General Circulation Model With Deep Learning. *Geophysical Research Letters*, *45*, 12616–12622. https://doi.org/10.1029/2018GL080704

Scibek, J., & Allen, D. M. (2006). Comparing modelled responses of two high-permeability, unconfined aquifers to predicted climate change. *Global and Planetary Change*, *50*(1–2), 50–62. https://doi.org/10.1016/J.GLOPLACHA.2005.10.002

Shen, C. (2018). A Transdisciplinary Review of Deep Learning Research and Its Relevance for Water Resources Scientists. *Water Resources Research*, *54*, 8558–8593. https://doi.org/10.1029/2018WR022643

Shen, C., Laloy, E., Elshorbagy, A., Albert, A., Bales, J., Chang, F.-J., et al. (2018). HESS Opinions: Incubating deep-learning-powered hydrologic science advances as a community. *Hydrology and Earth System Sciences*, *22*(11), 5639–5656. https://doi.org/10.5194/hess-22-5639-2018

Shortridge, J. E., Guikema, S. D., & Zaitchik, B. F. (2016). Machine learning methods for empirical streamflow simulation: a comparison of model accuracy, interpretability, and uncertainty in seasonal watersheds. *Hydrology and Earth System Sciences*, *20*(7), 2611–2628. https://doi.org/10.5194/hess-20-2611-2016

Simley, J. D., & Carswell, W. J. J. (2009). The National Map-Hydrography. *U.S. Geological Survey Fact Sheet*, *3054*(4). Retrieved from http://datagateway.nrcs.usda.gov/

Soranno, P. A., Bacon, L. C., Beauchene, M., Bednar, K. E., Bissell, E. G., Boudreau, C. K., et al. (2017). LAGOS-NE: a multi-scaled geospatial and temporal database of lake ecological context and water quality for thousands of US lakes. *GigaScience*, *6*(12), 1–22. https://doi.org/10.1093/gigascience/gix101

Stewart, R., & Ermon, S. (2017). Label-Free Supervision of Neural Networks with Physics and Domain Knowledge. Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17). arXiv preprint https://arxiv.org/abs/1609.05566

Toffolon, M., Piccolroaz, S., Majone, B., Soja, A. M., Peeters, F., Schmid, M., & Wüest, A. (2014). Prediction of surface temperature in lakes with different morphology using air temperature. *Limnology and Oceanography*, *59*(6), 2185–2202. https://doi.org/10.4319/lo.2014.59.6.2185

Tongal, H., & Booij, M. J. (2018). Simulation and forecasting of streamflows using machine learning models coupled with base flow separation. *Journal of Hydrology*, *564*, 266–282. https://doi.org/10.1016/J.JHYDROL.2018.07.004

Torbick, N., Hession, S., Hagen, S., Wiangwang, N., Becker, B., & Qi, J. (2013). Mapping inland lake water quality across the Lower Peninsula of Michigan using Landsat TM imagery. *International Journal of Remote Sensing*, *34*(21), 7607–7624. https://doi.org/10.1080/01431161.2013.822602

Wagener, T., Sivapalan, M., Troch, P. A., McGlynn, B. L., Harman, C. J., Gupta, H. V., et al. (2010). The future of hydrology: An evolving science for a changing world. *Water Resources Research*, *46*, W05301. https://doi.org/10.1029/2009WR008906

Werbos, P. J. (1988). Generalization of backpropagation with application to a recurrent gas market model. *Neural Networks*, *1*(4), 339–356. https://doi.org/10.1016/0893-6080(88)90007-X

Wetzel, R. G., & Likens, G. E. (2000). *The heat budget of lakes. In Limnological Analysis* (pp. 45–56). Springer.

Winslow, L. A., Hansen, G. J. A., Read, J. S., & Notaro, M. (2017). Large-scale modeled contemporary and future water temperature estimates for 10774 Midwestern U.S. Lakes. *Scientific Data*, *4*. https://doi.org/10.1038/sdata.2017.53

Worland, S. C., Farmer, W. H., & Kiang, J. E. (2018). Improving predictions of hydrological low-flow indices in ungaged basins using machine learning. *Environmental Modelling & Software*, *101*, 169–182. https://doi.org/10.1016/J.ENVSOFT.2017.12.021

Xia, Y., Mitchell, K., Ek, M., Sheffield, J., Cosgrove, B., Wood, E., et al. (2012). Continental-scale water and energy flux analysis and validation for the North American Land Data Assimilation System project phase 2 (NLDAS-2): 1. Intercomparison and application of model products. *Journal of Geophysical Research*, *117*, D03109. https://doi.org/10.1029/2011JD016048

Yan, S., & Minsker, B. (2006). Optimal groundwater remediation design using an Adaptive Neural Network Genetic Algorithm. *Water Resources Research*, *42*, W05407. https://doi.org/10.1029/2005WR004303