AGU ADVANCING EARTH AND SPACE SCIENCES

Water Resources Research

RESEARCH ARTICLE
10.1029/2023WR036570

Using System-Inspired Metrics to Improve Water Quality Prediction in Stratified Lakes

Kamilla Kurucz[1], Cayelan C. Carey[2], Peisheng Huang[1], Eduardo R. De Sousa[3], Jeremy T. White[3], and Matthew R. Hipsey[1]

[1]Centre for Water and Spatial Science, UWA School of Agriculture and Environment, The University of Western Australia, Perth, WA, Australia, [2]Department of Biological Sciences, Virginia Tech, Blacksburg, VA, USA, [3]INTERA Inc., Perth, WA, Australia

Key Points:

- We assessed the use of system-inspired metrics in a novel approach to calibrating Aquatic Ecosystem Models (AEMs)
- The use of system-inspired metrics in calibration improved model performance relative to traditional calibration methods
- Implementation of system-inspired metrics has the potential to greatly improve model prediction of complex ecosystem dynamics

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

K. Kurucz,
kamilla.kurucz@research.uwa.edu.au

Citation:

Kurucz, K., Carey, C. C., Huang, P., De Sousa, E. R., White, J. T., & Hipsey, M. R. (2024). Using system-inspired metrics to improve water quality prediction in stratified lakes. *Water Resources Research*, *60*, e2023WR036570. https://doi.org/10.1029/2023WR036570

Received 29 NOV 2023
Accepted 9 JUL 2024

Author Contributions:

Conceptualization: Kamilla Kurucz, Cayelan C. Carey, Matthew R. Hipsey
Data curation: Kamilla Kurucz, Cayelan C. Carey
Formal analysis: Kamilla Kurucz
Methodology: Kamilla Kurucz, Cayelan C. Carey, Peisheng Huang, Eduardo R. De Sousa, Jeremy T. White, Matthew R. Hipsey
Software: Jeremy T. White, Matthew R. Hipsey
Supervision: Matthew R. Hipsey
Visualization: Kamilla Kurucz
Writing – original draft: Kamilla Kurucz

Abstract Despite the growing use of Aquatic Ecosystem Models for lake modeling, there is currently no widely applicable framework for their configuration, calibration, and evaluation. Calibration is generally based on direct data comparison of observed versus modeled state variables using standard statistical techniques, however, this approach may not give a complete picture of the model's ability to capture system-scale behavior that is not easily perceivable in observations, but which may be important for resource management. The aim of this study is to compare the performance of "naïve" calibration and a "system-inspired" calibration, an approach that augments the standard state-based calibration with a range of system-inspired metrics (e.g., thermocline depth, metalimnetic oxygen minima), to increase the coherence between the simulated and natural ecosystems. A coupled physical-biogeochemical model was applied to a focal site to simulate two key state-variables: water temperature and dissolved oxygen. The model was calibrated according to the new system-inspired modeling convention, using formal calibration techniques. There was an improvement in the simulation using parameters optimized on the additional metrics, which helped to reduce uncertainty predicting aspects of the system relevant to reservoir management, such as the occurrence of the metalimnetic oxygen minima. Extending the use of system-inspired metrics when calibrating models has the potential to improve model fidelity for capturing more complex ecosystem dynamics.

1. Introduction

The use of process-based Aquatic Ecosystem Models (AEMs) for simulating the water quality of freshwater ecosystems has substantially increased over the past two decades for studying the effects of human activities and predicting future changes (Janssen et al., 2015; Soares & Calijuri, 2021). These models can be used for several different purposes across various spatiotemporal scales, making them useful decision-making tools for addressing the environmental issues affecting aquatic ecosystems (Mooij et al., 2010). For example, recent advancements have demonstrated their capabilities for the simulation of chemical and biological variables to investigate anoxia (Carey, Hanson, et al., 2022; Ladwig et al., 2021), eutrophication (Arhonditsis & Brett, 2005), greenhouse gas emissions (Stepanenko et al., 2016), and harmful algal blooms (Ranjbar et al., 2021). Moreover, they can be used for testing scenarios related to climate change and increased nutrient loading, which would not otherwise be feasible to study empirically at the system-scale (e.g., Elhabashy et al., 2023; Nielsen et al., 2014; Trolle et al., 2011).

Despite the advancement in the process descriptions within AEMs, the level of predictability they provide has not significantly improved since the 1990s (Arhonditsis & Brett, 2004; Soares & Calijuri, 2021). Moreover, some studies have argued that AEMs are mostly useful only in a heuristic way, which is a barrier for their uptake into policy and management applications (Kim et al., 2014; Kotamäki et al., 2024). While AEMs have become increasingly complex, we have yet to form consensus as to how best to configure, calibrate, and evaluate them, motivating the need for reliable and repeatable approaches for historical and future aquatic ecosystem prediction (Frassl et al., 2019). To date, these challenges have been discussed in the ecological modeling community, in which Grimm et al. (2005) highlighted the need for ecological models to capture the generative mechanisms of a system, which are responsible for producing system-level responses. Pattern-oriented modeling suggests that multi-criteria design, selection, and calibration of models is needed to reproduce a diverse set of patterns present in a complex system (Grimm & Railsback, 2012), with the implication that if a model cannot reproduce the relevant patterns, it cannot be trusted to make reliable predictions of the system. Similarly, the ability of AEMs to

Writing – review & editing: Kamilla Kurucz, Cayelan C. Carey, Peisheng Huang, Eduardo R. De Sousa, Jeremy T. White, Matthew R. Hipsey

reproduce system-scale patterns and responses and emergent behavior, in addition to capturing the basic trends in routinely observed monitoring data, is essential for reliably applying these models in novel conditions and increase confidence in their accuracy. Hence, there is an urgent need for implementing such a system-inspired approach in procedures used for AEM assessment.

The Concept/State/Process/System framework (CSPS; Hipsey et al., 2020) provides a multi-level approach for the evaluation of AEMs to extend the routinely applied model-data comparison method. The CSPS framework consists of four validation levels: evaluation of a model's adherence to ecological theory (Level 0; conceptual validation), comparison of modeled and observed state variables (Level 1; state validation), comparison of simulated fluxes with observed process rates (Level 2; process validation), and the assessment of the model's ability to capture system behavior (Level 3; system validation). The CSPS framework extends model evaluation beyond traditional state validation (Level 1) and evaluates whether the model is able to reproduce theoretically relevant, system-scale responses or patterns which are governed by complex non-linear interactions and feedbacks. It suggests the adoption of a suite of advanced metrics for model assessment that describe aquatic ecosystem structure and function. The advantage of the CSPS framework over traditional model evaluation is its ability to identify hidden deficiencies in the model which would not be detectable by comparing state variable predictions to corresponding observations (traditional validation). In recent case studies, the framework has been applied to assist validation of ecosystem models for Lake Kinneret (Israel) and the Great Barrier Reef (Australia), enabling improved assessment of each model's strengths and hidden deficiencies, which led to increased AEM reliability and accuracy (Regev et al., 2023; Robson et al., 2020).

Nonetheless, a major challenge in setting up and applying AEMs remains appropriately calibrating model parameters and conveying the degree of confidence we have in the predictions they provide. The traditional scope of calibration requires identifying the parameter set within the parameter space that best fits observations. For lake and reservoir models, this step is generally based on direct data comparison of observed versus modeled state variables using simple quantitative techniques such as the root-mean-square-error (RMSE; Soares & Calijuri, 2021), which in practice have limited utility. Consequently, the "success" of calibration remains subjective and dependent on noisy observations of the primary state variables, often limited in quantity, to adequately constrain the model inputs (Bennett et al., 2013). Additional sources of uncertainties in AEMs are introduced through model structural assumptions and simplifications, in the assignment of initial and boundary conditions, and through the error in the observed data used for calibration and validation (Beck, 1987). These uncertainties propagate into the model output, which hinders our ability to produce confident predictions, and there remains a need for more informative assessment frameworks able to provide critical estimates of reliability of model predictions (Huettmann & Arhonditsis, 2023).

It remains to be explored whether incorporating additional metrics during calibration and uncertainty assessment of AEMs, based on the CSPS framework, could potentially compensate for inadequate observed data sets and information deficits by introducing additional constraints to calibration. We herein refer to additional metrics as "system-inspired" metrics to signify that they represent theoretically important patterns and ecosystem responses that emerge from the non-linear interactions within the model. In this study, we selected four system-inspired metrics relevant to reservoir ecosystems that highlight relevant stratification and oxygen dynamics within the system of interest: the thermocline depth, Schmidt stability, metalimnetic oxygen minima, and the vertical extent of anoxia. Stratification is the key driver in lentic waterbodies controlling the redistribution of dissolved substances, and whilst models are often tested against thermal profiles, the evolution of thermocline depth is a key marker for the boundary where the maximum temperature gradient separates the epilimnion from the hypolimnion (Wilhelm & Adrian, 2008). From an ecosystem view, understanding the strength of thermal stratification also provides valuable information regarding the timing and tendency of mixing in a stratified water body. Strong stratification often leads to the development of low oxygen zones either in the metalimnion or hypolimnion, which restricts the abundance of oxygen-sensitive organisms (Gerling et al., 2016). The spatial and temporal pattern of anoxia has major implications for water quality, food webs, and ecosystem functioning (Carey, Hanson, et al., 2022). It is hypothesized that capturing specific patterns that emerge in the thermal and oxygen profiles via these metrics will improve the degree of confidence in our predictions and ultimately facilitate a more holistic understanding of aquatic system dynamics that can better support their management.

The aim of this study was therefore to answer two research questions: (a) Can applying non-traditional, system-inspired metrics based on the CSPS framework provide additional constraints that can improve the accuracy of

AEMs for water quality prediction? and (b) As system-inspired metrics have the potential to provide additional constraints to calibration, can they simultaneously reduce the uncertainty of model results? We compared two calibration approaches to address the questions. The first approach is naïve calibration, a frequently used approach based only on the statistical comparison of available observed versus modeled state variables. The second approach augments the more traditional naïve calibration with additional metrics, a new approach that explicitly includes a range of supplementary system metrics (e.g., thermocline depth, metalimnetic oxygen minima). This was undertaken by applying a coupled physical-biogeochemical model to a focus site to simulate water temperature and dissolved oxygen (DO), two key drivers of ecological functioning in lakes. Through an ensemble-based calibration analysis, the performance of the two distinct approaches was evaluated and the predictive uncertainty of the system-metrics of interest was assessed. With the use of system-inspired metrics in the analysis, we sought to provide a holistic approach for the calibration of complex AEMs transferable across a range of lentic systems.

2. Materials and Methods

2.1. Study Site

The focal site of this study was Falling Creek Reservoir (FCR), a small eutrophic reservoir located in Vinton, southwest Virginia, USA (Figure 1; 37.30, −79.84). FCR is a drinking water reservoir owned and operated by the Western Virginia Water Authority (WVWA; Carey, Hanson, et al., 2022). During construction in 1898, the dominant land use of the watershed was agriculture, however, the land is now covered by deciduous forest (Gerling et al., 2016). FCR has a maximum depth of 9.3 m and surface area of 0.119 km^2 (McClure et al., 2018). It is maintained at a constant level (full pond) by the WVWA and did not experience significant fluctuations throughout the duration of this study. The primary inflow to FCR is a tributary with a gauged weir, that receives water from the upgradient Beaverdam Reservoir (Gerling et al., 2016). FCR has a dimictic mixing regime and is thermally stratified between April and October, with intermittent ice cover between December and March (Carey & Breef-Pilz, 2022).

During the summer stratified period, FCR exhibits persistent hypolimnetic anoxia which has been causing water quality impairment (Carey, Hanson, et al., 2022). In order to mitigate the water quality problems, the WVWA deployed a side-stream hypolimnetic oxygenation system (HOx) in 2012, with the purpose of increasing the dissolved oxygen concentration in the hypolimnion without altering the thermal stratification of the water column (Gerling et al., 2014). Essentially, the HOx system extracts water from the hypolimnion at ∼8.5 m depth, injects DO into the water in a contact chamber, and returns it back to the reservoir at the withdrawal depth. Metalimnetic oxygen minimum zones (MOMs) commonly develop during the thermally-stratified period since the deployment of the HOx system (McClure et al., 2018). The HOx system was operational in summers between 2013 and 2021, with variable oxygen addition levels and durations. In-depth description of the system and operation details can be found in Gerling et al. (2014) and Carey, Hanson, et al. (2022), respectively. Due to the extensive monitoring of the physics, chemistry, and biology of the site in the last decade (e.g., Carey, Hanson, et al., 2022; Gerling et al., 2016; Lofton et al., 2019; McClure et al., 2018; McClure et al., 2021; Munger et al., 2019), sufficient empirical data for FCR were available for calibration.

2.2. Modeling Framework and Methodology

2.2.1. Modeling Framework and Overview

Our model framework composed a few stages during its development (Figure 2). A vertical 1D model was developed to simulate the hydrology (including mixing and thermal stratification) and dissolved oxygen variations in FCR. In this analysis, we built upon the model previously developed and described by Carey, Hanson, et al. (2022). We further improved the simulation by coupling the model with an independent Parameter ESTimation (PEST; Doherty, 2018a, 2018b) software package to optimize the model performance and compare two different calibration approaches: naïve and system-inspired calibration. We then tested the impact of different objective weighting strategies on the modeling results and assessed the predictive uncertainty of the system-metrics of interest. The details of model description, set up, and analysis methodologies are described in the following sections.

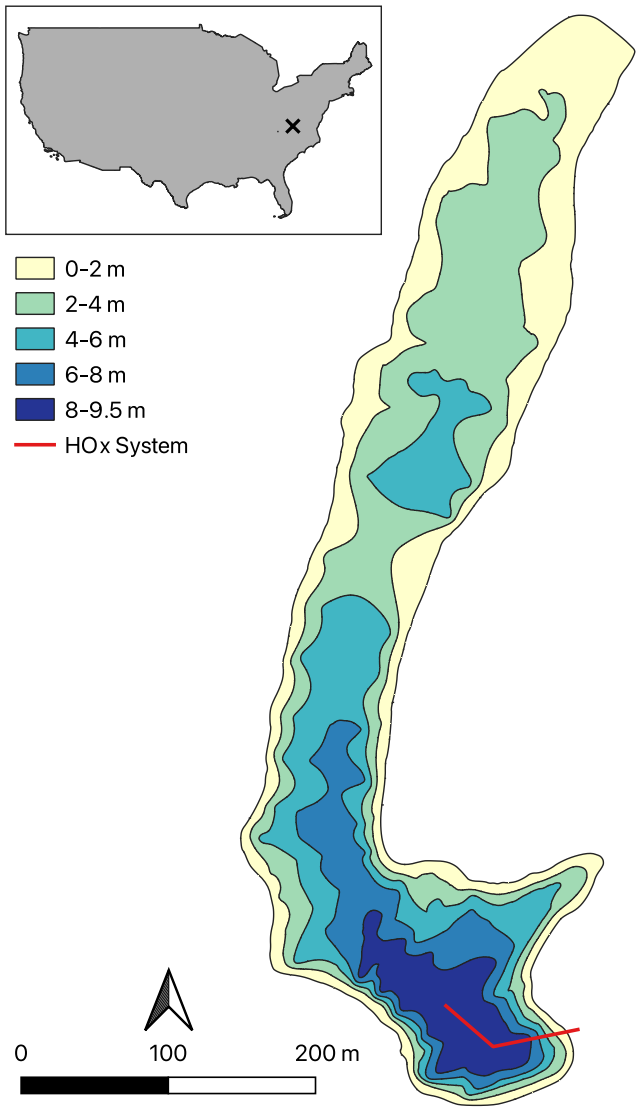

Figure 1. Map of the Falling Creek Reservoir, Vinton, Virginia, USA: Latitude: 37.30°, Longitude: −79.84°. The colored bands indicate the bathymetry contours of the reservoir, and the red line represents the location of the hypolimnetic oxygenation (HOx) system.

2.2.2. Model Description

We used the General Lake Model dynamically coupled to the Aquatic Eco-Dynamics Modules (GLM-AED; version 3.3.1a2) to simulate the physical and biogeochemical properties of FCR. GLM is a 1-D open-source model that can resolve the hydrodynamics and thermodynamics of enclosed water bodies including the water, ice and heat balance, vertical temperature distribution, transport, and mixing dynamics (Hipsey et al., 2019). The model has been applied to a range of different water body types across varying climatic regions for widespread validation and model assessment (Bruce et al., 2018). It requires meteorological, inflow and outflow driver data and incorporates a flexible Lagrangian layer scheme. In this approach, a series of horizontal layers contract or expand in response to water and heat fluxes. The sediment module allows for the setup of zone-specific sediment heating and biogeochemistry. GLM is able to simulate dominant FCR hydrodynamic processes, including summer stratification, ice formation, surface, and deep mixing (Carey, Hanson, et al., 2022). The in-depth description of GLM can be found in Hipsey et al. (2019).

The AED modeling library is an open-source project aimed at simulating aquatic ecosystem dynamics (In Hipsey, 2022). It consists of a number of modules such as DO, inorganic nutrients: C/N/P/Si, organic matter: DOM/POM, tracers, phytoplankton, zooplankton and others. Each module can work in isolation or combined with other modules, which makes AED suitable for the simulation of a range of aquatic ecosystems. In this application, the AED configuration was focused on DO, one of the most important indicators of water quality. In addition to the two core processes, atmospheric and sediment fluxes, the configuration included oxygen sources and sinks linked to the dynamics of C, N, P, Si, organic matter, and phytoplankton (see Kurucz et al., 2024, for the full model configuration and parameters).

The GLM-AED model setup for FCR by Carey, Hanson, et al. (2022) was used as the base model to build upon in this study. All GLM-AED model configuration files, parameters, and driver data for FCR were accessed from the Environmental Data Initiative repository (Carey, Thomas, & Hanson, 2022). In our configuration, the number of sediment zones was increased to four: zone 1 spanned 0–3 m, zone 2 covered 3–5 m, zone 3 included 5–7 m, and zone 4 extended beyond 7 m up to the water level height. Each sediment zone had different temperature dynamics, as well as different sediment flux rates for oxygen, to better capture the depth-specific sediment heating and biogeochemistry of the site. Additionally, the boundary condition for the HOx system deployed in FCR was configured to inject oxygenated water at varying depths in the hypolimnion. GLM-AED was run from 2015-07-20 to 2019-12-31 at an hourly timestep. The total simulation period was divided into calibration from 2016-12-02 to 2019-12-31 and validation from 2015-07-20 to 2016-11-30. Model performance was assessed by the Model Efficiency (MEF) error metric, also known as Nash-Sutcliffe Efficiency (NSE), based on the following equation (Nash & Sutcliffe, 1970):

$$\mathrm{MEF} = \mathrm{NSE} = 1 - \frac{\sum_{i=1}^{N} (P_i - O_i)^2}{\sum_{i=1}^{N} (O_i - \overline{O})^2} \quad (1)$$

where P_i is the model prediction at time i, O_i is the observed value at time i, and $\overline{O}$ is the mean of the observations.

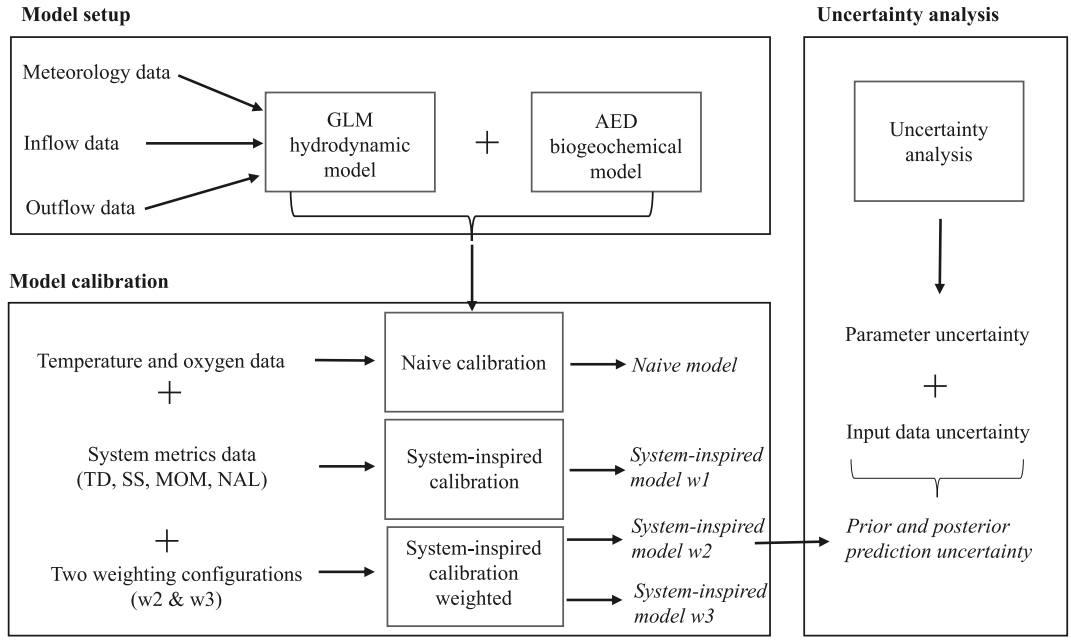

Figure 2. The modeling framework including the model setup, calibration, and uncertainty analysis. The system-inspired calibration includes additional data of the following extra metrics: thermocline depth (TD), Schmidt stability (SS), metalimnetic oxygen minima (MOM), the number of anoxic layers (NAL).

2.3. Driver (Boundary Condition) Data

GLM-AED driver data included hourly meteorological data, stream inflow data, HOx system inflow data and outflow data that were retrieved from the EDI repository (Carey, Thomas, & Hanson, 2022). The meteorological data set consisted of air temperature, relative humidity, shortwave and longwave radiation, wind speed, and precipitation data measured at the reservoir by a research-grade meteorological station (Carey & Breef-Pilz, 2023). The inflow data for the primary tributary consisted of daily discharge, water temperature, and chemistry observations from 2013 to 2021 (Carey, Hanson, et al., 2022). Discharge rates and water temperature of the primary inflow were measured at a gauged weir every 15 min and were aggregated to hourly data. Water chemistry inflow data were determined from manual grab sample observations that were collected at weekly to monthly intervals. The water temperature and chemistry data sets were linearly interpolated to a daily timestep. The HOx system was represented by a submerged inflow in the model, which matched the inflow volume and dissolved oxygen mass that was injected in the reservoir. The HOx system inflow data set included daily flow, elevation (the depth at which the oxygenated flow is injected in the reservoir), water temperature, and chemistry observations from 2013 to 2021. The daily outflow discharge was estimated to equal to the daily inflow discharge (following Carey, Hanson, et al., 2022), as the reservoir did not exhibit significant changes in water level throughout the duration of the study.

2.4. Calibration and Analysis Approach

2.4.1. State Variable Observations

Temperature and dissolved oxygen depth profiles were recorded in FCR from 2013 to 2021 at the reservoir's deepest site and were retrieved from the Environmental Data Initiative Repository (Carey, Lewis, et al., 2022). In short, temperature and dissolved oxygen depth profiles were collected with a CTD (Conductivity, Temperature, and Depth) profiler fitted with a SBE 43 Dissolved Oxygen sensor. In addition, discrete depth profiles of temperature and dissolved oxygen were also collected with YSI water quality probes at approximately 1-m intervals (Carey, Wander, et al., 2022). Samples were collected at the deepest site of FCR (near the dam), and other in-reservoir transects approximately monthly from October to February, fortnightly from March to May, and weekly from June to September. The YSI temperature profiles complement and fill in for missing CTD data. The observed temperature and dissolved oxygen profile data were spatially interpolated among depths on the data

collection days to fill in for missing data and to achieve higher spatial resolution for the calculation of system metrics. Data manipulation, analysis, visualization and computations were undertaken in R (version: 4.1.2).

2.4.2. Calibration

The GLM-AED model was coupled with an independent Parameter ESTimation (PEST; Doherty, 2018a) software package for calibration. PEST was run in estimation mode to minimize the objective function, which was defined as the sum of the weighted squared difference between measured observations and the corresponding model predictions. PEST implements the Gauss-Marquardt-Levenberg optimization algorithm for parameter estimation, which is able to rapidly find the best-fit parameter set in the user-defined parameter space. To accommodate varying observation types and frequency, the observed data was organized into different observation groups which were weighted based on different weighting strategies. Detailed description of the PEST++ software suite can be found in the PEST++ user manual (Doherty, 2018a).

2.4.3. Naïve Versus System-Inspired Calibration

The objective function of the naïve calibration (Φ_N) was based on direct comparison of the model predicted and observed temperature (T) and dissolved oxygen (DO) profiles at 0.1 m below the surface and every meter interval between 1 and 9 m depths below the surface, resulting in 20 depth-specific comparisons. The weights (w) of the T and DO observation groups were set to the reciprocal of the standard deviation of the corresponding measurements. The objective function was mathematically formulated as:

$$\Phi_N = \sum_i (w_T r_{T_i})^2 + \sum_i (w_O r_{O_i})^2 \tag{2}$$

where i denotes the number of observations in each observation group, w_T and w_O represent the weighting of the temperature and oxygen observation groups respectively and r_T and r_O denote the temperature and oxygen residuals respectively. The initial, minimum, maximum values and standard deviations of the parameters included in the adjustable parameter vector are listed in Table S1 in Supporting Information S1.

The objective function of the system-inspired calibration (Φ_S) was based on the comparison of a wide variety of system-based metrics along with the temperature (T) and dissolved oxygen (DO) profiles. The system metrics in the objective function included the thermocline depth (TD), Schmidt stability (SS), metalimnetic oxygen minima (MOM), and the number of anoxic layers per day (NAL), mathematically formulated as follows:

$$\begin{aligned}\Phi_S = &\sum_i (w_T r_{T_i})^2 + \sum_i (w_O r_{O_i})^2 + \\ &\sum_i (w_{TD} r_{TD_i})^2 + \sum_i (w_{SS} r_{SS_i})^2 + \sum_i (w_{MOM} r_{MOM_i})^2 + \sum_i (w_{NAL} r_{NAL_i})^2\end{aligned} \tag{3}$$

where w_{TD}, w_{SS}, w_{MOM}, w_{NAL}, represent the weighting of the TD, SS, MOM and NAL observation groups respectively, and r_{TD}, r_{SS}, r_{MOM}, r_{NAL} denote the TD, SS, MOM, and NAL residuals respectively.

SS is a stratification index that establishes the resistance of the system to mechanical mixing and is a good indicator of stratification strength (Idso, 1973). The SS indices were calculated from the observed temperature profiles on data collection days using the *ts.schmidt.stability* function in the rLakeAnalyzer package (Albers et al., 2018). The TD marks the upper boundary of the hypolimnion and is defined as the depth of the steepest temperature gradient in the water column during thermal stratification (Ladwig et al., 2021). The thermocline depths were calculated from the observed temperature profiles on data collection days in the stratification period (1 April–30 September) using the *ts.thermo.depth* function in the rLakeAnalyzer package with a minimum density gradient of 0.1 g/cm^3 (Albers et al., 2018). Comprehensive description of the thermocline depth and Schmidt stability index computations can be found in Read et al. (2011). The metalimnetic oxygen minimum is a zone of depleted dissolved oxygen in the middle of the water column, below the thermocline (McClure et al., 2018). It was expressed as the deviation from the expected oxygen concentration in the metalimnion, if a linear pattern in dissolved oxygen reduction is assumed from the epilimnion toward the hypolimnion. The MOM was calculated on each data collection day based on:

Table 1
Different Weighting Schemes for Incorporating System Metrics in the Objective Function

	Model w1	Model w2	Model w3
TD	1.8	0.917	1.834
SS	0.14	0.058	0.115
MOM	0.02	0.014	0.027
NAL	0.025	0.052	0.105

Note. The system metrics include the thermocline depth (TD), Schmidt stability (SS), metalimnetic oxygen minima (MOM), and the number of anoxic layers (NAL).

$$\text{MOM} = O_{2\ (\text{metalimnion})} - O_{2\ (\text{expected})} \tag{4}$$

where:

$$O_{2\ (expected)} = \frac{O_{2\ (\text{eplimnion})} + O_{2\ (\text{hypolimnion})}}{2} \tag{5}$$

The spatial and temporal pattern of anoxia in FCR was quantified by the number of anoxic layers per day. The observed NAL was calculated by temporally interpolating the observed DO data on a daily time step between 1 May and 30 November and spatially interpolating it by 0.1 m. The number of 0.1 m layers with DO concentrations below the anoxia threshold, set as 1 mg/L, were added up for each day resulting in a data set of daily count. In the system-inspired calibration process, the parameter vector and parameter transformations were equivalent to those of the naïve calibration.

Experiments with different objective function weighting schemes for incorporating the system-inspired metrics were undertaken to assess how weighting affects the calibration results (Table 1). In weighting scheme 1, hereafter referred to as Model w1, the extra metrics observation groups were given weights that resulted in an approximately equal contribution to the objective function by each advanced metric at the start of the calibration process (e.g., Wilsnack et al., 2012). Weighting scheme 2, hereafter referred to as Model w2, followed the practice of error-based weighting (e.g., Tiedeman et al., 2003), which was calculated as 1/standard deviation of the observation group (Doherty, 2018a), consistent with how state-variables were weighted. Lastly, in weighting scheme 3, hereafter referred to as Model w3, the weights were set to double that of Model w2. Moreover, the calibration process was repeated for two different deep mixing configuration sub-module options to evaluate their suitability for capturing the system-inspired metrics within FCR. One configuration adopted hypolimnetic mixing based on constant vertical diffusivity, hereafter referred to as DM 1, and the other configuration employed the Weinstock model, hereafter referred to as DM 2. In the latter, the diffusivity varies based on the strength of stratification and the depth-dependent rate of turbulent dissipation (Hipsey et al., 2019).

2.5. Uncertainty Analysis

Uncertainty analysis was carried out on the best performing system-inspired model (Model w2 with DM 2). This analysis was used to explore equifinal solutions by seeking an ensemble of parameter realizations that all acceptably reproduce both state measurements and the additional metrics (White, 2018). For this analysis, we used the iterative ensemble smoother algorithm of Chen and Oliver (2013) to express the prior and posterior parameter distributions. The iterative ensemble smoother algorithm can be seen as an approximate form of Bayes equation which is combined with subspace methods to perform ensemble parameter field adjustment (Chen & Oliver, 2013). The resulting ensemble can hence be considered to include samples from the posterior parameter distribution. By running the model for each member of the ensemble, the uncertainty in the model output, arising from the variability in parameter values, can be quantified. In this analysis, three iterations were undertaken with 300 prior parameter realizations. The prior parameter realizations were drawn from a multivariate Gaussian prior parameter distribution based on the initial parameter estimates and the specified standard deviation of each parameter. The standard deviation (σ) of each parameter was calculated using:

$$\sigma = \frac{\log_{10}(par_{max}) - \log_{10}(par_{min})}{4}, \tag{6}$$

and the corresponding values have been listed in Table S1 of Supporting Information S1.

The uncertainty arising from measurement noise was also accounted for. To quantify the measurement noise for each observation type, first, the observed data was linearly interpolated on a daily timestep. Second, the moving averages of the interpolated observations were calculated based on a 7-day window. Finally, the differences between the observed values and the corresponding moving averages were computed. The standard deviations of

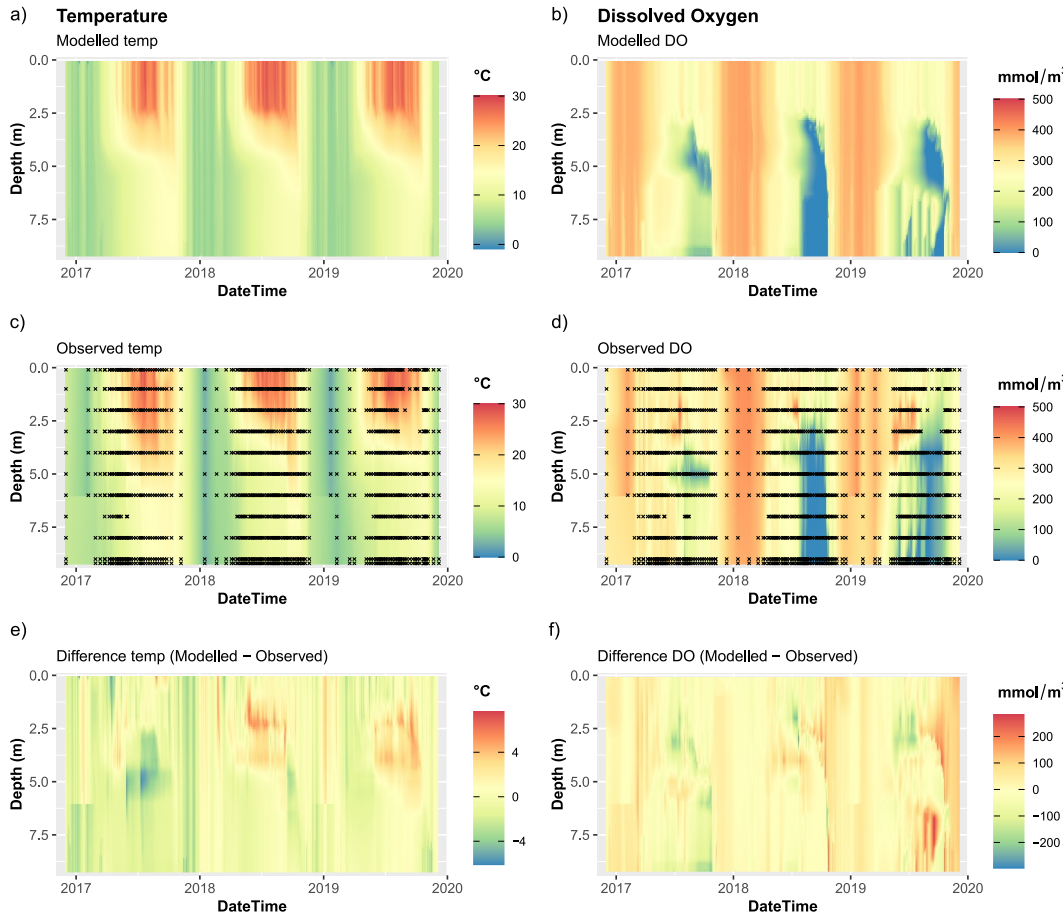

Figure 3. Contour plots of modeled (a, b), observed (c, d), and the difference of modeled and observed temperature and dissolved oxygen profiles (e, f) based on the naïve calibration model with DM 2. The black crosses on plots c and d represent the time and location of the temperature and dissolved oxygen observations respectively.

these differences for each observation type represent the noise in the measurements. For each realization, a differing calibration data set (as a result of the additive effect of measurement noise) was used to adjust each parameter field.

Table 2
Comparison of the DM 2 Naïve and System-Inspired Models' Performance in Simulating the State-Variables: Temperature (Temp), Dissolved Oxygen (DO) and the Extra Metrics: Thermocline Depth (TD), Schmidt Stability (SS), Metalimentic Oxygen Minima (MOM), Number of Anoxic Layers (NAL) During the Calibration Period Based on the Model Efficiency (MEF) Error Metric (Equation 1)

	Naïve	Model w1	Model w2	Model w3
Temp	**0.93**	**0.93**	**0.93**	0.92
DO	**0.65**	0.6	0.63	0.58
TD	0.14	**0.24**	0.18	0.15
SS	0.88	0.88	**0.89**	0.88
MOM	0.2	0.35	**0.37**	0.3
NAL	0.73	0.75	**0.77**	0.76

Note. The best performing model in simulating each variable was highlighted in bold text.

In the next step, the uncertainty analysis was rerun with updated noise. We added half of the maximum differences between observations and the posterior mean for temperature, dissolved oxygen, Schmidt stability, thermocline depth, metalimnetic oxygen minima, and the number of anoxic layers to the measurement noise applied in the first analysis run. This additional step inflated the noise to account for model error on the basis that the difference between observations and the posterior after the first uncertainty analysis run was due to model error.

3. Results

3.1. Temperature and Dissolved Oxygen Predicted by the Naïve Model

The naïve model successfully captured the dimictic mixing regime as observed in FCR, with some exceptions. Thermal stratification started to build in March, accompanied with the oxygen depletion in the bottom water (Figure 3). The modeled temperature profiles depicted the patterns and characteristics of the observed data reasonably well (Table 2). Modeled

10.1029/2023WR036570

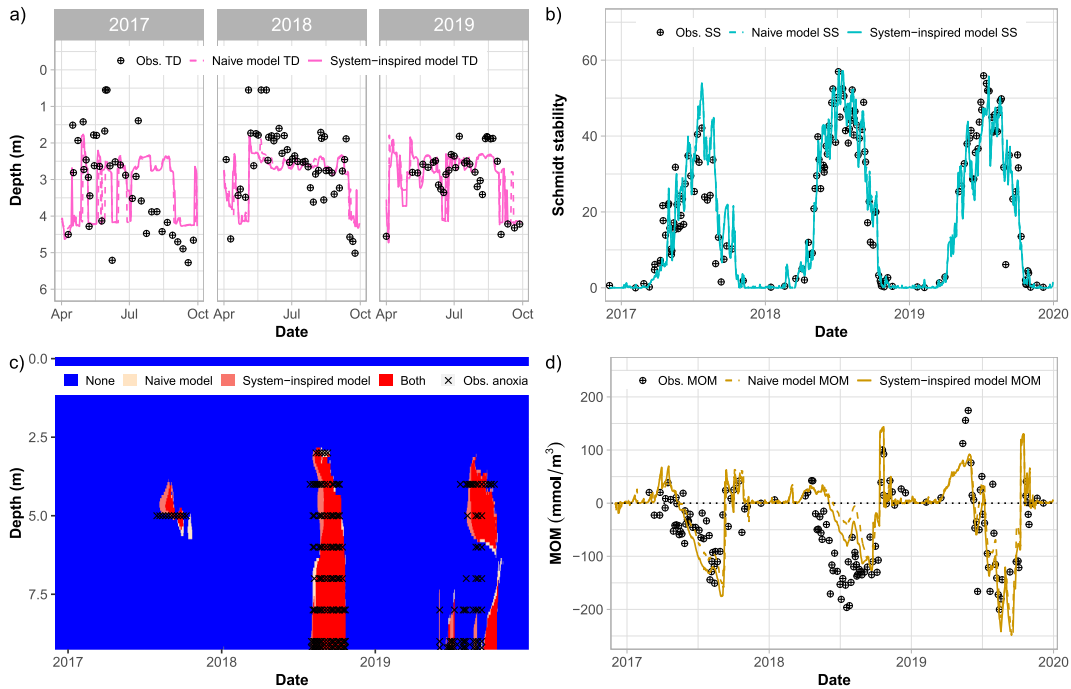

Figure 4. Comparison of observed (Obs.) system-metrics and system-metrics predicted by the naïve model and the best performing system-inspired model (Model w2) with DM 2. The metrics include the thermocline depth (TD) during the stratified period (a), Schmidt stability (b), spatial and temporal pattern of anoxia (c), and metalimnetic oxygen minimum (MOM, d).

hypolimnetic temperatures showed the greatest agreement with field measurements, relative to other layers. According to the difference plot (Figure 3e), the greatest deviation between observed and modeled temperatures occurred in the metalimnion. In the summer of 2017, the modeled metalimnetic temperatures were predicted to be 2–3° colder than the observed temperatures. However, in the summers of both 2018 and 2019, the metalimnetic temperatures were predicted to be approximately 2° warmer than the observations (Figure 3e). The modeled oxygen profiles showed a good agreement (MEF >0.5) with oxygen measurements for most of the time series (Table 2). In 2017 and 2018, the oxygen concentrations were reproduced well by the model, with moderate over-and under-estimations present. However, in 2019, the modeled hypolimnetic oxygen concentrations were higher than observations during the summer period, when the HOx system was in operation (Figure 3f).

3.2. Naïve Versus System-Inspired Approach

The augmentation of the objective function with system-inspired metrics generally led to more accurate simulation results, from a system performance perspective. The naïve model demonstrated a slightly better capability for simulating the DO profile than the system-inspired approach, while there was no discernible difference in the prediction of the water temperature profile between the two approaches (Table 2). The thermocline depth was better captured by the system-inspired model, particularly in 2018 and at the end of the stratified period in 2019 (Figure 4a). However, both models underestimated the thermocline depth in 2017 after July. Trends in the Schmidt stability were captured well by both models, which indicated that they were capable of reproducing the stratification strength of the reservoir (Figure 4b). The system-inspired model outperformed the naïve model in simulating the spatial and temporal pattern of anoxia in each simulation year (Figure 4c). However, neither model captured the observed anoxia pattern in 2019, when the oxygenation system was turned on and off multiple times during the stratified period. The system-inspired model significantly improved the simulation of the MOM compared to the naïve model, particularly in 2017 and 2018 (Figure 4d).

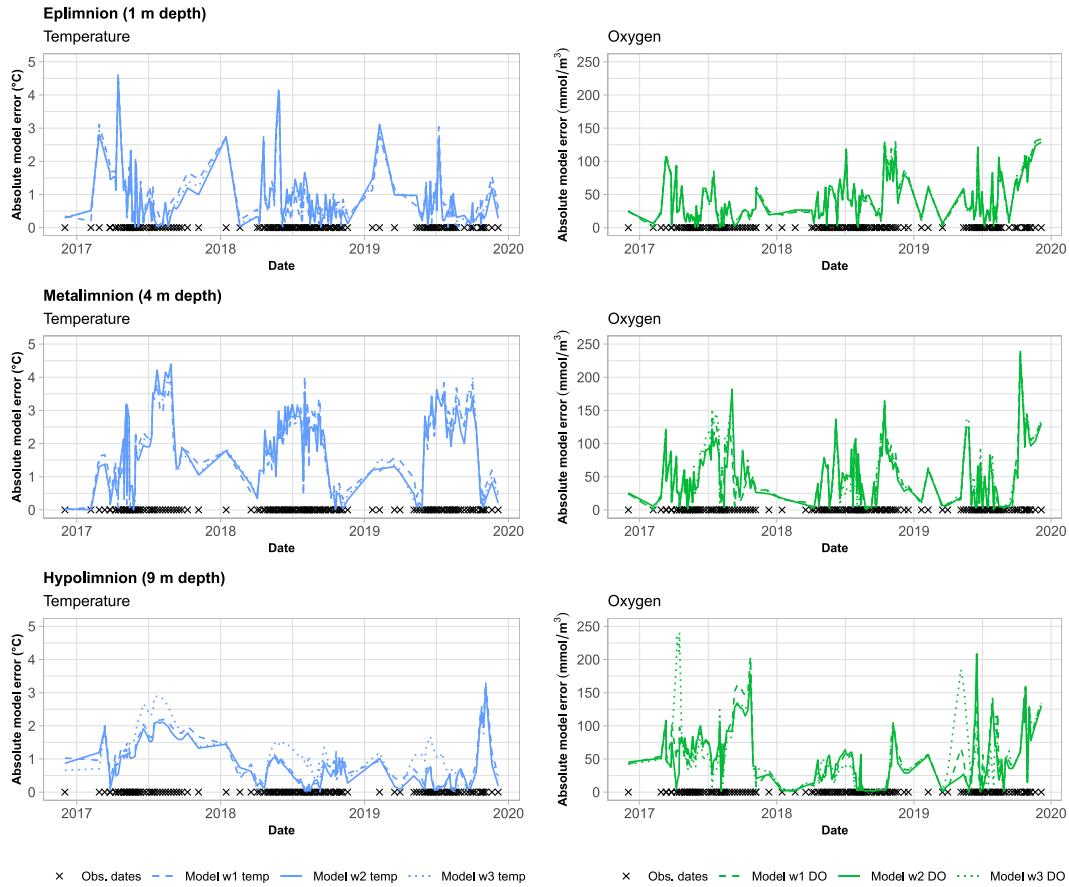

Figure 5. The absolute difference between the observed water temperature and dissolved oxygen concentration and the values predicted by the three models with different weighting schemes based on the system-inspired approach with DM 2 (Model w1, Model w2, Model w3). The observation dates show when observations were collected.

3.3. Weighting Strategies of the System-Inspired Metrics

The calibration results were, to a degree, sensitive to the weighting configuration applied to the system-inspired metrics (Table 2). The water temperature predictions were the least sensitive to the choice of weighting strategy (Figure 5). There were more significant differences present in the simulation of dissolved oxygen between the system-inspired models, particularly in the simulation of hypolimnetic oxygen concentrations (Figure 5). The greatest differences between the system-inspired models occurred in capturing the TD and MOM (Figure 6). Overall, Model w2 seemed to outcompete the other models in most aspects, including in simulating the MOM and the NAL. Model w1 outperformed the other models in reproducing the observed pattern of the TD. The worst performing model in all respects was Model w3, where the weights of the extra metrics observation groups were set to double that of Model w2.

3.4. Comparison of Calibration Approaches and Configurations

Compared to the reference model (Carey, Hanson, et al., 2022), the performance of the PEST calibrated GLM-AED models was substantially improved both in the calibration and validation period (Figure 7). The greatest improvement corresponded to the prediction of the DO profile and the oxygen related system metrics such as the MOM and the vertical extent of anoxia quantified by the number of anoxic layers per day. The difference in performance was less pronounced when moving from the naïve calibration to the system-inspired approach. When system metrics were added to the objective function, there was a clear improvement in the model's ability to capture the behavior of these extra metrics, which led to increased coherence between the system-scale

10.1029/2023WR036570

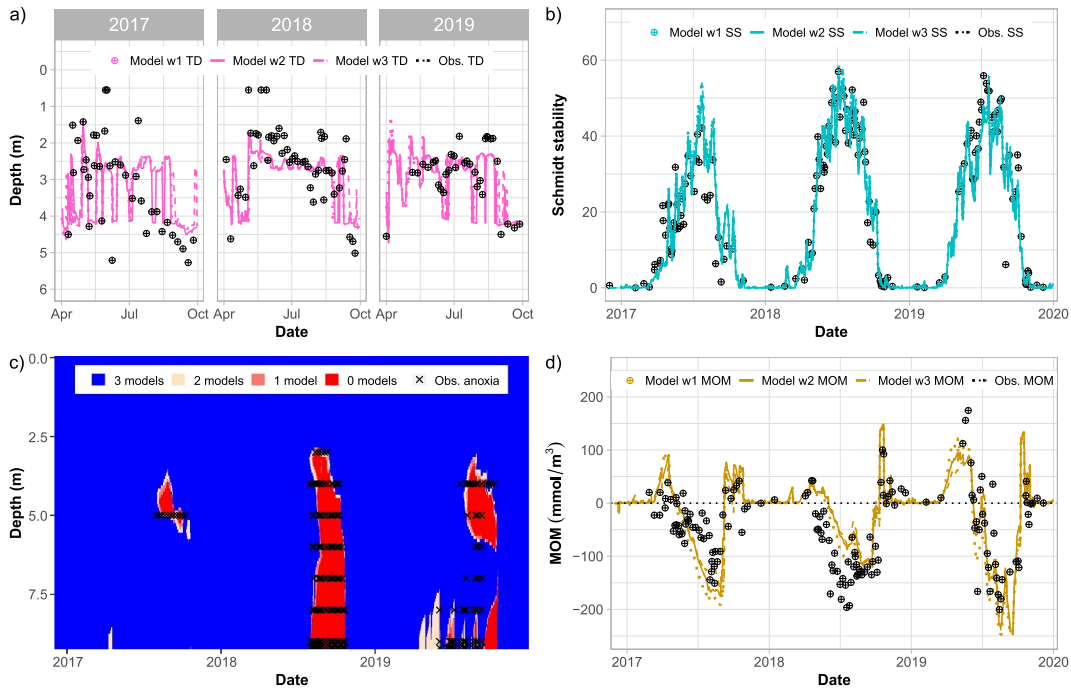

Figure 6. Comparison of the observed system-metrics (Obs.) and the system-metrics predicted by the three models with different weighting schemes based on the system-inspired approach with DM 2 (Model w1, Model w2, Model w3). The metrics include the thermocline depth during the stratified period (a), Schmidt stability (b), the spatial and temporal pattern of anoxia (c), and the metalimnetic oxygen minima (d). Figure 6c illustrates the number of models that predict a certain pixel to be anoxic within the water column.

dynamics of the simulated and natural ecosystem. However, there was a slight trade-off in accuracy between the simulation of extra metrics and the DO profile, while the accuracy of the temperature profile remained the same (Table 2). The loss in the MEF of the DO profile was less than 0.1 for all weighting strategies, while the gain in the MEF of the extra metrics was greater than 0.1 in the majority of cases. The choice of the deep mixing model structure had a significant effect on model performance. While hypolimnetic mixing with constant diffusivity was more suitable for the simulation of the TD and the MOM, the Weinstock model of diffusivity was able to better capture the vertical extent of anoxia in the reservoir during the calibration period. However, during the validation period, the models based on the Weinstock model of diffusivity also demonstrated superior performance in capturing the TD and MOM, in addition to the vertical extent of anoxia. In general, model performance was better in the calibration period except for simulating the MOM, which was better captured during the validation period.

3.5. Uncertainty Analysis

The uncertainty in predicting three system-inspired metrics of interest, the TD, SS and the MOM, was significantly reduced post-calibration compared to pre-calibration (Figure 8). The propagation of uncertainty prior to calibration was the greatest in predicting the MOM, which also exhibited the most substantial reduction in uncertainty post calibration. As expected, the posterior mean of each metric generally followed the pattern of the observed data, except for the TD in 2017 and the MOM in 2018. Hence it appears that the same parameter sets are not able to reproduce the patterns of the observed data each simulation year. The likely range of metrics outputs based on the prior distribution did not fully encompass all observation points (Figure 8). This phenomenon was most notable in the case of the TD and suggests that the maximum expected parameter uncertainty (i.e., the prior) did not include a wide enough range of model outputs to capture all observation points, which could be a manifestation of model structural error.

10.1029/2023WR036570

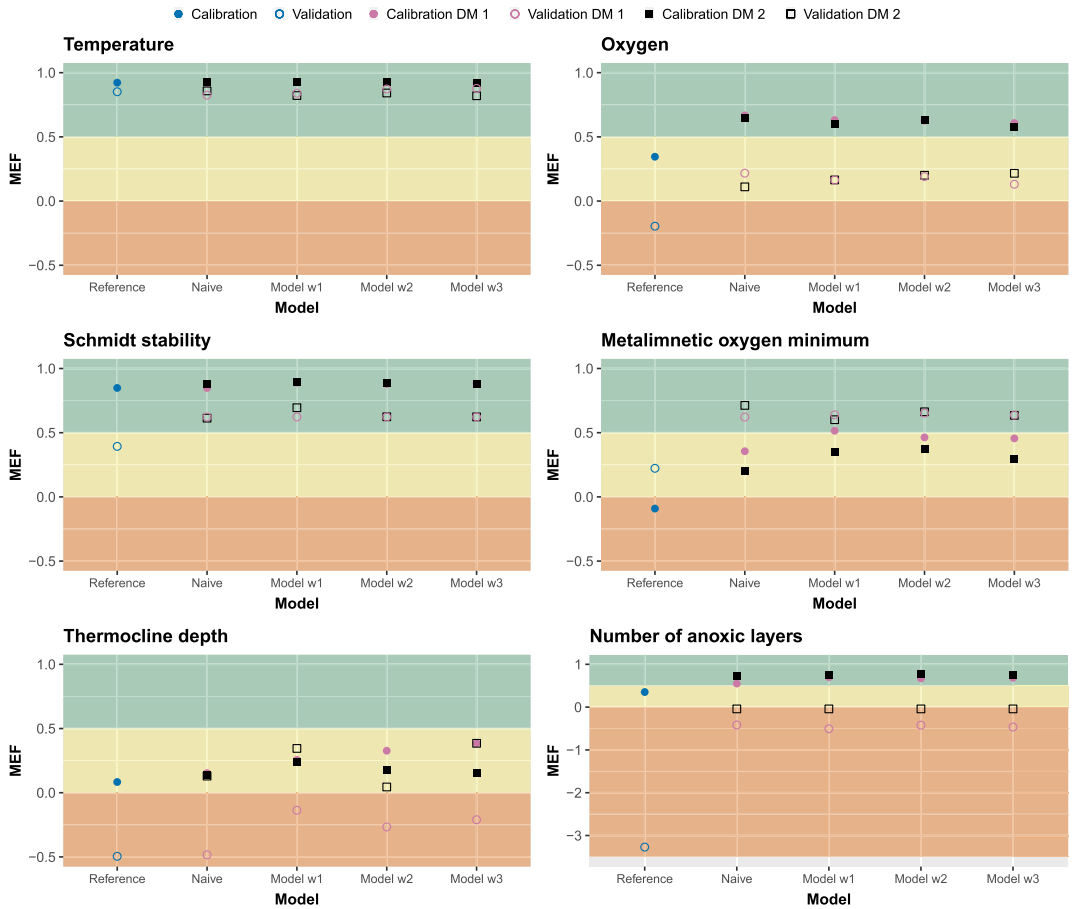

Figure 7. Comparison of model efficiency (MEF) between models with two different deep mixing configurations during the calibration and validation periods. One deep mixing configuration is based on constant diffusivity (DM 1) and the other configuration employs the Weinstock model to determine the diffusivity (DM 2). The orange coloring corresponds to poor model performance (MEF <0), the yellow coloring corresponds to acceptable model performance (0< MEF <0.5), and the green coloring corresponds to good model performance (MEF >0.5).

4. Discussion

This study aimed to answer the question of whether non-traditional, system-inspired metrics of ecosystem state or function can improve model performance and assist in characterizing uncertainty. By incorporating system-inspired metrics in the objective function, a highly targeted model calibration was achieved, leading to improved accuracy in resolving key ecosystem dynamics.

4.1. Model Accuracy

The implementation of system-inspired metrics focused calibration on specific ecosystem behaviors that the metrics were designed to represent. However, the simulated dissolved oxygen depth profile suffered a slight loss of accuracy relative to the naïve model, while the temperature profile remained similar between models. This trade-off is a result of achieving a greater overall objective function reduction by minimizing the error between system metrics model predictions and observations rather than the oxygen depth profile. As our focal AEM is one-dimensional (horizontally averaged), we deemed this trade-off to be acceptable, as it enabled refocusing the calibration on key patterns and ecosystem responses instead of individual observation points at a specific site that may not have represented the reservoir's average conditions.

The overall improvement in model prediction accuracy from the naïve approach to the system-inspired approach was less pronounced than anticipated. One contributing factor is that the observed data set of state-variable measurements had extremely comprehensive spatial and temporal resolution. Hence, adding extra metrics to

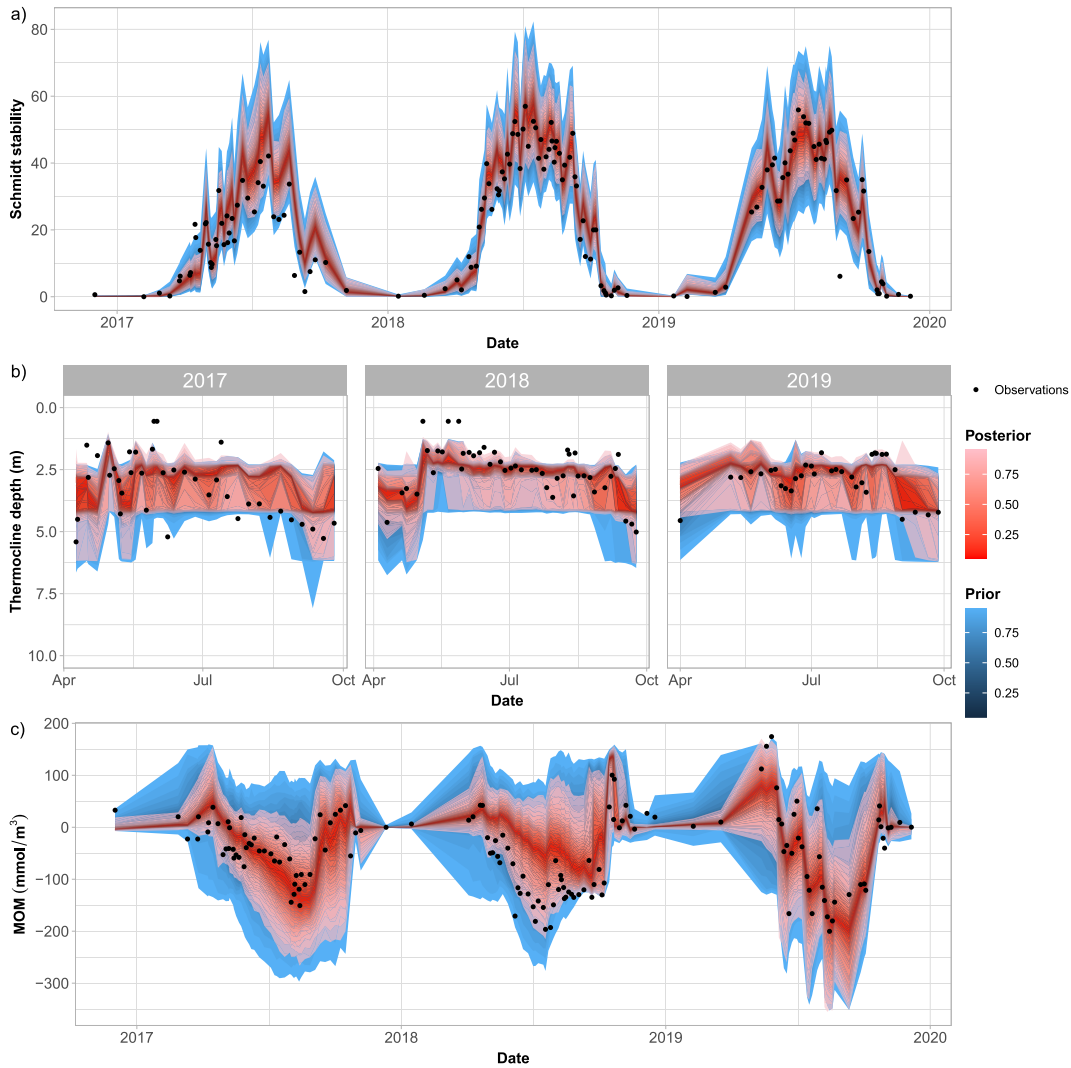

Figure 8. Prior and posterior probability distributions of the Schmidt stability (a), the thermocline depth (b), and metalimnetic oxygen minima (c) predictions compared to observations.

the calibration process may improve model prediction accuracy to a greater extent in other situations when there are more sparse monitoring data, there is more uncertainty, or the introduction of additional information to the model is more valuable (Sousa et al., 2023). Given that model prediction accuracy generally diminishes with increased model level and complexity (e.g., hydrodynamic—abiotic—biotic) (Soares & Calijuri, 2021), future efforts could explore the benefits of the calibration framework presented in this paper for more complex AEM configurations.

4.2. Modeling Procedure

Assigning weights to extra metrics remains a subjective element of the calibration process. In this study, we explored the sensitivity of the calibration results to different weighting schemes, which translate to the relative priorities given to certain objectives during the calibration process. Weighting scheme 1 represented a balanced approach to weighting, while weighting scheme 2 was focused on the uncertainty of the observations implemented in the objective function (Tiedeman et al., 2003), and weighting scheme 3 was focused on the system-inspired metrics. Weighting scheme 1 and 2 exhibited negligible differences in the calibration results, while the application of weighting scheme 3 led to diminishing model performance with respect to the prediction of both state-variables and system metrics. The loss in state variable accuracy was consistent with findings that increasing the weights of certain variables may reduce the prediction accuracy of others (Cho & Ha, 2010). However, in our

case it also led to poorer prediction of the metrics we intended to improve. Consequently, achieving an optimal balance in weighting is crucial as assigning unproportionate high weights to selected observations may cause overfitting, and degradation in overall model performance. To make the selection of weightings less subjective, the weightings of metrics could be considered as a hyper-parameter(s) in the model optimization framework, however, we note that it is important to balance both the computational costs and the potential benefits of any optimization approach.

Incorporating extra metrics in the calibration process can improve the evaluation of model structural decisions and eliminate the need for ad hoc selection. When two or more competing model structures have been identified to capture the study site, system-inspired metrics can be used in the context of comparing model structures. In cases when two different models perform equally well in predicting state-variables, comparing their ability to capture system behavior helps to shed light on previously hidden strengths and weaknesses (Hipsey et al., 2020). Comparing the two deep mixing sub-models based on the system-inspired metrics revealed that the constant diffusivity model performed better in simulating the TD and the MOM, while it did not capture the anoxic conditions in the hypolimnion well during the calibration period. The Weinstock model was more effective in depicting the vertical extent of anoxia, which aligned with our primary modeling aim. In pattern-oriented modeling, a similar approach is implemented when alternative sub-models are treated as hypotheses that can be tested based on whether they can reproduce the patterns of interest (Grimm & Railsback, 2012). Consequently, extra metrics can assist users in aligning model structural decisions with current and future modeling endeavor and can serve as a valuable tool guiding model development.

The implementation of extra metrics in the calibration process assists in evaluating when a model is "successfully" calibrated. Calibration is well-established as one of the essential steps of the modeling procedure (Refsgaard et al., 2007), however, what is regarded as a "successful" calibration is less clear. Finding the most suitable parameter set is an iterative process, whereby after each iteration, the calibration performance is examined (Mai, 2023). This is done by checking if the calibrated model accurately represents the features of the observed data (Jakeman et al., 2006). Whether the model is fit for purpose, and cannot be significantly improved by further calibration is based on expert knowledge, and often remains ad hoc depending on site specific data availability, model spatiotemporal scale, and model complexity. Identifying the point of diminishing returns in model fit is a challenging task, and using system-inspired metrics in the calibration process has much potential for providing reassurance that the study site is captured well on the system-level, and helps ensure the model is fit for purpose.

4.3. Management Benefits of Improved Model Reliability

Since management actions often have a system-level impact through interactions and feedbacks inherent within complex aquatic systems, a model that is able to reproduce important characteristics of ecosystem behavior can improve trust in the utility of complex AEMs in active decision making (Hipsey et al., 2015; Regev et al., 2023). The system-inspired metrics that were selected for this study represent important water quality indices that have tangible management applications. For reservoir operators, selective withdrawal and oxygenation system operations are among the limited management options available to control water quality and quantity in the reservoir, while avoiding negative downstream impacts (Gerling et al., 2016; Saadatpour et al., 2021; Weber et al., 2019). Managers can use system-inspired metrics to directly track whether their interventions maintain the thermal structure of the waterbody and an oxygenated bottom layer, which are critical objectives. For example, a shallower thermocline depth leads to a greater hypolimnetic volume, which is more costly to oxygenate (Gerling et al., 2014). Similarly, oxygen concentrations are a major control of nutrient and metal release from the sediments, which can in turn stimulate algal growth, intensify eutrophication, and deteriorate drinking water quality (Carey, 2023). Hence, the vertical extent of anoxia has a clear link to water quality impairment. The formation of metalimnetic oxygen minimum zones has been observed in number of reservoirs, particularly in those equipped with engineering systems (McClure et al., 2018). Epi-/metalimnetic withdrawal options for managers, which can be used to alleviate temperature pollution, are greatly limited in the presence of metalimnetic oxygen minimum zones, emphasizing the need for configuring AEMs to ensure that they are able to capture these system-level dynamics.

Our results suggest that implementing an extended list of system-inspired metrics in model assessment can assist in more accurately describing chemical and biological attributes of interest and demonstrate the system-level

benefits of management actions (Arhonditsis et al., 2019). For example, in the case of metabolism models, system-level understanding has been valuable in approximating crucial processes such as respiration and productivity in the face of equifinality (Appling et al., 2018), which presents a significant challenge for modelers and has important consequences for ecosystem functioning. Altogether, multi-objective optimization of lake ecosystem management to reduce water quality risks, improve habitat and biodiversity, and reduce greenhouse gas emissions are demanding more complex ecosystem models (Wu et al., 2023), requiring new approaches, such as system-inspired model calibration, to ensure that AEMs are up to the task.

5. Conclusion

Here, our use of system-metrics in calibration and uncertainty analysis workflows provides new insight into how to assess AEMs of stratified lakes. We found that introducing metrics relevant to the local system operation and modeling aim allowed for a targeted calibration. Marginal reduction in the accuracy of state-variables to improve the prediction of system-metrics was a worthwhile trade-off in our reservoir example. The calibration results were sensitive to the weighting scheme applied to the extra metrics, and over-weighting them led to degrading overall model performance. The use of uncertainty analysis for estimating the range of likely values of system-inspired metrics can assist in optimizing reservoir management. For example, quantifying the uncertainty in simulating metalimnetic oxygen minima dynamics can facilitate the improved operation of our focal reservoir's oxygenation system. The list of system-inspired metrics applied in this study can easily be extended over time for a wider diversity of applications. Altogether, developing system-metrics to assist in the calibration of complex AEMs has the potential to significantly improve their predictive accuracy and reliability, and is a step toward a more objective framework for AEM performance assessment.

Data Availability Statement

All model files, R scripts and model executable files are available in the Zenodo repository FCR-GLM-metrics (Kurucz et al., 2024). All observational data files used for model calibration, validation, and the calculation of system-metrics are available in the Environmental Data Initiative repository (Carey & Breef-Pilz, 2022, 2023; Carey, Lewis, et al., 2022; Carey, Wander, et al., 2022).

Acknowledgments

We thank the Reservoir Group at Virginia Tech for collecting the field observations used in this study during 2017–2019, especially Bethany Bookout, Alexandria Hounshell, Abby Lewis, Mary Lofton, Ryan McClure, and Heather Wander. Quinn Thomas helped with the naïve calibration of the GLM-AED model for FCR. This project was financially supported by a Robert and Maude Gledden Visiting Fellowship and Future Fulbright Fellowship to CCC, and U.S. National Science Foundation Grants 1933016, 2327030, 2330211, and 2318861, and MRH and PH were supported by funding from Australian Research Council Grants LP150100451, LP150100519, LP200200910, and LP220200882. Open access publishing facilitated by The University of Western Australia, as part of the Wiley - The University of Western Australia agreement via the Council of Australian University Librarians.

References

Albers, S., Winslow, L., Collinge, D., Read, J. S., Leach, T., Zwart, J., & Snortheim, C. (2018). rLakeAnalyzer: Lake physics tool (version 1.8.3.) [Software]. *Zenodo*. https://zenodo.org/record/1003169#.Ws82F9NubEY

Appling, A. P., Hall, R. O., Yackulic, C. B., & Arroita, M. (2018). Overcoming equifinality: Leveraging long time series for stream metabolism estimation. *Journal of Geophysical Research: Biogeosciences*, *123*(2), 624–645. https://doi.org/10.1002/2017jg004140

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

Arhonditsis, G. B., & Brett, M. T. (2005). Eutrophication model for Lake Washington (USA) Part I. Model description and sensitivity analysis. *Ecological Modelling*, *187*(2–3), 140–178. https://doi.org/10.1016/j.ecolmodel.2005.01.040

Arhonditsis, G. B., Neumann, A., Shimoda, Y., Javed, A., Blukacz-Richards, A., & Mugalingam, S. (2019). When can we declare a success? A Bayesian framework to assess the recovery rate of impaired freshwater ecosystems. *Environment International*, *130*, 104821. https://doi.org/10.1016/j.envint.2019.05.015

Beck, M. B. (1987). Water quality modeling: A review of the analysis of uncertainty. *Water Resources Research*, *23*(8), 1393–1442. https://doi.org/10.1029/wr023i008p01393

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

Bruce, L. C., Frassl, M. A., Arhonditsis, G. B., Gal, G., Hamilton, D. P., Hanson, P. C., et al. (2018). A multi-lake comparative analysis of the general lake model (GLM): Stress-testing across a global observatory network. *Environmental Modelling and Software*, *102*, 274–291. https://doi.org/10.1016/j.envsoft.2017.11.016

Carey, C. C. (2023). Causes and consequences of changing oxygen availability in lakes. *Inland Waters*, *13*(3), 316–326. https://doi.org/10.1080/20442041.2023.2239110

Carey, C. C., & Breef-Pilz, A. (2022). Ice cover data for falling Creek reservoir and Beaverdam reservoir, Vinton, Virginia, USA for 2013-2022 (version 4) [Dataset]. *Environmental Data Initiative*. https://doi.org/10.6073/pasta/917b3947d91470eecf979e9297ed4d2e

Carey, C. C., & Breef-Pilz, A. (2023). Time series of high-frequency meteorological data at falling creek reservoir, Virginia, USA 2015-2022 (version 7) [Dataset]. *Environmental Data Initiative*. https://doi.org/10.6073/pasta/f3f97c7fdd287c29084bf52fc759a801

Carey, C. C., Hanson, P. C., Thomas, R. Q., Gerling, A. B., Hounshell, A. G., Lewis, A. S. L., et al. (2022). Anoxia decreases the magnitude of the carbon, nitrogen, and phosphorus sink in freshwaters. *Global Change Biology*, *28*(16), 4861–4881. https://doi.org/10.1111/gcb.16228

Carey, C. C., Lewis, A. S., McClure, R. P., Gerling, A. B., Breef-Pilz, A., & Das, A. (2022). Time series of high-frequency profiles of depth, temperature, dissolved oxygen, conductivity, specific conductance, chlorophyll a, turbidity, pH, oxidation-reduction potential, photosynthetic active radiation, and descent rate for Beaverdam reservoir, Carvins Cove reservoir, falling creek reservoir, Gatewood reservoir, and spring hollow reservoir in Southwestern Virginia, USA 2013-2021 (version 12) [Dataset]. *Environmental Data Initiative*. https://doi.org/10.6073/pasta/c4c45b5b10b4cb4cd4b5e613c3effbd0

Carey, C. C., Thomas, R. Q., & Hanson, P. C. (2022). General Lake model-aquatic EcoDynamics model parameter set for falling creek reservoir, Vinton, Virginia, USA 2013-2019 (version 1) [Dataset]. *Environmental Data Initiative*. https://doi.org/10.6073/pasta/9f7d037d9a133076a0a0d123941c6396

Carey, C. C., Wander, H. L., McClure, R. P., Lofton, M. E., Hamre, K. D., Doubek, J. P., et al. (2022). Secchi depth data and discrete depth profiles of photosynthetically active radiation, temperature, dissolved oxygen, and pH for Beaverdam reservoir, Carvins cove reservoir, falling Creek reservoir, gatewood reservoir, and spring hollow reservoir in southwestern Virginia, USA 2013-2021 (version 10) [Dataset]. *Environmental Data Initiative*. https://doi.org/10.6073/pasta/887d8ab8c57fb8fdf3582507f3223cd6

Chen, Y., & Oliver, D. S. (2013). Levenberg–Marquardt forms of the iterative ensemble smoother for efficient history matching and uncertainty quantification. *Computational Geosciences*, *17*(4), 689–703. https://doi.org/10.1007/s10596-013-9351-5

Cho, J. H., & Ha, S. R. (2010). Parameter optimization of the QUAL2K model for a multiple-reach river using an influence coefficient algorithm. *The Science of the Total Environment*, *408*(8), 1985–1991. https://doi.org/10.1016/j.scitotenv.2010.01.025

Doherty, J. (2018a). *Manual for PEST: Model-independent parameter estimation. Part 1: PEST, SENSAN and global optimisers*. Watermark Numerical Computing.

Doherty, J. (2018b). *Manual for PEST: Model-independent parameter estimation. Part 2: PEST utility support software*. Watermark Numerical Computing.

Elhabashy, A., Li, J., & Sokolova, E. (2023). Water quality modeling of a eutrophic drinking water source: Impact of future climate on Cyanobacterial blooms. *Ecological Modelling*, *477*, 110275. https://doi.org/10.1016/j.ecolmodel.2023.110275

Frassl, M. A., Abell, J. M., Botelho, D. A., Cinque, K., Gibbes, B. R., Jöhnk, K. D., et al. (2019). A short review of contemporary developments in aquatic ecosystem modelling of lakes and reservoirs. *Environmental Modelling and Software*, *117*, 181–187. https://doi.org/10.1016/j.envsoft.2019.03.024

Gerling, A. B., Browne, R. G., Gantzer, P. A., Mobley, M. H., Little, J. C., & Carey, C. C. (2014). First report of the successful operation of a side stream supersaturation hypolimnetic oxygenation system in a eutrophic, shallow reservoir. *Water Research*, *67*, 129–143. https://doi.org/10.1016/j.watres.2014.09.002

Gerling, A. B., Munger, Z. W., Doubek, J. P., Hamre, K. D., Gantzer, P. A., Little, J. C., & Carey, C. C. (2016). Whole-catchment manipulations of internal and external loading reveal the sensitivity of a century-old reservoir to hypoxia. *Ecosystems*, *19*(3), 555–571. https://doi.org/10.1007/s10021-015-9951-0

Grimm, V., & Railsback, S. F. (2012). Pattern-oriented modelling: A 'multi-scope' for predictive systems ecology. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1586), 298–310. https://doi.org/10.1098/rstb.2011.0180

Grimm, V., Revilla, E., Berger, U., Jeltsch, F., Mooij, W. M., Railsback, S. F., et al. (2005). Pattern-oriented modeling of agent-based complex systems: Lessons from ecology. *Science*, *310*(5750), 987–991. https://doi.org/10.1126/science.1116681

Hipsey, M. R. (Ed.) (2022)., *Modelling aquatic eco-dynamics: Overview of the AED modular simulation platform*. Zenodo Repository. https://doi.org/10.5281/zenodo.6516222

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., Gal, G., Arhonditsis, G. B., Carey, C. C., Elliott, J. A., Frassl, M. A., et al. (2020). A system of metrics for the assessment and improvement of aquatic ecosystem models. *Environmental Modelling and Software*, *128*, 104697. https://doi.org/10.1016/j.envsoft.2020.104697

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*(9), 7023–7043. https://doi.org/10.1002/2015wr017175

Huettmann, F., & Arhonditsis, G. (2023). Towards an ecological informatics scholarship that is reflective, repeatable, transparent, and sharable. *Ecological Informatics*, *76*, 102132. https://doi.org/10.1016/j.ecoinf.2023.102132

Idso, S. B. (1973). On the concept of lake stability. *Limnology & Oceanography*, *18*(4), 681–683. https://doi.org/10.4319/lo.1973.18.4.0681

Jakeman, A. J., Letcher, R. A., & Norton, J. P. (2006). Ten iterative steps in development and evaluation of environmental models. *Environmental Modelling and Software*, *21*(5), 602–614. https://doi.org/10.1016/j.envsoft.2006.01.004

Janssen, A. B. G., Arhonditsis, G. B., Beusen, A., Bolding, K., Bruce, L., Bruggeman, J., et al. (2015). Exploring, exploiting and evolving diversity of aquatic ecosystem models: A community perspective. *Aquatic Ecology*, *49*(4), 513–548. https://doi.org/10.1007/s10452-015-9544-1

Kim, D.-K., Zhang, W., Watson, S., & Arhonditsis, G. B. (2014). A commentary on the modelling of the causal linkages among nutrient loading, harmful algal blooms, and hypoxia patterns in Lake Erie. *Journal of Great Lakes Research*, *40*, 117–129. https://doi.org/10.1016/j.jglr.2014.02.014

Kotamäki, N., Arhonditsis, G., Hjerppe, T., Hyytiäinen, K., Malve, O., Ovaskainen, O., et al. (2024). Strategies for integrating scientific evidence in water policy and law in the face of uncertainty. *The Science of the Total Environment*, *931*, 172855. https://doi.org/10.1016/j.scitotenv.2024.172855

Kurucz, K., Hipsey, M., Carey, C., Huang, P., De Sousa, E., & White, J. (2024). Data and software for the GLM-AED simulation and model calibration of the Falling Creek Reservoir (v1. 0. 1.) [Software]. *Zenodo*. https://doi.org/10.5281/zenodo.11183590

Ladwig, R., Hanson, P. C., Dugan, H. A., Carey, C. C., Zhang, Y., Shu, L., et al. (2021). Lake thermal structure drives interannual variability in summer anoxia dynamics in a eutrophic lake over 37 years. *Hydrology and Earth System Sciences*, *25*(2), 1009–1032. https://doi.org/10.5194/hess-25-1009-2021

Lofton, M. E., McClure, R. P., Chen, S., Little, J. C., & Carey, C. C. (2019). Whole-ecosystem experiments reveal varying responses of phytoplankton functional groups to epilimnetic mixing in a eutrophic reservoir. *Water*, *11*(2), 222. https://doi.org/10.3390/w11020222

Mai, J. (2023). Ten strategies towards successful calibration of environmental models. *Journal of Hydrology*, *620*, 129414. https://doi.org/10.1016/j.jhydrol.2023.129414

McClure, R. P., Hamre, K. D., Niederlehner, B. R., Munger, Z. W., Chen, S., Lofton, M. E., et al. (2018). Metalimnetic oxygen minima alter the vertical profiles of carbon dioxide and methane in a managed freshwater reservoir. *The Science of the Total Environment*, *636*, 610–620. https://doi.org/10.1016/j.scitotenv.2018.04.255

McClure, R. P., Schreiber, M. E., Lofton, M. E., Chen, S., Krueger, K. M., & Carey, C. C. (2021). Ecosystem-scale oxygen manipulations alter terminal electron acceptor pathways in a eutrophic reservoir. *Ecosystems*, *24*(6), 1281–1298. https://doi.org/10.1007/s10021-020-00582-9

Mooij, W. M., Trolle, D., Jeppesen, E., Arhonditsis, G., Belolipetsky, P. V., Chitamwebwa, D. B. R., et al. (2010). Challenges and opportunities for integrating Lake Ecosystem modelling approaches. *Aquatic Ecology*, *44*(3), 633–667. https://doi.org/10.1007/s10452-010-9339-3

Munger, Z. W., Carey, C. C., Gerling, A. B., Doubek, J. P., Hamre, K. D., McClure, R. P., & Schreiber, M. E. (2019). Oxygenation and hydrologic controls on iron and manganese mass budgets in a drinking-water reservoir. *Lake and Reservoir Management*, *35*(3), 1–15. https://doi.org/10.1080/10402381.2018.1545811

Nash, J. E., & Sutcliffe, J. V. (1970). River flow forecasting through conceptual models part I—A discussion of principles. *Journal of Hydrology*, *10*(3), 282–290. https://doi.org/10.1016/0022-1694(70)90255-6

Nielsen, A., Trolle, D., Bjerring, R., Søndergaard, M., Olesen, J. E., Janse, J. H., et al. (2014). Effects of climate and nutrient load on the water quality of shallow lakes assessed through ensemble runs by PCLake. *Ecological Applications*, *24*(8), 1926–1944. https://doi.org/10.1890/13-0790.1

Ranjbar, M. H., Hamilton, D. P., Etemad-Shahidi, A., & Helfer, F. (2021). Individual-based modelling of cyanobacteria blooms: Physical and physiological processes. *The Science of the Total Environment*, *792*, 148418. https://doi.org/10.1016/j.scitotenv.2021.148418

Read, J. S., Hamilton, D. P., Jones, I. D., Muraoka, K., Winslow, L. A., Kroiss, R., et al. (2011). Derivation of lake mixing and stratification indices from high-resolution lake buoy data. *Environmental Modelling and Software*, *26*(11), 1325–1336. https://doi.org/10.1016/j.envsoft.2011.05.006

Refsgaard, J. C., van der Sluijs, J. P., Højberg, A. L., & Vanrolleghem, P. A. (2007). Uncertainty in the environmental modelling process – a framework and guidance. *Environmental Modelling and Software*, *22*(11), 1543–1556. https://doi.org/10.1016/j.envsoft.2007.02.004

Regev, S., Carmel, Y., & Gal, G. (2023). Using high level validation to increase Lake Ecosystem model reliability. *Environmental Modelling and Software*, *162*, 105637. https://doi.org/10.1016/j.envsoft.2023.105637

Robson, B. J., Skerratt, J., Baird, M. E., Davies, C., Herzfeld, M., Jones, E. M., et al. (2020). Enhanced assessment of the eReefs biogeochemical model for the great barrier reef using the concept/state/process/system model evaluation framework. *Environmental Modelling and Software*, *129*, 104707. https://doi.org/10.1016/j.envsoft.2020.104707

Saadatpour, M., Javaheri, S., Afshar, A., & Solis, S. S. (2021). Optimization of selective withdrawal systems in hydropower reservoir considering water quality and quantity aspects. *Expert Systems with Applications*, *184*, 115474. https://doi.org/10.1016/j.eswa.2021.115474

Soares, L. M. V., & do Calijuri, M. C. (2021). Deterministic modelling of freshwater lakes and reservoirs: Current trends and recent progress. *Environmental Modelling and Software*, *144*, 105143. https://doi.org/10.1016/j.envsoft.2021.105143

Sousa, E. R. D., Hipsey, M. R., & Vogwill, R. I. J. (2023). Data assimilation, sensitivity analysis and uncertainty quantification in semi-arid terminal catchments subject to long-term rainfall decline. *Frontiers in Earth Science*, *10*, 886304. https://doi.org/10.3389/feart.2022.886304

Stepanenko, V., Mammarella, I., Ojala, A., Miettinen, H., Lykosov, V., & Vesala, T. (2016). Lake 2.0: A model for temperature, methane, carbon dioxide and oxygen dynamics in lakes. *Geoscientific Model Development*, *9*(5), 1977–2006. https://doi.org/10.5194/gmd-9-1977-2016

Tiedeman, C. R., Hill, M. C., D'Agnese, F. A., & Faunt, C. C. (2003). Methods for using groundwater model predictions to guide hydrogeologic data collection, with application to the Death Valley regional groundwater flow system. *Water Resources Research*, *39*(1). https://doi.org/10.1029/2001wr001255

Trolle, D., Hamilton, D. P., Pilditch, C. A., Duggan, I. C., & Jeppesen, E. (2011). Predicting the effects of climate change on trophic status of three morphologically varying lakes: Implications for lake restoration and management. *Environmental Modelling and Software*, *26*(4), 354–370. https://doi.org/10.1016/j.envsoft.2010.08.009

Weber, M., Boehrer, B., & Rinke, K. (2019). Minimizing environmental impact whilst securing drinking water quantity and quality demands from a reservoir. *River Research and Applications*, *35*(4), 365–374. https://doi.org/10.1002/rra.3406

White, J. T. (2018). A model-independent iterative ensemble smoother for efficient history-matching and uncertainty quantification in very high dimensions. *Environmental Modelling and Software*, *109*, 191–201. https://doi.org/10.1016/j.envsoft.2018.06.009

Wilhelm, S., & Adrian, R. (2008). Impact of summer warming on the thermal characteristics of a polymictic lake and consequences for oxygen, nutrients and phytoplankton. *Freshwater Biology*, *53*(2), 226–237. https://doi.org/10.1111/j.1365-2427.2007.01887.x

Wilsnack, M. M., Doherty, J. E., & Welter, D. E. (2012). Pareto-based methodology for the calibration and uncertainty analysis of gated culvert flows. *Journal of Irrigation and Drainage Engineering*, *138*(7), 675–684. https://doi.org/10.1061/(asce)ir.1943-4774.0000431

Wu, W., Eamen, L., Dandy, G., Razavi, S., Kuczera, G., & Maier, H. R. (2023). Beyond engineering: A review of reservoir management through the lens of wickedness, competing objectives and uncertainty. *Environmental Modelling and Software*, *167*, 105777. https://doi.org/10.1016/j.envsoft.2023.105777