Water Resources Research

RESEARCH ARTICLE
10.1029/2022WR033660

Bayesian Calibration Points to Misconceptions in Three-Dimensional Hydrodynamic Reservoir Modeling

Sebastian Schwindt[1], Sergio Callau Medrano[1], Kilian Mouris[1], Felix Beckers[1,2], Stefan Haun[1], Wolfgang Nowak[1], Silke Wieprecht[1], and Sergey Oladyshkin[1]

[1]Institute for Modelling Hydraulic and Environmental Systems, University of Stuttgart, Stuttgart, Germany, [2]Ministry of the Environment, Climate Protection and the Energy Sector Baden-Württemberg, Stuttgart, Germany

Key Points:

- Bayesian calibration efficiently and objectively fits constrained, case-specific model parameters and identifies remaining uncertainties
- Post-calibration uncertainty assessments help identify incorrect parameter combinations and constraints
- More constrained calibration leads to lower uncertainty, which is not detected by global statistics

Supporting Information:

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

Correspondence to:

S. Schwindt,
sebastian.schwindt@iws.uni-stuttgart.de

Citation:

Schwindt, S., Callau Medrano, S., Mouris, K., Beckers, F., Haun, S., Nowak, W., et al. (2023). Bayesian calibration points to misconceptions in three-dimensional hydrodynamic reservoir modeling. *Water Resources Research*, *59*, e2022WR033660. https://doi.org/10.1029/2022WR033660

Received 15 SEP 2022
Accepted 17 FEB 2023

Author Contributions:

Conceptualization: Sebastian Schwindt, Wolfgang Nowak, Sergey Oladyshkin
Data curation: Sergio Callau Medrano
Formal analysis: Sebastian Schwindt, Kilian Mouris
Funding acquisition: Silke Wieprecht, Sergey Oladyshkin
Investigation: Sebastian Schwindt, Sergio Callau Medrano
Methodology: Sebastian Schwindt, Sergio Callau Medrano, Felix Beckers, Sergey Oladyshkin
Project Administration: Stefan Haun, Silke Wieprecht, Sergey Oladyshkin

Abstract

Three-dimensional (3d) numerical models are state-of-the-art for investigating complex hydrodynamic flow patterns in reservoirs and lakes. Such full-complexity models are computationally demanding and their calibration is challenging regarding time, subjective decision-making, and measurement data availability. In addition, physically unrealistic model assumptions or combinations of calibration parameters may remain undetected and lead to overfitting. In this study, we investigate if and how so-called Bayesian calibration aids in characterizing faulty model setups driven by measurement data and calibration parameter combinations. Bayesian calibration builds on recent developments in machine learning and uses a Gaussian process emulator as a surrogate model, which runs considerably faster than a 3d numerical model. We Bayesian-calibrate a Delft3D-FLOW model of a pump-storage reservoir as a function of the background horizontal eddy viscosity and diffusivity, and initial water temperature profile. We consider three scenarios with varying degrees of faulty assumptions and different uses of flow velocity and water temperature measurements. One of the scenarios forces completely unrealistic, rapid lake stratification and still yields similarly good calibration accuracy as more correct scenarios regarding global statistics, such as the root-mean-square error. An uncertainty assessment resulting from the Bayesian calibration indicates that the completely unrealistic scenario forces fast lake stratification through highly uncertain mixing-related model parameters. Thus, Bayesian calibration describes the quality of calibration and correctness of model assumptions through geometric characteristics of posterior distributions. For instance, most likely calibration parameter values (posterior distribution maxima) at the calibration range limit or with widespread uncertainty characterize poor model assumptions and calibration.

Plain Language Summary

Software tools for replicating a real-world element, such as an artificial lake, need to account for many unknown parameters to create a physically sound conceptual computer model. Still, simplification assumptions are necessary to break down the complex reality into parameters that are easier to calculate. But the simplified parameters take on different values for each model and require specific adjustments. To perform these adjustments, a past event is typically reproduced with the conceptual model and different simplification parameter combinations. The simplification parameter combinations leading to the best possible replication of the past event are assumed to be valid to use the conceptual model for predictions of future events. Alas, many potentially false combinations can replicate a past event with very good results. Thus, a conceptual computer model can be overly adjusted regarding a particular phenomenon, such as heat transfer. Also, the number of possible adjustment tests is limited due to the long computing time of a conceptual model. For these reasons, we use a fast, simplified statistical model of a more complex conceptual model and machine learning for the adjustment process. We find that the statistic uncertainty increases with decreasing physical correctness of simplification parameter combinations.

1. Introduction

Water storage and supply reservoirs are highly dynamic systems with complex three-dimensional (3d) flow characteristics that can be modeled with computationally demanding numerical simulation software. Such numerical models are vital to predict and plan efforts to maintain the functionality of reservoirs (e.g., drinking water supply, irrigation, or hydropower; Woolway et al., 2021; Zarfl et al., 2015). Still, modeling complex 3d hydrodynamics is a great challenge because many processes and factors, such as thermal stratification, may alter hydrodynamics in a reservoir (Kerimoglu & Rinke, 2013; Li et al., 2010; Zhang et al., 2020). Thermal stratification occurs, for example, in monomictic, dimictic, or polymictic lakes and reservoirs with generally small flow velocities and

Software: Sergio Callau Medrano, Kilian Mouris, Felix Beckers, Wolfgang Nowak, Sergey Oladyshkin
Supervision: Sebastian Schwindt, Kilian Mouris, Felix Beckers, Stefan Haun, Silke Wieprecht, Sergey Oladyshkin
Validation: Sebastian Schwindt, Kilian Mouris, Stefan Haun, Wolfgang Nowak
Visualization: Sebastian Schwindt, Sergio Callau Medrano
Writing – original draft: Sebastian Schwindt, Sergio Callau Medrano, Sergey Oladyshkin
Writing – review & editing: Sebastian Schwindt, Kilian Mouris, Felix Beckers, Stefan Haun, Wolfgang Nowak, Sergey Oladyshkin

when the water temperature seasonally trespasses 4°C, where water has its maximum (temperature-dependent) density (Hutchinson & Löffler, 1956). With season-driven rising temperatures, thermal stratification happens when warm air and solar radiation heat the upper layers of a deep reservoir (Kirillin & Shatwell, 2016; Snucins & John, 2000). Thus, in spring and fall, slow temperature-driven mixing occurs in monomictic lakes when the ambient temperature heats the lake surface above 4°C or cools it below 4°C, respectively. Stratification is particularly pronounced at warmer water temperatures when density differences can become greater. In addition, many other factors such as salinity and organic processes affect heat absorption and transfer in reservoirs and lakes (Dong et al., 2020; Sommer et al., 1986; Thackeray et al., 2013). Still, the heat transfer between air and water or in water by advection and diffusion is a slow process, which often takes several weeks and seasonally shifts more and more due to climate change (Woolway et al., 2021). Thus, mixing because of other external forces can considerably counteract stratification. For instance, wind-driven mixing, or pump and turbine operations increase turbulence and promote temperature equalization (Müller et al., 2018).

In consequence, to simulate hydrodynamics in an artificial reservoir, a numerical model of a reservoir needs to account for many processes, and their implication requires substantial simplification hypotheses (Hodges et al., 2000; Wang et al., 2022). For instance, the simulation of internal waves and mixing (i.e., turbulence) requires simplifications in the form of bulk coefficients for turbulence closures, which ultimately, also drive heat transfer. Typically, heat transfer in reservoir models is simplified with the so-called Boussinesq approximation of advection and diffusion (Katopodes, 2019; Yang et al., 2018) as a function of bulk coefficients.

These bulk coefficients must be specifically fitted for every numerical case study by vetting model results against measurement data. This fitting process is called model calibration, which is defined as an inverse, multi-step problem aimed at reducing uncertainties and achieving good agreement between modeled and measured data with reasonable tolerance (Oberkampf et al., 2004). Thus, calibration involves the fitting of (bulk) model parameters within a physically reasonable range, updating the model, and comparing observations with model results (Soares et al., 2020; Wright et al., 2017).

Finding an optimum combination of values for multiple calibration parameters is typically addressed by a trial-and-error method that requires running the numerical model for every parameter variation. Calibration parameter values leading to better reproduction of measured data are saved and modified based on an expert opinion until the model performs satisfactorily. The goodness of agreement is typically measured by global (lump-sum) statistics (e.g., the Nash-Sutcliffe model efficiency or the root-mean-square error [RMSE]), or even visual inspection only. This subjective calibration procedure faces several substantial challenges. In particular, trial-and-error calibration is subjected to individual and subjective decisions (Li et al., 2015; Masoumi et al., 2021; Shoarinezhad et al., 2020), time-consuming because of computationally expensive numerical modeling (Afshar et al., 2013; Beckers et al., 2020; Lindim et al., 2011), and prone to equifinality (i.e., entirely different parameter combinations leading to similar model outcomes; Beven & Binley, 1992). For instance, unintentional, undetected flaws in the model simplification hypotheses or calibration setup may counterbalance each other and cancel out errors. In addition, if the available calibration data reflect only a fraction of reservoir hydrodynamics, then the unconstrained aspects of reservoir hydrodynamics will be arbitrarily (bad) simulated. The wrong calibration of unconstrained parameters represents overfitting (cf. Beven & Binley, 1992; Brodeur et al., 2020), where the global model performance regarding the constrained parameters may seem good, but forecast simulations of future scenarios may be poor because of the higher importance of the unconstrained parameters. Ultimately, trial-and-error calibration is poorly efficient with low informative value regarding model uncertainty, and unreliable for identifying a unique, meaningful calibration parameter value combination (Lindim et al., 2011).

Thus, better and more efficient calibration techniques are needed, and so-called Bayesian calibration has shown to be a powerful tool. A Bayesian calibration yields a probability distribution of relevant calibration parameter sets through a global optimum search across parameter value ranges and identifies the remaining (post-calibration) uncertainty of calibration parameters (Beckers et al., 2021; Camacho & Martin, 2013).

Bayesian calibration starts with defining physically meaningful ranges for relevant calibration parameters and associated pre-calibration parameter uncertainties. The parameter range definition corresponds to a so-called prior probability distribution of every calibration parameter (Kennedy & O'Hagan, 2001; Kim & Park, 2016). Then, measured data is used to define a so-called likelihood. The likelihood expresses how well a model fits the calibration data as a function of calibration parameter combinations. The likelihood also accounts for potential imprecisions in the calibration data. Finally, the prior distributions and likelihoods are combined and poor calibration results are rejected (e.g., through rejection sampling; Smith & Gelfand, 1992) to obtain so-called posterior distributions. These distributions encode both best-fit parameter sets and post-calibration uncertainty of model parameters.

However, the computing time of Bayesian calibration could quickly skyrocket, because it requires running the numerical model many times, for instance, with millions of Monte-Carlo samples from the defined ranges of prior distributions of calibration parameters. To bypass computationally expensive model runs, machine-learning-driven techniques with so-called surrogate models (also referred to as metamodels) have already demonstrated highly efficient performance (Beckers et al., 2021).

A surrogate model stochastically mimics the behavior of a full-complexity numerical model (especially how its simulation results change with changing calibration parameter values) at acceptable accuracy (Beckers et al., 2021; Leifsson et al., 2015; Oladyshkin & Nowak, 2012). The computing time of a surrogate model is typically in the order of $\mathcal{O}(10 \times 10^{-3}\text{s})$, and therefore, faster than a full-complexity model by orders of magnitude (Beckers et al., 2020; Forrester et al., 2008). A surrogate model can be constructed, for instance, using polynomial chaos expansion or Gaussian process emulators (GPEs) in combination with Bayesian optimization (Camacho et al., 2015; Jones et al., 1998; Kim et al., 2013; Mockus, 1994). The surrogate construction requires at least one run of the full complexity model for initial training, which can then be improved by successive training iterations. In this study, we use a specific surrogate model in the form of a GPE with an iterative Bayesian active learning (BAL) training algorithm.

While the surrogate-model technique addresses the time-inefficiency problem of subjective trial-and-error calibration, Bayesian calibration also promises to overcome other shortcomings of subjective calibration. To this end, in this study, we used Bayesian calibration for adjusting a hydrodynamic reservoir model with different physical correctness of model simplification hypotheses through varying the availability of measured data. In particular, in one of three test scenarios, we forced reservoir stratification through Bayesian calibration in an unrealistically short time with water temperature measurements. In two other scenarios, we provided Bayesian calibration with flow velocity measurements reflecting pump and turbine operations of a pump-storage reservoir. Finally, we compared the characteristics of posterior distributions to analyze how Bayesian calibration responds to different scenarios of data availability that concurrently determine the apparent nature of dominant physical processes to be simulated. Thus, we asked two research questions. First, *how does the selection of calibration data affect the (Bayesian) calibration results?* Second, *how does Bayesian calibration characterize physically unreasonable calibration results?* The first question focused on the physical significance of the calibration results, and the second question addressed statistical characteristics. Our test hypothesis was that *a reservoir model with a physically inappropriate calibration setting can appear to correctly reproduce measurement data*. In addition, we hypothesized that *the shape of posterior distributions points to physically unreasonable calibration results*.

To investigate and test the two hypotheses, we ran a Bayesian calibration of a large-scale, complex hydrodynamic Delft3D-FLOW (Deltares, 2022) model of the pump-storage Schwarzenbach reservoir (SR) in Germany (Encinas Fernández et al., 2020). The calibration considered two bulk parameters (background horizontal eddy diffusivity and viscosity) and an initial guess for the vertical water temperature profile in the reservoir (assumed to be a constant over depth). We used the initial water temperature as a calibration parameter, because of its strong influence on the model results and the generally high uncertainty involved in assigning the initial water temperature. By imposing physically nearly impossible fast reservoir stratification in one of the test scenarios, we investigated how supervised machine learning in the form of Bayesian calibration treated mixing-related bulk parameters to match the measured data.

Before we addressed the research questions to test the hypotheses in the results and discussion sections, we described the experimental design in the form of the available calibration data and models used in the next section.

2. Materials and Methods

2.1. The Schwarzenbach Reservoir (SR)

The SR is located in the Northern Black Forest (Germany, EPSG 3857: 48.654842°N, 8.329314°E) and serves as the upper reservoir of a hydroelectric pump-storage scheme (cf. Figure 1 and Encinas Fernández et al., 2020).

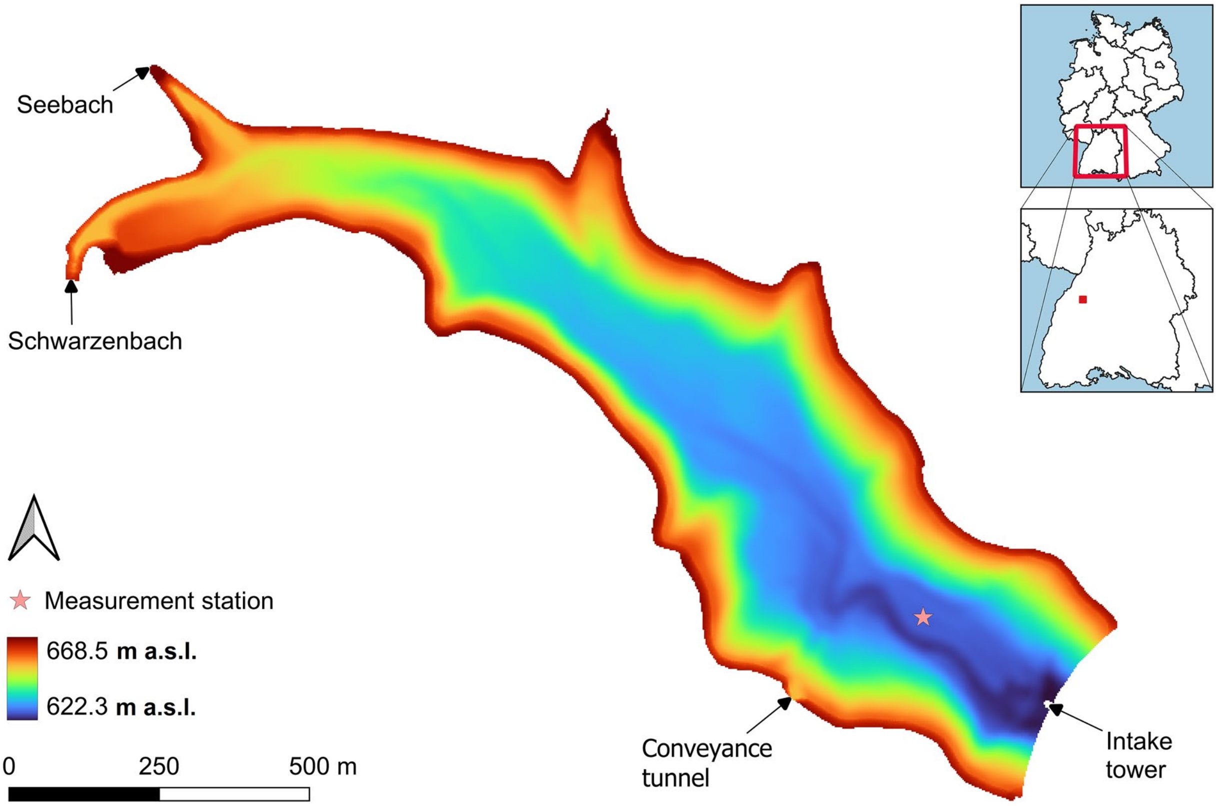

Figure 1. Location and bathymetry of the Schwarzenbach reservoir. The star indicates the measurement station where the acoustic Doppler current profiler -based flow velocity and boat-based water temperature were recorded.

The reservoir has a length of 2.2 km and a maximum depth of 47 m at its highest operating level of 668.5 m a.s.l., leading to a total storage capacity of 14.4 million m^3. It is fed by two rivers, the Schwarzenbach and the Seebach, and an artificial source in the form of the Raumuenzach conveyance tunnel, which transfers water from an adjacent catchment. In addition to the affluents, the hydrodynamic patterns of the reservoir are influenced by pump and turbine operations. Water is pumped from the Murg River at the valley bottom to an intake tower in the center of the dam that creates the reservoir. The Murg River has much higher temperature fluctuations than the inert water mass of the reservoir. Therefore, the currents close to the intake tower during pump operations are strongly driven by temperature-induced density differences. Furthermore, the reservoir is influenced by seasonal thermal stratification. Hence, the reservoir is separated into an epilimnion, metalimnion, and hypolimnion during summer, while turnovers in fall and spring cause full vertical mixing (Encinas Fernández et al., 2020).

2.2. Measurement Data

Meteorological data in the form of wind velocity, wind direction, relative humidity, air temperature, and total radiation was available from a nearby meteorological station at Freudenstadt (22 km from the SR at 48.463669°N, 8.407444°E, cf. DWD, 2021). The meteorological data constitute boundary conditions for the hydrodynamic model and do not vary in the calibration process.

Flow information was available in the form of hydrographs of the three inflows (Schwarzenbach, Seebach, and Raumuenzach conveyance tunnel) and a time series of flows resulting from pump and turbine operations at the intake tower (provided by the hydropower operator). The hydrograph that serves as liquid boundary (more detail in the next section) for the calibration at the intake tower starts on 1 August 2016, at 00:00 a.m. and ends on 7 August 2016, at 00:00 a.m. In addition, the reservoir operator provided water level recordings for this period. The inflow, outflow, and water level fluctuation time series reflect a typical scheme of pump and turbine operations, and Figure 2 shows the flows and water levels implemented in this study. The hydrographs and water levels were not used for model calibration in this study.

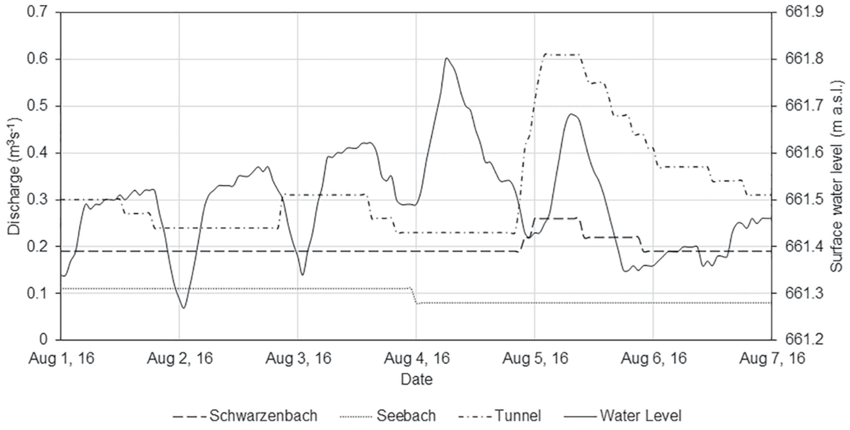

Figure 2. Flow series and water levels defined at the liquid boundaries of the numerical model.

Vertical profiles of flow velocity in the reservoir were measured with an acoustic Doppler current profiler (ADCP) between July and October 2016 (Encinas Fernández et al., 2020). The position of the ADCP is marked by a star at the center of the reservoir in Figure 1. In this study, we only used a share of these data in the first week of August for model calibration, corresponding to the available hydraulic data shown in Figure 2 and water temperature measurements.

Vertical profiles of water temperature were measured from a boat on 1 August 2016, and we used the water temperature profiles that were closest to the position of the ADCP. Both flow velocity and water temperature were measured approximately 270 m upstream of the dam. Since both measured quantities were recorded at almost the same location, we simply refer to this location as the measurement station in the following.

The reservoir bathymetry shown in Figure 1 corresponds to the 2012 situation and it stems from Landesanstalt für Umwelt Baden-Württemberg (LUBW) (2016). Thus, the bathymetry was recorded 4 yr earlier than the flow velocity and water temperature. However, topographic change in these 4 yr was negligible because the sediment yield of the catchment is small, and no major flood event occurred during that time.

More information on the available data is provided in Supporting Information S1 (see acknowledgments section).

2.3. Hydrodynamic Delft3D-FLOW Model

A 3d hydrodynamic numerical model of the SR was set up with the Delft3D-FLOW software. The software solves the Reynolds-averaged Navier Stokes equations and the continuity equation for incompressible fluids with the Boussinesq approximation for buoyancy-driven convection (Deltares, 2022). In the here used Delft3D-FLOW model, we took advantage of the software's computational efficiency with its z-layer model for 3d discretization in space based on a finite difference scheme (e.g., Platzek et al., 2014). The discretization in time uses the alternating direction implicit method (in line with Morgan et al., 2020). In contrast to CFD software (e.g., OpenFOAM) using hexahedral meshes for the representation of complex 3d structures, the vertical dimension of a 3d mesh in Delft3D-FLOW is based on multiple layers of a two-dimensional (2d) mesh.

The Boussinesq approximation simplifies the simulation of heat transfer in fluids (here: water) in which the temperature varies in space. The approach ignores changes in the fluid properties except for the fluid density, which only occurs as a multiplier of the gravitational acceleration. Moreover, the Boussinesq coefficient quantifies the momentum effect of the non-uniform velocity distribution over the water depth. Since the Boussinesq term appears in the advective acceleration in the shallow water equations, it can also be treated as a purely numerical calibration parameter to adjust the amount of fluid inertia to be considered in a simulation (Kundu & Cohen, 2008; Yang et al., 2018). The Boussinesq approximation (or hypothesis) is referred to as the (Boussinesq) eddy viscosity assumption in the following and according to the Delft3D-FLOW nomenclature.

With the eddy viscosity (Boussinesq) hypothesis, a $k-\epsilon$ turbulence closure was defined along with horizontal and vertical eddy viscosities. The eddy viscosities are proportionality factors for the turbulent energy transfer resulting from moving eddies and leading to tangential stresses (Blazek, 2005), which we considered drivers for the observed flow velocity and water temperature patterns.

The horizontal eddy viscosity ν_h varied in this study as a function of the vertical eddy viscosity ν_v, and the background horizontal eddy viscosity v_h^{back} (Deltares, 2022):

$$v_h = v_v + v_h^{back} \tag{1}$$

ν_v affects the 3d turbulence and Delft3D-FLOW calculates ν_v based on the $k - \epsilon$ turbulence closure (see also the Supporting Information S1).

The background horizontal eddy viscosity v_h^{back} was one of the calibration parameters in this study (see summary in the next section). It represents 2d turbulence and accounts for multiple hydrodynamic phenomena. In a stratified reservoir, turbulent eddy viscosity at stratification interfaces drops to zero and vertical mixing reduces to molecular diffusion. However, zero eddy viscosity is not physically correct because of the presence of internal waves, which are continuously generated by different sources and turbulence (i.e., vertical mixing; Hodges et al., 2000). Since internal waves are not explicitly considered in the Delft3D-FLOW turbulence model, they were added spatiotemporally using the background eddy viscosity v_h^{back} (Deltares, 2022).

While adjusting the eddy viscosity allowed us to mimic the mixing due to momentum, we additionally considered diffusivity that aids in representing the mixing of heat. Analogously to how Delft3D-FLOW calculates viscosities, the horizontal eddy diffusivity Δ_h is a function of the constant vertical eddy diffusivity of $\Delta_v = 10^{-6}$ m^2 s^{-1}, and the background horizontal eddy diffusivity Δ_h^{back} (Deltares, 2022):

$$\Delta_h = \Delta_v + \Delta_h^{back} \tag{2}$$

In addition to v_h^{back}, the background horizontal eddy diffusivity Δ_h^{back} was also a calibration parameter in this study (see also next section). The sub-grid scale horizontal eddy viscosity and diffusivity were zero in our model.

Regarding spatial abstraction, a boundary-fitted domain discretized the reservoir bathymetry (Figure 1) into 27 vertical layers with 5,299 tetrahedral cells each. In consequence, the average cell size was 9 × 15 × 1.7 m in the x, y, and z directions, respectively. We also tested the model with more (up to 40) and fewer (minimum 9) layers, which either led to unacceptably long computing time or could not represent variations in the vertical velocity profiles during pump operation. For the latter reason, we also used the z-layer (not σ-layer) model for vertical discretization.

The unsteady liquid boundaries were defined with the water level and the discharge time series provided for the three affluents and at the intake tower, corresponding to a one-week-long simulation (1 August at 00:00 a.m. through 7 August at 00:00 a.m., 2016). Figure 2 shows the prescribed liquid boundary characteristics. The bottom roughness was set to *Manning* with an equivalent global value of 0.015, which was not modified in the calibration process. Other physical processes affecting the flow pattern were considered in the form of air temperature and wind (processes module and the above-introduced meteorological data) because those are known to affect flow velocities in the upper layers of reservoirs, and therefore, density stratification (Dissanayake et al., 2019; Zhen-Gang, 2008).

The initial conditions corresponded to a water level of 661.34 m and a constant (in x, y, z) water temperature profile with possible values defined within a range between 5°C and 30°C. The water temperature profile was implemented in Delft3D-FLOW at the mesh boundary of the intake tower along with the discharge (i.e., liquid) boundary.

A run of the Delft3D-FLOW model took an average of 12.8 hr on a 6-core computer with 2.3 GHz per core and 16 GB memory. The simulations ran on all 6 cores. The output was written hourly. The Supporting Information S1 contains more details about the numerical model.

2.4. Relevant Calibration Parameters

The identification of relevant calibration parameters is the baseline for any model calibration. We identified three relevant calibration parameters based on other studies (e.g., Ahlfeld et al., 2003; Chanudet et al., 2012; Dissanayake et al., 2019; Goudsmit et al., 2002), our experience, and preliminary tests with the Delft3D-FLOW model of the SR.

Table 1
Calibration Parameters and Their Value Ranges (Min, Max) for Uniform Distributions $\mathcal{U}$(min, max) Considered in This Study

Parameter name	Symbol	Units	$\mathcal{U}$(min, max)
Background horizontal eddy viscosity	v_h^{back}	m^2s^{-1}	$\mathcal{U}(0.1, 5)$
Background horizontal eddy diffusivity	Δ_h^{back}	m^2s^{-1}	$\mathcal{U}(0.1, 5)$
Initial water temperature (intake tower)	T_{tow}	°C	$\mathcal{U}(5, 30)$

The three selected calibration parameters are listed in Table 1 and embraced background horizontal eddy viscosity v_h^{back}, background horizontal eddy diffusivity Δ_h^{back}, and the initial water temperature at the intake tower T_{tow} (i.e., the liquid boundary of the computational mesh). We combined the values ω of the three calibration parameters into a single vector $\mathbf{\Omega}$ that was required for the below-described Bayesian calibration:

$$\mathbf{\Omega} = \left\{ \omega_{v_h^{back}}, \omega_{\Delta_h^{back}}, \omega_{T_{tow}} \right\} \tag{3}$$

We chose the horizontal eddy viscosity and diffusivity because they have a strong influence on the turbulent viscosity terms of the advection-diffusion equations for heat transport (Boussinesq approximation). In addition, both parameters result from the simplification hypotheses of the mathematical equations, and hence, are typical candidates for calibration.

We chose the initial water temperature at the intake tower $\left(\omega_{T_{tow}}\right)$ as a calibration parameter because it has a considerable effect on the currents in the SR. Due to pump and turbine operations, thermal stratification, and missing monitoring data of water temperature directly at the intake tower (only available at the measurement station, cf. the star in Figure 1), the water temperature introduced considerable model uncertainty and it could reportedly not be well reproduced in other studies (e.g., Dissanayake et al., 2019). The initial water temperature profile was key to the physically unreasonable scenario explained in the test procedure section below.

Table 1 shows the considered value ranges (interpreted as uniform probability distributions) for the three calibration parameters, which stem from the literature (see Supporting Information S1 and Bermúdez et al., 2018; Dissanayake et al., 2019; Dong et al., 2020; Koşucu et al., 2019; Li et al., 2010; Salehi, 2017), the Delft3D user manual (Deltares, 2022), and our experience. These distributions formed the prior distributions to be updated to posterior distributions in the Bayesian calibration (see next section).

2.5. Bayesian Calibration Accelerated by Emulators and Active Learning

A Bayesian calibration is a specific form of Bayesian updating through Bayes' theorem:

$$p(\mathbf{\Omega}|\mathbf{D}, \mathcal{M}) = \frac{p(\mathbf{D}|\mathbf{\Omega}, \mathcal{M}) \cdot p(\mathbf{\Omega}|\mathcal{M})}{p(\mathbf{D}|\mathcal{M})} \tag{4}$$

where $p(\cdot)$ denotes probability density functions (pdfs). $\mathbf{\Omega}$ denotes the collection (vector) of calibration parameters for the deterministic model $\mathcal{M}$ according to Equation 3, and $p(\mathbf{\Omega}|\mathcal{M})$ denotes the prior pdf of the parameters in that model. It contains expert knowledge about the parameters, which is already available before the calibration. $\mathbf{D}$ denotes the collection (vector) of calibration data. $p(\mathbf{D}|\mathbf{\Omega}, \mathcal{M})$ is called *likelihood*. It expresses (as a function of the parameter values $\mathbf{\Omega}$) how statistically likely it is to observe the data $\mathbf{D}$ if the combination of $\mathbf{\Omega}$, $\mathcal{M}$ was true. The combination of prior information and calibration data updates the prior distribution $p(\mathbf{\Omega}|\mathcal{M})$ to the posterior distribution $p(\mathbf{\Omega}|\mathbf{D}, \mathcal{M})$. By definition, the posterior distribution is more informative than (or at least as informative as) the prior distribution (Box & Tiao, 1973; Oladyshkin & Nowak, 2019). The denominator $p(\mathbf{D}|\mathcal{M})$, often referred to as Bayesian model evidence (BME, see below), plays an important role in (competing) model selection problems. The BME is merely a normalizing constant in this equation but can assist in error diagnosis.

In this study, the model $\mathcal{M}$ was the full-complexity Delft3D-FLOW model (see above), calibration parameters were the background horizontal eddy viscosity and diffusivity as well as the initial water temperature (see Table 1), and measurement data $\mathbf{D}$ consisted of flow velocity and/or water temperature recordings (see above).

What remained to be defined is the functional form of the likelihood $p(\mathbf{D}|\mathbf{\Omega}, \mathcal{M})$ in Equation 4. A typical assumption is that there is a measurement error $\boldsymbol{\varepsilon}$ (i.e., residuals) between the model and observable reality, which can be derived from the following expression:

$$\mathbf{D} = \mathcal{M}(\mathbf{\Omega}) + \boldsymbol{\varepsilon} \tag{5}$$

The errors (stacked as a vector) typically have zero mean (no systematic errors) and a variance $\sigma^2_{\varepsilon,D}$ that resembles the imprecision of measurements. The most common distributional assumption for measurement errors is Gaussian, and errors for different data items are uncorrelated. We followed this common assumption and obtained:

$$p(\mathbf{D}|\mathbf{\Omega},\mathcal{M}) = 2\pi^{-\frac{n}{2}}|\mathbf{R}|^{-\frac{1}{2}}\exp\left[-\frac{1}{2}\left(\varepsilon^T\mathbf{R}^{-1}\varepsilon\right)\right] \tag{6}$$

where n is the number of data items (i.e., the number of measurements in $\mathbf{D}$ and of corresponding residuals in ε) and $\mathbf{R}$ is the (co-)variance matrix of measurement errors (sized $n \times n$). In our case, $\mathbf{R}$ was a diagonal matrix populated by the variances $\sigma^2_{\varepsilon,D}$ per data item. The error variance (imprecision) of the measurement data in this study had a standard deviation of 2°C for water temperature and 3 mm s^{-1} for flow velocity measurements.

Based on the posterior distributions, we extracted so-called maximum likelihoods of the calibration parameters (i.e., the location of an optimum in the likelihood function), which led to the best possible agreement with the measurement data (Beckers et al., 2020). Likewise, maximum-a-posteriori calibration parameter values can be extracted (i.e., the optimum of the joint posterior distribution), which correspond to the best possible compromise between priors and data. For uniform distributions as used in this study, both sets of parameter values coincide and the posterior distribution characteristics represent the post-calibration uncertainty of the parameters.

Typical assumption-free computational methods to evaluate the Bayesian theorem in Equation 4 are computationally inefficient Monte-Carlo or Markov-Chain Monte-Carlo methods (e.g., Smith & Gelfand, 1992). These methods are particularly inefficient when a model $\mathcal{M}$ is computationally expensive (e.g., if a Delft3D-FLOW run takes 12.8 hr on average), because these methods will make the model easily run millions of times to yield a good approximation of Equation 4. To bypass long computing times, a surrogate model response $\mathcal{S}(\mathbf{\Omega})$ was used to replace $\mathcal{M}$ in the computation of residuals (Equation 5) with a much faster approximation:

$$\mathcal{S}(\mathbf{\Omega}) \approx \mathcal{M}(\mathbf{\Omega}) \tag{7}$$

With a suitable surrogate model, speed is no longer an issue because it runs easily a million times faster than a full-complexity model (Beckers et al., 2020). That is, a surrogate model enabled us to run 10^6 random Monte Carlo realizations drawn from the prior pdfs of the calibration parameters. Then, we evaluated the surrogate model response $\mathcal{S}$ for every realization, computed the (approximate) residuals with Equation 5, calculated each realization's likelihood with Equation 6, and used rejection sampling (Smith & Gelfand, 1992) to approximate the posterior distribution in Bayes' theorem (i.e., the left side of Equation 4).

The approximation quality of the surrogate model had to be appropriate for the calibration task. The surrogate model's quality strongly depends on its construction method, including the selection of so-called training runs with the initial full-complexity model (Busby, 2009). This is where active learning becomes relevant: the training runs should concentrate mostly on high-likelihood areas, which are not known a priori. This is the starting point for an iterative process called active learning that incrementally improves the surrogate model in currently found high-likelihood regions, and subsequently refines the identification of high-likelihood regions.

Here, we used a GPE as a surrogate model and BAL for efficient training. The theoretical background for GPEs and the used BAL technique can be found, for example, in Oladyshkin et al. (2020) and Rasmussen and Williams (2006). To set up the GPE in this study, we applied a squared exponential kernel. This kernel required specifications of length scales and variance as hyperparameters, and we used the fitrgp function in Matlab (2018) to choose their values. For iterating between training and Bayesian updating, we implemented the iterative BAL technique described in Oladyshkin et al. (2020). The BAL builds on a combination of preliminary Bayesian updating and information theory (Oladyshkin & Nowak, 2019) to identify the next-best parameter set for training. Here, we used it to find the best out of 1,000 randomly proposed next-value candidates in every iteration. After every new training run, we re-iterated the hyperparameters and re-trained the GPE.

The BAL technique needs stopping criteria to identify that the iterations converge toward a final quality level. For this purpose, we tracked BME (see denominator in Equation 4) and the so-called relative entropy in the BAL iterations. Here, we approximated the full-complexity model $\mathrm{BME}_\mathcal{M}$ based on the GPE-surrogate model $\mathrm{BME}_\mathcal{S}$:

$$\mathbf{BME}_\mathcal{M} \equiv p_\Omega(\mathbf{D}|\mathcal{M}) \approx \mathbf{BME}_\mathcal{S} \equiv p_\Omega(\mathbf{D}|\mathcal{S}). \tag{8}$$

The BME expresses as an integral quantity the average goodness of fit of the model, relative to the prior distribution. If the surrogate model response $\mathcal{S}$ converges toward the full-complexity model response $\mathcal{M}$ during the BAL, then $\mathrm{BME}_{\mathcal{S}}$ converges toward $\mathrm{BME}_{\mathcal{M}}$ and stabilizes. In addition, relative entropy, also known as Kullback-Leibler divergence D_{KL}, quantifies the information gain from the prior to the posterior distributions of the calibration parameters, and it was approximated with:

$$D_{KL}[p(\Omega|\mathbf{D},\mathcal{M}), p(\Omega|\mathcal{M})] \approx D_{KL}[p(\Omega|\mathbf{D},\mathcal{S}), p(\Omega|\mathcal{S})]. \tag{9}$$

Also D_{KL} stabilizes when the surrogate model response converges toward the full-complexity model response. Thus, we continued the BAL iterations until both the BME and D_{KL} converged toward a constant. In addition, BME and D_{KL} were criteria for selecting training points in the BAL iterations. For details on the here-used approximation methods, in particular, regarding the GPE-surrogate model and BAL, see Oladyshkin et al. (2020).

2.6. Test Procedure

Initially, we ran the full-complexity, hydrodynamic, numerical 3d model (Delft3D-FLOW) of the SR 30 times to constitute the baseline for the BAL iterations. Next, we calibrated the three calibration parameters in the GPE-surrogate-assisted BAL iterations (see above). The prior calibration parameter distributions are listed in Table 1. We used the available calibration data in the form of depth-averaged horizontal flow velocity and water temperature profiles at the measurement station (see Figure 1) in the three below-described scenarios. We considered the calibration complete when additional training points in the BAL iterations did not yield an improvement in BME (Equation 8) and relative entropy (Equation 9). To verify convergence, we logged the BME and relative entropy at the end of every iteration. Finally, we implemented the maximum likelihoods resulting from the posterior distributions of the last BAL iteration to re-run the full-complexity and GPE-surrogate models at the end of each below-defined scenario. In particular, we defined three scenarios to test the two hypotheses we formulated at the end of the introduction:

1. A reservoir model with a physically inappropriate calibration setting can appear to correctly reproduce measurement data.
2. The shape of posterior distributions points to physically unreasonable calibration results.

To test the hypotheses, three calibration scenarios with varying physical correctness were defined. These three scenarios were one all-data scenario (0), and two single-data set scenarios (1 and 2), in which we applied the flow velocity and water temperature measurement data separately from each other for Bayesian calibration.

The all-data scenario 0 used both the flow velocity and water temperature measurement data from the measurement station to optimize the three calibration parameters. Computationally, this scenario was the most expensive because the number of training runs for the surrogate model scaled exponentially with the number of calibration parameters.

Scenario 1 only used the flow velocity data to calibrate all three parameters, although the data were primarily useful to calibrate the horizontal eddy viscosity and diffusivity. This scenario was aimed at a physically more correct calibration toward heat-independent, short-term flow velocity patterns driven by pump and turbine operations. Thus, scenario 1 was physically better conditioned than the other scenarios because it only calibrated toward data that can be reasonably well reproduced by the Delft3D-FLOW model.

Scenario 2 only used the water temperature data to calibrate all three parameters. Thus, the Bayesian calibration process had to attempt stratifying the reservoir (i.e., to match the measured water temperature profiles) based on an initially constant (over depth) water temperature profile at a distance of 270 m (i.e., at the intake tower).

According to the above-mentioned literature (e.g., Zhang et al., 2020), stratification within a simulation time of 6 days (real-time) is physically unlikely, which means that scenarios 0 and 2 corresponded to (partially) non-reasonable calibration frameworks because of too short simulation times. If the Bayesian calibration still succeeded (in terms of statistical scores) in calibrating the model seemingly well, we assumed that the model was poorly conditioned and evidence to support hypothesis (1) was provided.

In contrast to lump-sum statistics (e.g., the RMSE), we expected Bayesian calibration to indicate uncertainty related to the calibration results, which would not be visible after subjective trial-and-error calibration. Thus, if the physically more unreasonable scenarios 0 and 2 led to more uncertain posterior distributions and wrong bulk mixing parameters than the better-constrained scenario 1, evidence supporting hypothesis (2) was provided. In

Table 2
The Three Scenarios Defining the Frameworks for the Bayesian Calibration as a Function of Involved Measurements of Depth-Averaged Horizontal Flow Velocity U and Water Temperature T at the Measurement Station

Data scenario	# of BAL iterations	Measurement quantities involved	# of observations
Scenario 0	50	*U* and *T*	165
Scenario 1	10	*U*	145
Scenario 2	10	*T*	20

particular, we attempted to verify that high uncertainties in posterior distributions were consistent with the physical soundness of the calibration frameworks of the scenarios.

Table 2 summarizes the three test scenarios along with the number of BAL iterations, observations, and measurement data used per scenario. The number of BAL iterations anticipates the results and was driven by reaching the convergences of BME and relative entropy (see details in the results section). Note that we were using the depth-averaged horizontal flow velocity *U* for the Bayesian calibration, which was necessary to compare the results of the depth-averaged output of Delft3D-FLOW with the measurement data. In particular, the Bayesian framework (Oladyshkin et al., 2020) can currently only work with one measurement point per *x*-*y* coordinate. It cannot deal with vertically varying velocities because of the resulting dimensionality of calibration vectors. For instance, if Bayesian calibration was extended to the third dimension in space, the response surface would have an additional dimension. Because the total of posterior probabilities still needed to sum up to 1, the additional dimension would decrease the prior output space to infinitesimally small numbers, which would all be rejected in the rejection sampling step (Oladyshkin et al., 2020; Smith & Gelfand, 1992). The problem of too small (too close to zero) probabilities in the framework of active learning is also referred to as *curse of dimensionality* (Bellman, 1957), which is discussed in detail by Mouris et al. (2023).

In addition, we tested the lump-sum accuracy of the calibration results by running both the GPE and Delft3D-FLOW model with the final maximum likelihoods of the calibration parameters after each scenario. These additional runs allowed for vetting the calibrated GPE-surrogate and full-complexity models against the measurement data in a classical-deterministic manner that is commonly used with trial-and-error calibration methods. The additional runs also provided insights into the quality of final model results, and spatially explicit comparisons of the Delft3D-FLOW and the GPE-surrogate models in the entire SR, which we used for discussion of the physical relevance of calibration parameters.

3. Results

3.1. Convergence Speed

Every iteration step required on average 13 hr of computing time, including full-complexity model runs, updating the GPE, computing the likelihoods, and running the rejection sampling of 10^6 Monte Carlo candidates. In particular, a full-complexity model run took on average 12.8 hr and a GPE run took 32.5×10^{-3} s. Note that additional computing time on the order of a few seconds to minutes is required to train the GPE in each BAL iteration.

The BME and relative entropy converged at different BAL iteration steps for the three scenarios (see Table 2). In scenario 0, the BAL began converging after the 25th iteration. The BME converged toward a value of approximately 10^{-237}, and the relative entropy toward approximately 8.2. Afterward, the surrogate model produced statistically acceptable results that matched the full-complexity model results of the all-data scenario 0, and we ran a total of 50 iterations to confirm the convergence trend.

In scenario 1 (flow velocity data only), the surrogate model converged to a BME value of approximately 10^{-20} and relative entropy of approximately 2.5 after two iterations only, and we ran 10 iterations, which yielded similar BME and relative entropy values (i.e., the convergence confirmed). In scenario 2 (water temperature data only), the surrogate model converged to a BME of approximately 10^{-136}, and relative entropy of approximately 6.2 after three iterations, and again, we ran 10 iterations in total to confirm the convergence trend.

Scenarios 1 and 2 (cf. Table 2) thus converged considerably faster. The fast convergence can be explained by the smaller amount of measurement data to be matched compared with scenario 0. As a result, the high-likelihood regions are wider and less pronounced in scenarios 1 and 2, which facilitate their identification.

3.2. Marginal Posterior Distributions of Calibration Parameters

Figures 3a–3c illustrate the parameter-wise (marginal) posterior distributions obtained from the BAL-based GPE for the three scenarios 0, 1, and 2, respectively. These distributions indicate that the narrowest and clearest results

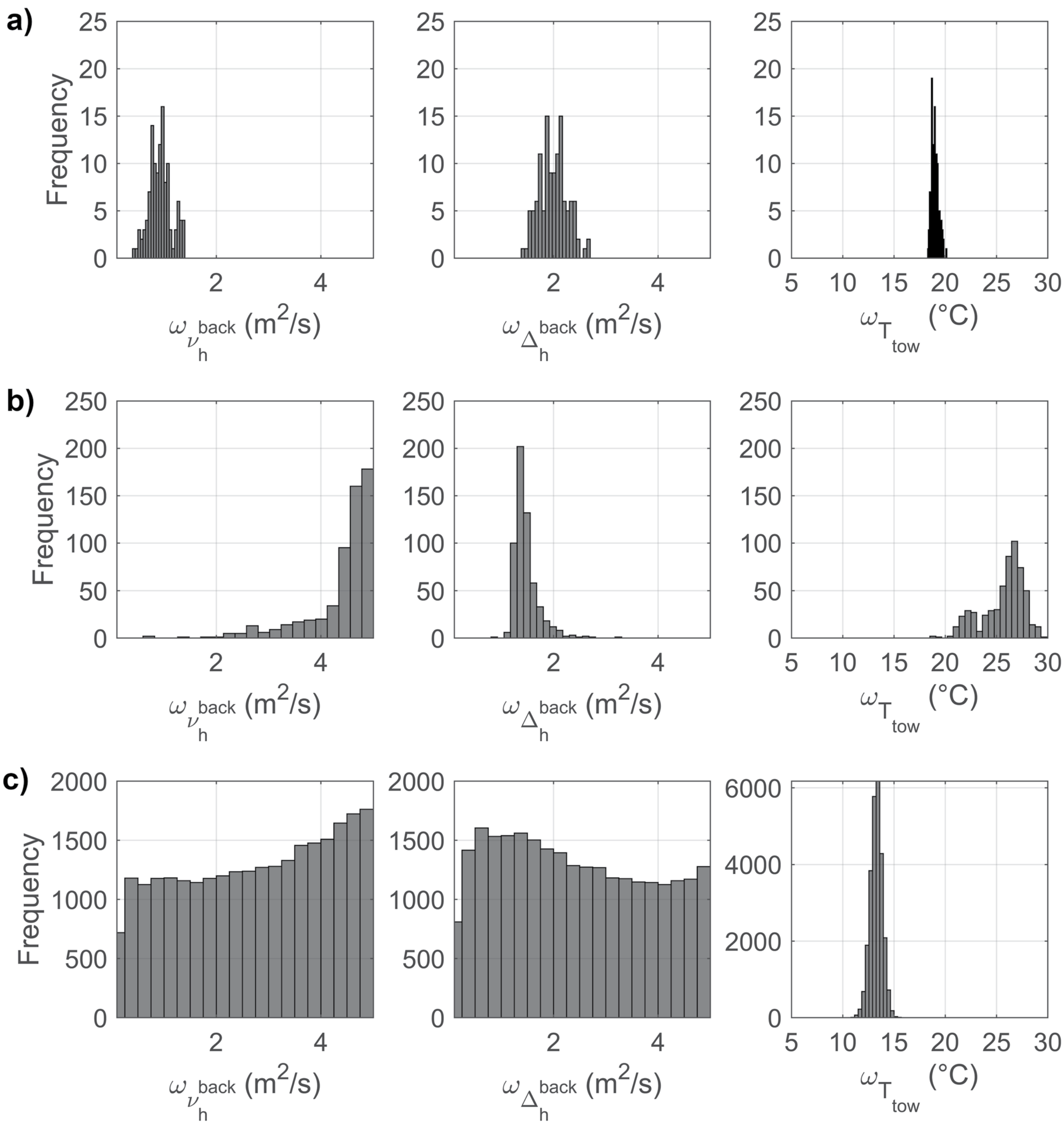

Figure 3. Posterior distributions of the calibration parameters after the iterative Bayesian updating in the framework of (a) scenario 0, (b) scenario 1, and (c) scenario 2.

were obtained in scenarios 0, tightly followed by scenario 1, and scenario 2 being substantially different. The combined consideration of both flow velocity and water temperature data (scenario 0 in Figure 3a) represents the most constrained scenario, which also reflects in the largest relative entropy of 8.2 (i.e., the largest information gain from prior to posterior).

Without the water temperature data in scenario 1, the posterior distribution of $\omega_{T_{tow}}$ is wider than in scenarios 0 and 2. Compared to scenario 0, the absence of the water temperature data also changes the maximum likelihoods of the three calibration parameters (see Table 3 below).

Table 3
Maximum-Likelihood Values of the Calibration Parameters per Scenario

Scenario		0	1	2
Background horizontal eddy viscosity	ν_h^{back} $\left(m^2 s^{-1}\right)$	0.91	4.83	4.93
Background horizontal eddy diffusivity	Δ_h^{back} $\left(m^2 s^{-1}\right)$	2.05	1.38	4.99
Initial water temperature (intake tower)	T_{tow} (°C)	18.86	26.63	13.42

In contrast, when the flow velocity data were excluded (scenario 2, Figure 3c), the posterior distributions of the background horizontal eddy viscosity $\omega_{\nu_h^{back}}$ and diffusivity $\omega_{\Delta_h^{back}}$ remain close to identical to the prior distributions (uniform between 0.1 and 5), indicating a poor calibration result regarding ν_h^{back} and Δ_h^{back}. Therefore, their maximum likelihood parameter values will be close to arbitrary in the marginal (single-parameter), and we only observe a considerably lower $\omega_{T_{tow}}$ maximum likelihood in scenario 2.

Figure 3 thus shows how leaving out either water temperature or flow velocity measurement data increases the uncertainties of calibration results through wider, less narrow, and therefore, less assertive posterior distributions. In scenario 1, missing water temperature data leads to increased uncertainty in $\omega_{T_{tow}}$. In scenario 2, missing flow velocity data leads to high uncertainty in $\omega_{\nu_h^{back}}$ and $\omega_{\Delta_h^{back}}$, which indicates that these two calibration parameters are less driven by water temperature than by flow velocity. In particular, measurements of flow velocity include patterns driven by hours of pump and turbine operation, which scenario 2 knows at its boundaries but it cannot evaluate the operation's importance to hydrodynamic patterns in the reservoir.

Fewer measurement data correspond to fewer constraints for the rejection sampling, which makes that the posterior distributions after scenarios 1 and 2 contain more samples than the more constraint scenario 0. The difference in posterior sample sizes between scenarios 1 and 2 can be explained by the higher number of velocity measurements, which makes scenario 1 more constrained than scenario 2. Therefore, Figure 3 also illustrates how more calibration constraints lead to statistically more distinct calibration results.

3.3. Maximum Likelihoods of Calibration Parameters

Table 3 lists the maximum likelihoods of the calibration parameters for the three scenarios. In scenario 0, the maximum likelihoods are in a statistically reasonable range (i.e., not at the limits of the prior limits given in Table 1). However, in scenario 1, the maximum likelihood for ν_h^{back} is 4.83, and therefore, close to the upper limit of the prior distribution. Also, the maximum likelihood of 26.63°C for the initial water temperature is close to the upper limit of the prior distribution and physically unlikely in deeper layers of the SR. Yet, this observation is not astonishing in the absence of measurement data on the water temperature. In scenario 2, the maximum likelihoods for both ν_h^{back} and Δ_h^{back} are even closer to the upper limits of the prior distribution test ranges, which indicates that the Bayesian calibration might have gone beyond the test ranges if we had allowed it (which we have intentionally not done). Therefore, ν_h^{back} and Δ_h^{back} in scenario 2 are poorly constrained by the water temperature-only measurement data. In addition, high maximum likelihoods for $\omega_{\nu_h^{back}}$ and $\omega_{\Delta_h^{back}}$, suggest that the calibration tries to adjust the model by making diffusion a key physical process, especially in scenario 2 for achieving stratification.

The maximum likelihoods do not necessarily coincide with the univariate marginal peaks shown in Figure 3, which is why we look at the nonlinear interdependence of the posterior distributions in the next section.

3.4. Joint Posteriors of Calibration Parameters

The marginal posterior distributions in Figure 3 do not visualize the dependence between the three calibration parameters. Therefore, we show the joint posterior distributions with 3d plots in Figure 4. A wide (or narrow) spread of the point cloud along a parameter's ω axis means high (or low) uncertainty. Diagonal structures imply a linear correlation, and curved shapes indicate nonlinear dependencies. The dark red color indicates the location of the maximum likelihoods listed in Table 1. For instance, comparing the location (red areas) and spread of the posterior distributions with respect to $\omega_{\nu_h^{back}}$ in Figures 4a and 4b (scenarios 0 and 1) highlights the substantial difference in its maximum likelihood (see Table 3).

The exclusive calibration toward water temperature data in scenario 2 (see Figures 4b and 4c) led to a wide spread of $\omega_{\nu_h^{back}}$ and $\omega_{\Delta_h^{back}}$, which corresponds to quasi-total independence. The spread of $\omega_{\nu_h^{back}}$ and $\omega_{\Delta_h^{back}}$ would probably

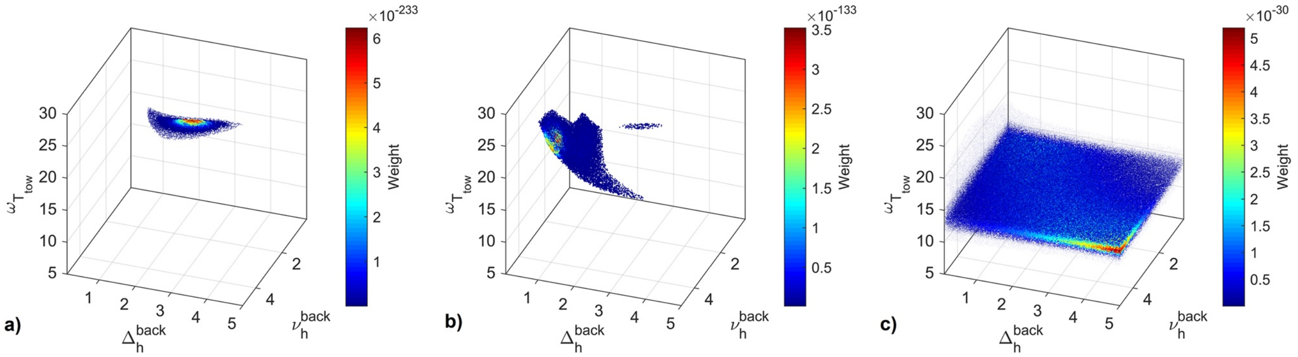

Figure 4. Nonlinear dependence of the calibration parameters after the last Bayesian active learning iterations of (a) scenario 0, (b) scenario 1, and (c) scenario 2.

have gone beyond the margins of Figure 4c, if we had also allowed physically non-meaningful value ranges. The only structure visible in Figure 4c for scenario 2 is imposed by the heavily constrained values of ωT_{tow}, which show the largest likelihood values in the front-right corner. This is additional (multivariate) evidence that the physically unreasonable forcing of water-temperature-driven stratification in scenario 2 caused high uncertainty regarding $\omega_{\nu_h^{back}}$ and $\omega_{\Delta_h^{back}}$. Still, if we were only looking at the initial water temperature as a calibration parameter result, an evident misconception would be that the calibration result was good, well-constrained, and plausible.

The interdependence of calibration parameters in scenarios 0 and 1 also demonstrates that there is more than one unique parameter combination that leads to a similar model fit. Thus, the Bayesian calibration reveals the uncertainty regarding a statistically well-fitting combination of calibration parameter values (i.e., maximum likelihoods). However, also the Bayesian approach cannot generally solve the issue of equifinality as a result of weak calibration data, model simplifications, or physically non-meaningful setups. It only enables to quantify the post-calibration uncertainty (Figures 3 and 4) indicating that the calibration result is likely to be subjected to equifinality.

3.5. Lump-Sum Statistics

We ran additional simulations with both Delft3D-FLOW and the GPE-surrogate models with the maximum-likelihood calibration parameter sets resulting from the three scenarios to calculate global lump-sum statistics characterizing a generalized model quality. The resulting lump-sum statistics in the form of the mean error (ē), standard deviation from the measurements (σ_e), and RMSE are listed in Table 4.

In the all-data scenario 0, the RMSE regarding the depth-averaged flow velocity U in Delft3D-FLOW is 2.56 mm s^{-1}, which represents a 17%-deviation compared with the time-averaged measurement data. One reason for the U deviations is the weaknesses in reproducing the magnitude and timing of peaks that stem from pump and turbine operations. The GPE-surrogate model reproduced the data slightly better, with a smaller RMSE of 1.44 mm s^{-1} and a slightly better representation of peaks in magnitude and time. This slightly better performance of the GPE indicates an imperfect surrogate approximation, as the maximum likelihood parameter sets identified with the GPE perform slightly less well when plugged into the full-complexity Delft3D-FLOW model. However, in light of the standard deviation of measurement errors of the flow velocity measurements (3 mm s^{-1}), the RMSEs represent an excellent performance of both models after scenario 0.

Table 4
Mean Error ē, Standard Deviation σ_e, and Root-Mean-Square Error (RMSE) of the Calibrated Full-Complexity and Gaussian Process Emulator (GPE)-Surrogate Model Runs With the Maximum Likelihoods of the Calibration Parameters After the Three Scenarios

Model	Parameter	Scenario	ē	σ_e	RMSE
Delft3D-FLOW	U (mm s^{-1})	0	0.44	2.53	2.56
		1	0.19	1.42	1.43
	T (°C)	0	−3.67	3.84	5.24
		2	0.16	1.80	1.76
GPE-surrogate	U (mm s^{-1})	0	0.31	1.41	1.44
		1	−0.03	1.25	1.24
	T (°C)	0	−3.77	3.85	5.31
		2	0.14	1.79	1.75

3.6. Bayesian Calibration Residuals and Posterior Predictions

The modeled U and T with the maximum likelihoods of the calibration parameters also enable us to evaluate the model improvement through the Bayesian calibration. For this purpose, we additionally consider the model residuals, here defined as the vertical (flow velocity) or horizontal (water temperature) distances between the measured and simulated values.

Figures 5a and 5b illustrate the depth-averaged flow velocity U at the measurement station for scenarios 0 and 1 (i.e., the two scenarios considering flow velocity measurements). The blue lines are the results of the Delft3D-FLOW (closed blue line) and surrogate (dashed blue line) models with the maximum likelihoods (Table 3). The gray lines are the results of the training runs, which constitute the standard deviation (uncertainty) of the posterior distributions (Figure 3). In addition, the closed black lines show the depth-averaged flow velocity measurements and enable us to identify the model residuals (i.e., vertical differences with the model results).

The posterior surrogate model outputs well fit the measurement data with small residuals after both scenarios. In addition, the posterior time-averages of U are 1.65 and 1.05 mm s^{-1} in scenarios 0 and 1, respectively. Thus, the calibration with flow velocity data yielded small calibration errors (i.e., residuals), which are close to zero on average. The calibration also reduced the model uncertainty in both scenarios, with, for instance, the time-averaged standard deviation in scenario 1 decreasing from 12.1 mm s^{-1} (prior according to Table 1) to 0.4 mm s^{-1} (posterior).

A comparison of the two scenarios 0 and 1 shows that the models perform especially better in scenario 1 regarding the magnitude and time of flow velocity peaks. In particular, after scenario 1, both models well simulate the shape and magnitudes of the measurement data, although with too small amplitudes of local extrema (lows and peaks). The improved accuracy after scenario 1 is also reflected in RMSEs of 1.43 mm s^{-1} (Delft3D-FLOW) and 1.24 mm s^{-1} (GPE-surrogate), which are both smaller than after scenario 0 (see also Table 4). Already Figure 3 indicated that the flow velocity data have higher importance for the hydrodynamic patterns in the reservoir than the water temperature measurements (differences between Figures 3b and 3c). Figure 5 confirms this observation by indicating that the water temperature data in scenario 0 lead to a worse replication of flow velocity measurements. The reason is that in scenario 0, the calibration had to meet two measurement targets that compete with each other due to imperfect model assumptions (i.e., too short time for stratification).

Figures 6a and 6b show the water temperature profiles T at the measurement station for scenarios 0 and 2 (i.e., the two scenarios considering water temperature data). The blue lines are the results of the Delft3D-FLOW (closed blue line) and surrogate (dashed blue line) models with the maximum likelihoods (Table 3). The gray lines are the results of the training runs, which constitute the standard deviation (uncertainty) of the posterior distributions (Figure 3). In addition, the closed black lines show the depth-averaged flow velocity measurements and enable us to identify the model residuals (i.e., vertical differences with the model results).

In scenario 0, the calibration was forced to compromise between fitting the water temperature and flow velocity data, which is reflected in the physical weaknesses of the results. For instance, the water temperature profile modeled with the posterior maximum likelihoods is close to constant over depth and far away from the measurement data. Thus, the calibrated model fails to reproduce the temperature stratification in the all-data scenario 0, which makes sense because of the too-short time frame for the onset of stratification. In addition, in scenario 0, the residuals of the prior and posterior model outputs have considerably high averages of 3.58°C and 4.93°C, respectively.

After scenario 2, the modeled posterior water temperature profiles show smaller residuals (i.e., it is generally closer to the measurements) than the prior profiles. The posterior profile also shows signs of temperature stratification, with water temperatures at the surface approximately 1°C higher than in deeper layers. The better simulation of water temperature profiles also reflects in the RMSE (Table 4) reducing from 5.3°C in scenario 0 to 1.75°C in scenario 2. Therefore, scenario 2 also performs better in terms of lump-sum statistics regarding T than scenario 0. However, the fast stratification achieved in scenario 2 is physically unlikely, and thus, represents a poor calibration even though the lump-sum statistics (Table 4) suggest good model performance. Thus, considering exclusively water temperature measurements in scenario 2 led to improved simulation of water temperature measurements, although the water temperature profile still does not correspond well to the measurements.

These observations show that scenario 0 simulates flow velocities better than water temperature. Still, scenario 0 simulates flow velocities less correctly than scenario 1, which does not know the water temperature data. Because pump and turbine operations are the main drivers for fluxes in the SR, the higher importance of flow velocity measurements indicates that hydrodynamic patterns are only secondarily controlled by heat-driven stratification.

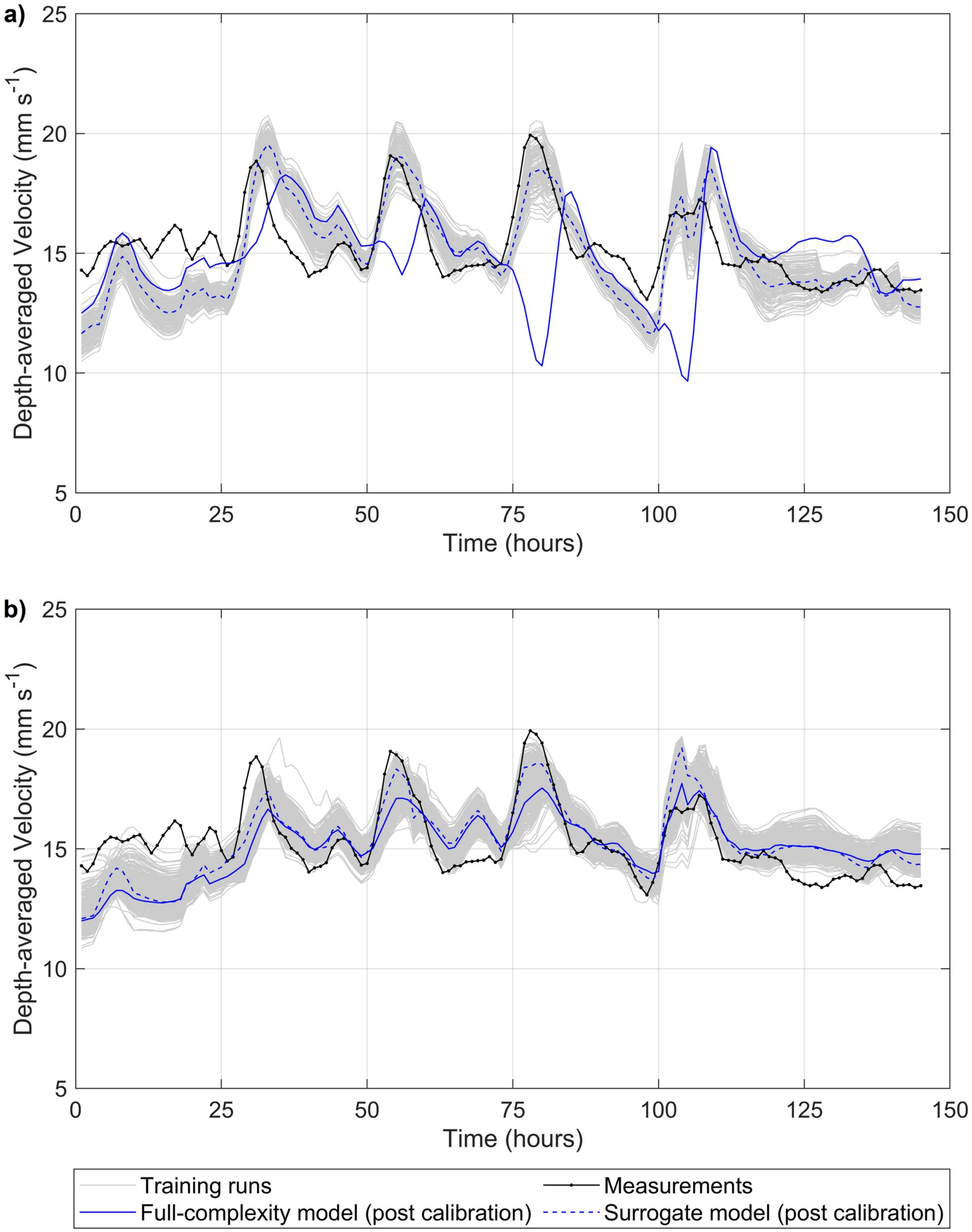

Figure 5. Simulated depth-averaged horizontal flow velocity *U* resulting from (a) scenario 0 and (b) scenario 1 in the considered period from August 1 (00:00 a.m.) through August 7 (00:00 a.m.), 2016. The graphs correspond to the location of the measurement station in the reservoir and feature results of the training runs, surrogate model, and full-complexity Delft3D FLOW model in addition to the measurements.

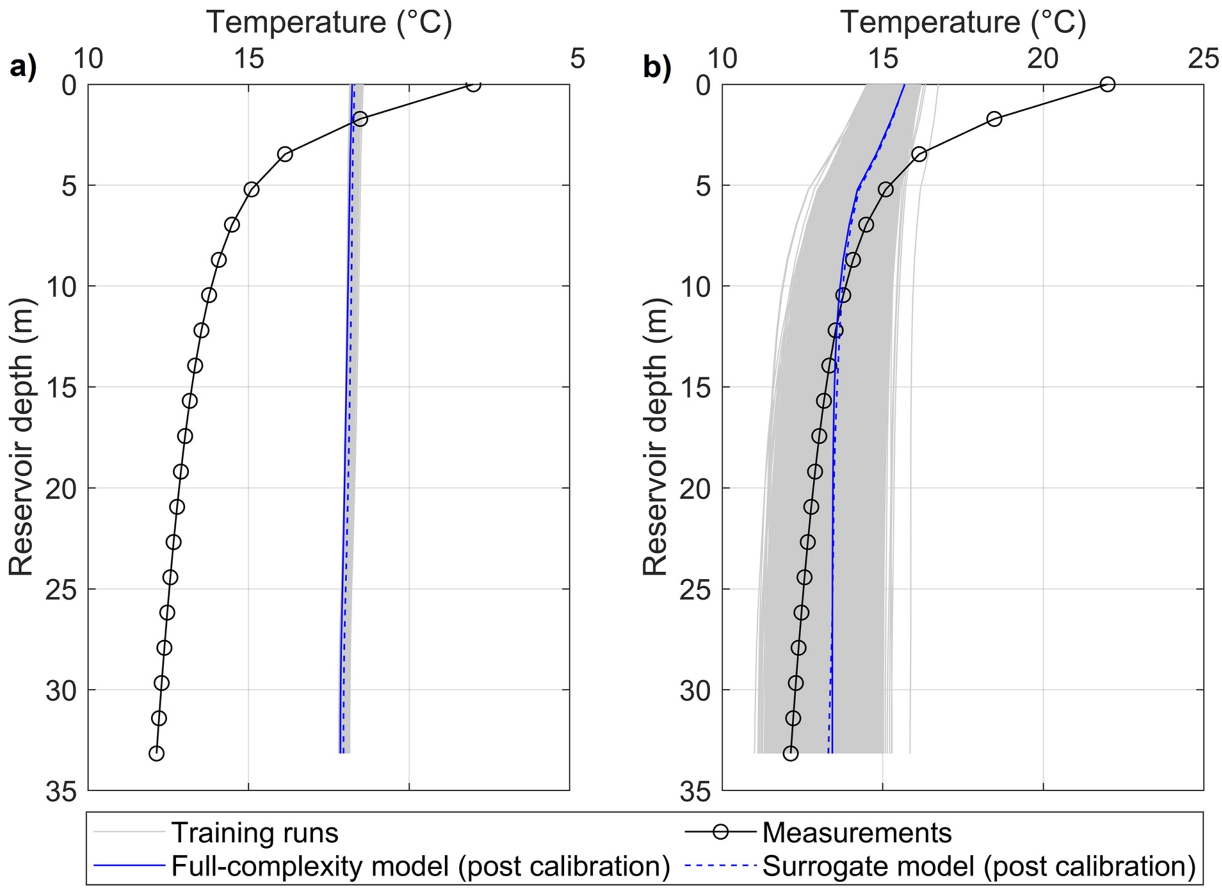

Figure 6. Simulated water temperature *T* profiles resulting from (a) the scenario 0 and (b) the scenario 2 calibrations at the measurement station. The graphs show the results of the training runs, surrogate model, and full-complexity Delft3D-FLOW model in addition to the measurements.

4. Discussion

4.1. Physical Relevance

Figures 7 and 8 map horizontal (non-depth-averaged) flow velocity calculated with the calibrated full-complexity Delft3D-FLOW and GPE-surrogate model. The maps show flow velocity magnitudes (not directions) in the entire SR during turbine and pump operation hours, respectively, and with the maximum likelihoods from the all-data scenario 0. The turbine operation map (Figure 7) represents a model snapshot on 2 August 2016, at 10:00 p.m., and the pump operation map (Figure 8) represents a snapshot on 3 August 2016, at 03:00 a.m. Both figures show the horizontal flow velocity magnitudes at different water depths of 3, 16, and 30 m below the maximum operation water level (668.5 m a.s.l.). The pump and turbine operation hours are particularly interesting in light of the above observations because they potentially counteract stratification.

In the case of turbine operation (Figure 7), the calibrated GPE-surrogate ($\text{GPE}_{\text{Adjusted}}$ in the plots) and full-complexity models show two currents in the reservoir center and near the dam at depths of 3 and 16 m. Those currents can be related to the inflows from the conveyance tunnel that is located at the South-West shore; approximately 400 m upstream of the dam (see also Figure 1). The highest flow velocities occur in the vicinity of the dam where turbine operation draws water from the reservoir.

The pump operations in Figure 8 show a similar pattern, but with inverse velocity vector directions (directions are not visible on the magnitude maps). At a water depth of 16 m, both turbine and pump operations form a large eddy pattern in the reservoir with slow horizontal flow velocities that might affect the stratification of the water temperature.

While scenario 0 replicates flow velocity at the measurement station acceptably well (Figure 5), it does not well reproduce temperature-driven stratification (Figure 6), supposedly because pump and turbine operations

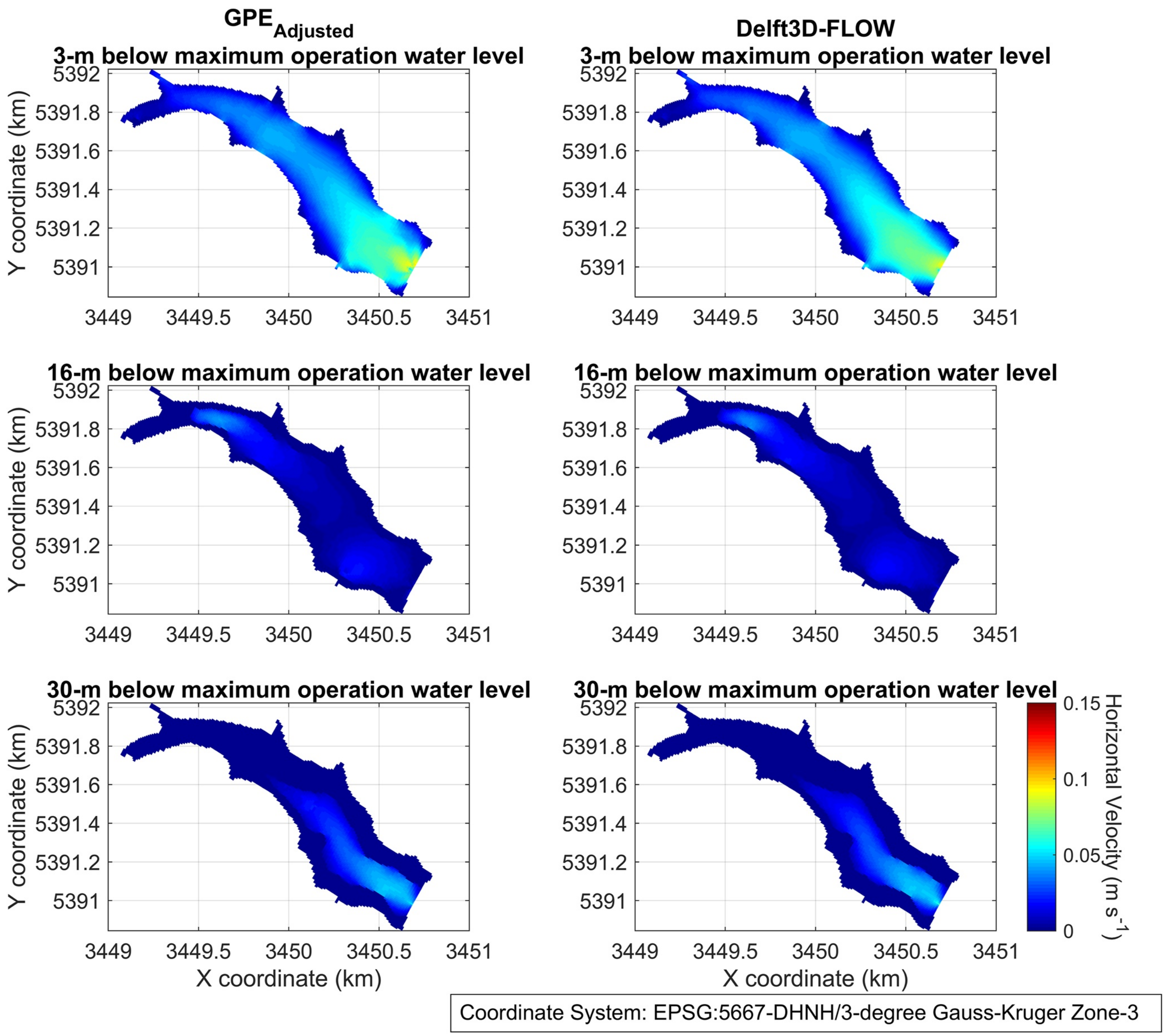
GPE Adjusted
Delft3D-FLOW
3-m below maximum operation water level
16-m below maximum operation water level
30-m below maximum operation water level
Y coordinate (km)
X coordinate (km)
5392
5391.8
5391.6
5391.4
5391.2
5391
3449
3449.5
3450
3450.5
3451
Horizontal Velocity (m s-1)
0.15
0.1
0.05
0
Coordinate System: EPSG:5667-DHNH/3-degree Gauss-Kruger Zone-3

Figure 7. Horizontal flow velocity magnitudes in the reservoir at 3 m (top row), 16 m (middle row), and 30 m (bottom row) below the water surface during turbine operation (on 2 August 2016, at 10:00 p.m.). The results are produced with the calibrated GPE ($\mathrm{GPE}_{\mathrm{Adjusted}}$) and Delft3D-FLOW model using the maximum likelihoods of calibration parameters according to scenario 0.

outweigh stratification. Also, the maps of the horizontal flow velocity magnitudes (not directions) in Figures 7 and 8 show that turbine and pump operations have a great influence on the reservoir hydrodynamics. That is, the primary importance for reservoir hydrodynamics in the presence of pump and turbine operations is measured flow velocity, which is in line with observations in other studies (Müller et al., 2018). Still, the high values for $\omega_{v_h^{back}}$ and $\omega_{\Delta_h^{back}}$ (Table 3), especially after scenario 2, indicate that the Bayesian calibration is trying to make diffusion a key process. However, the dominance of pump- and turbine-driven flow velocity patterns rather suggests that advection is the real key process.

The RMSEs of the calibrated GPE-surrogate and the full-complexity models, lump-summed over the entire simulation period, are small (cf. Table 4), but do not show the spatiotemporal variation that can be seen in Figures 7 and 8. Thus, the RMSEs compared with the absolute magnitudes suggest that the stochastic calibration yielded very good global model statistics, even though scenario 0 is a physical compromise that necessarily involves the wrong replication of flow velocity and water temperature patterns. In consequence and in light of Figures 5

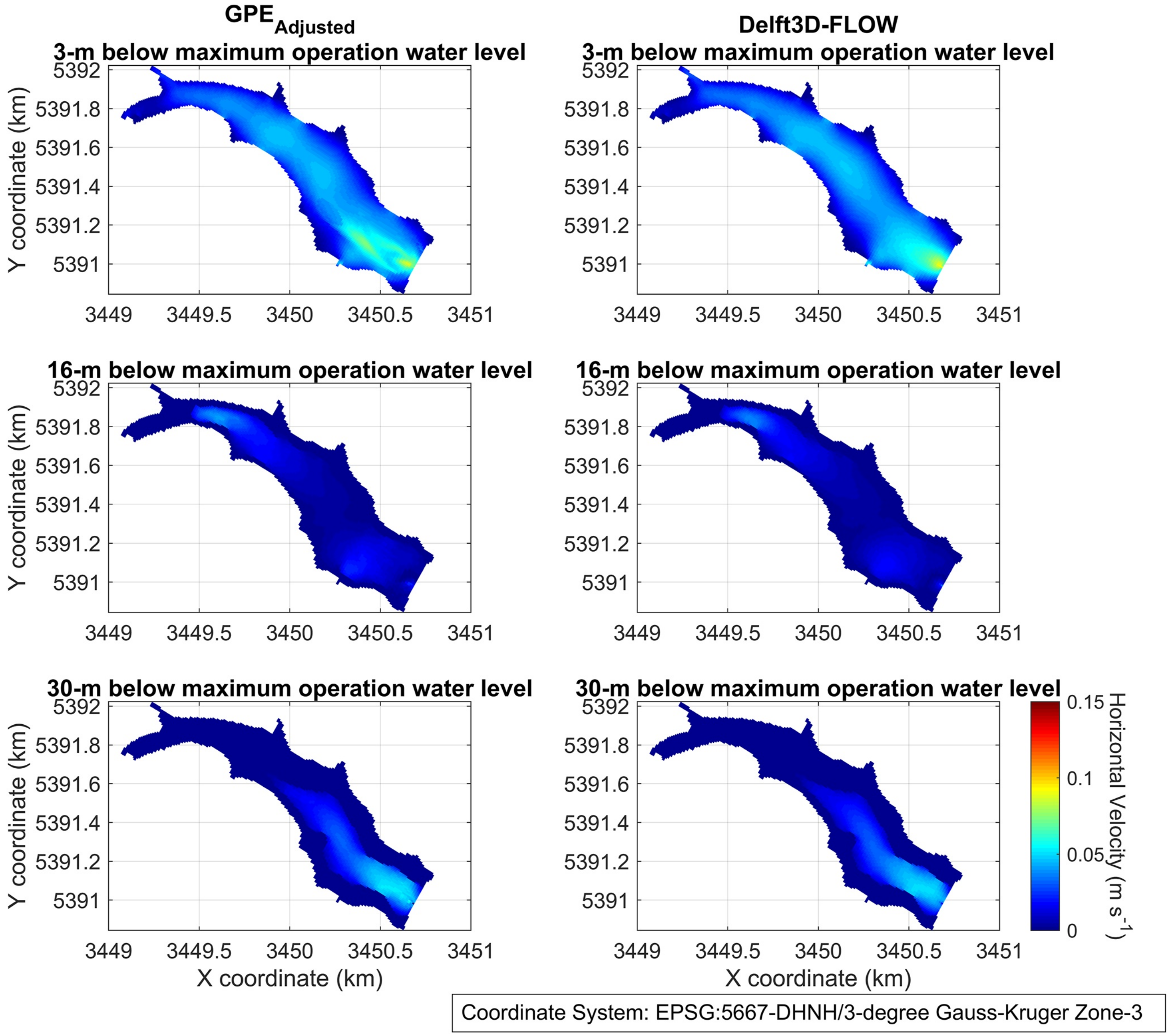
GPE_Adjusted
3-m below maximum operation water level
Delft3D-FLOW
3-m below maximum operation water level
16-m below maximum operation water level
16-m below maximum operation water level
30-m below maximum operation water level
30-m below maximum operation water level
Y coordinate (km)
5392
5391.8
5391.6
5391.4
5391.2
5391
3449
3449.5
3450
3450.5
3451
X coordinate (km)
0.15
0.1
0.05
0
Horizontal Velocity (m s-1)
Coordinate System: EPSG:5667-DHNH/3-degree Gauss-Kruger Zone-3

Figure 8. Horizontal flow velocity magnitudes in the reservoir at 3 m (top row), 16 m (middle row), and 30 m (bottom row) below the water surface during pump operation (on 3 August 2016, at 03:00 a.m.). The results are produced with the calibrated GPE ($\text{GPE}_{\text{Adjusted}}$) and Delft3D-FLOW model using the maximum likelihoods of calibration parameters according to scenario 0.

and 6, the horizontal flow velocity magnitude maps indicate that the small global errors may hide high errors at a detailed scale where local errors potentially cancel each other.

4.2. Efficiency of Bayesian Calibration With a GPE-Surrogate Model

The Bayesian calibration yielded statistically high global model accuracy in terms of RMSE (see Table 4), and the GPE-surrogate model increased computational efficiency almost by a factor of 10^6. Thus, the trained GPE-surrogate model enabled us to test as many as 10^6 parameter sets in the Monte-Carlo rejection sampling. In addition, the GPE-surrogate model reproduced the outcomes of the original model reasonably well and even outperformed the full-complexity model in some cases (see statistics in Table 4), which is in line with other studies using Bayesian calibration (e.g., Beckers et al., 2020). The better performance of the surrogate model is possible because its training builds on the results of the numerical model and measurement data. Still, the surrogate model can only mimic a snapshot state of the numerical model (e.g., the end of the simulation), not physical processes. Thus, the surrogate model cannot write results at a specific point in time, for example, during turbine or pump operation hours.

4.3. Choice of Calibration Parameters and Data

While the background horizontal eddy viscosity ν_h^{back} and diffusivity Δ_h^{back}, as well as the initial water temperature T_{tow} were chosen as constraining calibration parameters, there are many other unconstrained (constant) parameters. We made and had to make simplifications about environmental processes, which are necessary for modeling hydrodynamics of pump-storage reservoirs (Bermúdez et al., 2018; Encinas Fernández et al., 2020; Salehi, 2017). Other factors influencing lake stratification could be, for instance, biotic parameters such as nutrients, carbon, or oxygen concentration (Chanson, 2004). The importance of these factors can be considered small in this study because they are primarily relevant for long-term, multi-year processes (e.g., climate change predictions, cf. Wahl & Peeters, 2014; Woolway et al., 2021). However, other external physical factors, such as solar radiation and wind, are likely to have affected the measurements, but we did not consider their calibration. For instance, we did not consider the adaptation of model parameters accounting for potentially cooling precipitation (Zhang et al., 2020) and assumed the meteorological conditions for stratification (e.g., defined in Bermúdez et al., 2018; Salehi, 2017) as measured deterministic constants even though we interpolated them from a measuring station located 22 km away from the reservoir. The calibration of external meteorological parameters could additionally influence the lake stratification, but it would only be another possibility to achieve physically unlikely fast lake stratification within 6 days. For testing our hypotheses, however, it was sufficient to only calibrate reservoir-internal hydrodynamic parameters, where the environmental boundary conditions needed to allow for lake stratification. Ultimately, we emulated physically more reasonable boundary conditions for testing our hypotheses without calibrating parameters other than ν_h^{back}, Δ_h^{back}, and T_{tow}.

To broaden the range of validity of a reservoir model, the number of calibration parameters and measurement quantities (i.e., the number of columns in **D**) should be possibly high. In this context, the number of measurements per quantity (i.e., the number of rows in **D**) is less important where at least 20 training measurements per quantity were already enabled calibration toward water-temperature measurements, even though it was physically wrong. Thus, for a more holistic replication of system processes, more and different measurement quantities are better than more measurements per quantity at a single location. For instance, it was better to have multiple measurement devices to collect data about different quantities with a smaller number of measurements per device (quantity), rather than using only one measurement device to get a high number of measurements of one quantity only. Still, if a measurement quantity is influenced by the local environment, sufficient measurements should be recorded in every sub-environment. For instance, velocity measurement in a reservoir should be made in fast-flowing regions at the shallow reservoir head, and slow-flowing shallow regions (i.e., the shoreline) as well as in deep regions (center of the reservoir).

4.4. Effects of Varying Measurement Data Usage (Hypothesis (1))

The first hypothesis (1) was that a hydrodynamic model of a reservoir can correctly reproduce measurement data with calibration parameter combinations that are physically not meaningful. For instance, Reynolds-averaging in numerical simulations is known to result in physically non-meaningful representations of 3d flow characteristics, such as negative eddy viscosity (Booij, 2003). To test hypothesis (1) in light of Bayesian calibration, we first consider the global statistics in Table 4 with respect to the three scenarios and their varying calibration constraints through the number of measurement quantities involved.

The RMSE of depth-averaged horizontal flow velocities U is considerably smaller when calibrated with flow velocity data only (scenario 1) than when calibrated with both water temperature T and U in the all-data scenario 0. Likewise, the RMSE of water temperature T is substantially smaller when calibrated with water temperature data only (scenario 2) than when calibrated with both T and U in scenario 0. Thus, the water temperature data *statistically* improved calibration accuracy, even though scenario 2 was designed to be physically the most unreasonable. In this light, the lump-sum statistics in Table 4 suggest that adding information on either quantity improved the model quality more than adding information on both quantities. To unpack this observation, we recall the comparisons of measured and modeled U in Figure 5 and T in Figure 6 at the measurement station. Both quantities (U and T) are represented physically worse when the measurement data lead to contradicting adaptation trends of the three calibration parameters. For instance, the Bayesian calibration (cf. Table 3) suggests that the maximum likelihood for initial water temperature T_{tow} is 26.63°C when the flow velocity measurement data only was considered and 13.42°C when water temperature measurement data only was considered. When both measurement quantities were considered, the maximum likelihood is 18.86°C (i.e., somewhere between the single-quantity considerations).

A similar observation can be made regarding the background horizontal eddy diffusivity Δ_h^{back}. However, the combination of the somewhere-in-the-middle compromises for the maximum likelihoods of T_{tow} and Δ_h^{back} led to a maximum likelihood of background horizontal eddy viscosity ν_h^{back} of 0.91, which is very different than the single measurement data considerations that resulted in 4.81 and 4.93 in scenarios 1 and 2, respectively. An explanation for the differences in the ν_h^{back} maximum likelihoods is that the scenarios 1 and 2 fitted toward U and T based on U and T data only and respectively. Still, calibration with both U and T data resulted in the globally most robust model calibration (e.g., as indicated in the parameter interdependence plots in Figure 4), compared with using either U or T, even in the case of physically more unlikely conditions.

Thus, to mimic the measurement data, the calibration process in the T-only scenario 2 determined values for ν_h^{back} and Δ_h^{back} that are close to the upper limits of the physically relevant calibration ranges (Table 1 compared with Table 3) and subjected to high uncertainty (Figure 3). This perfection comes at the expense of unconstrained parameters that are not reflected in the measurement data. For instance, the exclusive consideration of water temperature measurements in scenario 2 made that T is well reproduced but with physically and statistically unreasonable calibration parameter values that will lead to the false representation of other, not measured quantities. Thus, the calibration of three constrained parameters in scenario 2 led to an overfitted model that is inclined toward the perfection of the simulation of the considered measurement quantity T. Yet, the scenario-2 calibration yielded good performance according to lump-sum statistics (Table 4), even though the calibration parameter results are not meaningful regarding physics and the uncertainty indicated by the posterior distributions (see Figure 4). Therefore, the influence of the data sets used for model calibration is considerable, which makes that there is evidence to not reject hypothesis (1).

In consequence, the answer to the research question of *how does the selection of calibration data affect the (Bayesian) calibration results?* Is that the measured quantities determine the physical processes that Bayesian calibration tries to emulate with the numerical model. In this study, the flow velocity measurements enabled us to replicate advective hydrodynamic patterns related to pump and turbine operations but water temperature data was not adequate to simulate physically unrealistic diffusive heat-driven stratification. In particular, the global physical robustness of the maximum likelihoods (i.e., optima) of calibration parameters, and therefore, the validity and accuracy of the models considering water temperature measurements for unrealistic process emulation were low.

4.5. Identification of Unreasonable Calibration Results (Hypothesis (2))

The maximum likelihoods of the calibration parameters yielded in this study can be considered as very good results regarding lump-sum statistics (Table 4). Good performance according to lump-sum statistics is not only the possible result of a Bayesian calibration but also of subjective trial-and-error calibration, even when the apparently best-fitting calibration parameters are physically not meaningful. However, we hypothesized that the *shape of posterior distributions points to physically unreasonable calibration results* (hypothesis (2)). To test this hypothesis, we recall that the Bayesian calibration goes beyond these lump-sum statistics (Table 4) and shows high calibration parameter interdependence (Figure 4c) with particularly high uncertainties for scenario 2. Also, the maxima of the posterior distributions are less narrow (i.e., more uncertain), and closer to the imposed boundaries of calibration parameters. For instance, ν_h^{back} and Δ_h^{back} are close to the upper limit of 5 in scenarios 1 and 2. As a consequence, the evidence for the rejection of hypothesis (2) is low, but Bayesian calibration can still be subjected to equifinality, and addressing this issue will require further improvement of Bayesian calibration strategies.

Ultimately, the answer to our second research question (*how does Bayesian calibration characterize physically unreasonable calibration results?*) is that the uncertainty expressed through wide-shaped posterior distributions can be linked with poor calibration. Thus, Bayesian calibration has considerable advantages over lump-sum statistics and individual, subjective calibration by indicating physically poor calibration through high statistical interdependence and uncertainty. However, more evidence is required to discern physical and statistical errors, in particular, regarding equifinality and overfitting issues. For instance, injecting physical information into the selection of calibration parameter values (and their combinations), or the early definition of required measurement quantities according to design-of-experiments standards (e.g., Box et al., 2005) might aid in identifying and reducing equifinality because of wrong physical assumptions. In addition, the convergence criteria (skill scores of BME and relative entropy) for the BAL iterations will require to be refined to account for physical relevance.

5. Conclusions

This study shows that Bayesian calibration (assisted by a GPE-surrogate model and BAL) yields statistically satisfactory fitting of a hydrodynamic numerical 3d model of a pump-storage reservoir. The Bayesian calibration strategy provides substantial advantages over expert-informed, subjective trial-and-error calibration regarding multiple aspects and unmasks physical disparities stemming from equifinality.

We calibrate three parameters, notably the background horizontal eddy viscosity, background horizontal eddy diffusivity, and initial water temperature in the reservoir. The calibration is informed by two measurement quantities, in particular, depth-averaged horizontal flow velocity and water temperature at a measurement station. Three scenarios consider both measured quantities in combination and individually. In an all-data scenario, we calibrate with both flow velocity and water temperature data, resulting in statistically meaningful calibration results. The next scenario exclusively considers flow velocity measurements and is physically more reasonable to simulate advective 2d currents due to pump and turbine operations. The last scenario considers the water temperature measurements only and is designed to not work physically correctly because of a too short modeling period for the development of thermal lake stratification. Yet, the last scenario yields good results regarding lump-sum statistics.

This study builds on Bayesian calibration indicating calibration parameter uncertainties and their interdependencies through characteristics of their posterior distributions. The posterior distributions of the calibration parameters resulting from the scenario only using water temperature data show particularly high uncertainties that we do not observe in the other two scenarios. Thus, physically incorrect overfitting is possible even with an objective calibration approach, but the Bayesian calibration strategy aids in detecting non-meaningful calibration results.

In addition, we show that the individual consideration of exclusively one measured quantity may yield better lump-sum statistics and lower global errors regarding the considered measurement quantity than the combined consideration of both quantities. While the overall model performance improved with fewer measurement quantities, using exclusively one measurement quantity (i.e., either depth-averaged flow velocity or water temperature) leads to the overfitting of the model toward the measured quantity. We also obtain statistically acceptable calibration results with as few as 20 training points for water temperature only. Thus, this study suggests that the number of measured quantities available for model calibration is more important to represent particular physical processes than the number of features (i.e., measurements) per quantity.

Finally, Bayesian calibration is an efficient and objective technique for calibrating hydrodynamic models of reservoirs. Within considerably reduced computing time, many possible combinations of calibration parameter values (here: 10^6) can be tested. However, to not only identify but also address equifinality, a future challenge will be to implement additional criteria for selecting calibration parameter value candidates and combinations into Bayesian calibration. For instance, skill scores considering the statistical quality and also physical relevance (i.e., beyond BME or relative entropy) could be a starting point. Such skill scores could infuse physical plausibility into Bayesian calibration.

Notation

Δ_h^{back}	background horizontal eddy diffusivity
Δ_h	horizontal eddy diffusivity
Δ_v	vertical eddy diffusivity
ε	measurement error/residual
ν_h^{back}	background horizontal eddy viscosity
ν_h	horizontal eddy viscosity
ν_v	vertical eddy viscosity

Ω	Vector of calibration parameter values
$\omega_{\Delta_h^{back}}$	parameter space of background horizontal eddy diffusivity
$\omega_{\nu_h^{back}}$	parameter space of background horizontal eddy viscosity
$\omega_{T_{tow}}$	parameter space of water temperature at the intake tower
$\sigma^2_{\varepsilon,D}$	error variance of measurement data
ADCP	acoustic Doppler current profiler
BAL	Bayesian active learning
BME	Bayesian model evidence
$\mathbf{D}$	measurement data set
D_{KL}	relative entropy (Kullback-Leibler divergence)
GPE	Gaussian process emulator
$\mathbb{E}()$	expected value of an expression or distribution
$\mathcal{M}$	full-complexity model response
n	number of measurements (i.e., number of rows in $\mathbf{D}$))
$\mathbf{R}$	diagonal (co-)variance matrix of measurement errors (sized $n \times n$)
RMSE	root-mean-square error
$\mathcal{S}$	response of the GPE-surrogate model
SR	Schwarzenbach reservoir
T_{tow}	water temperature at the intake tower
T	water temperature (in general)
t	time
U	depth-averaged flow velocity

Data Availability Statement

The methods for model reduction and Bayesian updating were published in the references Oladyshkin et al. (2013) and Oladyshkin and Nowak (2012). The codes for the Bayesian active learning with GPE are available at https://github.com/sergiocallau/ManuscriptSBT/releases/tag/v0.1 (Callau & Schwindt, 2022). Data at the numerical model boundaries and a keyhole markup language file indicating the location of the measuring station are provided at https://github.com/sschwindt/schwarzenbach-bc/archive/refs/tags/boundary-data.zip (Schwindt, 2022).

Acknowledgments

Stefan Haun is indebted to the Baden-Württemberg Stiftung for the financial support from the Elite program for Postdocs. The authors thank the many helpers involved in the fieldwork for acquiring the measurement data and preparing the numerical model, in particular, Frank Peeters, Pan Ei Ei Phyoe, and Jorge Encinas. The authors also thank the German Research Foundation (DFG) for the support of the project within the Cluster of Excellence "Data-Integrated Simulation Science" (EXC 2075) at the University of Stuttgart, and for supporting this work by SFB 1313, Project Number 327154368. In addition, the authors thank EnBW, which kindly supported this study with information. This study partially results from the CHARM (Challenges of Reservoir Management - Meeting Environmental and Social Requirements) project, which is part of the Water Research Network Baden-Württemberg. It is funded by the Ministry of Science, Research and Arts of the federal state of Baden-Württemberg, Germany. Open Access funding enabled and organized by Projekt DEAL.

References

Afshar, A., Shojaei, N., & Sagharjooghifarahani, M. (2013). Multiobjective calibration of reservoir water quality modeling using Multiobjective Particle Swarm Optimization (MOPSO). *Water Resources Management*, *27*(7), 1931–1947. https://doi.org/10.1007/s11269-013-0263-x

Ahlfeld, D., Joaquin, A., Tobiason, J., & Mas, D. (2003). Case study: Impact of reservoir stratification on interflow travel time. *Journal of Hydraulic Engineering*, *129*(12), 966–975. https://doi.org/10.1061/(ASCE)0733-9429(2003)129:12(966)

Beckers, F., Heredia, A., Noack, M., Nowak, W., Wieprecht, S., & Oladyshkin, S. (2020). Bayesian calibration and validation of a large-scale and time-demanding sediment transport model. *Water Resources Research*, *56*(7). https://doi.org/10.1029/2019WR026966

Beckers, F., Mohammadi, F., Heredia, A., Guthke, A., Noack, M., Wieprecht, S., et al. (2021). Machine learning and surrogate modeling in morphodynamic sediment transport research: Advantages and challenges. In M. Kalinowska, P. Rowinski, T. Okruszko, & M. Nones (Eds.), *IAHR 2020—Abstract book* (p. 2). https://doi.org/10.24425/136660

Bellman, R. (1957). *Dynamic programing*. Princeton University Press. https://doi.org/10.1515/9781400835386

Bermúdez, M., Cea, L., Puertas, J., Rodríguez, N., & Baztán, J. (2018). Numerical modeling of the impact of a pumped-storage hydroelectric power plant on the reservoirs' thermal stratification structure: A case study in NW Spain. *Environmental Modeling & Assessment*, *23*(1), 71–85. https://doi.org/10.1007/s10666-017-9557-3

Beven, K., & Binley, A. (1992). The future of distributed models: Model calibration and uncertainty prediction. *Hydrological Processes*, *6*(3), 279–298. https://doi.org/10.1002/hyp.3360060305

Blazek, J. (2005). Chapter 7—Turbulence modeling. In J. Blazek (Ed.), *Computational fluid dynamics: Principles and applications*, 2nd ed., (pp. 227–270). Elsevier Science. https://doi.org/10.1016/B978-008044506-9/50009-6

Booij, R. (2003). Measurements and large eddy simulations of the flows in some curved flumes. *Journal of Turbulence*, *4*, 1–16. https://doi.org/10.1088/1468-5248/4/1/008

Box, G. E. P., Hunter, S. J., & Hunter, W. G. (2005). *Statistics for experimenters: Design, innovation, and discovery* (2nd ed.). John Wiley & Sons.

Box, G. E. P., & Tiao, G. C. (1973). *Bayesian inference in statistical analysis*. Addison-Wesley. (OCLC: 585122).

Brodeur, Z. P., Herman, J. D., & Steinschneider, S. (2020). Bootstrap aggregation and cross-validation methods to reduce overfitting in reservoir control policy search. *Water Resources Research*, *56*(8), e2020WR027184. https://doi.org/10.1029/2020WR027184

Busby, D. (2009). Hierarchical adaptive experimental design for Gaussian process emulators. *Reliability Engineering & System Safety*, *94*(7), 1183–1193. https://doi.org/10.1016/j.ress.2008.07.007

Callau, S., & Schwindt, S. (2022). Manuscriptsbt (github repository). *Github*. Retrieved from https://github.com/sergiocallau/ManuscriptSBT/releases/tag/v0.1

Camacho, R. A., & Martin, J. L. (2013). Bayesian Monte Carlo for evaluation of uncertainty in hydrodynamic models of coastal systems. *Journal of Coastal Research*, *65*(10065), 886–891. https://doi.org/10.2112/SI65-150.1

Camacho, R. A., Martin, J. L., McAnally, W., Diaz-Ramirez, J., Rodriguez, H., Sucsy, P., & Zhang, S. (2015). A comparison of Bayesian methods for uncertainty analysis in hydraulic and hydrodynamic modeling. *Journal of the American Water Resources Association*, *51*(5), 1372–1393. https://doi.org/10.1111/1752-1688.12319

Chanson, H. (2004). Introduction to mixing and dispersion in natural waterways. In H. Chanson (Ed.), *Environmental hydraulics of open channel flows* (pp. 37–48). Butterworth-Heinemann. https://doi.org/10.1016/B978-075066165-2.50035-7

Chanudet, V., Fabre, V., & van der Kaaij, T. (2012). Application of a three-dimensional hydrodynamic model to the Nam Theun 2 Reservoir (Lao PDR). *Journal of Great Lakes Research*, *38*(2), 260–269. https://doi.org/10.1016/j.jglr.2012.01.008

Deltares. (2022). *Simulation of multi-dimensional hydrodynamic flows and transport phenomena, including sediments*. Deltares, Delft, The Netherland.

Dissanayake, P., Hofmann, H., & Peeters, F. (2019). Comparison of results from two 3D hydrodynamic models with field data: Internal seiches and horizontal currents. *Inland Waters*, *9*(2), 239–260. https://doi.org/10.1080/20442041.2019.1580079

Dong, F., Mi, C., Hupfer, M., Lindenschmidt, K.-E., Peng, W., Liu, X., & Rinke, K. (2020). Assessing vertical diffusion in a stratified lake using a three-dimensional hydrodynamic model. *Hydrological Processes*, *34*(5), 1131–1143. https://doi.org/10.1002/hyp.13653

DWD. (2021). *Historical hourly station observations of wind speed and wind direction for Germany, v21.3*. Climate Data Center (CDC). Retrieved from https://opendata.dwd.de/climate_environment/CDC/observations_germany/climate/hourly/wind/historical/DESCRIPTION_obsgermany_climate_hourly_wind_historical_en.pdf

Encinas Fernández, J., Hofmann, H., & Peeters, F. (2020). Diurnal pumped-storage operation minimizes methane ebullition fluxes from hydropower reservoirs. *Water Resources Research*, *56*(12). https://doi.org/10.1029/2020WR027221

Forrester, A. I. J., Sóbester, A., & Keane, A. J. (2008). *Engineering design via surrogate modeling: A practical guide* (1st ed.). Wiley. Retrieved 5 November 2020 from https://onlinelibrary.wiley.com/doi/book/10.1002/9780470770801

Goudsmit, G.-H., Burchard, H., Peeters, F., & Wüest, A. (2002). Application of $k - \epsilon$ turbulence models to enclosed basins: The role of internal seiches: Application of $k - \epsilon$ turbulence models. *Journal of Geophysical Research: Oceans*, *107*(C12), 1–13. https://doi.org/10.1029/2001JC000954

Hodges, B. R., Imberger, J., & Laval, B. (2000). Modeling the hydrodynamics of stratified lakes. In *Proceedings of the hydro-informatics conference* (p. 14). Iowa Institute of Hydraulic Research.

Hutchinson, G. E., & Löffler, H. (1956). The thermal classification of lakes. *Proceeding of the National Academy of Science*, *42*(2), 84–86. https://doi.org/10.1073/pnas.42.2.84

Jones, D. R., Schonlau, M., & Welch, W. J. (1998). Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, *13*(4), 455–492. https://doi.org/10.1023/A:1008306431147

Katopodes, N. D. (2019). Chapter 10—Geophysical effects. In N. D. Katopodes (Ed.), *Free-surface flow* (pp. 710–779). Butterworth-Heinemann. https://doi.org/10.1016/B978-0-12-815489-2.00010-1

Kennedy, M. C., & O'Hagan, A. (2001). Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B*, *63*(3), 425–464. https://doi.org/10.1111/1467-9868.00294

Kerimoglu, O., & Rinke, K. (2013). Stratification dynamics in a shallow reservoir under different hydro-meteorological scenarios and operational strategies. *Water Resources Research*, *49*(11), 7518–7527. https://doi.org/10.1002/2013WR013520

Kim, Y.-J., & Park, C.-S. (2016). Stepwise deterministic and stochastic calibration of an energy simulation model for an existing building. *Energy and Buildings*, *133*, 455–468. https://doi.org/10.1016/j.enbuild.2016.10.009

Kim, Y.-J., Yoon, S.-H., & Park, C.-S. (2013). Stochastic comparison between simplified energy calculation and dynamic simulation. *Energy and Buildings*, *64*, 332–342. https://doi.org/10.1016/j.enbuild.2013.05.026

Kirillin, G., & Shatwell, T. (2016). Generalized scaling of seasonal thermal stratification in lakes. *Earth-Science Reviews*, *161*, 179–190. https://doi.org/10.1016/j.earscirev.2016.08.008

Koşucu, M. M., Demirel, M. C., Kirca, V. S. O., & Özger, M. (2019). Hydrodynamic and hydrographic modeling of Istanbul strait. *Processes*, *7*(10), 710. https://doi.org/10.3390/pr7100710

Kundu, P. K., & Cohen, I. M. (2008). *Fluid mechanics* (4th ed.). Elsevier Inc. San Diego, USA.

Landesanstalt für Umwelt Baden-Württemberg (LUBW). (2016). INSPIRE Gewässernetz im UIS Baden-Württemberg. Retrieved from http://rips-dienste.lubw.baden-wuerttemberg.de/rips/ripsservices/apps/uis/metadaten/beschreibung.aspx?typ=0&uuid=7ef11b78-cd06-4cb8-8c26-9f45d410d09c

Leifsson, L., Hermannsson, E., & Koziel, S. (2015). Optimal shape design of multi-element trawl-doors using local surrogate models. *Journal of Computational Science*, *10*, 55–62. https://doi.org/10.1016/j.jocs.2015.01.006

Li, Y., Acharya, K., Chen, D., & Stone, M. (2010). Modeling water ages and thermal structure of Lake Mead under changing water levels. *Lake and Reservoir Management*, *26*(4), 258–272. https://doi.org/10.1080/07438141.2010.541326

Li, Y., Tang, C., Zhu, J., Pan, B., Anim, D. O., Ji, Y., et al. (2015). Parametric uncertainty and sensitivity analysis of hydrodynamic processes for a large shallow freshwater lake. *Hydrological Sciences Journal*, *60*(6), 1078–1095. https://doi.org/10.1080/02626667.2014.948444

Lindim, C., Pinho, J. L., & Vieira, J. M. P. (2011). Analysis of spatial and temporal patterns in a large reservoir using water quality and hydrodynamic modeling. *Ecological Modeling*, *222*(14), 2485–2494. https://doi.org/10.1016/j.ecolmodel.2010.07.019

Masoumi, F., Najjar-Ghabel, S., & Salimi, N. (2021). Automatic calibration of the two-dimensional hydrodynamic and water quality model using sequential uncertainty fitting approach. *Environmental Monitoring and Assessment*, *193*(2), 67. https://doi.org/10.1007/s10661-020-08831-z

Matlab. (2018). Fit a Gaussian process regression (GPR) model—MATLAB fitrgp. Retrieved from https://www.mathworks.com/help/stats/fitrgp.html

Mockus, J. (1994). Application of Bayesian approach to numerical methods of global and stochastic optimization. *Journal of Global Optimization*, *4*(4), 347–365. https://doi.org/10.1007/BF01099263

Morgan, J. A., Kumar, N., Horner-Devine, A. R., Ahrendt, S., Istanbullouglu, E., & Bandaragoda, C. (2020). The use of a morphological acceleration factor in the simulation of large-scale fluvial morphodynamics. *Geomorphology*, *356*, 107088. https://doi.org/10.1016/j.geomorph.2020.107088

Mouris, K., Acuna Espinoza, E., Schwindt, S., Mohammadi, F., Haun, S., Wieprecht, S., & Oladyshkin, S. (2023). Stability criteria for Bayesian calibration of reservoir sedimentation models. *Modeling Earth Systems and Environment*, 1–19. https://doi.org/10.1007/s40808-023-01712-7

Müller, M., De Cesare, G., & Schleiss, A. J. (2018). Flow field in a reservoir subject to pumped-storage operation—In situ measurement and numerical modeling. *Journal of Applied Water Engineering and Research*, *6*(2), 109–124. https://doi.org/10.1080/23249676.2016.1224692

Oberkampf, W. L., Trucano, T. G., & Hirsch, C. (2004). Verification, validation, and predictive capability in computational engineering and physics. *Applied Mechanics Reviews*, *57*(5), 345–384. https://doi.org/10.1115/1.1767847

Oladyshkin, S., Class, H., & Nowak, W. (2013). Bayesian updating via bootstrap filtering combined with data-driven polynomial chaos expansions: Methodology and application to history matching for carbon dioxide storage in geological formations. *Computational Geosciences*, *17*(4), 671–687. https://doi.org/10.1007/s10596-013-9350-6

Oladyshkin, S., Mohammadi, F., Kroeker, I., & Nowak, W. (2020). Bayesian active learning for the Gaussian process emulator using information theory. *Entropy*, *22*(8), 890. https://doi.org/10.3390/e22080890

Oladyshkin, S., & Nowak, W. (2012). Data-driven uncertainty quantification using the arbitrary polynomial chaos expansion. *Reliability Engineering & System Safety*, *106*, 179–190. https://doi.org/10.1016/j.ress.2012.05.002

Oladyshkin, S., & Nowak, W. (2019). The connection between Bayesian inference and information theory for model selection, information gain, and experimental design. *Entropy*, *21*(11), 1081. https://doi.org/10.3390/e21111081

Platzek, F. W., Stelling, G. S., Jankowski, J. A., & Pietrzak, J. D. (2014). Accurate vertical profiles of turbulent flow in *z*-layer models. *Water Resources Research*, *50*(3), 2191–2211. https://doi.org/10.1002/2013WR014411

Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian processes for machine learning*. MIT Press. Retrieved from https://gaussianprocess.org/gpml/chapters/RW.pdf

Salehi, M. (2017). Thermal recirculation modeling for power plants in an estuarine environment. *Journal of Marine Science and Engineering*, *5*(1), 5. https://doi.org/10.3390/jmse5010005

Schwindt, S. (2022). schwarzenbach-bc (github repository). Github. Retrieved from https://github.com/sschwindt/schwarzenbach-bc/releases/tag/boundary-data

Shoarinezhad, V., Wieprecht, S., & Haun, S. (2020). Comparison of local and global optimization methods for calibration of a 3D morphodynamic model of a curved channel. *Water*, *12*(5), 1333. https://doi.org/10.3390/w12051333

Smith, A. F. M., & Gelfand, A. E. (1992). Bayesian statistics without tears: A sampling-resampling perspective. *The American Statistician*, *46*(2), 84–88. https://doi.org/10.2307/2684170

Snucins, E., & John, G. (2000). Interannual variation in the thermal structure of clear and colored lakes. *Limnology & Oceanography*, *45*(7), 1639–1646. https://doi.org/10.4319/lo.2000.45.7.1639

Soares, L. M. V., do Carmo Calijuri, M., das Graças Silva, T. F., Moraes Novo, E. M. L., Cairo, C. T., & Barbosa, C. C. F. (2020). A parameterization strategy for hydrodynamic modeling of a cascade of poorly monitored reservoirs in Brazil. *Environmental Modeling & Software*, *134*, 104803. https://doi.org/10.1016/j.envsoft.2020.104803

Sommer, U., Gliwicz, Z. M., Lampert, W., & Duncan, A. (1986). The peg-model of seasonal succession of planktonic events in fresh waters. *Archiv für Hydrobiologie*, *106*(4), 433–471. https://doi.org/10.1127/archiv-hydrobiol/106/1986/433

Thackeray, S. J., Henrys, P. A., Feuchtmayr, H., Jones, I. D., Maberly, S. C., & Winfield, I. J. (2013). Food web de-synchronization in England's largest lake: An assessment based on multiple phenological metrics. *Global Change Biology*, *19*(12), 3568–3580. https://doi.org/10.1111/gcb.12326

Wahl, B., & Peeters, F. (2014). Effect of climatic changes on stratification and deep-water renewal in Lake Constance assessed by sensitivity studies with a 3D hydrodynamic model. *Limnology & Oceanography*, *59*(3), 1035–1052. https://doi.org/10.4319/lo.2014.59.3.1035

Wang, L., Xu, B., Zhang, C., Fu, G., Chen, X., Zheng, Y., & Zhang, J. (2022). Surface water temperature prediction in large-deep reservoirs using a long short-term memory model. *Ecological Indicators*, *134*, 108491. https://doi.org/10.1016/j.ecolind.2021.108491

Woolway, R. I., Sharma, S., Weyhenmeyer, G. A., Debolskiy, A., Golub, M., Mercado-Bettín, D., et al. (2021). Phenological shifts in lake stratification under climate change. *Nature Communications*, *12*(1), 2318. https://doi.org/10.1038/s41467-021-22657-4

Wright, K. A., Goodman, D. H., Som, N. A., Alvarez, J., Martin, A., & Hardy, T. B. (2017). Improving hydrodynamic modeling: An analytical framework for assessment of two-dimensional hydrodynamic models. *River Research and Applications*, *33*(1), 170–181. https://doi.org/10.1002/rra.3067

Yang, F., Liang, D., & Xiao, Y. (2018). Influence of Boussinesq coefficient on depth-averaged modeling of rapid flows. *Journal of Hydrology*, *559*, 909–919. https://doi.org/10.1016/j.jhydrol.2018.01.053

Zarfl, C., Lumsdon, A. E., Berlekamp, J., Tydecks, L., & Tockner, K. (2015). A global boom in hydropower dam construction. *Aquatic Sciences*, *77*(1), 161–170. https://doi.org/10.1007/s00027-014-0377-0

Zhang, F., Zhang, H., Bertone, E., Stewart, R., Lemckert, C., & Cinque, K. (2020). Numerical study of the thermal structure of a stratified temperate monomictic drinking water reservoir. *Journal of Hydrology: Regional Studies*, *30*, 100699. https://doi.org/10.1016/j.ejrh.2020.100699

Zhen-Gang, J. (2008). *Hydrodynamics and water quality: Modeling rivers, lakes, and estuarie*s (1st ed., Vol. 89). John Wiley & Sons, Inc. https://doi.org/10.1029/2008EO390008