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

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

[1]The University of Western Australia
[2]Virginia Tech
[3]INTERA

December 10, 2023

Abstract

Despite the growing use of Aquatic Ecosystem Models (AEMs) for lake modelling, there is currently no widely applicable framework for their configuration, calibration, and evaluation. To date, calibration is generally based on direct data comparison of observed vs. modelled state variables using standard statistical techniques, however, this approach may not give a complete picture of the model's ability to capture system-scale behaviour that is not prevalent in the state 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, a new approach that augments the standard state-based calibration with a range of system-inspired metrics (e.g. thermocline depth, metalimnetic oxygen minima), in an effort to increase the coherence between the simulated and natural ecosystems. This was achieved by applying a coupled physical-biogeochemical model to a focal site to simulate temperature and dissolved oxygen. The model was calibrated according to the new system-inspired modelling convention, using formal calibration techniques. There was a clear improvement in the simulation using parameters optimised on the additional metrics, which helped to focus calibration on aspects of the system relevant to reservoir management, such as the metalimnetic oxygen minima. Extending the use of system-inspired metrics for the calibration of models of nutrient cycling, algal blooms, and greenhouse gas emissions has the potential to greatly improve the prediction of complex ecosystem dynamics.

Hosted file

`977616_0_art_file_11609514_s4x0w0.docx` available at https://authorea.com/users/706108/articles/691583-using-system-inspired-metrics-to-improve-water-quality-prediction-in-stratified-lakes

Hosted file

`977616_0_supp_11609513_s4svtw.docx` available at https://authorea.com/users/706108/articles/691583-using-system-inspired-metrics-to-improve-water-quality-prediction-in-stratified-lakes

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, Virginia, USA.

[3]INTERA Inc., Perth, WA, Australia.

*Corresponding author: Kamilla Kurucz (kamilla.kurucz@research.uwa.edu.au)

Key Points:

- We assessed the use of system-inspired metrics in a novel approach to calibrating aquatic ecosystem models.
- The use of system-inspired metrics in calibration substantially 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.

Abstract

Despite the growing use of Aquatic Ecosystem Models (AEMs) for lake modelling, there is currently no widely applicable framework for their configuration, calibration, and evaluation. To date, calibration is generally based on direct data comparison of observed vs. modelled state variables using standard statistical techniques, however, this approach may not give a complete picture of the model's ability to capture system-scale behaviour that is not prevalent in the state 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, a new approach that augments the standard state-based calibration with a range of system-inspired metrics (e.g., thermocline depth, metalimnetic oxygen minima), in an effort to increase the coherence between the simulated and natural ecosystems. This was achieved by applying a coupled physical-biogeochemical model to a focal site to simulate temperature and dissolved oxygen. The model was calibrated according to the new system-inspired modelling convention, using formal calibration techniques. There was a clear improvement in the simulation using parameters optimised on the additional metrics, which helped to focus calibration on aspects of the system relevant to reservoir management, such as the metalimnetic oxygen minima. Extending the use of system-inspired metrics for the calibration of models of nutrient cycling, algal blooms, and greenhouse gas emissions has the potential to greatly improve the prediction of 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 (Jannsen et al., 2015; Soares & Calijuri, 2021). These models can be used for several different purposes across various time and spatial scales, making them useful decision-making tools for addressing the environmental issues affecting lentic 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 et al., 2022a; Ladwig et al., 2021), eutrophication (Arhonditsis & Brett, 2005), and greenhouse gas emissions (Stepanenko et al., 2016), and predict

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). However, despite their widespread use, there is no consensus as to how best to configure, calibrate, and evaluate AEMs for lake modelling, leading to the need for new approaches for historical and future aquatic ecosystem prediction (Frassl et al., 2019).

A major challenge in setting up and applying AEMs is appropriately calibrating model parameters. The values of model parameters are relatively unknown, in contrast to state variables, where information regarding values and variability is well established through empirical measurements (Hipsey et al., 2020). Hence, the scope of calibration requires identifying the parameter set within the parameter space that best fits observations. However, the prevalence of unknown model parameters combined with the lack of observed characterisation data results in equifinality, whereby distinct sets of parameters fit the observed state variable measurements equally well (Arhonditsis et al., 2008). The equifinality of model solutions can lead to instances whereby the model simulates the state variables of interest adequately, however it incorrectly resolves the relevant higher-level processes and system-scale dynamics (Arhonditsis et al., 2007). In addition to the equifinality of distinct parameter sets, several possible model structures might be acceptable simulators of the natural system (Janse et al., 2010). The complexity and formulation of process descriptions varies between models, and between in-model configuration options, which results in structural uncertainty (Refsgaard et al., 2007). These structural variations, whilst not being observable at the state variable level, may give rise to different process behaviours and system-scale dynamics (Anderson et al., 2010). The prevalence of equifinality in model solutions, raises the question: how can we better calibrate and constrain our water quality models?

To incentivise the implementation of all components of the modelling procedure, it is crucial to establish a common framework for improved calibration, validation, and uncertainty analysis. Despite the advancement in the process descriptions of AEMs, the level of predictability they provide has not significantly improved since the 1990s (Arhonditsis & Brett, 2004; Soares & Calijuri, 2021), among others. Like in all fields of environmental modelling, AEMs are simplifications of very complex real-world systems and comprise a significant number of uncertain model parameters whose true values are unknown and must therefore be estimated.

Additional sources of uncertainties in process-based models are introduced through model structural assumptions, and in the assignment of initial and boundary conditions (Beck, 1987), among others. In response to these issues, several researchers have proposed that the modelling procedure should include calibration, validation, and sensitivity/uncertainty analysis (e.g., Jørgensen, 1995; Refsgaard et al., 2007). However, according to a recent review on the state of process-based lentic aquatic systems modelling, the above-mentioned components of the modelling procedure were routinely neglected and were only applied in 67% (calibration), 53% (validation), and 34% (sensitivity/uncertainty analysis) of studies published between 2015 and 2020 (Soares & Calijuri, 2021).

A new framework for the evaluation of aquatic ecosystem models - the Concept/State/Process/System framework (CSPS; Hipsey et al., 2020) - proposes a system-inspired approach for model evaluation, as a way to extend the traditional model-data comparison method. The CSPS framework consists of four different validation levels (numbered 0-3) and suggests a suite of advanced metrics and system 'signatures' that can be adopted to assist in assessing the performance and suitability of an AEM simulation (Hipsey et al., 2020). In addition to the traditional 'state validation', the framework encourages targeted evaluation of process behaviour and system-scale dynamics that can give a more complete picture of the model's performance and whether it is fit-for-purpose. In recent case studies, the framework has been applied to validate ecosystem models of Lake Kinneret (Israel) and the Great Barrier Reef (Australia), enabling an assessment of each model's strengths and hidden deficiencies, highlighting the benefits of this systematic approach (Reger et al., 2023; Robson et al., 2020).

While the CSPS framework proposes a systematic approach to model evaluation, there is currently no widely applicable framework for the calibration of complex AEMs (e.g., Frassl et al., 2019; Janssen et al., 2015). To date, calibration is generally based on direct data comparison of observed vs. modelled state variables using simple quantitative techniques such as the root-mean-square-error (RMSE; Soares & Calijuri, 2021). Consequently, the success of calibration is dependent on noisy observations of the primary state variables, often limited in quantity, to adequately constrain the model inputs (Bennett et al., 2013). System-inspired metrics are applicable indicators of the broader behaviour of the ecosystem, including relevant dimensionless numbers, stoichiometric indicators, and a variety of relationships between variables of interest (Hipsey et al., 2020)–quantities that are important for maintaining plausible

simulated ecosystem behaviour during calibration. Incorporating system-inspired metrics in the calibration process, in addition to state variables, may compensate for low-quality datasets and information deficits. This can guide model calibration efforts to best capture ecosystem-scale dynamics. However, this has not been rigorously assessed relative to a non-system-inspired calibration approach to date.

Due to the high level of uncertainty of AEMs, incorporating uncertainty analysis in the modelling procedure is of increasing interest to provide critical estimates of reliability for the model outcomes. Uncertainty analysis is concerned with establishing bounds around point predictions to describe the degree of confidence we have in the model results. Herein, uncertainty analysis is performed by running the model multiple times with different inputs and configurations, referred to as single-model-ensembles (SMEs; Gal et al., 2014). SMEs can exploit the sensitivity of the model in question to different parameter sets, boundary conditions, initial values, and configuration options, and assess how these uncertainties propagate in the model output (Janssen et al., 2015). Bayesian-based calibration has been increasingly used as it allows direct assessment of parameter uncertainty (Janse et al., 2010). Instead of estimating one optimal parameter set, this approach seeks to determine the posterior probability distribution of model parameters, which convey the likelihood of certain parameter values (Arhonditsis et al., 2007), though this approach has yet to receive broad uptake within the AEM community (Soares & Calijuri, 2021).

The aim of this study was to answer two research questions: 1) 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 2) 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 vs. modelled state variables. The second approach augments the more traditional naïve calibration with additional metrics, a new approach that explicitly includes a wide range of supplementary system metrics (e.g., thermocline depth, metalimnetic oxygen minima) to help maintain the coherence of the posterior parameter ensemble. Additionally, we explore different objective function formulations in an effort to understand the interaction between matching historic measurements of system

state with these new system-inspired metrics. 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 reduce the equifinality of model solutions and provide a holistic approach for the calibration of complex aquatic ecosystem models. The new system-inspired calibration convention is scalable to a diversity of lentic systems, and is anticipated to aid model structural decisions and improve confidence in model predictions of complex AEMs.

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 et al., 2022a). 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 et al., 2022a). 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 et al. (2022a), respectively. Due to the extensive monitoring of the physics, chemistry, and biology of the site in the last decade, sufficient empirical data for FCR were available for calibration.

2.2 Modelling framework and methodology

2.2.1 Modelling 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 et al. (2022a). We further improved the simulation by coupling the model with an independent Parameter ESTimation (PEST; Doherty, 2018) software package to optimise the model performance and compare two different calibration approaches: naïve and system-inspired calibration. We then tested the impact of different weighting strategies on the modelling 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.

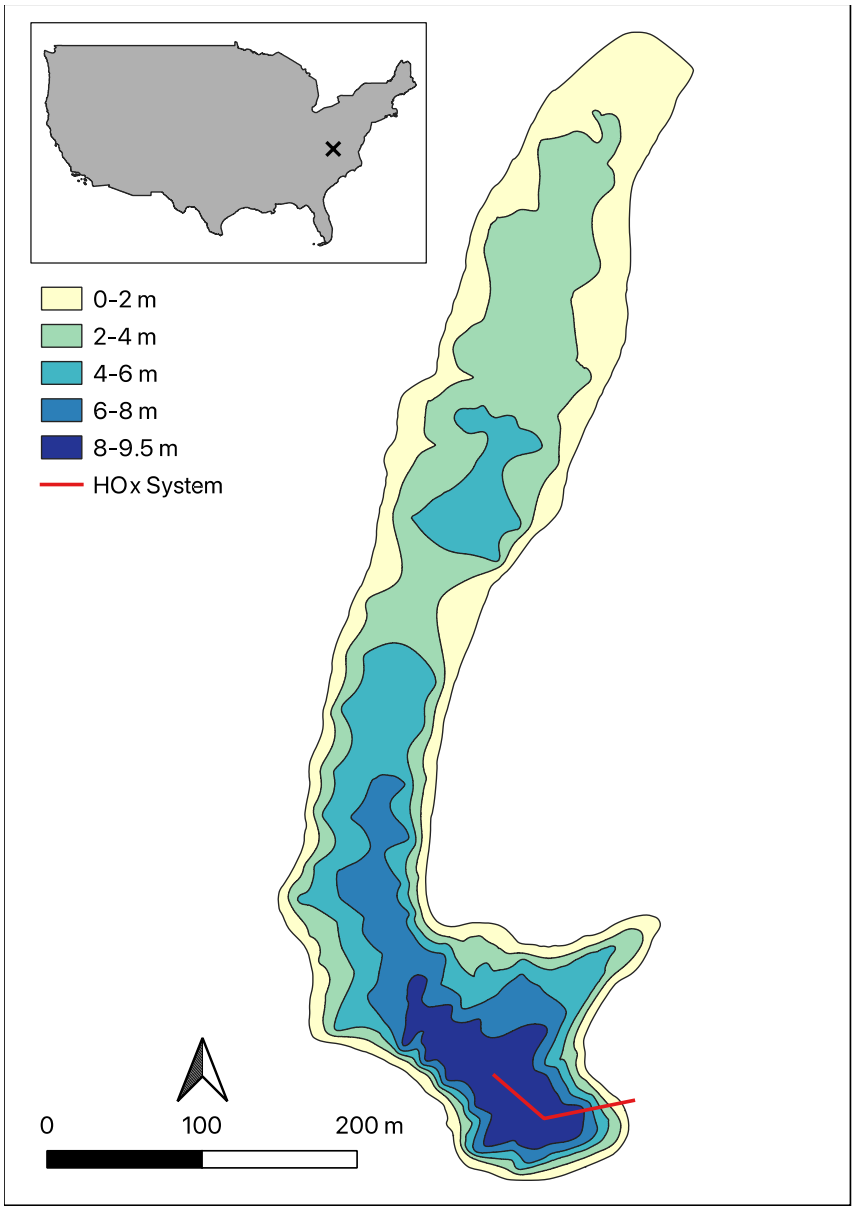

Figure 1. Map of the Falling Creek Reservoir, Vinton, Virginia, USA: Latitude: 37.30°, Longitude: -79.84°. The coloured bands indicate the bathymetry contours of the reservoir, and the red line represents the location of the hypolimnetic oxygenation (HOx) system.

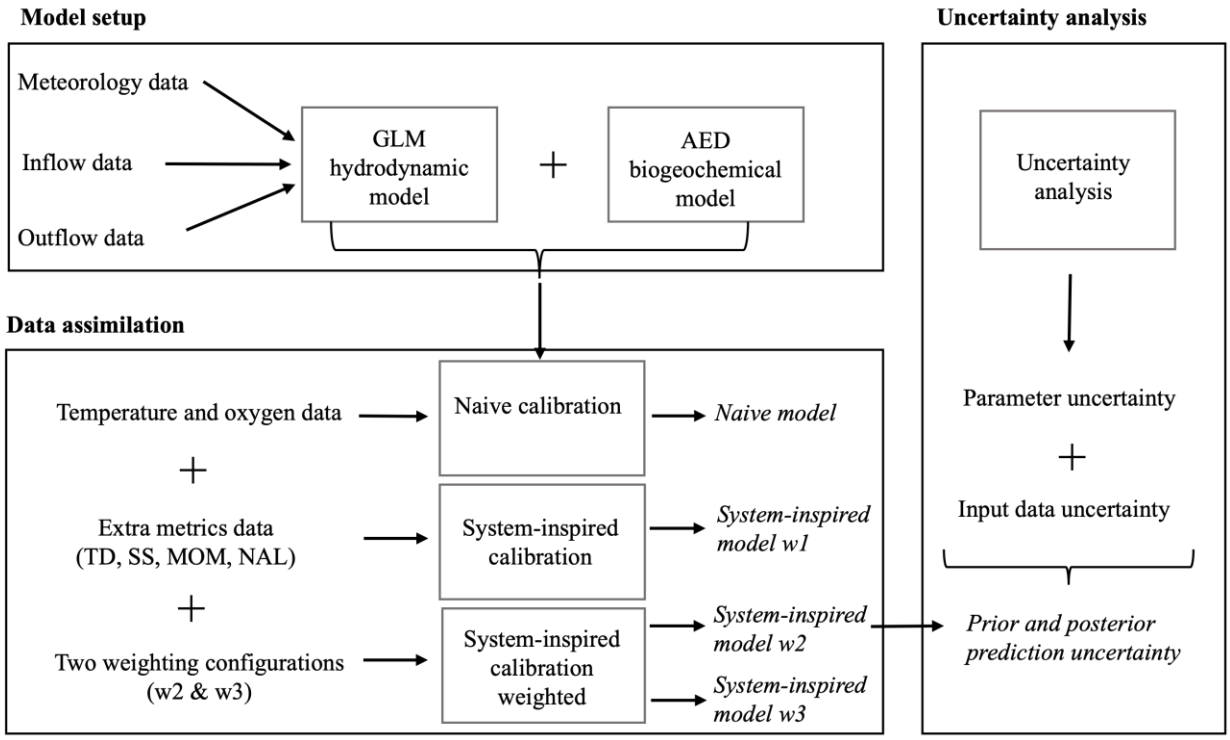

Figure 2. The modelling framework including the model setup, calibration, and uncertainty analysis. The system-inspired calibration includes additional data of following extra metrics: thermocline depth (TD), Schmidt stability (SS), metalimnetic oxygen minima (MOM), the number of anoxic layers (NAL).

2.2.2 Model description

We used the General Lake Model dynamically coupled to the Aquatic EcoDynamics 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 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 et al., 2022a). The in-depth description of GLM can be found in Hipsey et al. (2019).

The AED modelling library is an open-source project aimed at simulating aquatic ecosystem dynamics (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., 2023, for the full model configuration and parameters).

The GLM-AED model setup for FCR by Carey et al. (2022a) 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 et al., 2022c). In our configuration, the number of sediment zones was increased to four to better capture the depth-specific sediment heating and biogeochemistry. 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 2015-07-12 to 2019-12-31 at an hourly timestep. The total simulation period was divided into calibration from 2016-12-01 to 2019-12-31 and validation from 2015-07-12 to 2016-12-01.

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 et al., 2022c). The meteorological dataset consisted of air temperature, relative humidity, shortwave and longwave radiation, wind speed, and precipitation data from NASA's North American Land Data Assimilation System (Xia et al., 2012) from 2013-2021. The inflow data for the primary tributary consisted of daily discharge, water temperature and chemistry observations from 2013-

2021. The HOx system inflow included daily flow, elevation (the depth at which the oxygenated flow is injected in the reservoir), water temperature, and chemistry observations from 2013-2021. The daily outflow discharge was estimated to amount to the daily inflow discharge, 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 2013-2021 at the reservoir's deepest site and were retrieved from the Environmental Data Initiative Repository (Carey et al., 2022b). 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 et al., 2022d). 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, visualisation 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 minimise 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 organised 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 vs. 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 metre 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 follows:

$$\Phi_N = \sum_i (w_T r_{T_i})^2 + \sum_i (w_O r_{O_i})^2 \quad (1)$$

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.

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:

$$\Phi_S = \sum_i (w_T r_{T_i})^2 + \sum_i (w_O r_{O_i})^2 + \sum_i (w_{TD} r_{TD_i})^2 + \sum_i (w_{SS} r_{SS_i})^2 + \sum_i (w_{MOM} r_{MOM_i})^2 + \sum_i (w_{NAL} r_{NAL_i})^2 \quad (2)$$

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 towards the hypolimnion. The MOM was calculated on each data collection day based on equations (3) and (4).

$$MOM = O_{2\,(metalimnion)} - O_{2\,(expected)} \tag{3}$$

where:

$$O_{2\,(expected)} = \frac{O_{2\,(eplimnion)} + O_{2\,(hypolimnion)}}{2} \tag{4}$$

The spatial and temporal extent 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 dataset 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 thermocline 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).

Table 1. Different weighting schemes for incorporating system metrics in the objective function. The system metrics include the thermocline depth (TD), Schmidt stability (SS), metalimnetic oxygen minima (MOM), and the number of anoxic layers (NAL).

	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

2.5 Uncertainty analysis

Uncertainty analysis was carried out on the best performing model (Model w2 with DM 2). This analysis was used to explore equifinal solutions by seeking an ensemble of parameter realisations that all acceptably reproduce both state measurements and the additional metrics. 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 and Oliver, 2013). The resulting ensemble can hence be considered to include samples of 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 realisations. The prior parameter realisations 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 equation (5) and the corresponding values have been listed in Table S1.

$$\sigma = \frac{\log_{10}(par_{max}) - \log_{10}(par_{min})}{4} \quad (5)$$

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

3 Results

3.1 Naïve calibration

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 modelled temperature profiles depicted the patterns and characteristics of the observed data reasonably well (Table 2). Modelled 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 modelled temperatures occurred in the metalimnion. In the summer of 2017, the modelled metalimnetic temperatures were predicted to be 2-3 degrees colder than the observed temperatures. However, in the summers of both 2018 and 2019, the metalimnetic temperatures were predicted to be approximately 2 degrees warmer than the observations (Figure 3e). The modelled 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 modelled hypolimnetic oxygen concentrations were higher than observations during the summer period, when the HOx system was in operation (Figure 3f).

Relative to the temperature and DO state variables, the naïve model was less able to adequately recreate ecosystem-level behaviour, as represented by the system-inspired metrics. In 2017, the thermocline depth (TD) was underestimated by the model and did not follow the trend in the observed data well (Figure 4a). In 2018 and 2019, the model performed better and depicted the pattern in the observed TD data adequately. Interestingly, in the two years when a warm bias

was detected in the modelled temperature profiles, the TD was portrayed better than in the cold bias year of 2017. The same plot also illustrated the modelled and observed ice cover in the winter period. In general, the model simulated the presence of ice cover well, however, there were a few cases when it falsely predicted ice cover, predominantly in January 2018 (Figure 4a). Trends in the Schmidt stability were captured well, which indicated that the model was capable of reproducing the stratification strength of the reservoir (Figure 4b). The modelled sediment temperature in zone 2, which encompassed depths from ~4 m to ~6.5 m, and the modelled and observed water temperature at 5 m depth, are presented in Figure 4c. We chose this depth for analysis because it exhibited the greatest deviations between observed and modelled water temperatures, and was of interest to investigate the related sediment temperatures. While there was no available observed sediment temperature data for the reservoir, sediment temperatures are assumed to approximately follow the temperature of the water they are in contact with, which makes water temperature data a good alternative for comparison. Using the deepest water temperature data available for this comparison, modelled zone 2 sediment temperatures were in the range of observations and followed water temperature patterns adequately. In 2018, the spatial and temporal extent of anoxia was reproduced well by the model (Figure 4d). However, the extent of anoxia was underestimated by the model in 2017 and 2019. The best agreement between modelled and observed MOM occurred in 2017, while the deviation was underestimated in the next year, and a better agreement compared to 2018 was observed in the last year (Figure 4e).

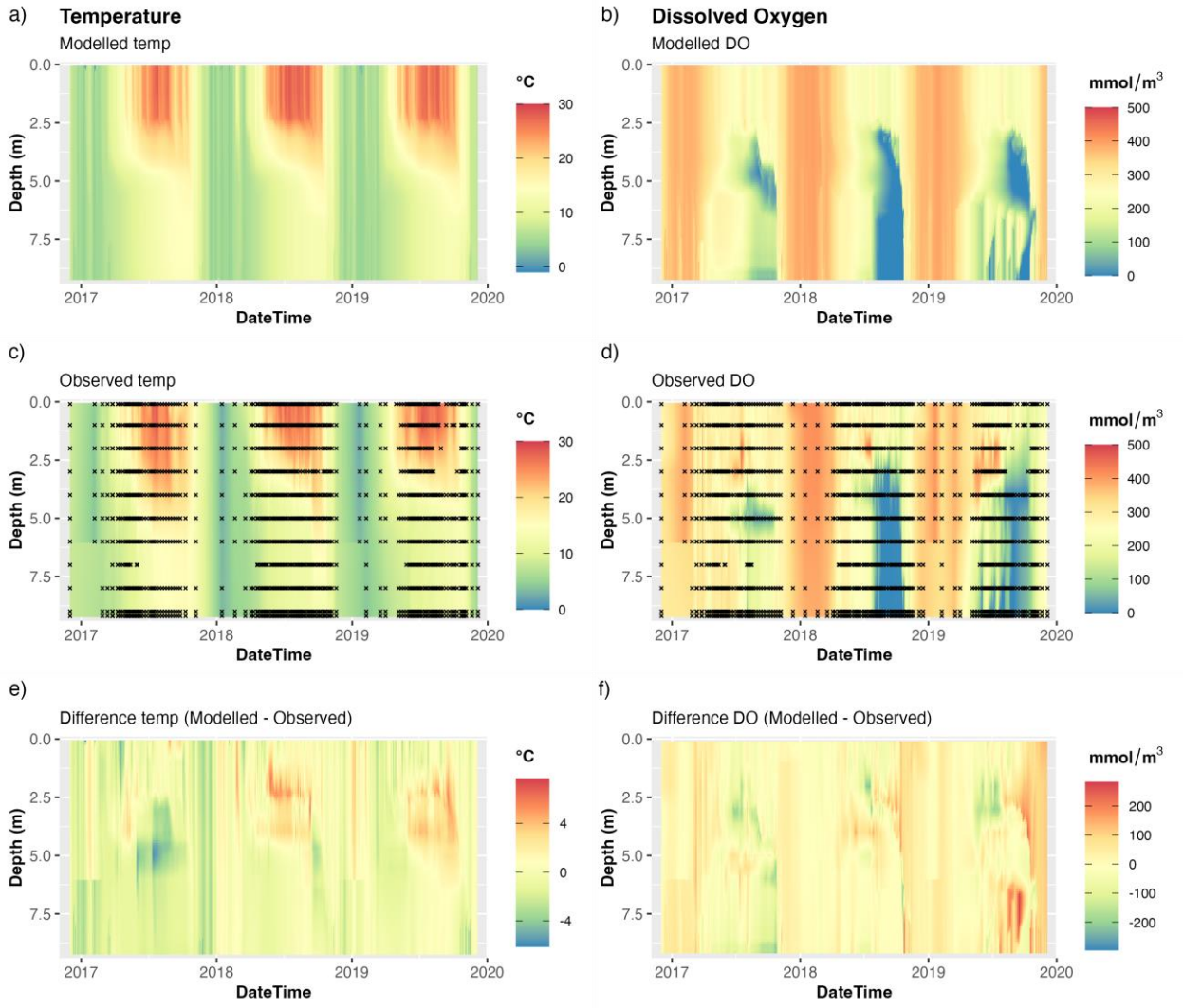

Figure 3. Contour plots of modelled (a, b), observed (c, d), and the difference of modelled and observed temperature and dissolved oxygen profiles (e, f) based on the naive calibration model with DM 2. The black crosses on plots c and d represent the time and location of the temperature and dissolved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

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

	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

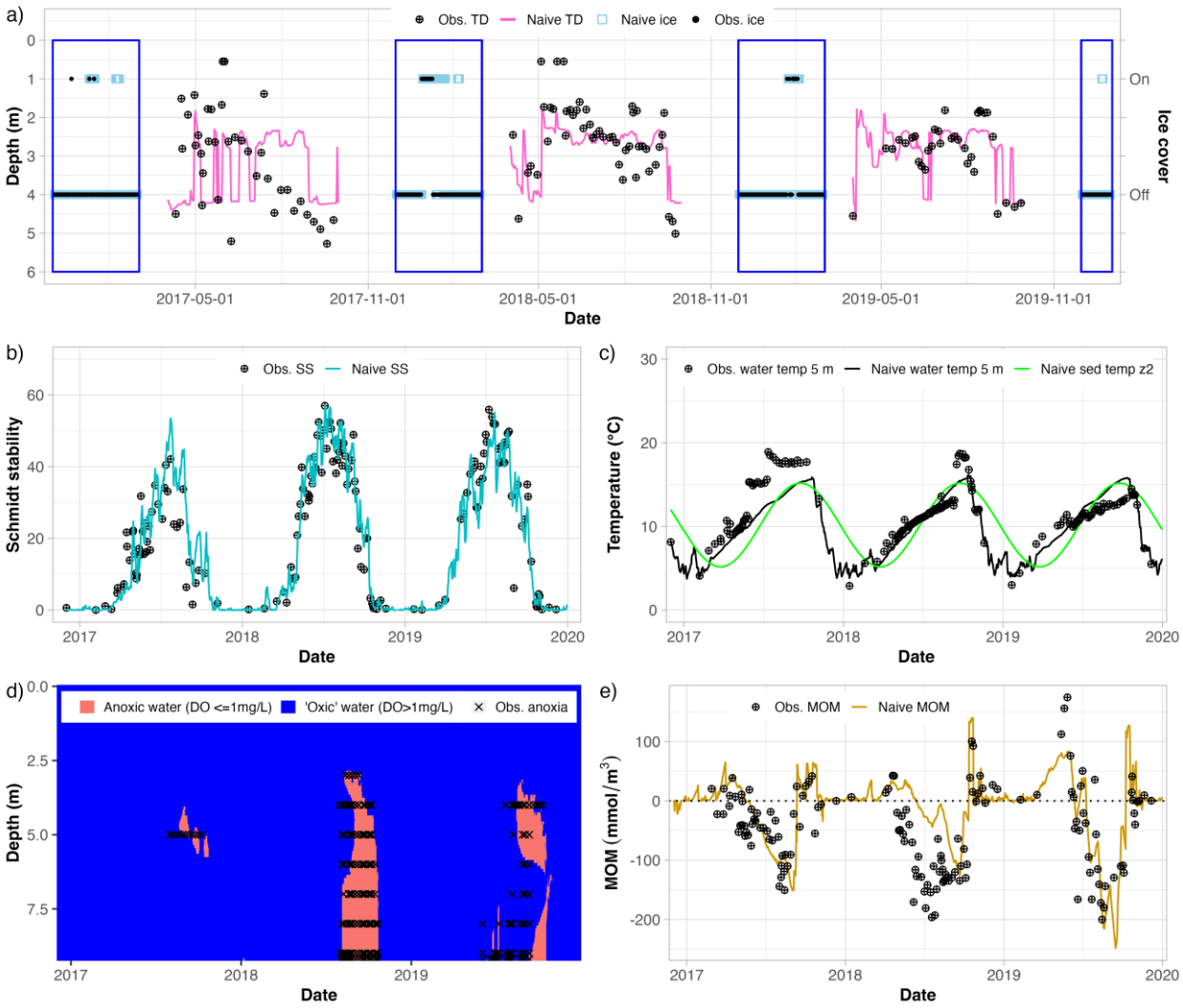

Figure 4. Comparison of observed (Obs.) system-metrics and system-metrics predicted by the naive calibration model with DM 2 (Naïve). The metrics include thermocline depth (TD) during the stratified period, ice cover presence and absence in the winter period (a), Schmidt stability (b), sediment temperature in zone 2 (visualised along with modelled and observed water temperatures at 5 m depth in zone 2) (c), spatial and temporal extent of anoxia (d), and metalimnetic oxygen minimum (MOM, e).

3.2 System-inspired calibration

The augmentation of the objective function with system-inspired metrics generally led to more realistic simulation results as viewed from a system performance perspective. There were

no significant differences in the simulation of temperature between the models calibrated with the system metrics (Table 2). The greatest deviations from the observed temperature data occurred in the metalimnion, while hypolimnetic temperatures were slightly underestimated. It appears that the temperature predictions were not 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 naïve calibration approach demonstrated a slightly better capability for simulating the DO profile than the system-inspired approach (Table 2). The calibration results were, to a degree, sensitive to the weighting configuration of the extra metrics observation groups (Table 2). It seems that the greatest differences occurred in the simulation of the TD and MOM between the models (Figure 6). Overall, Model w2 seemed to outcompete the other models in most aspects, however the differences were not significant. 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.

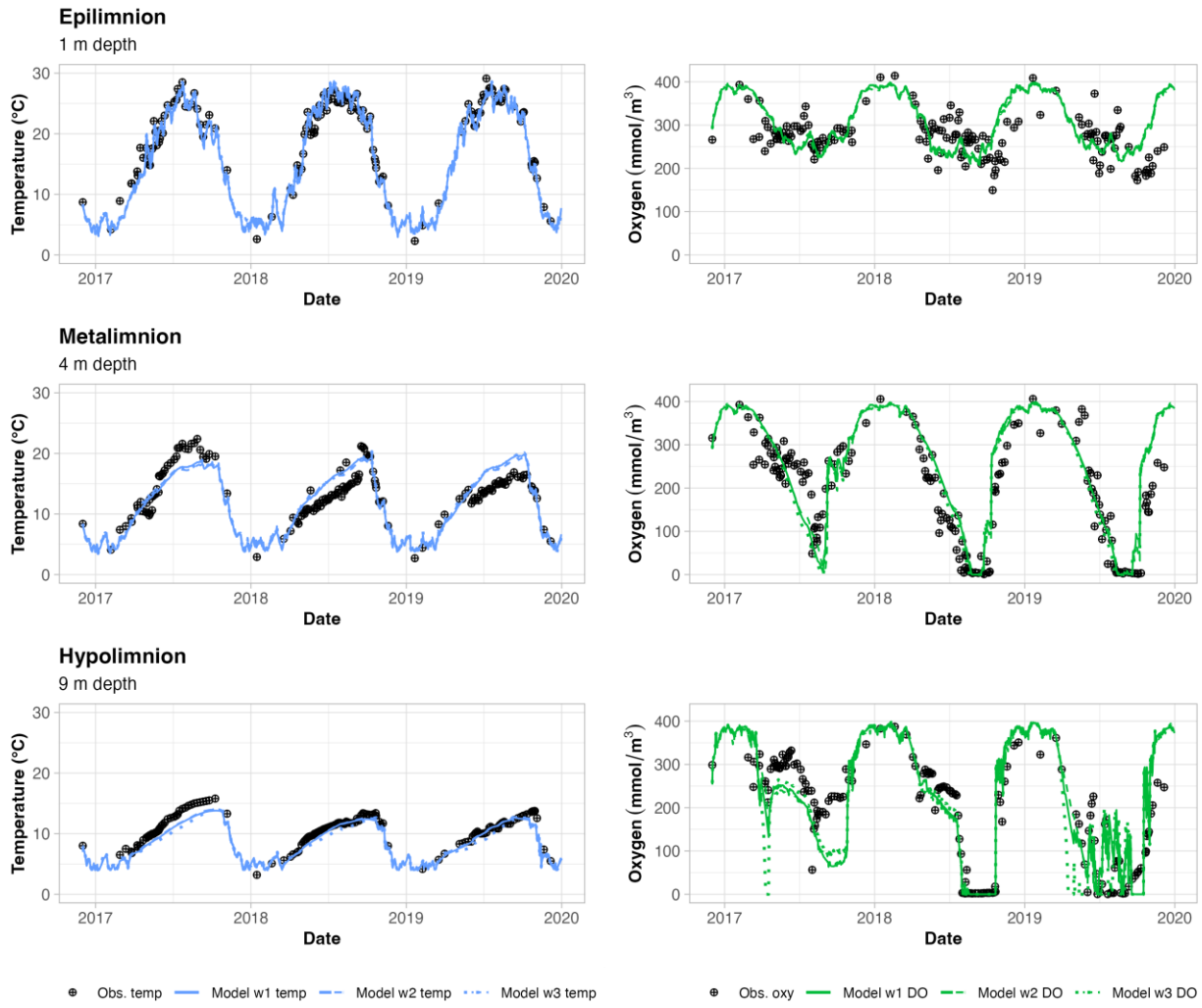

Figure 5. Water temperature (temp) and dissolved oxygen (DO) concentration in the epilimnion, metalimnion, and hypolimnion simulated by the three models with different weighting schemes based on the system-inspired approach with DM 2 (Model w1, Model w2, Model w3).

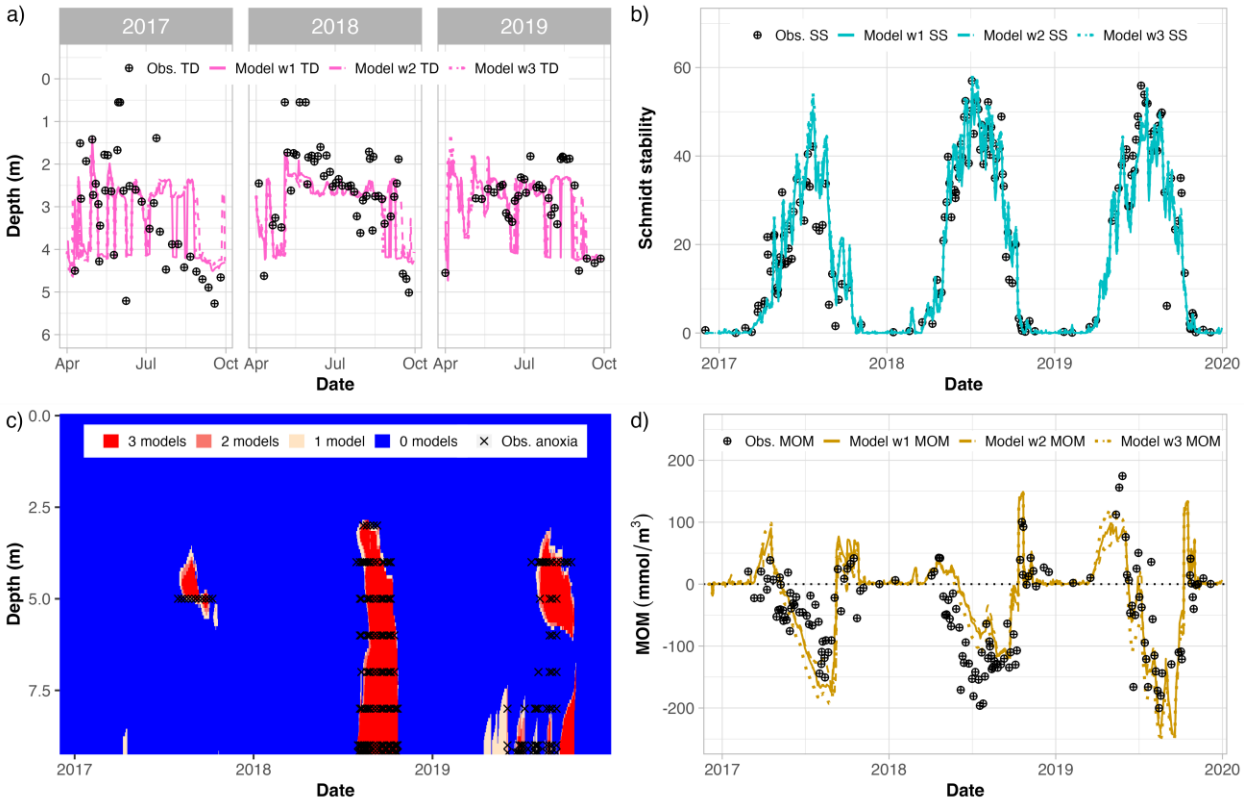

Figure 6. Comparison of the observed system-metrics (Obs.) and the system-metrics predicted by the three models with different weighting schemes based on the system-inspired approach with DM 2 (Model w1, Model w2, Model w3). The metrics include thermocline depth during the stratified period (a), Schmidt stability (b), the spatial and temporal extent 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.

3.3 Comparison of calibration approaches and configurations

Compared to the reference model (Carey et al., 2022a), 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 MOM and the spatial and temporal extent of anoxia quantified by the number of anoxic layers per day. Interestingly, 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

458 improvement in the model's ability to capture the behaviour of these extra metrics, which led to
459 increased coherence between the system-scale dynamics of the simulated and natural ecosystem.
460 However, there was a slight trade-off in accuracy between the simulation of extra metrics and the
461 DO profile, while the accuracy of the temperature profile remained the same (Table 2). The loss
462 in the MEF of the DO profile was less than 0.1 for all weighting strategies, while the gain in the
463 MEF of the extra metrics was greater than 0.1 in the majority of cases. The choice of the deep
464 mixing model structure had a significant effect on model performance. While hypolimnetic
465 mixing with constant diffusivity was more suitable for the simulation of the TD and the MOM,
466 the Weinstock model of diffusivity was able to better capture the spatial and temporal extent of
467 anoxia in the reservoir during the calibration period. However, during the validation period, the
468 models based on the Weinstock model of diffusivity also demonstrated superior performance in
469 capturing the TD and MOM, in addition to the extent of anoxia. In general, model performance
470 was better in the calibration period except for simulating the MOM, which was better captured
471 during the validation period.

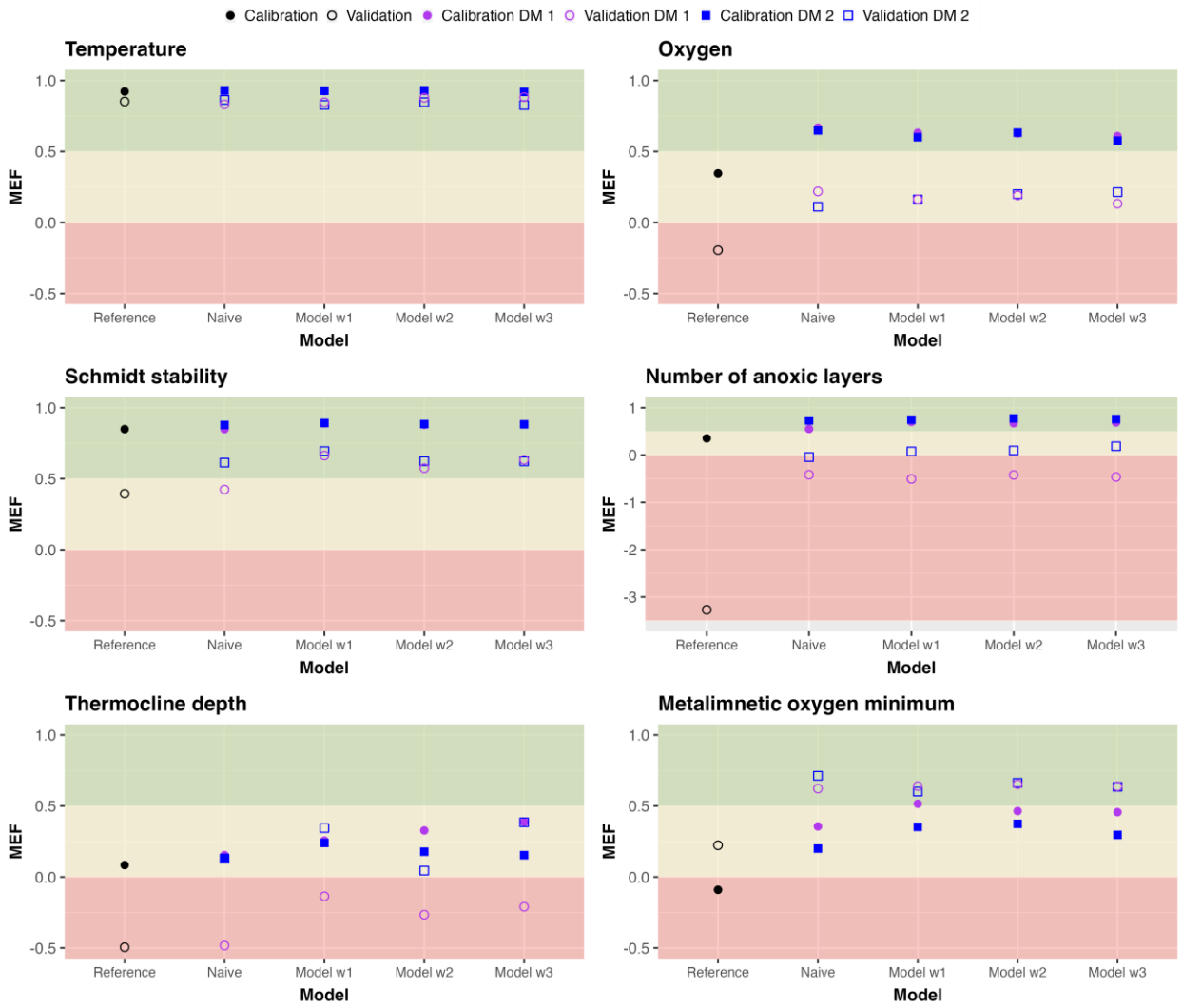

Figure 7. Comparison of model efficiency (MEF) between models with two different deep mixing configurations during the calibration and validation periods. One deep mixing configuration is based on constant diffusivity (DM 1) and the other configuration employs the Weinstock model to determine the diffusivity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

3.4 Uncertainty analysis

The uncertainty in predicting two system-inspired metrics of interest, the TD and the MOM, was significantly reduced post-calibration compared to pre-calibration (Figure 8). The

propagation of parameter uncertainty prior to calibration was greater in predicting the MOM, which also exhibited a more substantial reduction in uncertainty post calibration than the TD. However, the resulting narrow fan of the posterior distributions suggests high confidence in the prediction of both metrics. As expected, for both the TD and the MOM, the range of predictions based on the posterior distribution followed a similar pattern as the calibrated models illustrated in Figure 6a and Figure 6d. The likely range of thermocline depth outputs based on the prior distribution did not fully encompass all observation points (Figure 8a). This suggests that the maximum expected parameter uncertainty (i.e., the prior) doesn't include a wide enough range of model outputs to capture all observation points, which could be a manifestation of model structural error.

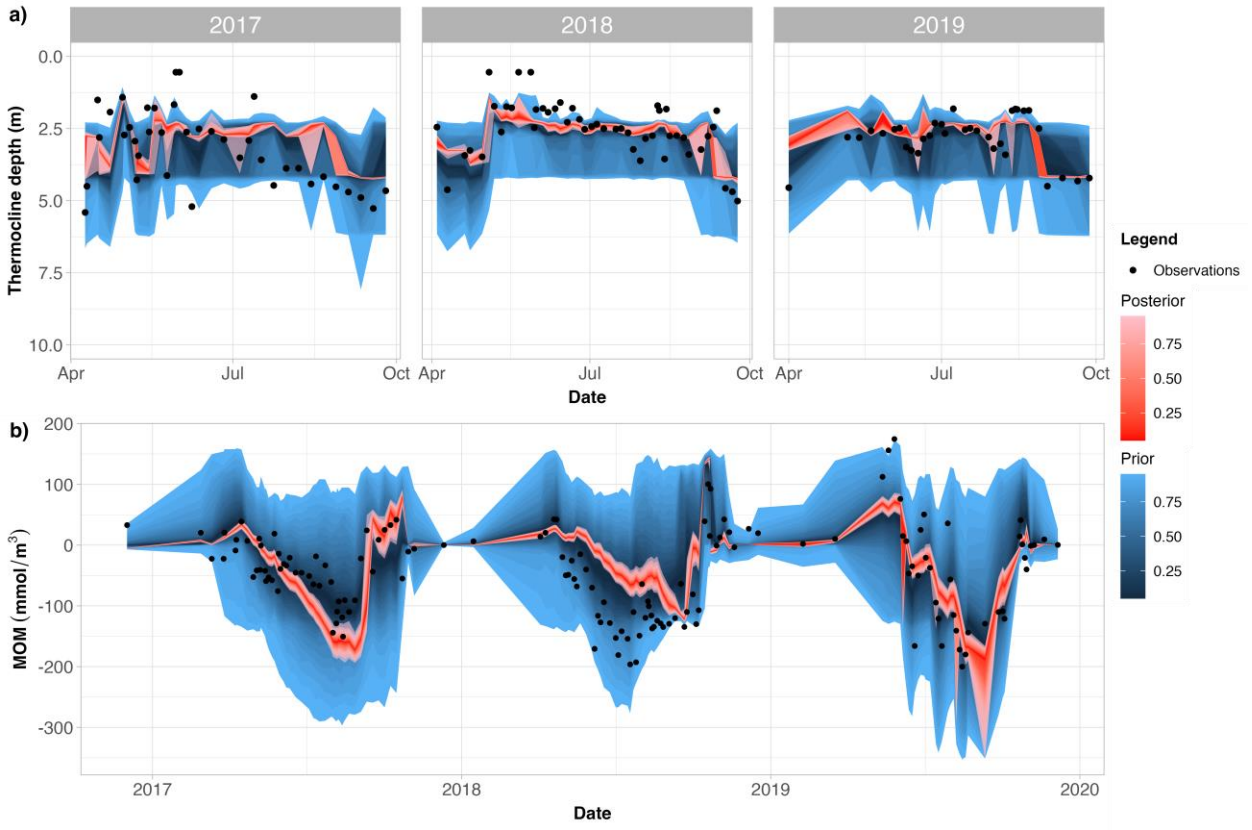

Figure 8. Prior and posterior probability distributions of the thermocline depth (a) and the metalimnetic oxygen minima (b) predictions compared to observations.

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 characterising uncertainty. By incorporating system-inspired metrics in the objective function, a highly targeted model calibration was achieved, leading to greater understanding of the ecosystem.

This new calibration approach augmented the objective function with non-traditional system-inspired metrics that amplify aspects of the state observations that represent theoretically important characteristics of overall ecosystem behaviour. Specifically, we chose thermocline depth (TD) and Schmidt stability as system-inspired metrics quantitatively summarizing the thermal structure of the reservoir because they integrate whole-ecosystem hydrodynamics (Wilhelm & Adrian, 2008). Since the deployment of the HOx system in FCR, metalimnetic oxygen minimum zones frequently arise during the stratified period (McClure et al., 2018). To quantify this particular property of the system, we developed a MOM metric to quantify how much the actual DO concentration in the metalimnion deviates from the expected concentration (equations (3) and (4)). Additionally, a metric linked to the model's ability to capture the extent of anoxia in the water column was defined as the number of anoxic layers per day (NAL). We were unable to use other metrics such as the Anoxic Factor (AF), which has been proposed to describe the spatial and temporal dimensions of anoxia per season (Nürnberg, 1995), because it is more suitable for quantifying long-term changes in anoxia in the hypolimnion (e.g., Ladwig et al., 2021), not short-term dynamics. Whilst ice cover was used for the post-calibration evaluation of the naïve model, its explicit implementation in the PEST objective function was abandoned due to the binary nature of ice cover data. Finally, although the parameters of the sediment model were included in the adjustable parameter vector, the calibration of zone-specific sediment temperature was not feasible due to a lack of observed sediment temperature data.

This case study presented four system-inspired metrics relevant to FCR, however, it is important to note that these metrics are not the only options available. When selecting metrics, researchers should consider their study site and choose advanced metrics that are most suitable to their specific research question. The two case studies that previously applied the CSPS framework for model validation provide numerous examples of system-metrics and characteristic

signatures both for marine (Robson et al., 2020) and freshwater applications (Regev et al., 2023). These additional metrics include N fixation, community respiration, phytoplankton community structure, along with others.

The implementation of extra metrics focused calibration on the most relevant aspects of the simulation. As expected, explicit inclusion of the metrics in the objective function resulted in a clear improvement in the simulation of the ecosystem behaviours that the metrics were designed to represent. However, the simulated dissolved oxygen profile suffered a slight loss of accuracy, while the temperature profile remained the same. This trade-off is likely a result of achieving a greater overall objective function reduction by minimising the error between system metrics model predictions and observations rather than the oxygen depth profile. Ultimately, this trade-off was deemed acceptable, as it enabled refocusing the calibration efforts on specific characteristics of the system, resulting in a more targeted approach. However, the overall model prediction accuracy has not improved significantly from the naïve approach to the system-inspired approach. One contributing factor is that the observed dataset of state-variable measurements exhibited great spatial and temporal resolution. Hence, adding extra metrics to the calibration process may improve model predictions to a greater extent where there is a lack of observed data, more uncertainty, and in this case 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 (hydrodynamic - abiotic - biotic) (Soares & Calijuri, 2021), the key elements presented in this paper have the potential to improve simulations of nutrient cycling, greenhouse gas emissions, and other higher-level processes.

Assigning weights to extra metrics added a subjective element to the calibration process, and overweighting these metrics resulted in degrading model performance. In this study, the sensitivity of the calibration results to different weighting schemes was explored (Figure 9). Weighting scheme 1 and 2 exhibited negligible differences, while the application of weighting scheme 3 led to diminishing model performance with respect to the prediction of both state-variables and system metrics. It was expected that greater weight added to the extra metrics observation groups may lead to a loss in state variable accuracy, however interestingly, it also led to poorer prediction of the extra metrics. Consequently, achieving an optimal balance in weighting is crucial, as giving unproportionate high weights to selected observation groups could result in overfitting that degrades the overall model performance.

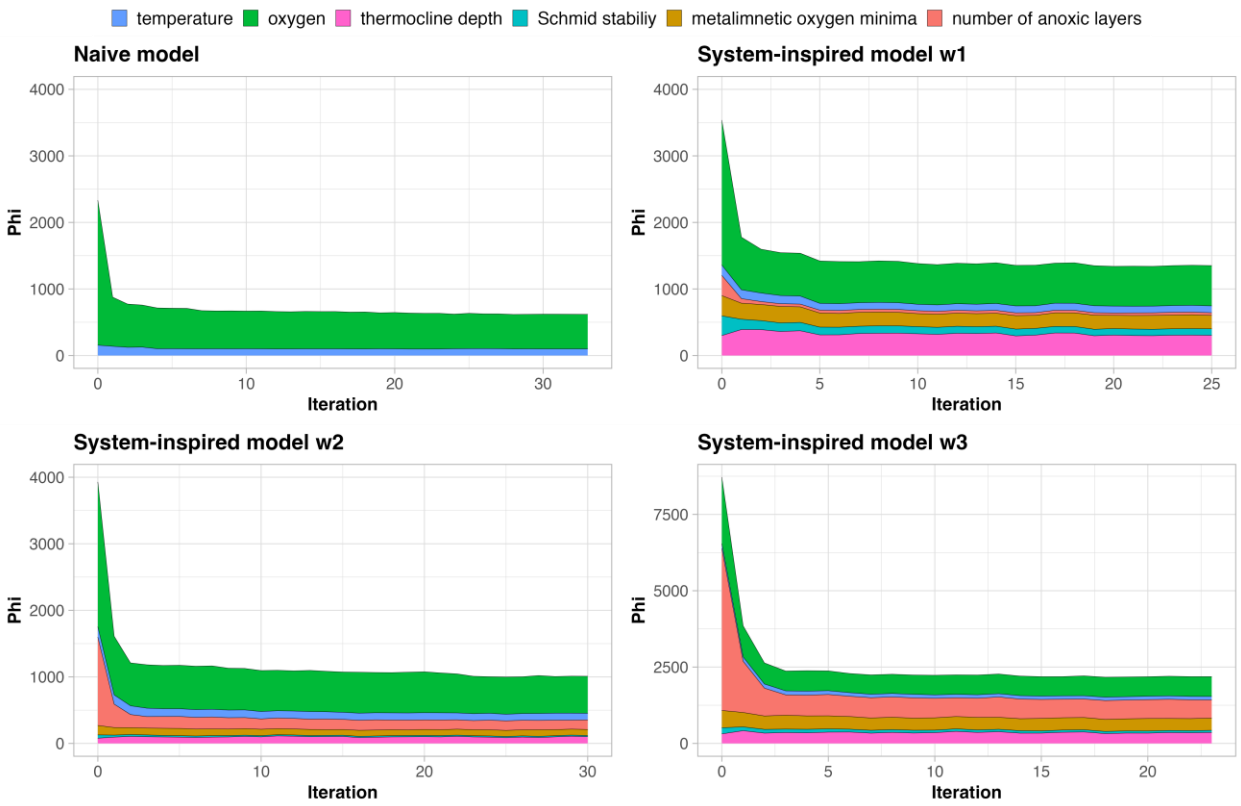

Figure 9. Convergence of the objective function in the case of the naïve model and the system-inspired models with different extra metrics weighting schemes during the calibration process. The deep mixing configuration of the models illustrated here is based on is based on the Weinstock model of diffusivity (DM 2).

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 possible 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 dynamics helps to shed light on previously hidden strengths and weaknesses (Hipsey et al., 2020). Comparing the two deep mixing 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 metalimnion well during the calibration period. The

Weinstock model was more effective in depicting the spatial and temporal extent of anoxia. As this study forms the basis for simulating methane in FCR, the Weinstock model configuration was preferable. The Weinstock model's strength was the simulation of anoxic conditions, a prerequisite of methane production (Borrel et al., 2011). Consequently, extra metrics can assist in aligning model structural decisions with current and future modelling endeavour and can serve as a valuable tool in model development.

System-inspired metrics provide valuable insights into prediction uncertainty. Both the TD and MOM prediction uncertainty was significantly reduced after calibration. This substantial reduction in parameter uncertainty is due to a well-determined inverse problem consisting of a great number of observations and a relatively small number of adjustable parameters. The parameters were highly identifiable from observations, which led to narrow posterior parameter probability distributions of the metrics. While prediction uncertainty included parameter uncertainty and measurement noise, model structure uncertainty was not accounted for. Model structure uncertainty is a significant source of prediction uncertainty; however, it is challenging to quantify and is often neglected (Refsgaard et al., 2006). The variation in the model output induced only by altering the deep mixing configuration indicates that model structure uncertainty could be a significant source of overall prediction uncertainty.

The implementation of extra metrics in the calibration process assists in evaluating when a model is 'successfully' calibrated. While calibration is established as one of the essential steps of the modelling procedure (Refsgaard et al., 2007), 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. However, identifying the point of diminishing returns in model fit is a challenging task. Using system-inspired metrics for evaluating the performance of a model in the calibration process has much potential for providing reassurance that the study site is captured well on the system-level, and help ensure the model is fit for purpose.

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 modelling 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 optimising reservoir management. For instance, quantifying the uncertainty in simulating the MOM, can facilitate the operation of the local reservoir oxygenation system. The list of system-inspired metrics applied in this study is to be extended over time for a number of applications. Altogether, developing system-metrics to assist in the calibration of nutrient cycling and greenhouse gas emission simulations has the potential to significantly improve the predictive accuracy of complex AEMs.

6 Acknowledgements

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.

7 Open research

All model files, R scripts and model executable files are available in the Zenodo repository FCR-GLM-metrics (Kurucz et al., 2023). 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; Carey et al., 2022b, 2022d).

8 Authorship contribution statement

KK conceptualised the problem with CCC and MRH. KK led the modelling procedure comprising calibration, validation, and uncertainty analysis. CCC, PH, and MRH provided guidance for model calibration and validation. JTW and EDS contributed to the uncertainty analysis component. KK and CCC carried out data curation and KK visualised and analysed the results, with MRH supervising. KK wrote the original draft of the manuscript and all authors reviewed and edited it.

642 **References**

643 Albers, S., Winslow, L., Collinge, D., Read, J. S., Leach, T., Zwart, J., & Snortheim., C.
644 (2018). rLakeAnalyzer: Lake physics tool (Version 1.8.3.) [Software]. Zenodo.

645 Anderson, T. R., Gentleman, W. C., & Sinha, B. (2010). Influence of grazing formulations on
646 the emergent properties of a complex ecosystem model in a global ocean general circulation
647 model. *Progress in Oceanography*, *87*(1–4), 201–213.
648 https://doi.org/10.1016/j.pocean.2010.06.003

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

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

655 Arhonditsis, G. B., Perhar, G., Zhang, W., Massos, E., Shi, M., & Das, A. (2008). Addressing
656 equifinality and uncertainty in eutrophication models. *Water Resources Research*, *44*(1).
657 https://doi.org/10.1029/2007wr005862

658 Arhonditsis, G. B., Qian, S. S., Stow, C. A., Lamon, E. C., & Reckhow, K. H. (2007).
659 Eutrophication risk assessment using Bayesian calibration of process-based models: Application
660 to a mesotrophic lake. *Ecological Modelling*, *208*(2–4), 215–229.
661 https://doi.org/10.1016/j.ecolmodel.2007.05.020

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

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

667 Beven, K. (2006). A manifesto for the equifinality thesis, *Journal of Hydrology*, *320*(1–2),
668 18–36. https://doi.org/10.1016/j.jhydrol.2005.07.007

669 Borrel, G., Jézéquel, D., Biderre-Petit, C., Morel-Desrosiers, N., Morel, J.-P., Peyret, P., et al.
670 (2011). Production and consumption of methane in freshwater lake ecosystems. *Research in*
671 *Microbiology*, *162*(9), 832–847. https://doi.org/10.1016/j.resmic.2011.06.004

672 Bruce, L. C., Frassl, M. A., Arhonditsis, G. B., Gal, G., Hamilton, D. P., Hanson, P. C., et al.
673 (2018). A multi-lake comparative analysis of the General Lake Model (GLM): Stress-testing
674 across a global observatory network. *Environmental Modelling & Software*, *102*, 274–291.
675 https://doi.org/10.1016/j.envsoft.2017.11.016

676 Carey, C.C., & Breef-Pilz, A. (2022). Ice cover data for Falling Creek Reservoir and
677 Beaverdam Reservoir, Vinton, Virginia, USA for 2013-2022 (Version 4) [Dataset].
678 Environmental Data Initiative.
679 https://doi.org/10.6073/pasta/917b3947d91470eecf979e9297ed4d2e

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

683 Carey, C. C., Lewis, A. S., McClure, R. P., Gerling, A. B., Breef-Pilz, A., & Das, A. (2022b).
684 Time series of high-frequency profiles of depth, temperature, dissolved oxygen, conductivity,
685 specific conductance, chlorophyll a, turbidity, pH, oxidation-reduction potential, photosynthetic
686 active radiation, and descent rate for Beaverdam Reservoir, Carvins Cove Reservoir, Falling
687 Creek Reservoir, Gatewood Reservoir, and Spring Hollow Reservoir in Southwestern Virginia,
688 USA 2013-2021 (Version 12) [Dataset]. Environmental Data Initiative.
689 https://doi.org/10.6073/pasta/c4c45b5b10b4cb4cd4b5e613c3effbd0

690 Carey, C. C., Thomas, R. Q., & Hanson P. C. (2022c). General Lake Model-Aquatic
691 EcoDynamics model parameter set for Falling Creek Reservoir, Vinton, Virginia, USA 2013-
692 2019 (Version 1) [Dataset]. Environmental Data Initiative.
693 https://doi.org/10.6073/pasta/9f7d037d9a133076a0a0d123941c6396

694 Carey, C. C., Wander, H. L., McClure, R. P., Lofton, M. E., Hamre, K. D., Doubek, J. P., et
695 al. (2022d). Secchi depth data and discrete depth profiles of photosynthetically active radiation,
696 temperature, dissolved oxygen, and pH for Beaverdam Reservoir, Carvins Cove Reservoir,
697 Falling Creek Reservoir, Gatewood Reservoir, and Spring Hollow Reservoir in southwestern

698 Virginia, USA 2013-2021 (Version 10) [Dataset]. Environmental Data Initiative.
699 https://doi.org/10.6073/pasta/887d8ab8c57fb8fdf3582507f3223cd6

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

703 Doherty, J. (2018a). Manual for PEST: Model-Independent Parameter Estimation. Part 1:
704 PEST, SENSAN and Global Optimisers. Brisbane, Australia: Watermark Numerical Computing.

705 Doherty, J. (2018b). Manual for PEST: Model-Independent Parameter Estimation. Part 2:
706 PEST Utility Support Software. Brisbane, Australia: Watermark Numerical Computing.

707 Elhabashy, A., Li, J., & Sokolova, E. (2023). Water quality modeling of a eutrophic drinking
708 water source: Impact of future climate on Cyanobacterial blooms. *Ecological Modelling*, *477*,
709 110275. https://doi.org/10.1016/j.ecolmodel.2023.110275

710 Frassl, M. A., Abell, J. M., Botelho, D. A., Cinque, K., Gibbes, B. R., Jöhnk, K. D., et al.
711 (2019). A short review of contemporary developments in aquatic ecosystem modelling of lakes
712 and reservoirs. *Environmental Modelling & Software*, *117*, 181–187.
713 https://doi.org/10.1016/j.envsoft.2019.03.024

714 Gal, G., Makler-Pick, V., & Shachar, N. (2014). Dealing with uncertainty in ecosystem model
715 scenarios: Application of the single-model ensemble approach. *Environmental Modelling &*
716 *Software*, *61*, 360–370. https://doi.org/10.1016/j.envsoft.2014.05.015

717 Gallagher, M., & Doherty, J. (2007). Parameter estimation and uncertainty analysis for a
718 watershed model. *Environmental Modelling & Software*, *22*(7), 1000–1020.
719 https://doi.org/10.1016/j.envsoft.2006.06.007

720 Gerling, A. B., Browne, R. G., Gantzer, P. A., Mobley, M. H., Little, J. C., & Carey, C. C.
721 (2014). First report of the successful operation of a side stream supersaturation hypolimnetic
722 oxygenation system in a eutrophic, shallow reservoir. *Water Research*, *67*, 129–143.
723 https://doi.org/10.1016/j.watres.2014.09.002

724 Gerling, A. B., Munger, Z. W., Doubek, J. P., Hamre, K. D., Gantzer, P. A., Little, J. C., &
725 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

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., ed. (2022). Modelling Aquatic Eco-Dynamics: Overview of the AED modular simulation platform. Zenodo. https://doi.org/10.5281/zenodo.6516222

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 & Software*, *128*, 104697. https://doi.org/10.1016/j.envsoft.2020.104697

Idso, S. B. (1973). On the concept of lake stability. *Limnology and 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 & Software*, *21*(5), 602–614. https://doi.org/10.1016/j.envsoft.2006.01.004

Janse, J. H., Scheffer, M., Lijklema, L., Liere, L. V., Sloot, J. S., & Mooij, W. M. (2010). Estimating the critical phosphorus loading of shallow lakes with the ecosystem model PCLake: Sensitivity, calibration and uncertainty. *Ecological Modelling*, *221*(4), 654–665. https://doi.org/10.1016/j.ecolmodel.2009.07.023

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

Jørgensen, S. E. (1995). State of the art of ecological modelling in limnology. *Ecological Modelling*, *78*(1–2), 101–115. https://doi.org/10.1016/0304-3800(94)00120-7

Kat, C.-J., & Els, P. S. (2012). Validation metric based on relative error. *Mathematical and Computer Modelling of Dynamical Systems*, *18*(5), 487–520. https://doi.org/10.1080/13873954.2012.663392

Kurucz, K., Hipsey M., Carey, C., Huang, P., De Sousa, E., & White J. (2023). Data and software for the GLM-AED simulation and model calibration of the Falling Creek Reservoir (v1. 0. 0.) [Software]. Zenodo. https://doi.org/10.5281/zenodo.10148689

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

Luo, L., Hamilton, D., Lan, J., McBride, C., & Trolle, D. (2018). Autocalibration of a one-dimensional hydrodynamic-ecological model (DYRESM 4.0-CAEDYM 3.1) using a Monte Carlo approach: simulations of hypoxic events in a polymictic lake. *Geoscientific Model Development*, *11*(3), 903–913. https://doi.org/10.5194/gmd-11-903-2018

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. *Science of The Total Environment*, *636*, 610–620. https://doi.org/10.1016/j.scitotenv.2018.04.255

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

Nürnberg, G. K. (1995). Quantifying anoxia in lakes. *Limnology and Oceanography*, *40*(6), 1100–1111. https://doi.org/10.4319/lo.1995.40.6.1100

Omlin, M., & Reichert, P. (1999). A comparison of techniques for the estimation of model prediction uncertainty. *Ecological Modelling*, *115*(1), 45–59. https://doi.org/10.1016/s0304-3800(98)00174-4

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

784 Read, J. S., Hamilton, D. P., Jones, I. D., Muraoka, K., Winslow, L. A., Kroiss, R., et al.
785 (2011). Derivation of lake mixing and stratification indices from high-resolution lake buoy data.
786 *Environmental Modelling & Software*, *26*(11), 1325–1336.
787 https://doi.org/10.1016/j.envsoft.2011.05.006

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

792 Refsgaard, J. C., Sluijs, J. P. van der, Brown, J., & Keur, P. van der. (2006). A framework for
793 dealing with uncertainty due to model structure error. *Advances in Water Resources*, *29*(11),
794 1586–1597. https://doi.org/10.1016/j.advwatres.2005.11.013

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

798 Reichert, P., & Omlin, M. (1997). On the usefulness of overparameterized ecological models.
799 *Ecological Modelling*, *95*(2–3), 289–299. https://doi.org/10.1016/s0304-3800(96)00043-9

800 Robson, B. J., Skerratt, J., Baird, M. E., Davies, C., Herzfeld, M., Jones, E. M., et al. (2020).
801 Enhanced assessment of the eReefs biogeochemical model for the Great Barrier Reef using the
802 Concept/State/Process/System model evaluation framework. *Environmental Modelling &*
803 *Software*, *129*, 104707. https://doi.org/10.1016/j.envsoft.2020.104707

804 Soares, L. M. V., & Calijuri, M. do C. (2021). Deterministic modelling of freshwater lakes
805 and reservoirs: current trends and recent progress. *Environmental Modelling & Software*, *144*,
806 105143. https://doi.org/10.1016/j.envsoft.2021.105143

807 Sousa, E. R. D., Hipsey, M. R., & Vogwill, R. I. J. (2023). Data assimilation, sensitivity
808 analysis and uncertainty quantification in semi-arid terminal catchments subject to long-term
809 rainfall decline. *Frontiers in Earth Science*, *10*, 886304.
810 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

Tan, J., Duan, Q., Gong, W., & Di, Z. (2022). Differences in parameter estimates derived from various methods for the ORYZA (v3) Model. *Journal of Integrative Agriculture*, *21*(2), 375–388. https://doi.org/10.1016/s2095-3119(20)63437-2

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 & Software*, *26*(4), 354–370. https://doi.org/10.1016/j.envsoft.2010.08.009

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

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

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: Atmospheres*, *117*(D3). https://doi.org/10.1029/2011jd016048