Calibration of a complex hydro-ecological model through Approximate Bayesian Computation and Random Forest combined with sensitivity analysis

Francesco Piccioni, Céline Casenave, Meïli Baragatti, Bertrand Cloez, Vinçon-Leite Brigitte

▶ **To cite this version:**

Francesco Piccioni, Céline Casenave, Meïli Baragatti, Bertrand Cloez, Vinçon-Leite Brigitte. Calibration of a complex hydro-ecological model through Approximate Bayesian Computation and Random Forest combined with sensitivity analysis. Ecological Informatics, 2022, 71, 23p. 10.1016/j.ecoinf.2022.101764 . hal-03819623

HAL Id: hal-03819623
https://hal.inrae.fr/hal-03819623

Submitted on 7 Sep 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Calibration of a complex hydro-ecological model through Approximate Bayesian Computation and Random Forest combined with sensitivity analysis

Francesco Piccioni[a], Céline Casenave[b,**], Meïli Baragatti[b], Bertrand Cloez[b] and Brigitte Vinçon-Leite[a]

[a]LEESU, Ecole des Ponts ParisTech, Univ Paris Est Créteil, Marne-la-Vallée, France
[b]MISTEA, Université Montpellier, INRAE, Institut Agro, Montpellier, France

ARTICLE INFO

Keywords:
automated calibration
hydro-ecological modelling
Approximate Bayesian Computation (ABC)
random forest
sensitivity analysis

ABSTRACT

An automated calibration method is proposed and applied to the complex hydro-ecological model Delft3D-BLOOM which is calibrated from monitoring data of the lake Champs-sur-Marne, a small shallow urban lake in the Paris region (France). This method (ABC-RF-SA) combines Approximate Bayesian Computation (ABC) with the machine learning algorithm Random Forest (RF) and a Sensitivity Analysis (SA) of the model parameters. Three target variables are used (total chlorophyll, cyanobacteria and dissolved oxygen concentration) to calibrate 133 parameters. ABC-RF-SA is first applied on a set of simulated observations to validate the methodology. It is then applied on a real set of high-frequency observations recorded during about two weeks on the lake Champs-sur-Marne. The methodology is also compared to standard ABC and ABC-RF formulations. Only ABC-RF-SA allowed the model to reproduce the observed biogeochemical dynamics. The coupling of ABC with RF and SA thus appears crucial for its application to complex hydro-ecological models.

1. Introduction

Modelling biogeochemical cycling and phytoplankton dynamics in aquatic ecosystems is a complex task. It implies taking into account many processes that belong to different scientific fields, ranging from physics to biology to chemistry. Mechanistic hydro-ecological models, which seek to include all these processes, are often very complex and over-parameterized [36, 27, 37, 49, 25]. In addition, most parameters are difficult to measure directly by field observations. Reference values for key model parameters can be found in scientific literature but they are uncertain and often have a wide range of variability [e.g. 27, 18], which affects the reliability of the models.

For these reasons, sensitivity analysis, calibration and validation of complex hydro-ecological models are important tasks. However, Shimoda and Arhonditsis [38] showed that only half of the publications published between 1980 and 2012 include a proper sensitivity analysis, and when calibration is performed, it is mostly done by trial-and-error. Yet, while the results of trial-and-error calibration depend heavily on the skill and knowledge of the modeler [25], automated calibration can reduce model uncertainty and simultaneously allow to carry out a sensitivity analysis of the model parameters. However, it is rarely applied for complex hydro-ecological models, especially when they are three-dimensional. In the literature, automated calibration is only applied on 0D or 1D models, most often by optimization or Monte Carlo and Bayesian inference [49, and references therein].

There are several reasons for this. First, automated calibration strategies are generally computationally expensive. They often require a large number of model runs and their computational cost increases with the number of parameters to be estimated, which hinders their application to complex hydro-ecological models. Moreover, in limnological studies, data traditionally come from field campaigns which, although regular, lead at best to sparse datasets that are not well suited to automated calibration strategies.

If the available data set is rich enough, a wide range of approaches and techniques can be applied for automated calibration. This includes various optimization algorithms, such as Newton's algorithms and genetic algorithms (e.g., Particle Swarm Optimization), as well as Bayesian parameter inference algorithms [26].

However, classical Bayesian parameter inference is often problematic for complex mechanistic models. For such models, the likelihood function is analytically intractable and its evaluation by computational methods is extremely computationally demanding. Approximate Bayesian Computation (ABC) is an innovative and promising technique for parameter inference, rooted in Bayesian statistics, which has the great advantage of bypassing the computation of the likelihood function. It requires a large number of model runs with different sets of parameters obtained by random sampling according to user-defined prior distributions. This set of simulations is used as a training dataset, in order to approximate the posterior probability distribution function of the parameters. Different methods can be used for this purpose, among which machine learning techniques. For example, the use of random forests has been recently proposed and seems to be particularly advantageous [35].

Corresponding author
francesco.piccioni@enpc.fr (F. Piccioni); celine.casenave@inrae.fr (C. Casenave); meili.baragatti@supagro.fr (M. Baragatti); bertrand.cloez@inrae.fr (B. Cloez); b.vincon-leite@enpc.fr (B. Vinçon-Leite)
ORCID(s): 0000-0003-0489-4665 (C. Casenave)

Like most calibration techniques, ABC is expensive in terms of computational effort. However, it offers a good compromise between the number of parameters to identify and the number of model evaluations [26]. Moreover, compared to other calibration techniques such as evolutionary algorithms, it has the advantage of not being iterative. This allows the model evaluations to be performed in parallel, which is particularly interesting in the case of complex hydro-ecological models with high simulation times. ABC has already been applied to complex statistical models [29] and individual-based ecological models (e.g. [24, 47, 16]) with a few dozen parameters, but, to the best of our knowledge, never to a complex process-based model with more than 100 parameters.

In this work, an innovative method for automated calibration is proposed and applied to the complex hydro-ecological model Delft3D-BLOOM [14]. This method (called ABC-RF with SA hereafter) is based on the ABC-RF (Approximate Bayesian Computation - Random Forest) method proposed in [35] which is combined with a sensitivity analysis (SA) of the model parameters.

The main computational cost of ABC is the large number of model simulations that must be performed in order to build a robust training dataset to apply the ABC. In this study, the availability of high-frequency data aggregated to an hourly time step, allowed the calibration effort to be focused on a 16-day simulation, greatly reducing the computational time while focusing on the model's ability to simulate short-term variations. The aim of this study is to test the ability of the ABC-RF with SA to reproduce a series of observations with a complex biogeochemical model that involves a large number (133) of parameters. Three target variables are considered in this calibration procedure: total chlorophyll, phycocyanin and dissolved oxygen. These variables are representative of biological processes in aquatic ecosystems. Total chlorophyll is an indicator of total phytoplankton biomass and is the variable on which most alert guidelines for monitoring harmful algal blooms are based. Phycocyanin is a pigment specific to cyanobacteria that can be considered an indicator of their abundance. Finally, dissolved oxygen concentration, especially in a eutrophic environment, can be considered as a resultant variable of various processes: growth, mortality, decomposition of organic matter, and nutrient recycling.

The method ABC-RF with SA is first applied on a set of simulated data to validate the method and test its ability to reproduce both the simulated data and the parameter values. It is then applied on a real observation dataset of the lake Champs-sur-Marne, a small shallow lake of the Paris region. The standard ABC method and the ABC-Random Forest (ABC-RF) method are also applied for comparison.

2. Materials and methods

2.1. Dataset and study site

The lake Champs-sur-Marne is a small and shallow lake located in the Great Paris region. Its surface area is of 0.12 km^2, and the average and maximum depths are about 2.5 m and 4 m respectively. As shown in Fig. 1, the lake has no inflow or outflow and is fed primarily by groundwater from the Marne River that flows north of the water body.

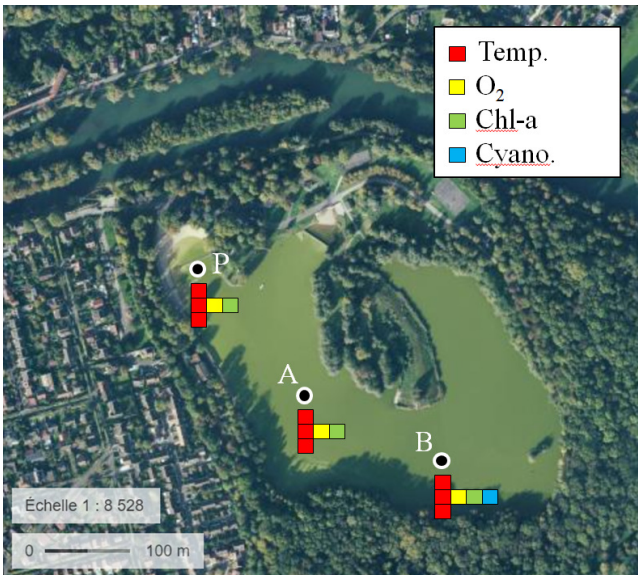

Figure 1: Satellite picture of the lake Champs-sur-Marne (source: *géoportail.fr*) and sketch of the measuring system at the three locations (A, B and P).

The lake Champs-sur-Marne suffers from strong eutrophication conditions that lead to a succession of serious harmful algal blooms between the months of February and October. The lake is a valuable recreational area for the neighbourhood. However, especially during the summer months, the dominance of toxic cyanobacteria such as *Microcystis* and *Aphanizomenon* often leads to bathing bans and access restrictions to the lake.

For these reasons, the lake is monitored by both periodic surveys and high-frequency automated *in situ* measurements of relevant physico-chemical variables at three measuring sites [46]. Each measurement site is equipped with sensors located at three depths: at the surface (0.5 m depth), in the middle (1.5 m depth) and at the bottom (2.5 m depth) layers (see Fig. 1). Water temperature is measured by the SP2T10 sensor (nke INSTRUMENT®) at the surface and bottom layers, with a precision of 0.02 °C and a resolution of 0.05 °C. At the middle of the water column, a multi-parameter sensor (MPx, nke INSTRUMENT®) measures, in addition to the water temperature, the values of oxygen, total chlorophyll and, at site B only, phycocyanin concentration. Phycocyanin is a pigment specific to cyanobacteria that is commonly used as a proxy for their biomass [7]. All measurements are collected every 10 minutes. The resolution and precision of the multi-parameter sensor are given in table 1.

2.2. Model configuration

Delft3D is a well established and reliable modelling tool for hydrodynamic and water quality simulations. It includes different modules that cover a wide range of applications.

Table 1
Resolution and precision of the high-frequency Mpx multi-parameter sensor implemented at site B [46].

Variable	Resolution	Precision
Temperature	0.01°C	0.05°C
Dissolved Oxygen	0.01%	1%
Total chlorophyll	<0.008 μg.L^{-1}	0.03 μg.L^{-1}
Phycocyanin	<0.008 μg.L^{-1}	2 μg.L^{-1}

Phytoplankton concentration is modelled in Delft3D as the result of two distinct processes, transport and biomass production, handled by the FLOW and BLOOM models respectively. FLOW is a hydrodynamic model that solves the Reynolds averaged Navier-Stokes equations for an incompressible fluid. It has already been tested in various scenarios and is extremely reliable [31, 10, 42]. BLOOM is a biogeochemical model that computes phytoplankton biomass in aquatic ecosystems, based on a linear programming algorithm designed to optimize biomass production as a function of local nutrient, light and temperature conditions.

The 3D hydrodynamic model Deltf3D-FLOW has been configured on the lake Champs-sur-Marne. The bathymetry was interpolated from *in situ* measurements. The horizontal mesh is composed of 813 square cells of 10 m side. Twelve horizontal layers with a fixed thickness of 27 cm were used for the discretization of the vertical axis.The choice of parallel horizontal layers (rather than σ-layers) avoids artificial mixing, and improves the model results in terms of water temperature distribution [21]. The k-ϵ turbulence closure model was used for the computation of turbulent eddy viscosity and diffusivity. Background values for horizontal viscosity and diffusivity were set to 0.0025 m^2.s^{-1}, according to literature values [42] and mesh cell size. Background values were set to zero [m^2.s^{-1}] for vertical viscosity and diffusivity. The heat budget at the air-water interface was calculated using the Ocean model. It requires as inputs time series of relative humidity [-], air temperature [°C], net solar radiation [J.s^{-1}.m^{-2}], sky cloudiness [-], wind speed [m.s^{-1}] and wind direction [°N].

The water temperature simulated with the hydrodynamic model was compared to the high-frequency observations recorded at measurement site B for the surface and bottom layers. The model correctly reproduced the water temperature at both layers. The computed root mean square error (RMSE) between model results and high-frequency observations is only 0.5°C for the surface layer, and 0.6°C for the bottom layer.

The BLOOM module uses the simulation results from the FLOW module (current, water temperature), but is run separately from the hydrodynamic simulation. Four main modules are activated in the configuration implemented in this study: oxygen and Biological Oxygen Demand (BOD), dissolved inorganic matter, organic matter and phytoplankton. Each module contains many variables, which are listed in table 2. In particular, the phytoplankton module includes four algal groups commonly present in the lake Champs-sur-Marne: green algae, diatoms, flagellates and cyanobacteria.

Table 2
Modules and variables activated in the configuration of the biogeochemical model.

Module	Variables
Oxygen-BOD	Dissolved oxygen
Particulate and dissolved inorganic matter	Inorganic matter (IM1)
	Ammonium
	Nitrate
	Ortho-phosphate
	Adsorbed ortho-phosphate
	Dissolved Silica
	Opal-Si
Organic matter	POC, fractions 1,2,3,4
	PON, fractions 1,2,3,4
	POP, fractions 1,2,3,4
	DOC
	DON
	DOP
	Detritus C in sediment layer
	Detritus N in sediment layer
	Detritus P in sediment layer
Phytoplankton	Cyanobacteria
	Freshwater diatoms
	Freshwater flagellates
	Green algae

In the biogeochemical cycle, the activated variables depend on each other through a large number of processes, simulated by the BLOOM module. A complete description of these processes can be found in the user manual [15]. Biogeochemical models often include a large number of parameters, which may be site-dependent. In our case study, the activated processes and variables lead to a set of 144 modifiable parameters.

2.3. Formulation of the calibration problem

In this work, we are interested in the automated calibration of the complex biogeochemical model BLOOM applied to the case of the lake Champs-sur-Marne. The objective is to find one or more parameter sets that lead to simulated values of the variables of interest that are close to the observed data over a chosen time period.

Among the 144 parameters of the chosen configuration of the BLOOM model (presented in Section 2.2), 114 were selected to be estimated by the calibration process, along with 19 initial conditions. Ultimately, this leads to 133 parameters and initial conditions to be estimated by the calibration. The other parameters and initial conditions were not included in the calibration. Either their values were considered to be known with a sufficiently low uncertainty, or it was demonstrated by previous tests that they have a negligible influence on the model outcomes.

A 16-days high-frequency monitoring period from July 25 to August 10, 2018, was selected for the automated calibration of the biogeochemical model. The variables of interest considered for calibration are total chlorophyll, phycocyanin, and dissolved oxygen. Before being used in the automated calibration procedure, the raw measurements of the

three variables had to be converted to the appropriate units used in the model simulations, namely gC.m^{-3} for cyanobacteria and gO.m^{-3} for oxygen concentration, while total chlorophyll already had the correct unit (μg Chl.L^{-1}). Oxygen solubility in water is temperature dependent and was therefore converted from a percentage of saturation to gO.m^{-3} using the empirical equation proposed by Weiss [51] together with high-frequency water temperature data from the MPx sensor. Phycocyanin was first converted to the equivalent of chlorophyll using a conversion factor deduced by comparison with monthly profiles taken *in situ* with the BBE FluoroProbe profiler, and finally to carbon content using the stoichiometric ratio Chl:C value of 0.03, often found in scientific literature [e.g. 19, 18]. In addition, profiles taken on July 25, 2018, with the BBE FluoroProbe were used to set and validate the initial conditions of the model in terms of μg Chl-a.L^{-1}.

The methodology used for the automated calibration is based on a recently developed approach combining Approximate Bayesian Computation (ABC) and Random Forests (RF), hereafter referred to as ABC-RF. This approach is described in section 2.4. In section 2.5, we then introduce a calibration procedure that combines the ABC-RF with a sensitivity analysis of the model outputs.

2.4. Calibration method

2.4.1. Approximate Bayesian Computation

Approximate Bayesian Computation (ABC) is a class of computational methods rooted in Bayesian statistics first proposed by Beaumont in 2002 [5]. It allows for parameter inference without the need to explicitly compute the likelihood function [44]. Developed in the field of population genetics, it has quickly grown as a solid alternative to likelihood-based methods for model calibration and it has already been applied in evolutionary biology and ecology [12].

Given a model $\mu(x, \theta)$ where x is the vector of variables and θ the vector of parameters, and D a vector of observed values of x, the posterior probability of the model parameters can be obtained through the Bayes' theorem:

$$\pi(\theta|D) = \frac{\pi(D|\theta)\pi(\theta)}{\pi(D)} \tag{1}$$

where $\pi(\theta|D)$ is the conditional probability of the model parameters given the observations D (the posterior probability), $\pi(D|\theta)$ is the conditional probability of the observations given the parameter values (the likelihood function), $\pi(\theta)$ is the prior distribution of θ, and $\pi(D)$ is the marginal probability of the data. The marginal probability can be considered a normalizing constant and is often neglected in applications where model intercomparison is not involved. In Bayesian inference, the desired posterior probability can therefore be described through the prior distribution and the likelihood function.

However, for the present application, the likelihood function is analytically intractable making the estimation of the posterior distribution through standard computational methods (such as Markov Chain Monte Carlo algorithms) impossible. The idea at the core of ABC is to bypass the explicit evaluation of the likelihood function, directly obtaining an approximation of the posterior probability distribution. To do so, prior probability distributions are first defined for the model parameters. The model μ is then used to generate a large set of simulations by randomly sampling the parameter values according to their priors [47]. From these simulations a set of relevant summary statistics, which summarize the information contained in the model runs, is computed and stored, along with the corresponding parameter values, in a dataset called "reference table" [1]. The posterior distributions can eventually be estimated using this reference table through the application of a rejection algorithm or of machine learning techniques (see Figure 2 for a conceptual overview of ABC).

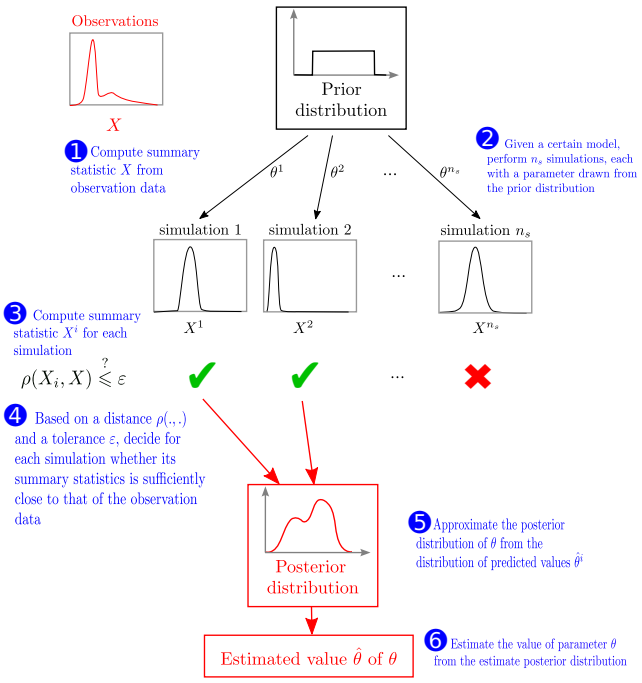

Figure 2: "Parameter estimation by Approximate Bayesian Computation: a conceptual overview." Figure adapted from [43].

2.4.2. ABC random forest

In its standard form, ABC retrieves the posterior parameters distribution starting from the reference table through a rejection algorithm [9]. This entails the definition of a distance and of a tolerance level separating acceptance from rejection. However, such threshold is arbitrary and should be calibrated for each particular application [39]. Moreover, to use the standard ABC algorithm, a small number of summary statistics should be used, and it is often not easy to find out which ones are the most relevant for the available data.

To overcome these issues, Raynal et al. [35] proposed to substitute the distance-based rejection algorithm with a machine learning technique, namely the random forests (RF).

[1] In practice, the reference table is only a table having as many rows as model simulations, and in the columns of which are stored the values of the parameters and of the summary statistics of each of these simulations.

At the expense of introducing a few parameters defining the structure of a RF, this allows overcoming the definition of a distance and a tolerance level and enables the user to take into account a large number of summary statistics. Moreover, the RF algorithm is proven to be numerically more efficient than the rejection algorithm used in the standard ABC [8].

A regression tree is a structure made of binary nodes that can either be internal nodes or terminal nodes (the leaves). It can be automatically built by iteratively dividing a training dataset into subsets of increasing uniformity until a certain condition is satisfied. Namely, the process of growth of the tree continues until all terminal nodes either (a) have less than n data points (with n possibly equal to 1), or (b) are "pure", that is all elements in a node have (almost) the same outcome.

With such a process, we can build a regression tree to get an estimation of the value of $\theta_i \in \mathbb{R}$, the i^{th} component of the model parameters vector θ. This tree will be trained on a training set of M summary statistics (X_k with $k = 1 : M$, see section 2.4.3), which are computed from the set of model simulations and which constitute the reference table. Once the tree is trained, we can apply it on the observed dataset D and get the estimated value of the parameter θ_i.

A random forest consists in aggregating (or bagging) randomized regression trees. A large number of trees (n_{tree}) are trained each on a different bootstrap subsample taken from the complete available reference table. Furthermore, only a subset of m_{try} summary statistics among the M available are randomly considered at each node for splitting [35]. The estimations obtained by the n_{tree} regression trees can be treated and used to obtain an approximation of the posterior probability distribution for the parameters θ_i [e.g. 35]. Eventually, once the random forest is grown, different choices can be made for the inference of the parameter value (see section 2.4.5). For example, the final estimated value of θ_i can be determined by averaging all the n_{tree} predictions obtained in the random forest, or by taking the most probable value from the posterior distribution. A conceptual overview of the process of parameter estimation by ABC-RF is given in figure 3.

In the present paper, we tested a first calibration procedure that relies on the assumption that the model parameters can be considered independent of each other. In that case, the ABC-RF method is applied separately once for each parameter of the model. This way, one RF is built for each parameter and the associated approximate posterior distributions can be plotted, from which an estimate value of the parameters are deduced (see section 2.4.5). The different steps for the application of the ABC-RF to multiple parameters are outlined in Algorithm 1.

For the implementation of the ABC-RF, we make use of a *R* package (*regAbcrf*), developed by Raynal et al. [35] for ABC-RF parameter inference. To configure the ABC-RF, we need to choose two main arguments: (i) the number of trees to grow in the random forest (n_{tree}), and (ii) the number of variables among which to choose for splitting at each

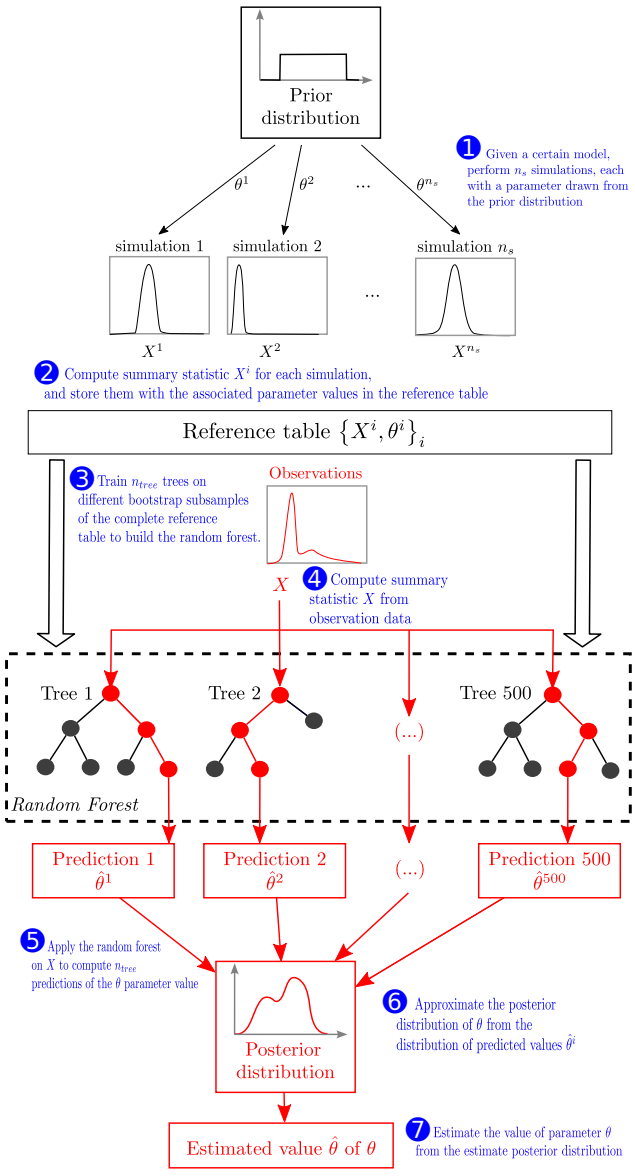

Figure 3: Parameter estimation by Approximate Bayesian Computation with Random Forest: a conceptual overview. Figure adapted from [43].

node (m_{try}). For the former, different values have been tested (see section 2.6) while the latter was set to its default value, which is the maximum between 1 and the number of variables divided by 3.

2.4.3. *Prior distributions and summary statistics*

For parameter inference, BLOOM was run 30000 times to create the reference table, by drawing parameters randomly from the user-assigned prior distributions. Either a uniform or a gamma distribution was assigned to each model parameter. Such distributions were defined on the basis of the modeller expertise, derived from default and literature parameter values and previous "trial and error" calibration tests. Gamma distributions were defined for most of the parameters, using values from previous trial-and-error calibration tests to set the mean value of the distribution, and setting

Algorithm 1: ABC-RF (multidimensional)

Data: N: number of parameters to be estimated
M: number of summary statistics
X_k: summary statistics ($k = 1{:}M$)

Result: estimated value $\hat{\theta}_i$ of θ_i for $i = 1{:}N$

```
for i=1:N do
    Application of ABC-RF to estimate the parameter θ_i from the set of summary statistics {X_k, k = 1 : M} → posterior distribution of θ_i
    Determination of the estimated value θ̂_i of θ_i from the approximate posterior distribution of θ_i
end
```

the standard deviation to 20%. The choice of the gamma distribution is motivated by the fact that, unlike normal distributions, it is defined only on positive values. A uniform density function was assigned to those parameters (such as the initial conditions) for which very little information was available from literature or previous studies. The uniform distributions were built using default values or values from previous calibration tests as a central value and taking 0 as a left limit. For parameters with default values very close to 0, the upper limit was set to 1.

In the application of ABC, model results and observations are summarized into a set of user-chosen summary statistics. The ABC-RF allows the use of a large number of summary statistics without incurring in the curse of dimensionality [33], as the relevant summary statistics will be automatically selected by the ABC-RF. A set of summary statistics is a set of metrics that summarizes the most relevant characteristics of model results. For each model run originated by the model μ, the summary statistics are computed and stored in the *reference table* together with the corresponding parameter set. Ultimately, the reference table constitutes the training data set on which ABC is applied.

The summary statistics therefore replace the raw model runs in the calibration procedure, and their definition is crucial. They should minimize information loss and maximize dimension reduction [12]. However, their choice is also related to the processes subject of the study. Here, we are mainly interested in the time evolution of a phytoplankton community. Summary statistics are calculated on the complete set of model runs as well as on the observed data for the three variables considered for calibration, which are some time series of total chlorophyll, cyanobacteria and dissolved oxygen concentrations. Two different summary statistics were tested: (1) the normalized square of residuals (R) between each model run ($\hat{D}$) and the observation series (D) and (2) the normalized mean square error ($NMSE$) between $\hat{D}$ and D. In this work, the summary statistics have the particularity to be dependent on the observation series D; this choice will be further discussed in section 4.

Consider two time series data D and $\hat{D}$, the first one corresponding to the set of measured values D_i of a given variable at different time instants t_i for $i = 1 : n_t$, and the second one to the set of simulated values $\hat{D}_i$ of the same variable at the same time instants t_i.
The normalized square of residuals (R) between D and $\hat{D}$ is defined as follows:

$$R = \frac{I_n}{I_d} \tag{2}$$

where:

- I_n is the numerical integration (over time) of the time series data $\left(D_i - \hat{D}_i\right)^2$ which is an approximation of $\int (\hat{D}(t) - D(t))^2 dt$, where $D(t)$ and $\hat{D}(t)$ are the time-continuous variables associated with D and $\hat{D}$ respectively,
- I_d is the numerical integration (over time) of the time series data $\left(D_i\right)^2$ which is an approximation of $\int (D(t))^2 dt$, where $D(t)$ is the time-continuous variable associated with D,

The numerical integration in I_n and I_d has been performed with the function *integrate.xy* of the *R* package *sfsmisc*.

The normalized mean square error ($NMSE$) was computed as defined in [32], and normalized over the product between the mean $\bar{\hat{D}}$ of the time series of simulated data and the mean $\bar{D}$ of the time series of observation data:

$$NMSE = \frac{1}{n_t} \sum_{i=1}^{n_t} (\hat{D}_i - D_i)^2 \cdot \left(\bar{\hat{D}} \cdot \bar{D}\right)^{-1} \tag{3}$$

2.4.4. Preselection of a subset of simulations

The ABC-RF method can be applied to the whole set of 30000 simulations or on a subset of simulations. This would reduce the dimension of the reference table and, consequently, the computational time. Such subset can be chosen in different ways. For instance, as the goal is to find some values of model parameters that correspond to the observed data, we can focus on simulations close to the observations, that is on simulations with small values of the summary statistics R and $NMSE$. For this purpose, we calculated for each simulation the sum of the summary statistics (that is the sum of the R or $NMSE$ values of the variables): this value is hereafter called "total R" or "total $NMSE$". The set of simulations with the smallest values of total R or total $NMSE$ were selected for the subset. Different sizes were tested for the subsets, as described in section 2.6.

2.4.5. Estimation of the parameter value from the posterior distribution

In a classical Bayesian framework, the estimated values of the parameters are determined from the posterior probability distributions, which can be obtained using the Bayes rule (1). However, as mentioned previously the likelihood is intractable here. Hence, instead of using exact posterior distributions, we use approximate posterior distributions, obtained thanks to the ABC approach (the ABC-RF algorithm

in our case). These distributions might present various local maxima which makes the estimation of the optimal value for the parameters difficult. Three options were therefore considered and tested in this study.
Consider a parameter θ and an approximation π^ϵ of the posterior distribution $\theta \mapsto \pi(\theta|D)$ of the form $\pi^\epsilon = \sum_{i=1}^{N} \omega_i \delta_{\theta_i}$, where $(\omega_i)_i$ is a sequence of non-negative numbers summing to 1, $(\theta_i)_i$ is an increasing sequence and $N \in \mathbb{N}^*$. We can choose an estimated value $\hat{\theta}$ of the parameter as follows:

• **option** P_{max}: the estimated parameter value is the one with the maximal value of the approximate posterior probability distribution:

$$\hat{\theta} = \theta_X \tag{4}$$

where θ_X is such that $\omega_X > \omega_i$ for all $i \neq X$.

• **option** P_{med}: the estimated parameter value is the median of the approximate posterior probability distribution:

$$\hat{\theta} = \theta_d \tag{5}$$

where θ_d is such that $\sum_{i=1}^{d} \omega_i \leq 1/2$ and $\sum_{i=1}^{d+1} \omega_i \geq 1/2$.

• **option** $P_{mix,k}$: it is a compromise between the first two options. Depending on a criterion, $\hat{\theta}$ will be equal either to the most probable value (option P_{max}) or the median (option P_{med}):

$$\hat{\theta} = \begin{cases} \theta_X & \text{if } \omega_X > \frac{k}{\theta_N - \theta_1} \\ \theta_d & \text{elsewhere} \end{cases} \tag{6}$$

where θ_1 and θ_N are the lower and upper bounds of the support of the approximate posterior distribution π^ϵ (which is determined numerically and has therefore a finite support), and k is a constant value that has to be chosen. In this study, we will test the values $k = 2, 3$.

The option P_{max} is well adapted to the case where the posterior distribution is peaked, whereas the option P_{med} is more suitable for flat distributions (see figure 7 as an example). The option $P_{mix,k}$ introduces a threshold to switch between the first two options depending on the shape of the posterior distribution.

2.5. Including sensitivity analysis in ABC-RF

When the model parameters are considered independent from each other, the ABC-RF can be applied to each parameter independently from each other as in algorithm 1. However, the value of some of the model parameters might have a non-negligible influence on the remaining ones. In order to take into account the possible mutual influence, we set up a different calibration procedure, which includes a sensitivity analysis of the model parameters.

2.5.1. General procedure

Before applying ABC-RF to the model parameters, a sensitivity analysis is performed using the set of 30000 available simulations. This allows us to identify the parameters that have the greatest influence on the simulated model outputs and to rank them from most to least important. We can then apply the ABC-RF method iteratively, starting with the most important parameter, and at each iteration adding the previously estimated parameters to the set of summary statistics. The steps of this calibration procedure are summarized in the Algorithm 2 and will be discussed in paragraphs 2.5.2 and 2.5.3.

Hereafter, this calibration procedure will be referred to as ABC-RF with SA or ABC-RF SA.

Algorithm 2: ABC-RF with SA

Data: N: number of parameters to be estimated
M: number of summary statistics
X_k: summary statistics ($k = 1{:}M$)
$S_{i,k}$: sensitivity indices ($i = 1{:}N, k = 1{:}M$)

Result: estimated value $\hat{\theta}_i$ of θ_i for $i = 1{:}N$

Sorting of the parameters according to the values of the sensitivity indices $S_{i,k}$
→ vector σ of sorted sensitivity indices
for *i=1:N* **do**
  Selection of a subset Y of summary statistics X_k
  Application of ABC-RF to estimate the parameter $\theta_{\sigma(i)}$ from the set of summary statistics $\{Y, \theta_{\sigma(1:i-1)}\}$ → posterior distribution of $\theta_{\sigma(i)}$
  Determination of the estimated value $\hat{\theta}_{\sigma(i)}$ of $\theta_{\sigma(i)}$ from the posterior distribution of $\theta_{\sigma(i)}$
end

2.5.2. Sensitivity indices and sorting

Performing a standard sensitivity analysis (based on Sobol or FAST methods) directly to the set of 30000 simulations was not possible because the model parameters are possibly not independent from one another. To overcome this issue, the methodology for models with correlated inputs proposed in [48] was adopted. Details about the computation of the sensitivity indices following this methodology are given in the appendix A.

Following this methodology, a sensitivity index was calculated for each of the 133 parameters and for each summary statistic (namely the R and the $NMSE$ values) of the three variables (chlorophyll, phycocyanin and dissolved oxygen). These indices represent the sensitivity of the summary statistic of a variable to the variation of the parameter. For each summary statistic and each parameter, three sensitivity indices have therefore been obtained, one for each variable.

For a given summary statistic, two options were then tested to sort the parameters based on the sensitivity indices obtained for the 133 parameters and the three variables (step 1 of Algorithm 2):

- **option "Max"**: sort the parameters following the value of the largest sensitivity index among the three;

- **option "Sum"**: sort the parameters following the value of the sum of the three sensitivity indices.

2.5.3. Selection of a subset of summary statistics

In Algorithm 2, at each iteration of the loop on the model parameters to be estimated, we can choose a subset of summary statistics on which the calibration will be applied (step 3 of Algorithm 2). If the parameters are sorted following the values of the sum of the sensitivity indices, we chose to apply the calibration on the whole set of summary statistics at each iteration (3 summary statistics) to which we add the previously estimated parameter values. In the case where the parameters are sorted following the value of the largest sensitivity index, two options were considered: the calibration was performed (1) either on the whole set of summary statistics (3 summary statistics + the previously estimated parameter values), (2) or only on the summary statistics of the variable for which the sensitivity index of the current parameter is the largest (1 summary statistic + the previously estimated parameter values).

2.6. Preliminary tests

The calibration methodology was first applied to a set of simulated observations, that is some data issued from a model simulation. The use of simulated observations instead of real observations ensures the existence of a known parameter set with whom the model will reproduce the data correctly. This allows us to test the capability of the calibration methodology to reproduce both the simulated observations and the parameter values in an ideal case where the model is exact.

Here, the simulation with the lowest total $NMSE$ (i.e. the closest one to the real observations, see section 2.4.4) among the 30000 model runs was selected, and the results in terms of total chlorophyll, dissolved oxygen and cyanobacteria concentration were used as simulated observations. Namely, the best simulation is simulation number 4022, and the associated summary statistics were discarded from the reference table before the application of the calibration methodology.

In order to choose the main characteristics of the ABC-RF (e.g. the number of simulations constituting the reference table, and the number of trees used to build the random forests) of the ABC-RF, a series of preliminary tests were performed. The tests were performed with the classic ABC-RF formulation only (see Algorithm 1). Namely, the tests investigated the influence on the calibration outcomes of the number of simulations, the randomness inherent to the ABC-RF procedure, the use of a preselected subset of simulations (according to section 2.4.4), the number of trees in the random forests, and the different options for the estimation of the parameter values from the posterior distribution (see section 2.4.5). For these preliminary tests, only one summary statistic was considered ($NMSE$). The tests are detailed hereafter.

In order to test the influence of the number of simulations used in the ABC-RF, various calibrations were carried out using subsets of simulations of increasing size. Namely, the number of simulations was varied between 2000 and 30000, with a 2000 step. The subsets were chosen in two ways. Either the elements of each subset were chosen randomly or only the best simulations (in terms of values of total NMSE) were selected as proposed in section 2.4.4. For these tests, the number of trees was set to 500, and each calibration was performed ten times (with the same parameters configuration) to test the variability of the results inherent to the randomness of the method.

Similarly, eight calibrations were carried out with an increasing value of the number of trees. The number of trees was varied from 250 to 2000 with a 250 step and tested for all simulation subsets between 5000 and 30000 with a 5000 step. Each calibration was performed only once.

For each of these tests, once the posterior distribution was obtained for each model parameter, the four options defined in section 2.4.5 (options P_{max}, P_{med}, $P_{mix,2}$ and $P_{mix,3}$) were applied. This provided several sets of estimated parameter values. For each of these estimated parameter sets, the model was then run. The so-obtained simulations were finally compared with one another by calculating the total $NMSE$ between calibrated model results and simulated observations.

A summary of all the calibration runs performed for the preliminary tests is given in Table 3.

2.7. Validation on simulated observations

Following the results of the preliminary tests presented in section 3.1.1, the ABC-RF (algorithm 1) and the ABC-RF with SA (algorithm 2) were tested and compared under two configurations: using a subset of either (i) 10000 or (ii) 25000 preselected simulations. In both cases, the number of trees used to build the random forests was set to 500, and both R and $NMSE$ were tested as summary statistics. In the case of ABC-RF with SA, the three possible combinations of parameters sorting options (section 2.5.2) and options for the selection of a subset of summary statistics (section 2.5.3) were tested. Furthermore, the four options for the choice of the parameter values (options P_{max}, P_{med}, $P_{mix,2}$ and $P_{mix,3}$) described in section 2.4.5 were also examined for each calibration.

The combination of all the methodologies (ABC-RF and ABC-RF with SA), configurations and above-described options, results, for each summary statistic, in a set of 8 calibration runs with the ABC-RF, and 24 calibration runs with the ABC-RF with SA, that is (8+24)×2 summary statistics = 64 calibration runs, which are summarized in Table 4. Model simulations were then performed with the 64 estimated parameters sets, and the model outputs were compared to the simulated observations through the value of total R or total $NMSE$, coherently with the choice of the summary statistics.

The ABC-RF (algorithm 1) and ABC-RF with SA (algorithm 2) were first tested using the set of simulated observations (as described in section 2.6). In that case, the estimated parameter values were also compared with the known parameter values used to generate the simulated observations. To do so, the error (e) between estimated and known param-

Table 3
Summary of the calibration runs performed for the preliminary tests.

Method of calibration	Summary statistic	Reference table size	Simulation preselection	Number of trees	Repetition number	Parameter value estim.	Number of calib. runs
ABC-RF	$NMSE$	2000:2000:30000	random	500	10	P_{max}, P_{med}	1200
ABC-RF	$NMSE$	5000:5000:30000	& closest	250:250:2000	1	$P_{mix,2}$, $P_{mix,3}$	384

Table 4
Summary of the calibration runs performed for the validation on simulated data and the application on real data with (i) preselection of the closest simulations for the reference table, (ii) 500 trees per random forest, (iii) no repetition, (iv) the three variables

Method of calibration	Summary statistic	Reference table size	Parameter estimation	Sensitivity index sorting	Variables used in calib.	Number of calib. runs
ABC-RF	$NMSE$ & R	10000 & 25000	P_{max}, P_{med}, $P_{mix,2}$, $P_{mix,3}$	-	-	16
ABC-RF SA	$NMSE$ & R	10000 & 25000	P_{max}, P_{med}, $P_{mix,2}$, $P_{mix,3}$	Max	One	48
				Max	All	
				Sum	All	

eters was calculated, normalized over the range of variability allowed for each parameter, and converted into a percentage:

$$e = \frac{|\theta_{estim} - \theta_{true}|}{\theta_{max} - \theta_{min}} \cdot 100 \quad (7)$$

where θ_{estim} is the value of the estimated parameter, θ_{true} is the known parameter value used to generate the simulated observations, and θ_{max} and θ_{min} are the values of θ above and below which the prior distribution takes values smaller than 0.05.

2.8. Application on real data

After validation of the calibration procedures on the simulated observations, the ABC-RF (algorithm 1) and ABC-RF with SA (algorithm 2) were tested using the set of real data as observations. The 64 calibration runs described in section 2.7 in the case of simulated observations (and summarized in Table 4) were applied to the real data set.

Finally, we tested the ABC-RF with SA using only one of the three variables (i.e. as if observations were recorded for only one variable). To do so, we decided to focus on total chlorophyll, one of the variables most commonly measured in the framework of freshwater ecological studies. This was done with two main objectives: (i) to assess the quality of the calibration when only one variable is targeted; (ii) to test the capacity of the calibrated model to simulate the two remaining variables when they are not included in the calibration. For this test, only the ABC-RF with SA was applied and the summary statistics relative to total chlorophyll was the only one used. Similarly to the previous calibration runs, two preselected subsets of 10000 and 25000 simulations were tested, but the selection was based on the total R values of the total chlorophyll only. The four options described in section 2.4.5 for parameter values estimation were considered. The two options for parameters sorting (see section 2.5.2) and the two options for the selection of a subset of summary statistics (see section 2.5.3) being the same when only one variable is considered, we finally performed 8 calibration runs that are summarized in Table 5.

3. Results

3.1. Validation of the methodology

The ABC methodology was first validated on a set of simulated observations issued from the best model run among the complete set of simulations in terms of total $NMSE$ value (run number 4022, see Fig. 8).

3.1.1. Preliminary tests

The most relevant results of the tests that investigate the influence on the calibrated model outputs of (i) the number of simulations used to build the reference table, (ii) the uncertainty deriving from the inherent randomness of the ABC-RF, and (iii) the use of preselected simulations to build the reference table are highlighted in Fig. 4. The figure shows the evolution of the total $NMSE$ between the calibrated model outcomes and the simulated observations, according to the size of the simulations subset used for the cal-

Table 5
Summary of the calibration runs performed for the application on real data with (i) preselection of the closest simulations for the reference table, (ii) 500 trees per random forest, (iii) no repetition, (iv) only real chlorophyll data

Method of calibration	Summary statistic	Reference Table size	Parameter estimation	Sensitivity index sorting	Variables used in calib.	Number of calib. runs
ABC-RF SA	R	10000 & 25000	P_{max}, P_{med}, $P_{mix,2}$, $P_{mix,3}$	Max=Sum	One=All	8

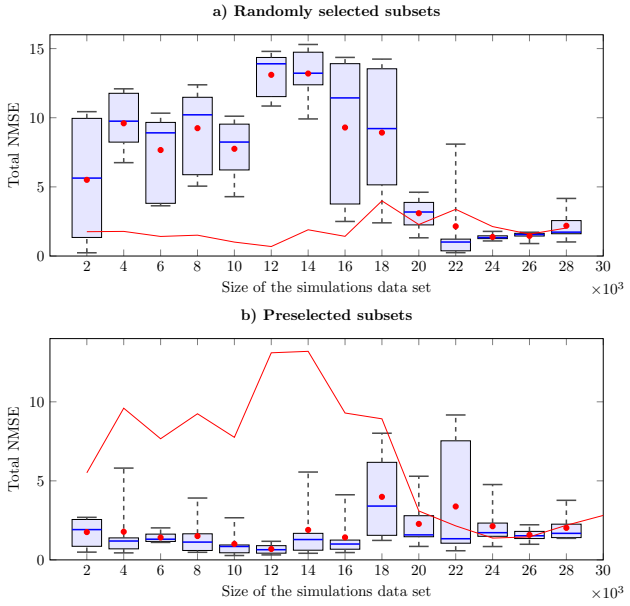

Figure 4: Preliminary tests. Total $NMSE$ between the simulated observations and the model results calibrated through the ABC-RF calibration, according to the size of the simulations subset used to generate the reference table. Panel a): randomly selected subsets; panel b): subsets of preselected simulations. The boxplots represent relevant statistical characteristics of the sets of ten equivalent runs of ABC-RF calibration. The blue lines and red points indicate the median and the mean values of the set, respectively. The bottom and top edges of the boxes mark the 25th and 75th percentiles respectively, and the maximal and minimal values of the ensemble are marked through the whiskers extent. Eventually, the solid red line in panel a) represents, for a direct comparison, the mean values of the boxplots in panel b). Vice versa for the solid red line in panel b). The number of trees was set to 500 and the option for parameter values estimation is $P_{mix,2}$.

ibration. Panel a), is relative to subsets with randomly selected simulations, while panel b) is relative to preselected simulations (see section 2.4.4). For each subset size, ten calibration runs were carried out. The resulting values of total $NMSE$ are plotted in Fig. 4 as a series of boxplots. On the boxplots, the blue lines and red points indicate the median and the mean values of the set, respectively; the bottom and top edges of the box mark the 25th and 75th percentiles, respectively. Eventually, the maximal and minimal values of the ensemble are marked through the whiskers extent. Fig. 4 shows the calibration results obtained with the option $P_{mix,2}$ only; similar results have been obtained for the other options for parameter values estimation.

The results in Figure 4 show that, for the case of randomly selected simulations (panel a), the mean value and variability (i.e., the range of boxplots) of the calibration error decrease sharply for subsets of at least 20000 simulations. In particular, the minimum values of mean value and variability of total $NMSE$ are obtained for the case of a set of 24000 simulations. Panel (b) of the same figure shows that preselecting the best simulations largely improves the results of the calibration procedure, both in terms of mean value and variability of total NMSE. In the case of preselected subsets, the mean value and variability seem rather independent of the number of simulations used for calibration, as they do not vary strongly with increasing size of the set of simulations (except for the cases of 18000 and 22000 simulations where the mean values and variability are larger, possibly due to the randomness of the method and the insufficiently large number of repetitions). Minimum values are obtained for 10000 and 12000 simulations. With randomly selected subsets, at least 20000 simulations are needed to obtain performances comparable to the case of the preselected subsets.

The tests performed on the number of trees used to construct the random forests did not show any significant influence on the calibration results, which did not improve as the number of trees increased, neither in terms of total $NMSE$ nor in terms of variability (see Appendix B for some detailed numerical results).

The way the parameter values are chosen from the posterior distribution (options described in section 2.4.5) has an impact on the calibration results. However, the preliminary tests did not show clear and conclusive results, and it was not possible to identify an option that consistently outperformed the others for all three variables at once. For this reason, all four options (P_{max}, P_{med}, $P_{mix,2}$ and $P_{mix,3}$) have been tested in subsequent applications.

In conclusion, the tests described above indicate that, for the model considered, a reference table built from at least 24000 randomly selected simulations is necessary to minimize the mean value and the variability of the calibration error obtained with the ABC-RF method. In the case of preselected simulations, comparable results can be obtained with a smaller reference table of about 10000 simulations. The number of trees did not show a significant effect on the calibration error, and none of the four options for choosing parameter values could be preferred over the others.

3.1.2. Application on simulated observations

Following the results of the preliminary tests (section 3.1.1), ABC-RF (algorithm 1) and ABC-RF with SA (algorithm 2) were tested and compared to the simulated observations, under two main configurations: with subsets of either (i) 10000 or (ii) 25000 preselected simulations. The set of calibration runs summarized in Table 4 were performed, resulting, for each summary statistic, in eight calibration runs for ABC-RF and 24 runs for ABC-RF with SA.

In general, the calibration results were of similar quality whether using R or $NMSE$ as a summary statistic. Since the best calibration was obtained with R as a summary statistic, the results using $NMSE$ will not be discussed.

The eight best calibration runs (in terms of total R value) for ABC-RF with SA with R as the summary statistic are listed in Table 6, along with the two best calibration runs for ABC-RF without SA. In the table, for each calibration run, the size of the reference table built from the preselected set of simulations (10000 or 25000), the name of the method

Table 6
Application to the simulated observations. List of the calibration runs with the lowest total R values for the ABC-RF with SA (eight best runs) and without SA (two best runs), sorted according to the total R value. The characteristics of each calibration run are detailed in terms of: size of the reference table, calibration method, sorting of the sensitivity indices, variables used in the calibration, and option for the estimation of the parameter value (see sections 2.5.2, 2.5.3 and 2.4.5). The total R value of the closest simulation to the simulated observations is also provided as a reference.

Reference table size	Method of calibration	Sensitivity index sorting	Variables used in calib.	Parameter estimation	Total R
25000	ABC-RF SA	Max	One	P_{max}	0.057
10000	ABC-RF SA	Sum	All	$P_{mix,3}$	0.109
25000	ABC-RF SA	Sum	All	P_{med}	0.115
10000	ABC-RF SA	Max	All	P_{max}	0.125
10000	ABC-RF SA	Sum	All	$P_{mix,2}$	0.127
25000	ABC-RF SA	Sum	All	P_{max}	0.137
10000	ABC-RF SA	Max	All	$P_{mix,2}$	0.137
25000	ABC-RF SA	Max	All	P_{max}	0.138
25000	ABC-RF	-	-	P_{med}	0.594
10000	ABC-RF	-	-	P_{med}	0.671
				Sim. number	
			Best simulation:	12936	0.144

(ABC-RF or ABC-RF with SA), the options used for the calibration, and the total R value are specified. The different options for the ABC-RF with SA method are described in the sections 2.5.2, 2.5.3 and 2.4.5. The calibration runs are sorted according to the total R value. The total R value of the simulation that is closest to the simulated data (sim. n. 12936) is also provided as a reference in the table.

The best calibration run among all (the one with the lowest value of total R) is obtained with the larger set of 25000 simulations. Its total R value (0.057) is sensibly lower than all the other calibration runs. However, the remaining seven calibration runs presented in table 6 for the ABC-RF with SA also show good model performances. In particular, the use of the smaller preselected set of simulations to generate the reference table does not deteriorate model performances. Notably, the second best calibration (total $R = 0.109$) is obtained with a set of 10000 preselected simulations.

The integration of sensitivity analysis into the ABC-RF with SA method greatly improves the calibration results when compared to those obtained with ABC-RF and to the closest simulation to the data. All calibration runs listed in Table 6 that use SA show total R values five to ten times lower than those obtained by ABC-RF without SA. Compared to simulation n. 12936, the calibration error is reduced by about 60% compared to the best calibration run, and by about 25% compared to the second best calibration run. Thus, the integration of SA in the ABC-RF framework seems to be crucial for the application of ABC to a complex process-based model.

Simulations from the two best calibration runs are plotted for the three variables of interest in Fig. 5, and compared with both simulated observations and the simulation closest to the data (simulation n. 12936).

The total chlorophyll value (Fig. 5-a) shows the strongest variations from one calibration run to another. The simulation from the best calibration run (red line) follows the simulated observations very closely, reproducing the daily oscillations correctly. The simulation from the second best calibration run (blue line) shows an early and slightly overestimated peak of chlorophyll, while simulation n. 12936 shows a delayed and still slightly overestimated peak of chlorophyll.

Simulations from both calibration runs give good results for cyanobacteria (Fig. 5-b), with a slight underestimation with the second best calibration run (blue line) in the last days of the bloom. For oxygen concentration (Fig. 5-c), the results obtained with both calibration runs are significantly better than the simulation that is closest to the data (black lines).

As shown in Table 6, the two best calibration runs shown in Fig. 5 are obtained by different configurations of ABC-RF with SA. Indeed, the best calibration run, which is performed with a reference table built from a set of 25000 simulations, was obtained by sorting the parameters according to the value of the largest sensitivity index among the three variables of interest (option "Max" in section 2.5.2), and by calibrating each parameter only using the summary statistic of the corresponding most influential variable (see section 2.5.3). The second best calibration run (reference table built from a subset of 10000 simulations) is obtained by using the opposite options for sorting the parameters and selecting the summary statistics used for calibration (see sections 2.5.2 and 2.5.3). The options for estimating parameter values from the posterior distribution are also different for these two calibrations (P_{max} and $P_{mix,3}$).

The use of simulated observations ensures that there exists a known set of parameter values to replicate the data. The difference between the estimated parameter sets and the reference parameter set (the one used to generate the simu-

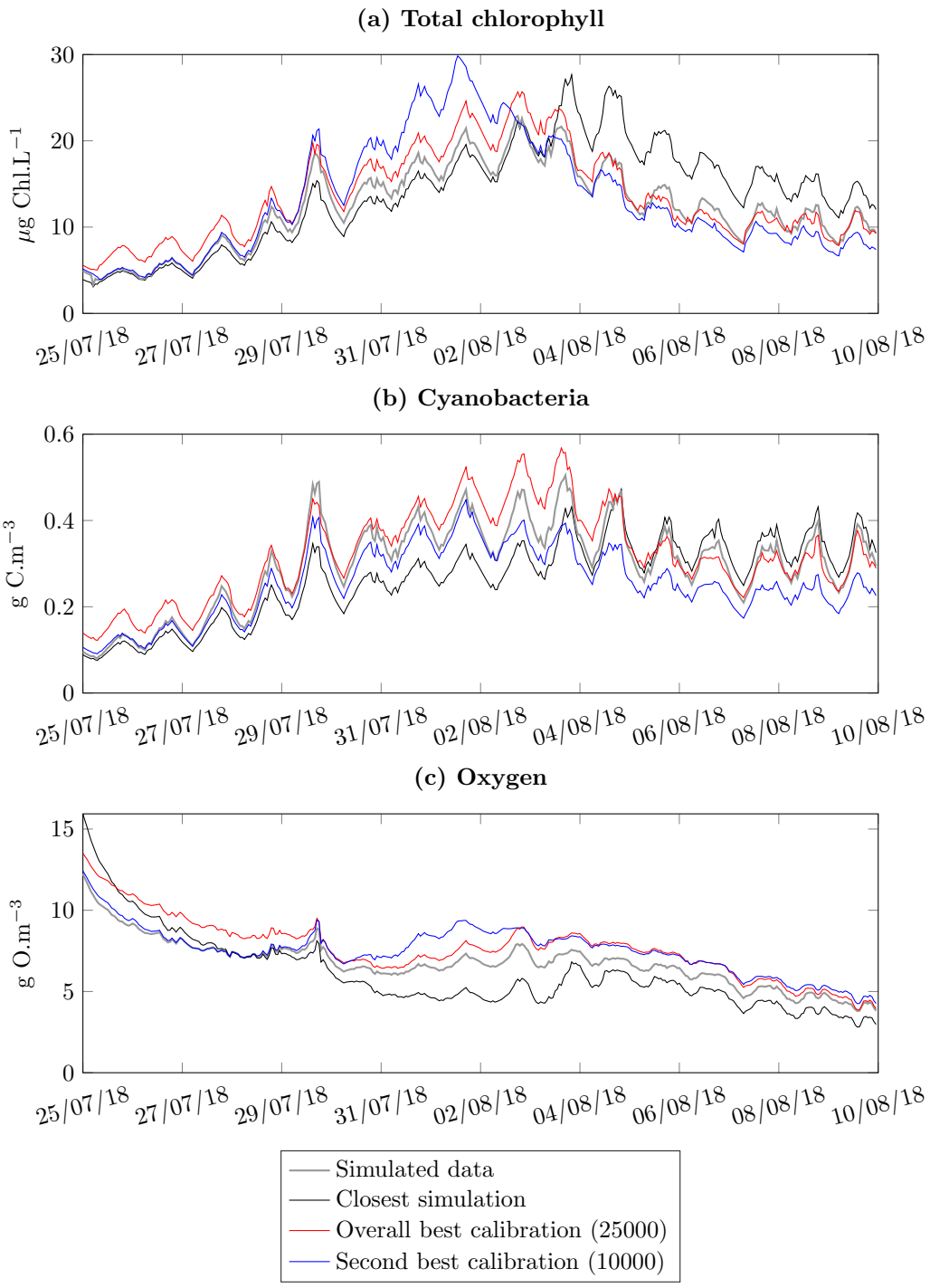

Figure 5: Time series of total chlorophyll (a), cyanobacteria (b) and dissolved oxygen (c) concentrations for the two best calibration runs (red and blue lines), the simulated observations (grey lines) and the closest simulation to the data (simulation n. 12936, black lines). The best calibration run among all (red lines) is obtained with a set of 25000 simulations, while the second best (blue lines) with a set of 10000 preselected simulations.

lated observations) was quantified using the error e defined in equation (7). Figure 6 shows the values of this error e for the two best calibration runs (i.e. those plotted in Fig. 5): the best calibration run among all (obtained with a subset of 25000 simulations, Fig. 6-a) and the second best calibration run (obtained with a subset of 10000 simulations, Fig. 6-b).

The parameters whose errors are shown in Fig. 6 are sorted according to the value of the sensitivity indices and the option chosen: option "Max" for the best calibration run among all, option "Sum" for the second best calibration run (see section 2.5.2 and table 6 for the details). Only the first 30 parameters are shown. The values (based on the sensi-

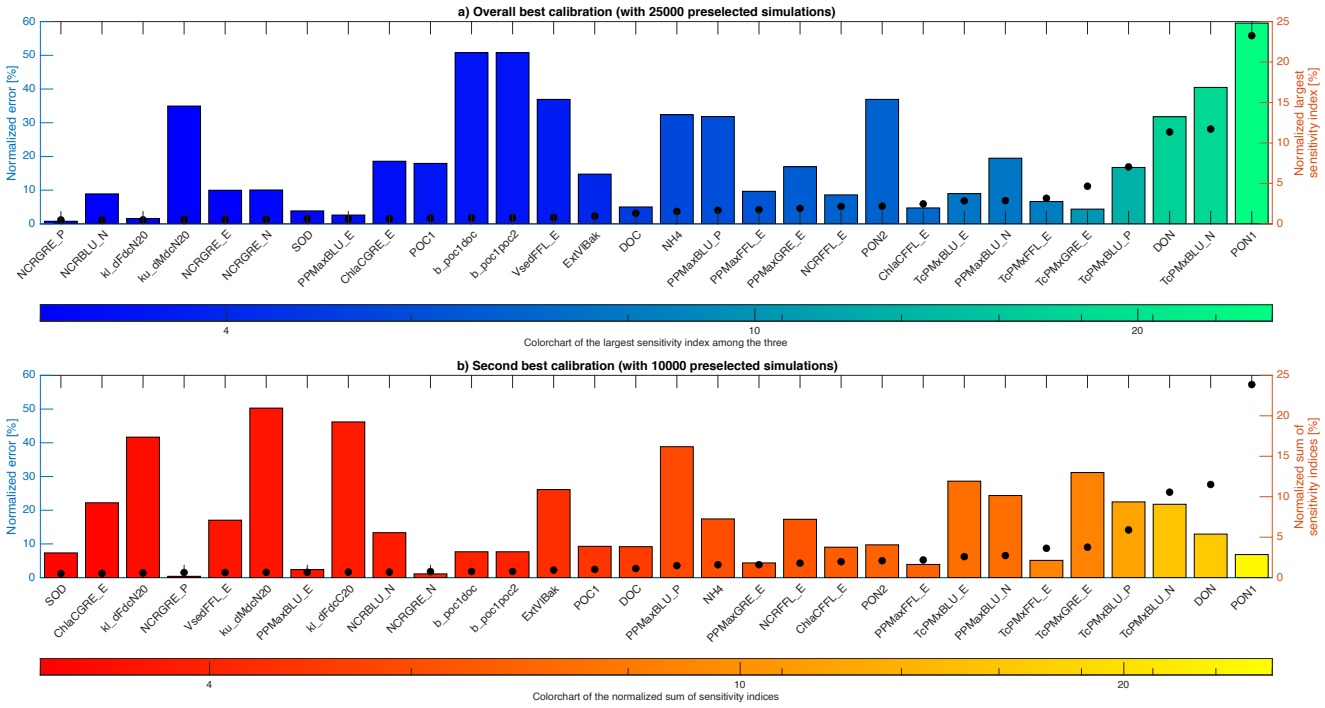

Figure 6: Normalized error e (see equation (7)) between the estimated parameters set and the one used to generate the simulated observations and associated color charts showing the sensitivity indices of the 30 most influential parameters. Panel a) is relative to the best calibration run among all (obtained with a subset of 25000 simulations). The parameters were sorted according to the value of the largest sensitivity index among the three. Panel b) is relative to the second best calibration run (obtained with a subset of 10000 simulations). The parameters were sorted according to the value of the sum of the three sensitivity indices. Both color charts are plotted with a logarithmic scale.

tivity indices) used to sort the parameters are also shown in Figure 6 (black dots and color chart); they are normalized over the sum of the values and converted to percentage of importance.

As shown in Figures 6-a) and 6-b), the order of importance of parameters varies only slightly between the two calibration runs. Regarding the ten most important parameters, the main differences between the two calibration runs are the permutation of the second and third most important parameters and the ninth parameter, which changes from the growth rate of flagellates (for the run with 10000 preselected simulations) to the Chl/C ratio for flagellates (for the run with 25000 simulations). Figure 6 also shows that the percentage of importance of the parameters decreases rapidly: for both calibration runs, it is about 25% for the most important parameter, and drops below 5% after the tenth parameter in importance order.

In terms of errors on parameter values, the two calibration runs in Fig. 6 show different behaviours. The second best calibration run (panel b) has rather small errors, less than 25%, for the ten most important parameters. Errors exceed 40% only for parameters with a very low impact on the model outputs.

On the other hand, the best calibration run among all (panel a) shows considerable errors for the most important parameters (about 60% for the most important, and about 30% for the second and third). Low errors are then found for the other parameters until the tenth position. After the tenth position, the errors increase without a specific pattern, similar to what was found for the other calibration run. Despite these large errors in the estimation of the most relevant parameters, the simulations from this calibration run give the best overall results (see Figure 6-a).

The application of ABC-RF with SA influences the shape of the posterior distributions of the parameters. As the number of summary statistics used for calibration increases with each iteration of the algorithm 2, the resulting posterior distributions tend to be smoother and less irregular compared to those obtained with the ABC RF method, in which the same set of summary statistics is used to calibrate each of the parameters independently. This smoothing effect increases with the number of iterations of the algorithm 2, but can already be seen after only a few iterations and thus on most parameters.

Figure 7 shows, as an example, the posterior probability densities of one calibration run with ABC-RF and those of the two best calibration runs of ABC-RF with SA for three parameters, namely the first (PON1), the fourth (TcPMxBLU_P) and the twentieth (b_poc1doc) parameters in order of importance according to the sensitivity analysis presented in Fig. 6. By the fourth iteration, the posterior distribution appears to be significantly smoother in the case of ABC-RF with SA (Fig. 7-b). This is even more striking for parameters further down the calibration loop, such as for the

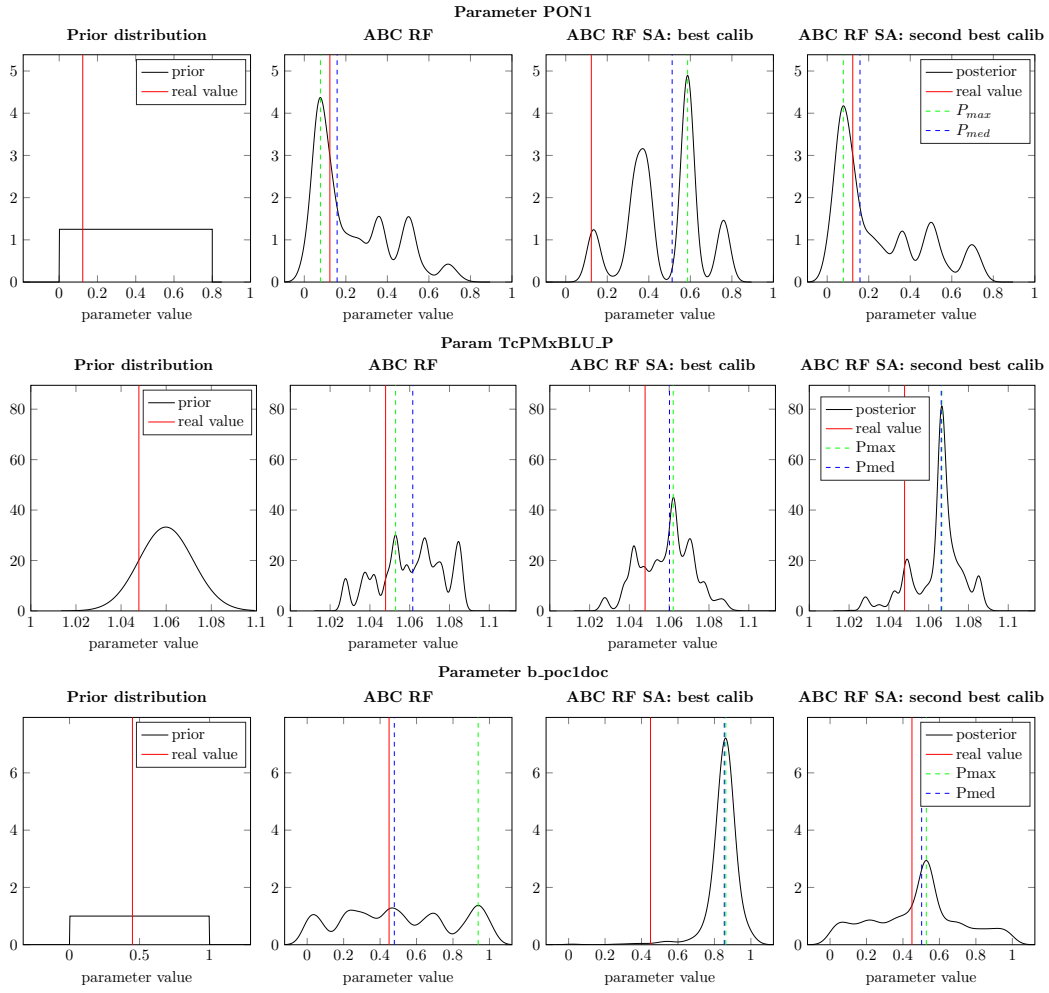

Figure 7: Examples of prior and posterior probability densities of three parameters: PON1, TcPMxBLU_P and b_poc1doc. The corresponding values for P_{max} and P_{med} are also compared. The posterior densities obtained with ABC-RF, and ABC-RF with SA for the two best calibration runs are given for comparison.

twentieth parameter.

This smoothing effect has a strong impact on the estimation of the parameter values. With the ABC-RF method, we see in Fig. 7 that at least three significantly different parameter values correspond to local maxima of the posterior distribution of TcPMxBLU_P that have nearly the same value. Although local maxima are also present on the posterior distributions obtained with ABC-RF with SA, it is easier to determine the most likely parameter value.

The shape of the posterior distribution also modifies the results of the four options described in section 2.4.5. For the posterior distribution in Fig. 7 obtained with ABC-RF with SA, the P_{max} and P_{med} options lead to values that are very close to each other, whereas in the case of classical ABC-RF, the two values are very different.

3.2. Application on real observations

The ABC RF and ABC RF with SA were applied to the real observation dataset (see Table 4). As for the application on simulated observations, the calibration results were of similar quality whether R or $NMSE$ was used as summary statistics. As the best calibration run among all is obtained with R, only the results obtained with this summary statistic will be presented hereafter.

Table 7 lists the best calibration runs obtained with ABC-RF with SA (eight runs) and ABC-RF without SA (two runs) on the real observation dataset, using R as the summary statistic. In the table, the calibration runs are sorted according to the total R value, which is also shown in the table with the characteristics of each calibration run. The total R value of the simulation that is closest to the real observations (simulation number 4022, see section 2.6) is also provided as a reference.

As already pointed out in the section 3.1.2, the use of sensitivity analysis via the implementation of ABC-RF with SA (algorithm 2) is crucial to make ABC work on a complex model such as the one examined. Indeed, only the calibration runs with ABC-RF with SA lead to total R values lower than the one of the best simulation (simulation number 4022). The best calibration runs of the ABC-RF without SA are reported in table 7 for subsets of 10000 and 25000 simulations, and show poor performances, with a total R value

Table 7
Application to the real observations. List of the calibration runs with the lowest total R values for the ABC-RF with SA (first eight runs) and without SA (first two runs), sorted according to the total R value. The characteristics of each calibration run are detailed in terms of: size of the reference table, calibration method, sorting of the sensitivity indices, variables used in the calibration, and option for the estimation of the parameter value (see sections 2.5.2, 2.5.3 and 2.4.5). The total R value for the closest simulation to the real observations is also provided as a reference.

Reference table size	Method of calibration	Sensitivity index sorting	Variables used in calib.	Parameter estimation	Total R
10000	ABC-RF with SA	Sum	All	$P_{mix,3}$	0.282
25000	ABC-RF with SA	Sum	All	P_{max}	0.327
10000	ABC-RF with SA	Sum	All	P_{max}	0.341
10000	ABC-RF with SA	Sum	All	$P_{mix,2}$	0.342
25000	ABC-RF with SA	Sum	All	$P_{mix,2}$	0.343
10000	ABC-RF with SA	Max	All	P_{max}	0.344
25000	ABC-RF with SA	Max	All	$P_{mix,2}$	0.345
10000	ABC-RF with SA	Max	All	$P_{mix,2}$	0.346
25000	ABC-RF	-	-	P_{med}	0.594
10000	ABC-RF	-	-	P_{med}	0.671
				Sim. number	
			Best simulation:	4022	0.368

about two times larger than that of the best calibration run.

The best calibration run among all is obtained here with the smallest set of 10000 simulations, and by implementing the algorithm 2 (ABC-RF with SA). With a total R value of 0.282, it is significantly better than the other calibration runs. Indeed, the results of the second best calibration run, obtained with a larger set of 25000 simulations, lead to a total R value 16% larger (R=0.327). This total R value is 21% larger for the third best calibration run.

In terms of configurations, the five best calibration runs are all obtained by sorting the parameters on the basis of the value of the sum of the sensitivity indices of the three variables (option "Sum" in section 2.5.2). But none of the four options for estimating parameter values from the posterior distribution seems to be preferable to the others.

The simulations from the best calibration run are plotted in Figure 8 (red lines) and compared with the real observations (gray lines) and with the simulation that is closest to the observations (simulation number 4022, black lines). Compared to the simulation n. 4022, the model simulations after calibration with ABC-RF with SA improve by about 24%. In particular, the peak of chlorophyll is more correctly modelled after calibration, both in terms of timing and maximum concentration; the cyanobacteria concentration is also more accurately simulated, especially during the growth phase. Concerning the oxygen concentration, identifying the initial condition avoided underestimation in the second half of the simulation. After calibration with the ABC-RF with SA, the model seems to be able to reproduce the general behaviour for the three variables correctly. However, the amplitude of the observed daily variations is strongly reduced in the simulations, especially for total chlorophyll and oxygen concentration.

ABC-RF with SA was finally applied using only total chlorophyll data. Several calibration runs with different configurations were performed, as presented in section 2.8 and summarized in table 5. The best calibration run, in terms of R value for total chlorophyll only, was obtained with a subset of 10000 preselected simulations and is shown in Fig. 8 (purple lines). In terms of total chlorophyll (panel a), the simulations from this calibration are better than those from the best calibration on the full dataset (red lines), with R values (for total chlorophyll) equal to 0.0764 and 0.0929, respectively. However, the improvement obtained by focusing only on one variable is marginal (about 15%), especially when looking at the other two variables. Panels b and c of Figure 8 show how phytoplankton growth is attributed to species other than cyanobacteria, and how anoxic conditions are poorly simulated in this case by the model. In particular, the total R value (over the three variables) of the simulation calibrated only on total chlorophyll is equal to 1.377, while that of the best overall calibration is equal to 0.282.

Finally, note that in this application, the alternatives for estimating the parameters $P_{mix,2}$ and $P_{mix,3}$ always coincided with P_{max}, suggesting particularly peaked posterior distributions. Calibration on total chlorophyll data alone was not significantly improved by using a larger subset of 25000 simulations, leading to an R value of 0.0761 for total chlorophyll and a total R value (over the three variables) of 1.377.

4. Discussion

In this paper, Approximate Bayesian Computation with Random Forest (ABC-RF) has been tested for the calibration of a highly parametrized complex biogeochemical model. The calibration procedure focuses on three variables that are particularly relevant to aquatic ecology and water resource management: total chlorophyll, cyanobacteria and dissolved oxygen concentrations.

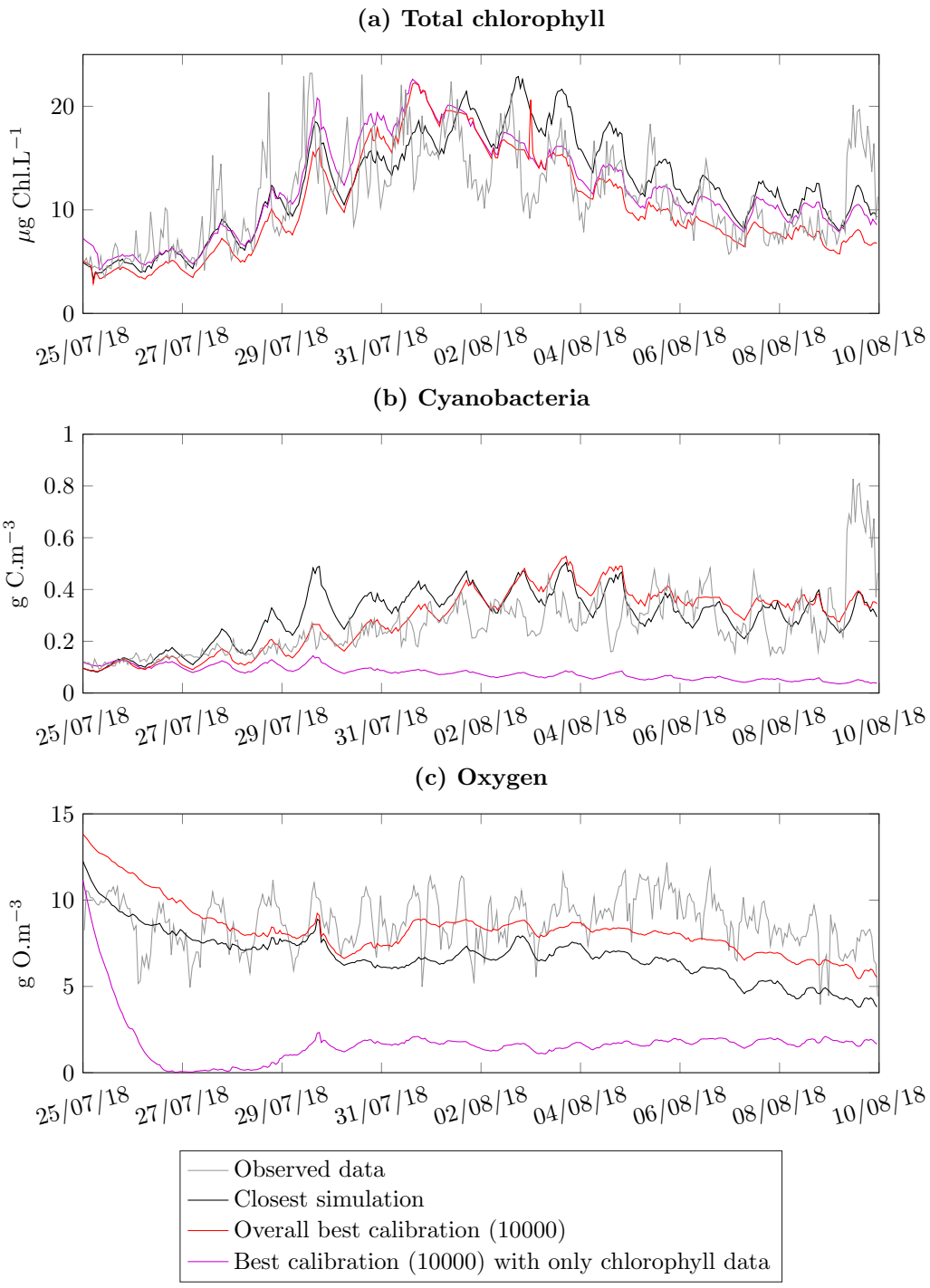

Figure 8: Time series of total chlorophyll (a), cyanobacteria (b) and dissolved oxygen (c) concentrations for: real observations (grey lines), closest simulation to the data (simulation n. 4022, blak lines), best overall calibration run (red lines), and the best calibration considering only total chlorophyll (purple lines).

4.1. A novel approach for parameter inference

Approximate Bayesian Computation (ABC) is a methodology that has quickly become a standard technique for parameter inference [5, 45, 35]. Although ABC theoretically allows Bayesian inference for models of almost arbitrary complexity [47, 3, 12], both deterministic and non-deterministic [45], ABC has not yet been tested on highly parameterized deterministic models. To our knowledge, our application to the Delft3D-BLOOM biogeochemical model is the first to test an ABC methodology for the calibration of a complex physically based model with so many parameters to be estimated.

To date, most applications of ABC for parameter inference either use its standard formulation, in which the pos-

terior distribution is estimated by a rejection algorithm [e.g. 24, 47, 16], or focus on methods that improve the robustness and efficiency of the rejection algorithm (e.g. Markov chain Monte Carlo [28] or sequential Monte Carlo methods [40, 45, 4]).

The ABC random forest (ABC-RF), proposed in 2019, replaces the rejection algorithm with the random forest machine learning technique [35]. There are currently only a few applications of this recent methodology in the scientific literature outside of [35] (e.g. [11, 17]). ABC-RF was tested here in its original form as proposed by Raynal et al. [35], and in a new framework where the results of a sensitivity analysis are integrated into the calibration procedure. This allow taking into account the mutual influence between model parameters and their relative importance with respect to the model outputs.

Preliminary tests were performed on ABC-RF to define a robust configuration in terms of the number of simulations used to generate the reference table and the number of trees in the random forests. These tests show that a reference table generated from at least 25000 simulations is necessary to obtain a good calibration of the three variables under consideration. A comparable value of 20000 simulations was found in a similar application for a model with fewer parameters [24]. In [17], where the authors applied ABC-RF not for parameter inference, but for model selection, a set of 10000 simulations was used. In our application, we also found that the number of simulations could be reduced to 10000 without deteriorating the overall performance of the calibration, provided that the simulations closest to the observations were selected. On the other hand, the test performed on the influence of the number of trees shows that this variable does not have a great impact on the results of the calibration: it has therefore been set at its default value as recommended in [17].

Due to the novelty of the methods, ABC-RF and ABC-RF with SA were tested in different configurations. Both methods were then validated on simulated observations before being applied on the real observations dataset.

Regardless of the number of simulations, the application of the ABC-RF method (without SA) did not allow us to calibrate the model correctly. On the other hand, the calibration procedure greatly benefits from the coupling with the sensitivity analysis proposed here in the algorithm 2. After calibration with the ABC-RF with SA, the simulations resulting from the model have improved considerably, fully justifying the use of this methodology.

4.2. Computational effort

4.2.1. The stock of simulations

The main computational cost in ABC is the generation of the set of simulations from which the reference table is built. The preliminary tests showed that, for the model BLOOM, a set of at least 25000 simulations can drastically reduce both the overall error and the uncertainty of model outcomes. Depending on the model under consideration, this might be a relatively high number of model runs.

However, differently from other popular techniques for automated calibration (e.g. Newton or genetic algorithms), the computational cost of the methods based on ABC resides mainly in the generation of the set of simulations. Once this task is completed, the calibration itself is computationally inexpensive. This allows, for instance, to carry out numerous calibration runs under different configurations. This represents a great advantage, especially in relation to a young methodology such as ABC that still lacks a structured working framework.

4.2.2. The computational impact of coupling ABC-RF and SA

With the implementation of the algorithm 2 (ABC-RF with SA), the computational cost increases: at each iteration the number of summary statistics is incremented, inducing an increase in computation time. Moreover, the calibration of each of the parameters depends on those of the preceding ones and their calculation can therefore not be parallelized. Generally speaking, with a set of 25000 simulations, the ABC-RF with SA can take up to ten times longer than ABC-RF to complete the estimation of all 133 parameters.

In this respect, using a subset of preselected simulations to build the reference table can significantly reduce the computation time taken by the ABC-RF with SA. Our results showed that the use of a smaller number of preselected simulations (the 10000 simulations that are the closest to the observation data) did not deteriorate the calibration results and made it possible to reduce the calculation time by approximately 2/3 compared to a calibration made on 25000 simulations.

Finally, the results of the sensitivity analysis could be exploited to select a reduced number of parameters to include in the calibration, further reducing the computational cost of ABC-RF with SA. Such an approach has not yet been tested in this work. For this, it would be necessary to set a threshold of significance for the parameters according to their sensitivity index, which would subsequently make it possible to discard the parameters of lesser importance.

4.3. Parameter estimation is improved by coupling ABC-RF and SA

4.3.1. The uncertainty in parameter estimation

In Bayesian parameter inference, once the posterior probability is retrieved, estimating the value of the parameter is not always straightforward [23]. Some distributions can be very peaked which facilitates the determination of the value of the parameter. But others will instead show multiple local maxima or will be rather flat [23]. The $P_{mix,2}$ and $P_{mix,3}$ options described in section 2.4.5 were designed to discriminate automatically between peaked and flat distributions, making it easier to estimate the value of the parameter. Among the four options tested for estimating the values of the parameters from the posterior distributions, the results of the application of the calibration on simulated and real observations data showed that the option P_{med} was the least suitable.

In the algorithm 2, new information is added to each iteration of the calibration routine (namely, the values of the previously estimated parameters). After only a few iterations, this has a marked effect on the shape of the posterior distributions which appear smoother and more peaked. This clearly reduces the uncertainty in the estimation of the parameters and ultimately represents a clear advantage for the calibration procedure.

4.3.2. Equifinality

The benefit of reducing uncertainty in parameter estimation is negligible during the very first iterations of the algorithm that deal with the most relevant parameters. This is evident when looking at Fig. 6 which shows the error between real and estimated parameters (calibration performed on simulated data). For some parameters, a considerable error is made. The error is very low for the parameters of medium-high importance (between the 3rd and the 10th position approximately), whereas it increases again for certain parameters having a low sensitivity index. Despite the differences in parameter values between the simulations (set of parameters used to generate the simulated observations and those obtained with the two calibration runs in Fig. 6), the simulations are globally comparable. The total R values given in tables 6 and 7 indeed show that different sets of parameters, resulting from calibration runs with different characteristics, can lead to comparable model performances.

The objective of this calibration work was not to recover the real values of the model parameters, but rather to identify sets of parameters that lead to model simulations close to the observations, for the three variables of interest. Our results suggest that several sets of distinct parameters can thus be obtained. This is known as equifinality: because model variables are related to each other by complex relationships in the model, different sets of parameters can produce equivalent model outputs [2, 20].

The non-uniqueness of the parameterization of complex hydro-ecological models is a known problem [6, 2, 20]. It derives from the fact that the dimension of the observations $\boldsymbol{D}$ is much smaller than those of the vectors of state variables and parameters (x, θ) [6]. Automated calibration procedures therefore seek to optimize certain dynamics which are of a significantly higher order than what can be observed to describe the system [6, 20].

The ability to run multiple calibrations, thanks to the rather low computational cost of ABC methodologies, once the reference table has been built, has highlighted the existence of multiple sets of parameters with equivalent model outputs. This is certainly more related to the complex structure of the model than to the proposed calibration methodology, partly calling into question the idea of seeking an optimal set of parameters for a complex biogeochemical model.

4.4. A specific working framework

In this article, the ABC-RF method has been applied to a specific framework, particularly in terms of available dataset, parameters to be estimated and summary statistics.

4.4.1. Use of high-frequency data

High-frequency measurements of three variables relevant to aquatic ecology and environmental modelling were available for this work. This allowed the calibration effort to be concentrated in a period of 16 days, relatively short for typical hydro-ecological modelling applications, which often extend over a few months at least. The main objective here was to test the ability of a complex biogeochemical model to reproduce a bloom event correctly, and to discriminate the biomass between cyanobacteria and other algal species.

Events of this duration are often completely ignored by traditional limnological monitoring, which is based on periodic sampling or profiling. However, these events are extremely important for the management of water resources in general and for that of our study site in particular, where bathing bans must be issued quickly in the event of the presence of cyanobacteria. Short-term reliable model simulations could be a great advantage in this regard.

Moreover, the choice of a short simulation period also made it possible to contain the computational cost of each model simulation, considerably alleviating the application of ABC-based methods from a computational point of view.

4.4.2. Choice of the parameters

A set of 133 parameters is considered in this work for the calibration. It includes 114 model parameters and 19 initial conditions. The total number of model parameters is considerably higher. However, previous trial-and-error tests have shown that many of these parameters have little influence on the model outputs, at least with respect to the three target variables. This is the case for parameters involved in processes that do not directly affect the target variables. Three main physical processes were targeted in this calibration: those related to algal physiology (e.g. growth, mortality, and sedimentation), oxygen consumption, and nutrient and organic matter evolution. The 114 model parameters included in the calibration were selected based on their physical significance and direct association with the processes of interest.

The choice to include certain initial conditions in the calibration is explained by the fact that the 34 variables listed in table 2 must be initialized. However, some of them, such as the four fractions of particulate organic matter for example, are extremely difficult to measure or estimate, despite their importance in the model. The presence of nutrients in readily available forms or in less accessible compounds clearly influences the model results in terms of phytoplankton dynamics and, subsequently, in terms of oxygen concentration.

In our application, the available data did not allow the concentrations of the different nutrient fractions to be estimated without uncertainty; these concentrations were therefore included in the parameters to be calibrated. Even when measurements are available, they may be affected by a degree of uncertainty justifying their calibration. For example, the scattering that characterizes high-frequency measurements of oxygen and cyanobacteria concentrations intro-

duces uncertainties in the measurements. This is why these two initial conditions have also been included in the list of parameters to be estimated.

However, the sensitivity analysis shows that most of the considered parameters have negligible influence on the model outputs, with 10 (20, respectively) parameters accounting for about 70% (80%, respectively) of the overall variability. According to the sensitivity analysis, nitrogen is the most important nutrient in the system. In particular, its distribution among dissolved and fast-decomposing particulate organic fractions was particularly important for model simulations. The calibration of the initial conditions, when their values are uncertain, can therefore significantly improve the results of the model, and, in the light of the formulation of the model, give new information on the functioning of the system.

The most important physiological parameters were those directly involved in the equation of phytoplankton growth (i.e. the coefficients for temperature dependence of growth and the potential growth rates).

4.4.3. Choice of the summary statistics

The choice of summary statistics is crucial for the ABC. To our knowledge, in all applications of the ABC, summary statistics do not depend on observations. Generally, the set of simulations is used by ABC approaches to generate an inverse model which is intended to be applied to several sets of observations to estimate the associated parameters. This is a great advantage of ABC approaches. However, this cannot be applied to any model. In our case, for example, the set of simulations depends on the meteorological conditions that are specific to the period under consideration. In this case, the inverse model generated by the ABC method is also specific and only remains valid for the period considered.

This is the reason why it was possible in our case to use summary statistics that depend directly on the observations, such as R and $NMSE$. Other choices of summary statistics independent of the observations have been tested to describe the time series (e.g. series of successive means, spline projection coefficients). However, using error measures as summary statistics has proven to be the most effective.

Finally, this particular framework also justifies the use of a subset of preselected simulations to run the ABC, as we have suggested. Since the subset of simulations can only be used to calibrate the model on a specific set of observations, simulations that deviate from the target behaviour do not provide any useful information and can therefore be removed from the reference table.

4.4.4. Analysis of the performance of the model BLOOM

The model calibrated on the set of real observation data reproduces very well the general behaviour of the three target variables over the selected period. However, the observations show a strong sub-daily variability which is not entirely reproduced by the model. This is probably due to the structure of the model rather than the calibration methodology. Complex biogeochemical models are generally designed to represent dynamics that extend over longer time periods than those simulated here (i.e., monthly to seasonal), and often do not explicitly model processes at a sub-daily scale [e.g. 14]. Moreover, the configuration set up for this work, i.e. the set of substances and processes activated in the model, might not be optimal. A large number of models characterized by different degrees of complexity are available to simulate the biogeochemical cycle in aquatic ecosystems. A large literature has already addressed their advantages and disadvantages [e.g. 1, 34, 50, 22], highlighting the impossibility of fully validating such models due to the complexity of the biogeochemical cycle and the lack of commonly available observations [50, 22]. In this respect, our configuration, although complex, describes only part of the real natural ecosystem. For example, benthic processes, macrophytes and zooplankton are not explicitly included. However, as with all modelling efforts, the challenge is to find the right level of complexity for the dynamics of interest.

In this work, the comparison with three variables measured at high-frequency shows that short-term phytoplankton blooms can be simulated with a model integrating relatively basic processes that can be easily measured (i.e. growth, mortality, nutrient uptake, oxygen production and decomposition of organic matter). However, without additional data, it is not possible to assess accurately the importance of certain processes such as the mineralization of organic matter, for example. This problem was highlighted with the results of the model calibration which was carried out using only total chlorophyll data (this variable was chosen because of its importance for the management of aquatic ecosystems [52]). After calibration with total chlorophyll data only, the model results were only slightly improved in terms of the target variable compared to the best calibration using data from the three available variables. On the other hand, the dynamics in terms of oxygen and cyanobacteria concentration were extremely inaccurate. This result highlights the importance of gathering the widest range of data possible to assess the performance of a model aimed at describing a complex natural system. Only the comparison with several variables makes it possible to determine whether the overall functioning of the system considered is apprehended as a whole or not [e.g. 50, 22].

5. Conclusion

Biogeochemical models are often highly parameterized and complex. Their calibration is difficult and often neglected in the scientific literature. Our study shows that, among the various techniques available for automated calibration, ABC-RF can be successfully applied to calibrate a complex and highly parameterized biogeochemical model. Our work focuses on a short-term algal bloom, an event that could possibly be missed by a traditional periodic survey. After calibration, the model was able to reproduce the rapid biogeochemical dynamics that extend over a relatively short period. The growth and mortality of phytoplankton, as well

as the evolution of the concentrations of cyanobacteria and dissolved oxygen, were correctly simulated.

To obtain such results, the coupling of the ABC-RF with a sensitivity analysis (SA) via the algorithm 2 was crucial, as well as the availability of high-frequency data. Indeed, the main computational effort required by the ABC is dedicated to the generation of a set of simulations, which must be composed of at least 25000 simulations for the optimization of more than 100 parameters. The computational cost of the ABC algorithm itself (once the simulations have been performed) can be reduced by preselecting 10000 simulations among the simulations available to build the reference table.

The summary statistics have been defined here based on the expertise of the modeller, an approach followed in most ABC applications so far (eg [24, 16]). This highlights the importance of the modeller's experience and knowledge, which remains an essential feature of the Bayesian approach for parameter inference, and which should not be ignored also when applying automated calibration methods.

In optimization techniques such as local gradient-based methods, the exploration of the parameter space depends on the initial values of the parameters chosen by the user. This is not the case for ABC where the parameter space is explored from user-defined prior distributions. In this regard, ABC could be a useful technique to define appropriate initial parameter values for the application of other calibration algorithms.

Finally, attention should also be paid to the data sets available when approaching the calibration of a complex hydro-ecological model. We have indeed shown that the use of measurements of several variables considerably improves the overall performance of the model.

6. Replication of results

6.1. Software availability

Version 5.01.03.000000 of the model Delft3D-DELWAQ and version 6.01.06.62914 of model Delft3D-FLOW2D3D of the modelling suite software called Delft3D has been used for the simulation of the concentration of the total chlorophyll, the phycocyanin and the dissolved oxygen in the lake Champs-sur-Marne. This software is open-source and the version used in this study can be downloaded at https://svn.oss.deltares.nl/repos/delft3d/tags/delft3d4/3426.

The *R* and Matlab scripts used to implement the calibration methods (standard ABC, ABC-RF and ABC-RF with SA) on the case of lake Champs-sur-Marne and perform the analysis of calibration results presented in this paper are preserved in a dataset (https://doi.org/10.15454/QSR3YO, [30]) published on the French repository "Recherche Data Gouv" (https://entrepot.recherche.data.gouv.fr/).

A project called *Calibration_ABC-RF-SA* has been created in the software repository Gitlab of the INRAE institute with all the scripts and functions (written only in R) of the calibration methods ABC, ABC-RF and ABC-RF-SA, and some scripts (in R) to apply these methods on a "toy example". This repository is accessible at: https://forgemia.inra.fr/simlake/calibration_abc-rf-sa

6.2. Data availability

The real observation data, the simulated data of the 30000 simulations that have been performed to make the calibration, the values of the model parameters of the 30000 simulations, and the files with the results of the calibration runs performed in this study (preliminary tests and application of the calibration methods on simulated and real observation data) are also stored in the dataset (https://doi.org/10.15454/QSR3YO, [30]) published on the French repository "Recherche Data Gouv" (https://entrepot.recherche.data.gouv.fr/).

7. Acknowledgement

This work was supported by the French National Research Agency [ANSWER research project, grant number ANR-16-CE32-0009-02].

The authors would like to thank Max Zinsou Debaly for his internship work, Sébastien Roux for his help in choosing the sensitivity analysis method and Isabelle Sanchez who helped us with the sharing of the datasets and simulation codes.

A. Computation of the sensitivity indices

In this section, the methodology proposed in [48] for calculating sensitivity indices for models with correlated input parameters is presented.

Consider the following model:

$$Y = \eta(X) \tag{8}$$

where $\eta : \mathbb{R}^d \to \mathbb{R}$ is a given continuous function, $X = (X_1, ..., X_d)$ is the vector of model parameters and $Y \in \mathbb{R}$ is the model output. Both Y and X are considered as random variables. In variance-based methods for sensitivity analysis, the impact of variations in the input X_i on the variance of Y (Var(Y)) is evaluated by calculating the first order sensitivity index S_i:

$$S_i = \frac{\mathrm{Var}(\mathbb{E}(Y|X_i))}{\mathrm{Var}(Y)} \tag{9}$$

Under the assumption of independent inputs, various techniques are available to estimate S_i [41, 13]. However, this assumption is often not verified in practice, as in our case. For this reason we rely here on the methodology proposed by Da Veiga *et al* [48] for models with correlated inputs X_i, to compute an estimate of S_i based on local polynomial approximation of the conditional statistical moment $\mathbb{E}(Y|X_i)$. The methodology is described hereafter.

Consider $\left(X_i^k\right)_{k=1,...,n}$ and $\left(\tilde{X}_i^l\right)_{l=1,...,n'}$ two parameters samples from the joint distribution of the d-dimensional input $X = (X_1, ..., X_d)$. The methodology we have used to compute an estimate of S_i is the one proposed by Da Veiga *et al* [48] that is composed of 4 steps:

1. For each parameter set $X^k = (X_1^k, ..., X_d^k)$, $k = 1 : n$, computation of the model outputs $Y^k = \eta(X^k)$

2. Estimation of the variance $\mathrm{Var}(Y)$ with the classical unbiased estimator:

$$\hat{\sigma}_Y^2 = \frac{1}{n-1}\sum_{k=1}^{n}(Y^k - \bar{Y})^2 \tag{10}$$

3. Estimation of the conditional statistical moment $m_i(x_i) = \mathbb{E}(Y|X_i = x_i)$ which can be locally approximated by a polynomial of order p of the form:

$$\sum_{j=0}^{p}\beta_j(x_i^0)(x_i - x_i^0)^j \tag{11}$$

for any x_i in the neighborhood of x_i^0. To obtain the local polynomial approximation denoted $\hat{m}_i(x_i)$ in the sequel, and determine the parameters $\beta_j(x_0)$, the function 'loess' of the R-package 'nprobust' has been used.

4. Estimation of the variance $\mathrm{Var}(\mathbb{E}(Y|X_i)) = \mathrm{Var}(m_i(X_i))$ from the sample $\left(\tilde{X}_i^l\right)_{l=1,..,n'}$ using the classical empirical variance expression:

$$\hat{T}_i = \frac{1}{n'-1}\sum_{l=1}^{n'}\left(\hat{m}_i(\tilde{X}_i^l) - \hat{\bar{m}}_i\right)^2 \tag{12}$$

where $\hat{\bar{m}}_i = \frac{1}{n'}\sum_{l=1}^{n'}\hat{m}_i(\tilde{X}_i^l)$.

5. Computation of the estimate $\hat{S}_i$ of the sensitivity index S_i by the following formula:

$$\hat{S}_i = \frac{\hat{T}_i}{\hat{\sigma}_Y^2} \tag{13}$$

B. Complementary results about the preliminary tests

In this appendix, some numerical results are presented to illustrate the conclusions given in section 3.1.1 on the influence (on the calibration results) of the number of trees and the options for estimating the value of the parameter from the posterior distribution.

In table 8, the value of the total $NMSE$ between the simulated observations and the results of the model calibrated with ABC-RF is given, depending on the size of the subset of simulations used to generate the reference table and the number of trees used in the random forest. The results given in this table correspond to calibration runs where the closest simulations are preselected to build the reference table, and where the parameter value estimation option is $P_{mix,2}$. Similar results are obtained with the other options for estimating the values of the parameters and when the simulations are randomly selected from the set of available simulations. As we can see, the number of trees used to build the random forests did not show any particular influence on the calibration results.

In table 9, the value of the total $NMSE$ between the simulated observations and the results of the model calibrated with ABC-RF is given, depending on the number of simulations used to generate the reference table and the options chosen for estimating the value of the parameter from the posterior distribution. The number of trees used to build the random forest is here equal to 500. As can be seen, the option that leads to the smallest value of the total $NMSE$ varies according to the number of simulations used to generate the reference table and the options for the preselection of the subset of simulations.

References

[1] Anderson, T.R., 2005. Plankton functional type modelling: running before we can walk? J Plankton Res 27, 1073–1081. doi:10.1093/plankt/fbi076.

[2] Arhonditsis, G.B., Perhar, G., Zhang, W., Massos, E., Shi, M., Das, A., 2008. Addressing equifinality and uncertainty in eutrophication models. Water Resources Research 44. doi:https://doi.org/10.1029/2007WR005862.

[3] Beaumont, M.A., 2010. Approximate Bayesian Computation in Evolution and Ecology. Annual Review of Ecology, Evolution, and Systematics 41, 379–406. doi:10.1146/annurev-ecolsys-102209-144621.

[4] Beaumont, M.A., Cornuet, J.M., Marin, J.M., Robert, C.P., 2009. Adaptive approximate Bayesian computation. Biometrika 96, 983–990. doi:10.1093/biomet/asp052. arXiv: 0805.2256.

[5] Beaumont, M.A., Zhang, W., Balding, D.J., 2002. Approximate Bayesian Computation in Population Genetics. Genetics 162, 2025–2035. Publisher: Genetics Section: Investigations.

[6] Beck, M.B., 1987. Water quality modeling: A review of the analysis of uncertainty. Water Resources Research 23, 1393–1442. doi:https://doi.org/10.1029/WR023i008p01393.

[7] Beck, R., Xu, M., Zhan, S., Liu, H., Johansen, R.A., Tong, S., Yang, B., Shu, S., Wu, Q., Wang, S., Berling, K., Murray, A., Emery, E., Reif, M., Harwood, J., Young, J., Martin, M., Stillings, G., Stumpf, R., Su, H., Ye, Z., Huang, Y., 2017. Comparison of Satellite Reflectance Algorithms for Estimating Phycocyanin Values and Cyanobacterial Total Biovolume in a Temperate Reservoir Using Coincident Hyperspectral Aircraft Imagery and Dense Coincident Surface Observations. Remote Sensing 9, 538. doi:10.3390/rs9060538.

[8] Biau, G., Devroye, L., 2010. On the layered nearest neighbour estimate, the bagged nearest neighbour estimate and the random forest method in regression and classification. Journal of Multivariate Analysis 101, 2499–2518.

[9] Burr, T., Skurikhin, A., 2013. Selecting Summary Statistics in Approximate Bayesian Computation for Calibrating Stochastic Models. BioMed Research International doi:https://doi.org/10.1155/2013/210646.

[10] 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, 260–269. doi:10.1016/j.jglr.2012.01.008.

[11] Chapuis, M.P., Raynal, L., Plantamp, C., Meynard, C.N., Blondin, L., Marin, J.M., Estoup, A., 2020. A young age of subspecific divergence in the desert locust inferred by ABC random forest. Molecular Ecology 29, 4542–4558. doi:https://doi.org/10.1111/mec.15663.

[12] Csilléry, K., Blum, M.G.B., Gaggiotti, O.E., François, O., 2010. Approximate Bayesian Computation (ABC) in practice. Trends in Ecology & Evolution 25, 410–418. doi:10.1016/j.tree.2010.04.001.

[13] Cukier, R., Fortuin, C., Shuler, K.E., Petschek, A., Schaibly, J.H., 1973. Study of the sensitivity of coupled reaction systems to uncertainties in rate coefficients. i theory. The Journal of chemical physics 59, 3873–3878.

[14] Deltares, 2018a. D-Water Quality User Manual. Delft Hydraulics, Delft.

[15] Deltares, 2018b. D-Water Quality User Manual. Delft Hydraulics, Delft.

[16] Dominguez Almela, V., Palmer, S.C.F., Gillingham, P.K., Travis, J.M.J., Britton, J.R., 2020. Integrating an individual-based model

Number of simulations	Number of trees								Mean value	standard deviation
	250	500	750	1000	1250	1500	1750	5000		
2000	5.1229	6.8319	1.6552	1.6181	2.0003	1.6721	1.6324	1.4634	2.7495	2.0494
10000	8.0599	0.9985	0.4979	0.3859	0.7645	0.5888	0.9296	5.0760	2.1626	2.8410
15000	1.1205	0.7562	0.9313	2.0830	3.0540	1.2979	1.1802	1.2144	1.4547	0.7544
20000	1.0874	0.3675	0.3970	0.7469	1.2964	2.4549	1.2106	3.3308	1.3614	1.0331
25000	1.4088	1.8528	1.3224	1.3768	2.1155	1.7682	1.4167	2.1665	1.6785	0.3437
30000	1.3480	1.6790	1.9491	1.9163	2.2417	1.5650	1.9026	1.9620	1.8205	0.2773

Table 8
Preliminary tests (complementary results). Total $NMSE$ between the simulated observations and the model results calibrated with ABC-RF, according to the number of simulations used to generate the reference table and the number of trees used in the forest. The closest simulations are preselected to build the reference table, and the option for parameter values estimation is $P_{mix,2}$.

Option for Parameter estimation	Number of simulations					
	5000	10000	15000	20000	25000	30000
	randomly selected					
P_{max}	3.6606	3.7882	5.8984	5.7134	1.5134	1.4637
P_{med}	7.1250	5.4202	**2.5318**	4.8222	**0.4389**	**1.1439**
$P_{mix,2}$	3.6606	3.8847	5.9867	5.2575	1.5134	1.6790
$P_{mix,3}$	**2.3667**	**3.4322**	4.1556	**0.7414**	1.7087	1.4220
	preselected					
P_{max}	7.1570	0.8627	**0.7548**	0.4209	1.8528	
P_{med}	3.8034	3.4905	3.5417	1.9831	**1.2346**	
$P_{mix,2}$	6.8319	0.9985	0.7562	0.3675	1.8528	
$P_{mix,3}$	**1.6672**	**0.7376**	1.5152	**0.3150**	1.5677	

Table 9
Preliminary tests (complementary results). Total $NMSE$ between the simulated observations and the model results calibrated with ABC-RF, according to the number of simulations used to generate the reference table and the options chosen for the estimation of the parameter value from the posterior distribution. The number of trees used to build the random forest is equal to 500. The values in bold correspond to the smallest values obtained among the four options for parameter estimation.

with approximate Bayesian computation to predict the invasion of a freshwater fish provides insights into dispersal and range expansion dynamics. Biol Invasions 22, 1461–1480. doi:10.1007/s10530-020-02197-6.

[17] Estoup, A., Raynal, L., Verdu, P., Marin, J.M., 2018. Model choice using approximate bayesian computation and random forests: analyses based on model grouping to make inferences about the genetic history of pygmy human populations. Journal de la Société Française de Statistique 159, 167–190.

[18] Fenocchi, A., Rogora, M., Morabito, G., Marchetto, A., Sibilla, S., Dresti, C., 2019. Applicability of a one-dimensional coupled ecological-hydrodynamic numerical model to future projections in a very deep large lake (Lake Maggiore, Northern Italy/Southern Switzerland). Ecological Modelling 392, 38–51. doi:10.1016/j.ecolmodel.2018.11.005.

[19] Geider, R.J., MacIntyre, H.L., Kana, T.M., 1997. Dynamic model of phytoplankton growth and acclimation: responses of the balanced growth rate and the chlorophyll a: carbon ratio to light, nutrient-limitation and temperature. Marine Ecology Progress Series 148, 187–200. Publisher: Inter-Research Science Center.

[20] Hipsey, M.R., Gal, G., Arhonditsis, G.B., Carey, C.C., Elliott, J.A., Frassl, M.A., Janse, J.H., de Mora, L., Robson, B.J., 2020. A system of metrics for the assessment and improvement of aquatic ecosystem models. Environmental Modelling & Software 128, 104697. doi:10.1016/j.envsoft.2020.104697.

[21] Hodges, B., 2014. Hydrodynamical Modeling, in: Encyclopedia of Inland Waters, p. 22 pgs. doi:10.1016/B978-0-12-409548-9.09123-5.

[22] Kriest, I., 2017. Calibration of a simple and a complex model of global marine biogeochemistry. Biogeosciences 14, 4965–4984. doi:10.5194/bg-14-4965-2017.

[23] Kruschke, J.K., 2018. Rejecting or Accepting Parameter Values in Bayesian Estimation. Advances in Methods and Practices in Psychological Science 1, 270–280. doi:10.1177/2515245918771304. publisher: SAGE Publications Inc.

[24] Lagarrigues, G., Jabot, F., Lafond, V., Courbaud, B., 2015. Approximate Bayesian computation to recalibrate individual-based models with population data: Illustration with a forest simulation model. Ecological Modelling 306, 278–286. doi:10.1016/j.ecolmodel.2014.09.023.

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

[26] Mahevas, S., Picheny, V., Lambert, P., Dumoulin, N., Rouan, L., Soulié, J.C., Brockhoff, D., Lehuta, S., Riche, R.L., Faivre, R., Drouineau, H., 2019. A Practical Guide for Conducting Calibration and Decision-Making Optimisation with Complex Ecological Models. Publisher: Preprints doi:10.20944/preprints201912.0249.v1.

[27] Makler-Pick, V., Gal, G., Gorfine, M., Hipsey, M.R., Carmel, Y., 2011. Sensitivity analysis for complex ecological models – A new approach. Environmental Modelling & Software 26, 124–134. doi:10.1016/j.envsoft.2010.06.010.

[28] Marjoram, P., Molitor, J., Plagnol, V., Tavaré, S., 2003. Markov chain Monte Carlo without likelihoods. PNAS 100, 15324–15328. doi:10.1073/pnas.0306899100. publisher: National Academy of Sciences Section: Physical Sciences.

[29] Nott, D.J., Ong, V.M.H., Fan, Y., Sisson, S., 2018. High-dimensional abc, in: Handbook of approximate Bayesian computation. Chapman and Hall/CRC, pp. 211–241.

[30] Piccioni, F., Casenave, C., Baragatti, M., Cloez, B., Vinçon-Leite, B., 2022. Approximate Bayesian Computation with Random Forest and Sensitivity Analysis for the calibration of a complex aquatic ecological model (dataset V1). Recherche Data Gouv doi:10.15454/QSR3YO.

[31] Piccolroaz, S., Amadori, M., Toffolon, M., Dijkstra, H.A., 2019. Importance of planetary rotation for ventilation processes in deep elongated lakes: Evidence from Lake Garda (Italy). Scientific Reports 9, 1–11. doi:10.1038/s41598-019-44730-1.

[32] Poli, A.A., Cirillo, M.C., 1993. On the use of the normalized mean square error in evaluating dispersion model performance. Atmospheric Environment. Part A. General Topics 27, 2427–2434. doi:10.1016/0960-1686(93)90410-Z.

[33] Prangle, D., 2015. Summary Statistics in Approximate Bayesian Computation. arXiv:1512.05633 [math, stat] ArXiv: 1512.05633.

[34] Raick, C., Soetaert, K., Grégoire, M., 2006. Model complexity and performance: How far can we simplify? Progress in Oceanography 70, 27–57. doi:10.1016/j.pocean.2006.03.001.

[35] Raynal, L., Marin, J.M., Pudlo, P., Ribatet, M., Robert, C.P., Estoup, A., 2019. ABC random forests for Bayesian parameter inference. Bioinformatics 35, 1720–1728. doi:10.1093/bioinformatics/bty867. publisher: Oxford Academic.

[36] Reichert, P., Omlin, M., 1997. On the usefulness of overparameterized ecological models. Ecological Modelling 95, 289–299. doi:10.1016/S0304-3800(96)00043-9.

[37] Rigosi, A., Marcé, R., Escot, C., Rueda, F.J., 2011. A calibration strategy for dynamic succession models including several phytoplankton groups. Environmental Modelling & Software 26, 697–710. doi:10.1016/j.envsoft.2011.01.007.

[38] Shimoda, Y., Arhonditsis, G.B., 2016. Phytoplankton functional type modelling: Running before we can walk? A critical evaluation of the current state of knowledge. Ecological Modelling 320, 29–43. doi:10.1016/j.ecolmodel.2015.08.029.

[39] Simola, U., Cisewski-Kehe, J., Gutmann, M.U., Corander, J., 2021. Adaptive Approximate Bayesian Computation Tolerance Selection. Bayesian Analysis -1, 1–27. doi:10.1214/20-BA1211. publisher: International Society for Bayesian Analysis.

[40] Sisson, S.A., Fan, Y., Tanaka, M.M., 2007. Sequential Monte Carlo without likelihoods. PNAS 104, 1760–1765. doi:10.1073/pnas.0607208104.

[41] Sobol', I., 1993. Sensitivity estimates for nonlinear mathematical models. Math. Model. Comput. Exp 1, 407–414.

[42] Soulignac, F., Vinçon-Leite, B., Lemaire, B.J., Scarati Martins, J.R., Bonhomme, C., Dubois, P., Mezemate, Y., Tchiguirinskaia, I., Schertzer, D., Tassin, B., 2017. Performance Assessment of a 3D Hydrodynamic Model Using High Temporal Resolution Measurements in a Shallow Urban Lake. Environ Model Assess 22, 309–322. doi:10.1007/s10666-017-9548-4.

[43] Sunnåker, M., Busetto, A.G., Numminen, E., Corander, J., Foll, M., Dessimoz, C., 2013. Parameter estimation by approximate bayesian computation: a conceptual overview. URL: https://doi.org/10.1371/journal.pcbi.1002803.g001. figure.

[44] Sunnåker, M., Busetto, A.G., Numminen, E., Corander, J., Foll, M., Dessimoz, C., 2013. Approximate Bayesian Computation. PLOS Computational Biology 9, e1002803. doi:10.1371/journal.pcbi.1002803. publisher: Public Library of Science.

[45] Toni, T., Welch, D., Strelkowa, N., Ipsen, A., Stumpf, M.P., 2009. Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. Journal of The Royal Society Interface 6, 187–202. doi:10.1098/rsif.2008.0172. publisher: Royal Society.

[46] Tran Khac, V., Hong, Y., Plec, D., Lemaire, B.J., Dubois, P., Saad, M., Vinçon-Leite, B., 2018. An Automatic Monitoring System for High-Frequency Measuring and Real-Time Management of Cyanobacterial Blooms in Urban Water Bodies. Processes 6, 11. doi:10.3390/pr6020011.

[47] van der Vaart, E., Beaumont, M.A., Johnston, A.S.A., Sibly, R.M., 2015. Calibration and evaluation of individual-based models using Approximate Bayesian Computation. Ecological Modelling 312, 182–190. doi:10.1016/j.ecolmodel.2015.05.020.

[48] Veiga, S.D., Wahl, F., Gamboa, F., 2009. Local Polynomial Estimation for Sensitivity Analysis on Models With Correlated Inputs. Technometrics 51, 452–463. doi:10.1198/TECH.2009.08124.

[49] Vinçon-Leite, B., Casenave, C., 2019. Modelling eutrophication in lake ecosystems: A review. Science of The Total Environment 651, 2985–3001. doi:10.1016/j.scitotenv.2018.09.320.

[50] Ward, B.A., Schartau, M., Oschlies, A., Martin, A.P., Follows, M.J., Anderson, T.R., 2013. When is a biogeochemical model too complex? Objective model reduction and selection for North Atlantic time-series sites. Progress in Oceanography 116, 49–65. doi:10.1016/j.pocean.2013.06.002.

[51] Weiss, R.F., 1970. The solubility of nitrogen, oxygen and argon in water and seawater. Deep Sea Research and Oceanographic Abstracts 17, 721–735. doi:10.1016/0011-7471(70)90037-9.

[52] World Health Organization, 2003. Guidelines for safe recreational water environments. Volume 1, Coastal and fresh waters. Technical Report. Publisher: World Health Organization.