Oxygen data assimilation for estimating micro-organism communities' parameters in river systems

Shuaitao Wang, Nicolas Flipo, Thomas Romary

Version of Record: https://www.sciencedirect.com/science/article/pii/S004313541930795X
Manuscript_8a2a4b10f4d7e3bc441b842e6225b0fc

Oxygen data assimilation for estimating micro-organism communities' parameters in river systems

Shuaitao Wang[a,*], Nicolas Flipo[a], Thomas Romary[a]

[a]*Geosciences and Geoengineering Department, MINES ParisTech, PSL University, 35 Rue Saint-Honoré 77300 Fontainebleau, France*

Abstract

The coupling of high frequency data of water quality with physically based models of river systems is of great interest for the management of urban socio-ecosystems. One approach to exploit high frequency data is data assimilation which has received an increasing attention in the field of hydrology, but not for water quality modeling so far. We present here a first implementation of a particle filtering algorithm into a community-centered hydro-biogeochemical model to assimilate high frequency dissolved oxygen data and to estimate metabolism parameters in the Seine River system. The procedure is designed based on the results of a former sensitivity analysis of the model (Wang et al., 2018) that allows for the identification of the twelve most sensible parameters all over the year. Those parameters are both physical and related to micro-organisms (reaeration coefficient, photosynthetic parameters, growth rates, respiration rates and optimal temperature). The performances of the approach are assessed on a synthetic case study that mimics 66 km of the Seine River. Virtual dissolved oxygen data are generated using time varying parameters. This paper aims at retrieving the predefined parameters by assimilating those data. The simulated dissolved oxygen concentrations match the reference concentrations. The identification of the parameters depends on the hydrological and trophic contexts and more surprisingly on the thermal state of the river. The physical, bacterial and phytoplanktonic parameters can be retrieved properly, leading to the differentiation of two successive algal blooms by comparing the estimated posterior distribution of the optimal temperature for phytoplankton growth. Finally, photosynthetic parameters' distributions following circadian cycles during algal blooms are discussed.

Keywords: Data assimilation, Dissolved oxygen, Parameter estimation, Particle filter, ProSe-PA

1. Introduction

The coupling of high frequency data of water quality with physically based model of river systems is of great interest for the management of urban socio-ecosystems. One approach to exploit high frequency data is data assimilation

*Corresponding author
Email addresses: shuaitao.wang@mines-paristech.fr, shuaitaowang@outlook.com (Shuaitao Wang), nicolas.flipo@mines-paristech.fr (Nicolas Flipo), thomas.romary@mines-paristech.fr (Thomas Romary)

which combines observations and short-range forecasts to estimate the distribution of the true state of a process (Wikle and Berliner, 2007). Assimilating high frequency data allows for the identification of the multiple sources of model uncertainty relative to parameters, model structure, forcing data (e.g. temperature, wind speed and solar radiation) and observations (Evensen, 2003), which limit the validation and application of water quality models (Beven, 1989; Polus et al., 2011).

A lot of data assimilation techniques exist in literature such as the variational methods (Sasaki, 1955, 1958), the Kalman filter (KF, (Kalman, 1960)), the extended Kalman filter (EKF, (Beck, 1987)), the ensemble Kalman filter (EnKF, (Evensen, 1994)) and the particle filter (PF, (Doucet et al., 2001)). A review of these methods can be found in the literature (Wikle and Berliner, 2007; Cappe et al., 2007; Särkkä, 2013). The variational methods minimizing a cost function have been widely used in numerical weather prediction (Courtier et al., 1994, 1998; Gauthier et al., 2007; Kleist and Ide, 2015; Yucel et al., 2015). The major drawbacks of variational methods are the numerical implementation complexity and the computational cost. The KF handles only linear models with Gaussian errors, which are not consistent with the hydrologic and water quality models.

To extend to nonlinear models, the EKF uses the linearized formula tangent to the forward model, which is an approximation and would be very costly to implement for high-dimensional systems (Evensen, 2003). The EnKF uses Monte Carlo samples to approximate the forecast distribution and then applies linear update formulas to obtain the posterior distribution (Evensen, 2003), while the particle filter based methods estimate the forecast and posterior distributions using discrete probability densities (Arulampalam et al., 2002) obtained via Bayes' theorem (Bayes, 1763). Since the EnKF and particle filter resolve the two major drawbacks related to the use of EKF and handle highly nonlinear models, the EnKF and particle filter based methods have become the most commonly used sequential data assimilation techniques in hydrologic modeling for state-parameter estimation (Moradkhani et al., 2005; Weerts and El Serafy, 2006; Andreadis et al., 2007; Salamon and Feyen, 2009; Plaza et al., 2012; DeChant and Moradkhani, 2012; Vrugt et al., 2013; Shi et al., 2014; Abbaszadeh et al., 2018).

In biogeochemical oceanography, the efficiency of the ensemble-based Kalman filters has been assessed on ocean models for ocean biogeochemical state and parameter estimation (Simon and Bertino, 2012; Simon et al., 2012; Gharamti et al., 2017; Yu et al., 2018). Due to the lack of data, the difficulties in representing complex hydro-ecosystems and the computational cost, there are still relatively few applications about the state-parameter estimation by assimilating high-frequency dissolved oxygen (DO) concentrations in river systems. Pastres et al. (2003) have applied the EKF to update three parameters of a simple DO-chlorophyll model in the lagoon of Venice. Mao et al. (2009) used EKF to forecast algal blooms and dissolved oxygen dynamics in a marine ecosystem. The recent studies focus particularly on forecasting algal bloom dynamics using EnKF in river system (Kim et al., 2014) or in lakes

(Huang et al., 2013; Page et al., 2018). Xue et al. (2012) applied also EnKF in the Massachusetts Bay to design the optimal monitoring sites for DO measurements.

However, as stated by Wikle and Berliner (2007, p. 10-11), "We assume that the forecast distibution can be characterized by its first two moments (or, equivalently, that it is Gaussian with mean and (estimated) variance/covariance matrix). ... However, in nonlinear cases, since Gaussianity cannot hold for all time, the EnKF must yield biased samples and estimates, even for unlimited sample sizes". The recent investigations concluded also that the assumption of a Gaussian error structure for the forecast distribution may not be realistic for hydrologic systems and phytoplankton dynamics and the authors suggested testing the particle filter based methods to overcome this problem (Plaza et al., 2012; Pasetto et al., 2012; DeChant and Moradkhani, 2012; Huang et al., 2013). The feasibility of Bayesian inference with a Particle Markov Chain Monte Carlo algorithm is tested on a simple predator-prey model in ecological research (Kattwinkel and Reichert, 2017).

In this paper, we present the ProSe-PA model, which consists in a first implementation of a particle filtering algorithm into the community-centered hydro-biogeochemical model ProSe (Even et al., 1998, 2007b; Flipo et al., 2004; Vilmin et al., 2015b) in order to assimilate 15 min-DO data and estimate metabolisms' parameters in the Seine River system. The procedure is designed from the results of a sensitivity analysis of the biogeochemical module, C-RIVE , of the ProSe model that identified the twelve most sensible parameters of the model (Wang et al., 2018). Those parameters are both physical (water re-aeration by fluvial navigation) and physiological (for instance growth rate of heterotrophic bacteria and photosynthetic parameters of phytoplankton). The performances of ProSe-PA are assessed on a synthetic case study that mimics 66 km of the Seine River and generates virtual "observation" data of DO. The objectives of this study are to retrieve the predefined parameters used to generate the "observation" data and to distinguish two successive algal blooms by identifying different physiological properties.

The manuscript is organized as follows. The section 2 presents the ProSe model and the study area, followed by the ProSe-PA approach including the mathematical formulations of the particle filter, the resampling algorithm as well as the full numerical algorithm. The considered parameters, the "observation" data and the input data of ProSe model are described in section 2.3. The numerical settings and the computational cost are given in section 3.1. Then, we evaluate the simulated DO concentrations by four statistical criteria (section 3.2) and show the identification of metabolisms' parameters in frame of state classification (section 3.3). The real time parameter identification is discussed in section 4.1, followed by the identifiability of two successive algal blooms by comparing the different physiological properties. The photosynthetic parameters' distributions following circadian cycles are described in section 4.3. The sensitivity of ProSe-PA performances to the observation error is discussed in the section 4.4. To finish, a brief conclusion is given in section 5.

2. Material and methods

2.1. Synthetic case study

2.1.1. The ProSe model

The ProSe model (Even et al., 1998, 2007b; Flipo et al., 2004; Vilmin et al., 2015b), couples three libraries (a hydrodynamic library, a transport library and a biogeochemical library), that simulate together the hydro-biogeochemical functioning of a river system. The hydraulic module solves the 1D shallow water equations with a finite volume scheme. The transport module simulates the advection and dispersion of both particulate and dissolved substances, including also water re-oxygenation by overflowing over hydraulic works. The biogeochemical library is based on the RIVE conceptual model which is a community-centered model (Billen et al., 1994; Garnier et al., 1995) (https://www.fire.upmc.fr/rive). The cycles of carbon, nutrients and dissolved oxygen are simulated in both water column and sediment layer. The compound exchanges between water column and benthic layers have been successively developed (Even et al., 2004, 2007a; Flipo et al., 2004, 2007; Vilmin et al., 2015b). The ProSe model has been well validated and largely applied in the Seine River system (Even et al., 1998, 2004, 2007a; Polus et al., 2011; Raimonet et al., 2015; Vilmin et al., 2015a,b, 2016, 2018).

2.1.2. Study area

The study area is located downstream Paris city and consists of 66 km of the Seine River (Fig. 1). The bathymetry data of the channel is extracted from the full 220 km Seine model used in recent studies (Vilmin et al., 2016, 2018). This area is highly impacted by human activities including two Waste Water Treatment Plants (WWTPs, Seine Aval and Seine Centre) and two major Combined Sewer Overflows (CSOs). The Seine Aval (SAV) is the largest WWTP of Europe, which treats the effluents of over 6.5 million equivalent inhabitants (Rocher et al., 2011). During rain events, CSOs discharge large amount of suspended solids, organic matters and nutrients into the Seine River (Even et al., 2007b). Assimilating DO is therefore of great interest for the decision makers in this area.

2.2. Data assimilation framework using particle filtering, the ProSe-PA approach

2.2.1. State-space model

To begin with, we represent the ProSe model as a state-space model (Kalman, 1960). A state-space model uses inputs, state variables, outputs to describe the evolution of a system over time. In our case, it uses three equations (Eq. (1), (2), (3)). Let $\mathbf{y}$ represents the state variable vector of the system in terms of physics (DO concentrations in this study) and $\mathbf{x}$ be the vector of model parameters. The state variable and the model parameters are both modelled as random variables $\mathbf{Y}$, $\mathbf{X}$ characterized by their probability distribution functions (pdf). $\mathbf{x}$ and $\mathbf{y}$ denote the realizations

of random variables $\mathbf{X}$, $\mathbf{Y}$. The observation vector $\mathbf{y}^*$ is a realization of the random variable $\mathbf{Y}^*$. The state-space model explains the temporal evolution of the system by assuming that it behaves as a Markov process (Markov, 1906). This means that the values at t depend on the values at $t-1$ only:

$$\mathbf{x}_t = \mathbf{x}_{t-1} + \boldsymbol{\eta}_t \tag{1}$$

$$\mathbf{y}_t = M(\mathbf{y}_{t-1}, \mu_t, \mathbf{x}_t) + \boldsymbol{\nu}_t \tag{2}$$

$$\mathbf{y}_t^* = h(\mathbf{y}_t) + \boldsymbol{\epsilon}_t \tag{3}$$

The transition of model parameter $\mathbf{x}$ from $t-1$ to t is described by a Gaussian perturbation ($\boldsymbol{\eta}_t$, Eq. (1)). In the above equations, M is the forward model (in our case ProSe) and h is the observation operator relating the forecasted state variable ($\mathbf{y}_t$) to the observation ($\mathbf{y}_t^*$). In other words, h denotes the selection of model cells where observations are available. $\mathbf{y}_{t-1}$ stands for the posterior state variable at previous time step ($t-1$) or equivalently for the prior state variable at current time step (t). The symbols $\boldsymbol{\mu}_t$ and $\mathbf{x}_t$ represent respectively the prior forcing data (e.g. temperature, wind speed and solar radiation) and the prior model parameters at time step t. The variables $\boldsymbol{\nu}_t$ and $\boldsymbol{\epsilon}_t$ characterize the unknown model and measurement errors respectively. The ProSe model aims at conserving the mass balance. The model does it with accuracy, the relative errors being in the order 10^{-5} or 10^{-6}. This is far below observation errors. Therefore, no model errors are considered ($\boldsymbol{\nu}_t = 0$) in our case.

Note that the nomenclature we use here is not canonical. In dynamical system literature, the parameters would have been called the "states" because they are the hidden dynamic variables that govern the system's behaviour. In forward modelling and especially hydrology, the state variables designate the variable that are computed by the model, while the parameters represents the time varying functional inputs of the numerical model. We consider the latter as the reader is more likely to be accustomed to this definition. Please note that, with this definition, the input fluxes at the boundary of the model are not called parameters, but boundary conditions.

2.2.2. Bayesian inference

We seek to estimate the posterior pdf of both the random variable $\mathbf{Y}$ and the parameter $\mathbf{X}$ conditionally to the observations $\mathbf{Y}^*$. To simplify the writing, we define the state vector $\mathbf{Z} = [\mathbf{Y}^{\mathrm{T}}, \mathbf{X}^{\mathrm{T}}]^{\mathrm{T}}$ and its realizations $\mathbf{z} = [\mathbf{y}^{\mathrm{T}}, \mathbf{x}^{\mathrm{T}}]^{\mathrm{T}}$. Therefore, our goal can be reformulated in finding the conditional pdf of $\mathbf{Z}$ knowing $\mathbf{y}^*$ noted $f_{\mathbf{Z}}(\mathbf{z}|\mathbf{y}^*)$. Through Bayes' theorem (Bayes, 1763), the posterior pdf $f_{\mathbf{Z}}(\mathbf{z}|\mathbf{y}^*)$ can be deduced from the product of the prior pdf ($f_{\mathbf{Z}}(\mathbf{z})$) and the likelihood distribution ($f_{\mathbf{Y}^*}(\mathbf{y}^*|\mathbf{z})$), up to a normalizing constant independent from $\mathbf{Z}$:

$$f_{\mathbf{Z}}(\mathbf{z}|\mathbf{y}^*) \propto f_{\mathbf{Y}^*}(\mathbf{y}^*|\mathbf{z}) f_{\mathbf{Z}}(\mathbf{z}) \tag{4}$$

The prior pdf $f_{\mathbf{Z}}(\mathbf{z})$ gives the prior knowledge on $\mathbf{z}$ before the measurements are taken. The likelihood $f_{\mathbf{Y}^*}(\mathbf{y}^*|\mathbf{z})$ describes the pdf of the observation data given the model predictions.

2.2.3. *Sequential form*

The formulation (4) does not relate to time. To introduce a sequential form for data assimilation, we define the temporal trajectories of the random variables $\mathbf{Z}$ and $\mathbf{Y}^*$, $\mathbf{Z}_{1:t} \equiv \{\mathbf{Z}_1, \cdots, \mathbf{Z}_t\}$ and $\mathbf{Y}^*_{1:t} \equiv \{\mathbf{Y}^*_1, \cdots, \mathbf{Y}^*_t\}$, discretized over time by sequences of state $(\mathbf{Z}_i, \mathbf{Y}^*_i, i = 1, \cdots, t)$, as well as their realizations $\mathbf{z}_{1:t}$, $\mathbf{y}^*_{1:t}$. The posterior pdf of the trajectory $\mathbf{Z}_{1:t}$ given observation $\mathbf{y}^*_{1:t}$ writes $f_{\mathbf{Z}_{1:t}}(\mathbf{z}_{1:t}|\mathbf{y}^*_{1:t})$.

Next, we can write the posterior pdf $f_{\mathbf{Z}_{1:t}}(\mathbf{z}_{1:t}|\mathbf{y}^*_{1:t})$ under a recursive form (Arulampalam et al., 2002; Doucet et al., 2001; Wikle and Berliner, 2007; Särkkä, 2013) owing to the Markovian nature of the process. We simplify the notations rewriting a pdf $f_{\mathbf{Z}}(\mathbf{z})$ into $f(\mathbf{z})$.

$$\begin{aligned} f(\mathbf{z}_{1:t}|\mathbf{y}^*_{1:t}) &\propto f(\mathbf{y}^*_t|\mathbf{z}_{1:t}, \mathbf{y}^*_{1:t-1}) f(\mathbf{z}_{1:t}|y^*_{1:t-1}) && (5)\\ &= f(\mathbf{y}^*_t|\mathbf{z}_t) f(\mathbf{z}_t|\mathbf{z}_{1:t-1}, y^*_{1:t-1}) f(\mathbf{z}_{1:t-1}|y^*_{1:t-1}) && (6)\\ &= f(\mathbf{y}^*_t|\mathbf{z}_t) f(\mathbf{z}_t|\mathbf{z}_{t-1}) f(\mathbf{z}_{1:t-1}|y^*_{1:t-1}), && (7) \end{aligned}$$

where we use first the Bayes' theorem to get equation (5), then the Bayes' theorem and the Markov property for equation (6) and the Markov property to obtain the sequential relation (7). We propagate $\mathbf{z}_t$ using the evolution equations (1) and (2) in order to model the distribution $f(\mathbf{z}_t|\mathbf{z}_{t-1})$. Note in case of a deterministic formulation ($(\boldsymbol{\nu}_t, \boldsymbol{\eta}_t) = (0, 0)$ in equations (1) and (2)), $f(\mathbf{z}_t|\mathbf{z}_{t-1})$ is a Dirac measure. Consequently, we can access the posterior pdf $f(\mathbf{z}_{1:t}|\mathbf{y}^*_{1:t})$ of the trajectory $\mathbf{Z}_{1:t}$ by updating it at each time step. This will be done through the use of sampling techniques.

2.2.4. *The particle filter*

The particle filter aims at approximating the posterior pdf $f(\mathbf{z}_{1:t}|\mathbf{y}^*_{1:t})$ of the trajectory $\mathbf{Z}_{1:t}$ knowing $\mathbf{y}^*_{1:t}$ by a set of particles (simulations) associated with weights (ω_t). It is generally impossible to sample from $f(\mathbf{z}_{1:t}|y^*_{1:t})$ directly. To address this problem, sequential importance sampling has been suggested (Doucet et al., 2000; Liu, 2001). For the ease of reading, the principle of the importance sampling and the definition of the importance weights are detailed

in Appendix B. We give here the weight update formula for each particle directly, which is a consequence of the sequential decomposition of the posterior of the trajectory obtained in (Eq. (7)) .

$$\omega_t^i = f(\mathbf{y}_t^*|\mathbf{z}_t^i)\omega_{t-1}^i \tag{8}$$

$$\hat{\omega}_t^i = \frac{\omega_t^i}{\sum \omega_t^i} \tag{9}$$

where $f(\mathbf{y}_t^*|\mathbf{z}_t^i)$ denotes the likelihood function, which quantifies how $\mathbf{y}_t^*$ is likely to be observed given $\mathbf{z}_t^i$ at time t. ω_t^i and ω_{t-1}^i stand for the posterior and prior weights at time t. In other words, the posterior weight at time $t-1$ serves as a prior weight at time t. We also define the normalized importance weight, $\hat{\omega}_t^i$, associated with the particle i.

In practice, it is usually not necessary to estimate the posterior pdf $f(\mathbf{z}_{1:t}|\mathbf{y}_{1:t}^*)$ of the trajectory $\mathbf{Z}_{1:t}$ knowing $\mathbf{y}_{1:t}^*$. We are interested here in the marginal distribution of $f(\mathbf{z}_{1:t}|\mathbf{y}_{1:t}^*)$, called filtering distribution $f(\mathbf{z}_t|\mathbf{y}_{1:t}^*)$. It represents the distribution of the state variables knowing all the past observations as well as the current one. In this paper, we note the filtering posterior pdf at time t as $f(\mathbf{z}_t|\mathbf{y}_{1:t}^*)$ which can be approximated as,

$$f(\mathbf{z}_t|y_{1:t}^*) \approx \sum_{i=1}^{N} \hat{\omega}_t^i \delta\mathbf{z}_t^i \tag{10}$$

where $\delta(\cdot)$ is a Dirac delta function and N denotes the number of particles. That is the filtering distribution is approximated by a discrete distribution, whose probability mass function is defined by the normalized importance weights and charges the particles.

If we assume that the observation error is Gaussian, then the likelihood probability of each particle can be calculated using the pdf of the multivariate normal distribution:

$$\ln L(\mathbf{y}_t^*|\mathbf{z}_t^i) = -\frac{m}{2}\ln(2\pi) - \frac{1}{2}\ln(|\Sigma|) - \frac{1}{2}(\mathbf{y}_t^* - h(\mathbf{y}_t^i))^{\mathrm{T}}\Sigma^{-1}(\mathbf{y}_t^* - h(\mathbf{y}_t^i)) \tag{11}$$

$$f(\mathbf{y}_t^*|\mathbf{z}_t^i) = \frac{L(\mathbf{y}_t^*|\mathbf{z}_t^i)}{\sum_{i=1}^{N} L(\mathbf{y}_t^*|\mathbf{z}_t^i)}$$

where m is the number of observation sites and Σ is the error covariance matrix of the observations.

2.2.5. *Resampling algorithm*

A common problem when using particle filtering is the degeneracy phenomena. This occurs when almost all the particles have zero or very small importance weights. Only a few particles have significant importance weights.

Therefore, the discrete probability densities cannot represent the posterior pdf of a given state adequately. A resampling procedure reduces the degeneracy effect. The basic idea of the resampling is to discard particles that have a small weight and to duplicate particles with a large weight. Generally, it is not necessary to perform resampling at every time step, but only when necessary. A way to monitor the need of resampling is to estimate the "effective" number of particles by the effective sample size (N_{eff}). The N_{eff} cannot be evaluated exactly but it can be estimated it by (Kong et al., 1994; Doucet et al., 2000):

$$\widehat{N_{eff}} = \frac{1}{\sum_{i=1}^{N} (\hat{\omega}_t^i)^2} \tag{12}$$

When N_{eff} is below a threshold ($N_{thres} = \alpha \cdot N$) predefined by the user, resampling is performed. After the resampling step, all weights are reset to $1/N$. As the particles having a important weight may be copied many times, which results in sample impoverishment problem. To maintain the diversity of the ensemble, a random perturbation is added to the parameters' value after the resampling step (eq. (13)).

$$\mathbf{x}_{t+1}^i = \mathbf{x}_{t,resampling}^i + \eta_t^i \quad \eta_t^i \sim N(0, (s \cdot \mathbf{\Phi})^2) \tag{13}$$

where s is a percent perturbation predefined by the user (0.1 in this study) and $\mathbf{\Phi}$ denotes the parameter range.

The resampling technique used in this work is referred to as systematic resampling. The procedure of systematic resampling is thoroughly described in the literature (Kitagawa, 1996; Moradkhani et al., 2005; Li et al., 2015).

2.2.6. Numerical algorithm

The coupling of the particle filtering algorithm with ProSe model is called ProSe-PA. PA stands for Parallel computing and data Assimilation. A full description of the ProSe-PA approach is given below (Fig. 2).

2.3. Description of the synthetic case study

2.3.1. Parameters considered and virtual "observation" data

Twelve parameters of the ProSe model were identified in different hydrological and trophic contexts in a previous work (Wang et al., 2018). Those parameters are both physical and physiological (Table 1).

Five monitoring stations are located in the study area (Suresnes, Chatou, Bougival, Sartrouville, and Andresy, from upstream to downstream respectively. Fig. 1). The assimilation period is the year 2011 when algal blooms occurred in March and July in the Seine River (Vilmin et al., 2016). The two successive algal blooms are characterized by

different physiological properties, especially the optimal temperature for growth of phytoplankton (Vilmin, 2014). A high-frequency DO (dissolved oxygen) dataset ($\mathbf{y}^{ref}$, every 15 minutes), which corresponds to the frequency of the real measurement, is generated using predefined parameters (Table 1). Three phytoplanktonic parameters are modified manually at day 139 in order to represent the two different algal blooms ($P_{max,pp}$, $R_{m,pp}$ and $T_{opt,pp}$, see Tab. 1 for parameter definition). In this case study, the "observation" data are obtained adding a Gaussian error to the reference dataset with a mean of zero and a standard deviation of $0.01 \times \mathbf{y}^{ref}$ ($\mathbf{y}_t^* = \mathbf{y}_t^{ref} + \boldsymbol{\epsilon}_t, \quad \boldsymbol{\epsilon}_t \sim N(0, (0.01 \times \mathbf{y}_t^{ref})^2)$). A standard deviation of $0.01 \times \mathbf{y}^{ref}$ ensures that the 95% observation errors are less than $0.02 \times \mathbf{y}^{ref}$, which is coherent with the sensors used in the Seine River system (Garnier et al., 2019). The reference dataset and predefined parameters are assumed to be true and are assessed via data assimilation. In this study, no model errors are considered ($\boldsymbol{\nu}_t = 0$, see Eq. (2)). The major aims of this work are to retrieve the reference DO concentrations and the dominant predefined parameters identified by Wang et al. (2018) over time.

2.3.2. *Input data for* ProSe *model*

Apart from hydraulic (time varying river discharge and CSOs data), geometric data (river channels), the ProSe model also requires time varying concentrations of the biogeochemical variables (micro-organism biomass, dissolved oxygen, nutrients, organic matters, suspended solids) entering the system and the meteorological data (solar irradiance, water temperature and wind speed). These data are used as boundary conditions of the ProSe model. The quantification of uncertainties on boundary conditions is an ambitious topic that extent far beyond the scope of this paper and will require the study of real systems. Here we only investigate uncertainties on parameters.

2.3.3. *Qualitative description of the simulation period: state classification*

According to the former sensitivity analysis of the biogeochemical module, C-RIVE , of the ProSe model (Wang et al., 2018), we can classify the year 2011 into different periods (Fig. 3). The polygons at the bottom show this classification (Fig. 3). The water temperature increases from black to white (color gradient) and the line shaded polygons represent algal bloom periods ($C_{\text{chl}a} > 10\,\mu$g/L). Wang et al. (2018) concluded that at low temperatures (Interbloom T $< 6\,^{\circ}$C, black polygons), the river system is controlled by the reaeration process (K_{navig}). At moderate and high water temperatures, the maximum growth rate of bacteria ($\mu_{max,hb}$) is most sensitive to variation of DO concentrations when no algal bloom occurs (Interbloom T $> 6\,^{\circ}$C, gray polygons). Once algal bloom occurs, phytoplanktonic parameters related to respiration ($R_{m,pp}$) and to photosynthesis (α_{pp}, $P_{max,pp}$ etc.) are dominant. A supplementary sensitivity analysis by Sobol method (Sobol, 1993) during algal bloom shows that the optimal temperature ($T_{opt,pp}$) plays an important role on the growth of phytoplankton and that the sensitivity of $R_{m,pp}$ (respiration of maintenance) depends on the water temperature (Fig. A.1). When the water temperature exceeds 20 °C, the total sensitivity index of $R_{m,pp}$

decreases dramatically (Fig. A.1).

2.4. Numerical setting and computational cost

According to a test of the number of particles (not shown here), we select 500 particles in this work which is sufficient to i) match the observed oxygen concentrations and ii) identify the posterior pdfs of sensible parameters. The OpenMP Application Programming Interface (API) is implemented in the code to simulate the 500 particles in parallel. For a 1 year-simulation period (365 days) at a 15-min time step, the computation takes 1.13 days with 20 processors (Intel(R) Xeon(R) CPU E5-2640 v4 @ 2.40 GHz). A resampling threshold of $\alpha = 0.3$ is chosen, corresponding to a minimum effective sample size of 150.

3. Results: Oxygen simulation and parameter identification

3.1. Effective sample size for resampling procedure

Although the effective sample size (N_{eff}) reduce fast after two months, synchronously with the start of the first algal bloom (day 64), the parameter perturbation allows the restoration of particle diversity (Fig. 4, $N_{eff} > 450$ after resampling), which indicates a proper exploration of the parameter space by the algorithm. This is achieved by selecting a configuration of the Gaussian random walk (Pearson, 1905) through its standard deviation. After several trials (not shown here), the standard deviation of the random walk has been set to 0.1 times the parameters' range (eq. (13)).

3.2. Almost perfect DO simulation with ProSe-PA

Figure 5 shows simulated DO concentrations at three downstream monitoring stations. Those stations are selected because they are less sensitive to input boundary conditions and therefore permit to evaluate the conceptual structure of the code and the conceptualization of processes. The model performances are evaluated by four statistical criteria (RMSE, Root-Mean-Square-Error; MAE, Maximum Absolute Error; NSE, Nash-Sutcliffe Efficiency; R, Correlation coefficient).

The ensemble weighted average DO concentrations at all stations match the reference DO concentrations. Two algal blooms are well retrieved. The first starts at 64th day and the second around 171st day (Fig. 5). The maximum RMSE and MAE values between ensemble weighted mean and reference DO data are obtained at the Suresnes station with an error of 0.035 mgO_2/L and an error of 0.321 mgO_2/L respectively (Tab. 2). Furthermore, all NSEs and correlation coefficients (R) are close to 1, which signifies a perfect match of simulated DO concentration to the reference DO data. In addition, the 95 percentile confidence intervals are very narrow, which means a perfect simulation also.

3.3. Parameter identification in the frame of the state classification

In this section, the identification of physical, bacterial and phytoplanktonic parameters is shown in plots displaying normalised weights over time (Fig. 6,7,8) for each parameter of interest. The daily normalised weights over time permits to represent the evolution of the posterior pdf of each parameter. Each parameter range is divided in 20 intervals. The sum of daily normalised weights in each interval which approximates the posterior pdf are shown by a image plot. The dashed black line represents the predefined value of parameter used to generate "observation" data.

3.3.1. Physical parameters

Two physical parameters are assimilated in this paper, the light extinction coefficient for pure water (η_{water}) and the reaeration coefficient related to the navigation (K_{navig}). Albeit light extinction by clear water (η_{water}) is a physical parameter, it is an important control for phytoplankton growth (Wang et al., 2018). It is not very well determined most of the time, except during algal blooms when the posterior pdfs focus around the reference value of 0.32 m^{-1} (see the line shaded polygons in Fig. 6A). K_{navig} controls the reaeration process and is the most sensitive parameter to variation of DO in winter when the water temperature is below 6 °C (Wang et al., 2018). In this period (see the second black polygon in Fig.6B), the particle filter achieves a very fine identification of its pdf. Seeing other periods of the year, its effect remains negligible and logically it remains unidentified.

3.3.2. Bacterial parameters

In the previous sensitivity analysis (Wang et al., 2018), the maximum growth rate of bacteria ($\mu_{max,hb}$) is the first ranked parameter out of algal bloom with a moderate water temperature (T > 6 °C) and the bacterial growth yield (Y_{hb}) is identified as the second sensitive parameter for inter algal bloom periods (both low and moderate water temperatures). $\mu_{max,hb}$ can be retrieved well for several periods. These periods match globally the gray polygons which correspond to inter algal bloom periods with a water temperature T > 6 °C (Fig. 7). Y_{hb} is slightly overestimated during the simulation, except for the algal blooms periods (see the line shaded polygons in Fig. 7). The optimal temperature for bacterial growth ($T_{opt,hb}$) can be determined at the beginning of simulation. However, the maximal mortality rate of bacteria ($mort_{hb}$) spreads uniformly over the parameter range during the data assimilation period.

3.3.3. Phytoplanktonic parameters

It can be observed that the respiration of maintenance ($R_{m,pp}$), the photosynthetic capacity (α_{pp}), the light extinction coefficient by algal sell-shading ($\eta_{chla,pp}$), and the optimal temperature for growth of phytoplankton ($T_{opt,pp}$) are well estimated during the first algal bloom (see the gray line shaded polygon in Fig 8), while the maximum photosynthesis rate ($P_{max,pp}$) is overestimated (Fig. 8). Similarly, almost all the phytoplanktonic parameters can be retrieved

during the second algal bloom (see the clear line shaded polygons in Fig. 8), but not for $R_{m,pp}$ and Chl$a$/C$_{pp}$. The ratio of chlorophyll a to carbon (Chl$a$/C$_{pp}$) is unidentified during the simulation, because Chl$a$/C$_{pp}$ has little effect on the variation of DO concentrations (Wang et al., 2018). The particle filter captures the change of phytoplanktonic properties after 139 day for $P_{max,pp}$ and $T_{opt,pp}$. During algal blooms, five among seven parameters including η_{water} can be retrieved. Even though the two other parameters are not well estimated, we obtain satisfying simulated DO concentrations during algal blooms (Fig. 5). Although α_{pp}, η_{water}, $\eta_{chla,pp}$ and $T_{opt,pp}$, are always identified by the algorithm during blooms, this is not the case for $R_{m,pp}$ and $P_{max,pp}$ which are supposed to be important during algal bloom (Wang et al., 2018). This is further discussed in the next section.

4. Discussion

4.1. Specifications for real time parameter identification

As shown by the results, the identification of the different parameters is not always ensured during the data assimilation period. This result can be explained by the parameters' sensitivity in contrasted hydrological and trophic contexts. One parameter that is dominant on DO concentration can be well estimated, while the other insensitive parameters have a relatively large uncertainty. In this section, we discussed the real time parameter identification.

4.1.1. Parameters identified between algal blooms

Out of algal blooms, the identified parameters are $\mu_{max,hb}$ and K_{navig}, which is coherent with the previous sensitivity analysis (Wang et al., 2018). At low temperatures (T $<$ 6 °C), bacterial and phytoplanktonic activities are limited, the reaeration (K_{navig}) controls DO concentration within water column. The river system is governed by physical processes. The daily posterior pdfs of K_{navig} at low temperature period (day 21-38) are shown (Fig. 9A). Although the pdfs of K_{navig} from day 21 to day 28 have more dispersion than those for day 29-38, their modes are centred on the reference value (Fig. 9A). However, the water temperature is below 6 °C at the beginning of the simulation (days 0-6), K_{navig} is not very well estimated (see the first black polygon in Fig. 6B). This can be explained by the fact that the prior pdf of K_{navig} is uniform at the start of the simulation and this period is too short.

When the water temperature increases (T $>$ 6 °C), the heterotrophic bacteria develops and degrades organic matters. $\mu_{max,hb}$ becomes the most important parameter governing the DO concentration in river system (Wang et al., 2018). The identification of $\mu_{max,hb}$ in those periods can be expected (see the gray polygons in Fig. 7). The pdfs of $\mu_{max,hb}$ are displayed for day 45-62 when the water temperature increases (Fig. 9B). It can be clearly seen that the modes of the pdfs of $\mu_{max,hb}$ move towards the reference value (Fig. 9B). The particle filter captures the change from physical control (K_{navig}) to bacterial control ($\mu_{max,hb}$).

Next, the mortality of phytoplankton at the end of algal blooms supplies organic matters for bacterial growth. $\mu_{max,hb}$ is thus well estimated after algal blooms (see the gray polygons day 91-140 and day 280-300 in Fig. 7). This indicates that the algorithm adapts the evolution of the regime of river system from autotroph to heterotroph. However, $\mu_{max,hb}$ is not well estimated for day 300-325, even if this period corresponds to inter algal blooms. The correlation between the water temperature and the DO concentration in this period (Fig . 3) signifies that the DO concentration is controlled by water temperature and then colsed to saturation. The reaeration and micro-organism activities are not important in this period. Therefore, no parameter can be identified in this period.

4.1.2. Parameters identified during algal blooms

The maximum photosynthesis rate ($P_{max,pp}$) is third ranked parameter during algal blooms in the sensitivity analysis (Wang et al., 2018). However, $P_{max,pp}$ is overestimated with a stable posterior pdf during the first algal bloom (see line shaded polygon day 64-90 in Fig. 8). The particles with $P_{max,pp}$ close to 0.2 h^{-1} (predefined value) are under weighted. That means the compensation of $P_{max,pp}$ with the other parameters exist. The parameter interaction of $P_{max,pp}$ has been shown by the difference between the total sensitivity index and the first order sensitivity index (Wang et al., 2018). Nonetheless, the DO concentrations have been well estimated.

The former sensitivity analysis which used almost constant water temperatures didn't highlight the sensitivity of $T_{opt,pp}$. The supplementary result shows that the sensitivity of $T_{opt,pp}$ depends on the water temperature (Fig. A.1). Therefore, the water temperature is a crucial factor for algal blooms when $T_{opt,pp}$ is defined. This is the reason why $T_{opt,pp}$ can be identified before algal blooms, but not after algal blooms (Fig. 8).

Contrarily to the first algal bloom, the identification of $R_{m,pp}$ fails for the second algal bloom (Fig. 8). The supplementary sensitivity analysis shows that when the water temperature exceeds 20 °C, the total sensitivity index of $R_{m,pp}$ reduces dramatically (Fig. A.1). The water temperatures are over 21 °C during the second bloom (Fig. 3). Therefore, $R_{m,pp}$ is insensitive during the second algal bloom and remains unidentified. The other identified parameters match the sensitivity analysis ($\alpha_{pp}, \eta_{water}, \eta_{chla,pp}$).

In conclusion, the identification of the parameters depends on not only the hydrological and trophic contexts but also on the thermal state of the river system. The detailed identification of the parameters is resumed in figure 10.

4.2. Possibility to identify different phytoplanktonic communities

In this paper, we define two different phytoplanktonic properties to distinguish blooms in March and in July ($R_{m,pp}$, $P_{max,pp}$ and $T_{opt,pp}$). The particle filter detects well the change of phytoplanktonic properties (Fig. 8), especially the optimal temperature for the growth of phytoplankton ($T_{opt,pp}$). The posterior pdfs of $T_{opt,pp}$ are stable during the two algal blooms (Fig. 11). It is therefore possible to differentiate phytoplanktonic communities in real time by

comparing the posterior distributions of $T_{opt,pp}$. The different optimal temperatures that have been used for modeling the phytoplanktonic communities during spring and summer are the ones determined in the Loire and Seine rivers (Descy et al., 2012; Garnier et al., 1995; Vilmin, 2014).

4.3. Circadian rhythm

At night, the photosynthetic parameters are absolutely insensitive and thus should remain unidentified. Therefore, a circadian rhythm (day and night) can be expected for the photosynthetic parameters. To study the circadian rhythm of photosynthetic parameters, the posterior pdfs of $P_{max,pp}$, α_{pp} and $T_{opt,pp}$ are displayed every 3 hours during days 184-186 (Fig. 12). It can be clearly noticed that the posterior pdfs of the three parameters have more dispersion at night (00:00-6:00) than those during the day (9:00-18:00). Their modes don't match the reference values at night, while the three parameters are well estimated during the day (Fig. 12). The ProSe-PA approach retrieves well the circadian rhythm of the photosynthetic parameters' sensitivity.

4.4. Performances' sensitivity to the observation error

The proof of concept of the particle filter has been achieved assuming a relative error on observations of 1 %. In this section, we assess the impact of the relative error on both ProSe-PA ability to simulate oxygen concentrations and to identify parameter posterior pdfs. To this aim, various relative observation errors are tested from 1% to 10%. The Kling-Gupta Efficiencies (KGE) (Gupta et al., 2009; Kling et al., 2012) are calculated to evaluate the model performances. KGE is based on the decomposition of the mean squared error and NSE performance criteria (Gupta et al., 2009). KGEs range from -Inf to 1. Essentially, the closer to 1, the more accurate the model is.

The results show that the model performance decreases slightly with the increase of the observation relative error. Nevertheless, ProSe-PA retrieves in average the reference DO concentrations for all observation relative errors tested. KGEs are always larger than 0.96 (not shown here) whatever the relative error. The identifiability of the bacterial parameters ($\mu_{max,hb}$, Y_{hb} and $T_{opt,pp}$) and the reaeration coefficient related to navigation (K_{navig}) is not ensured when the observation relative error exceeds 5% while ProSe-PA is able to capture phytoplanktonic properties for all tested relative errors (1%-10%), except for the ratio of chlorophyll *a* to carbon (Chla/C_{pp}) which has little influence on the variation of DO concentrations (Wang et al., 2018).

5. Conclusions

In this work, we present a first implementation of particle filter into a hydro-biogeochemical model for metabolism's parameter estimation. The assimilation of a 15-min "observation" DO data is realized in the Seine River system on a synthetic case study.

- It can be concluded that the particle filter is an efficient method for the biogeochemical data assimilation and for the metabolism's parameter estimation in urban river systems.
- The ProSe-PA approach is capable to retrieve perfectly the reference DO concentrations at all stations.
- The identification of the metabolism parameters depends on the hydrological an trophic conetxts and more surprisingly on the thermal state of the river system.
- The data assimilation method adapts to the trophic state's change of the Seine River system and the circadian cycle of photosynthetic parameters is well captured by the ProSe-PA approach.
- It is possible to distinguish phytoplanktonic species by identifying the different physiological properties via data assimilation.
- The ProSe-PA model is operational and transferable to waste water manager for assessing the impact of their practices now. However, it is necessary to quantify input uncertainties (forcing data, reject water of WWTPs and CSOs etc.), which are not considered in this paper.

Software availability

A virtual machine including the ProSe-PA0.30 executable and a tutorial based on the synthetic dataset will be accessible online. If one wants to access the source code of the ProSe-PA model, please contact Dr. Nicolas Flipo. The software license is currently under discussion.

Declarations of interest

The authors report that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome. The authors alone are responsible for the content and writing of this article.

Acknowledgements

This work is a contribution to the PIREN-Seine research program, part of the french Long Term Socio-Ecological Research (LTSER) site "Zone Atelier Seine". We thank the editor and the three anonymous reviewers for their fruitful remarks which helped improve the overall quality of the paper.

401 **List of figures**

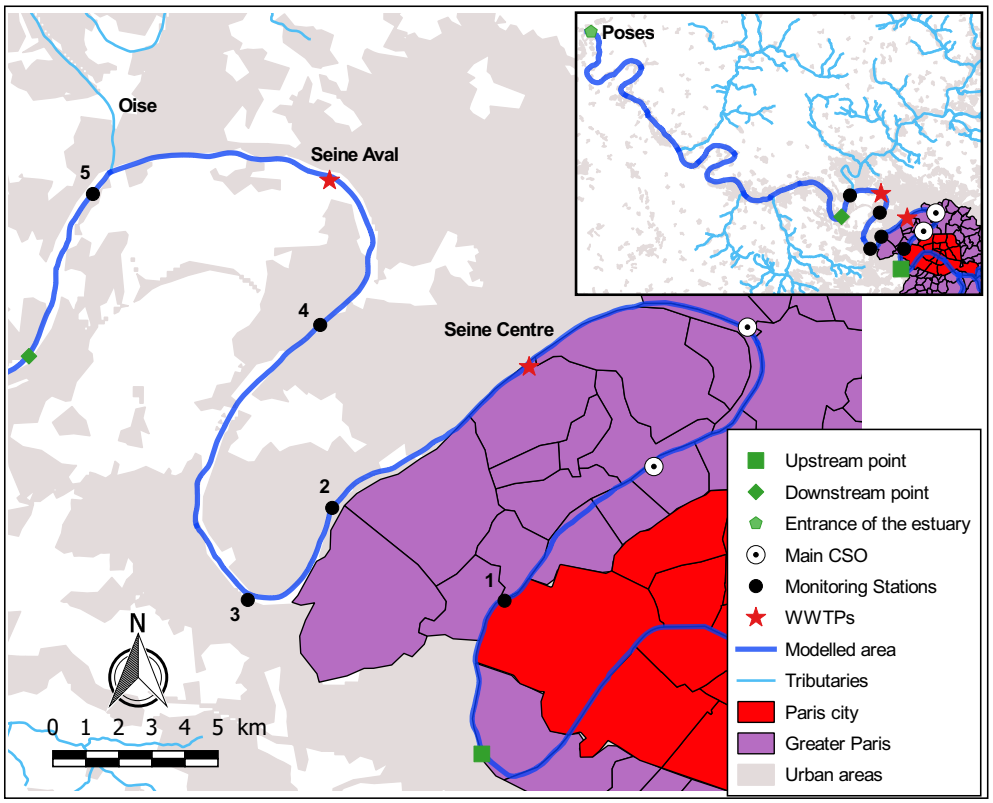


Fig. 1: The description of study area and monitoring sites. Full Seine model is shown at upper-right corner; Monitoring stations from upstream to downstream: 1. Suresnes, 2. Chatou, 3. Bougival, 4. Sartrouville, 5. Andresy.

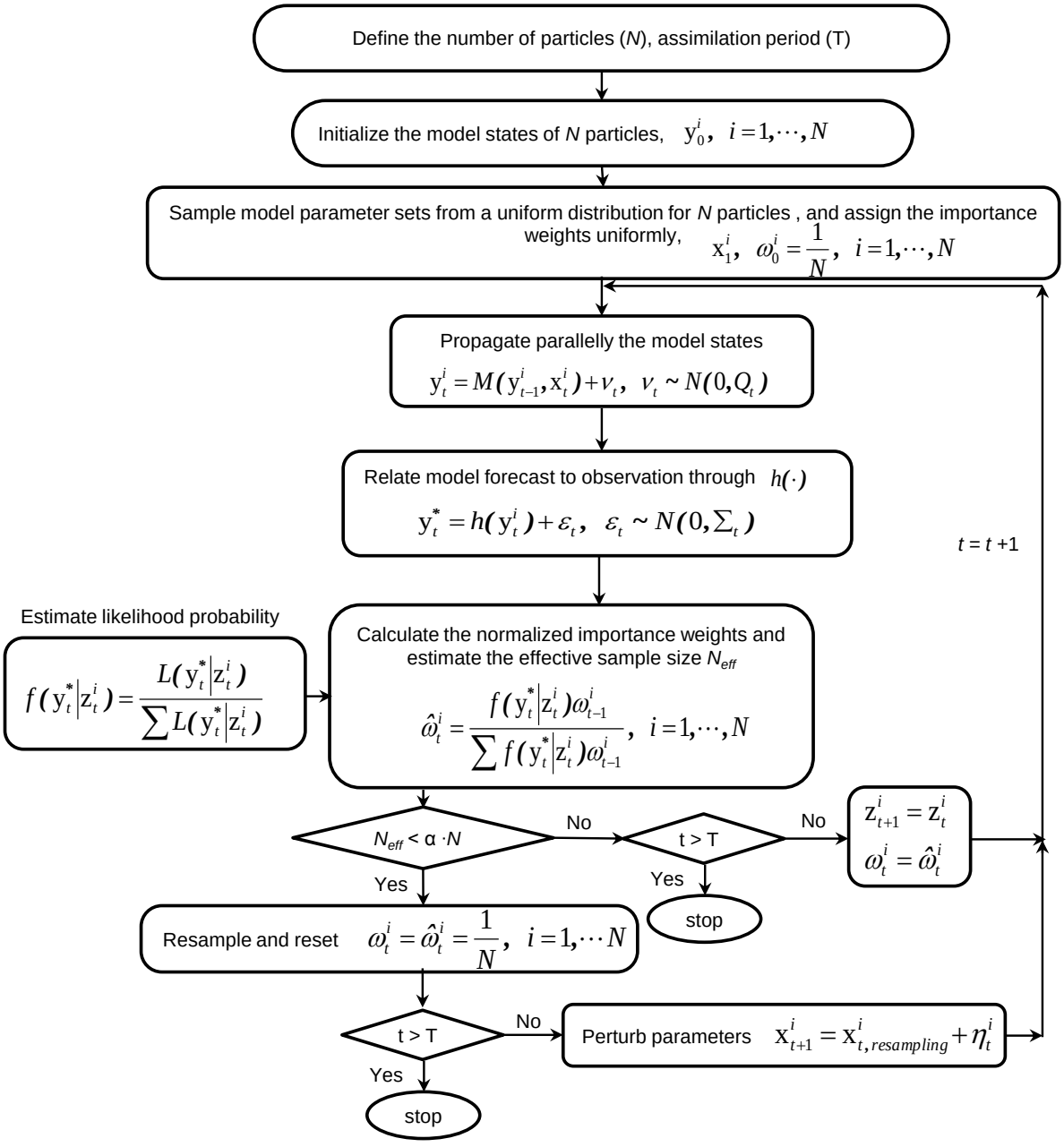


Fig. 2: The flowchart of ProSe-PA approach. The state $\mathbf{z}^i$ denotes ith particle or a realization of random variable $\mathbf{Z}=[\mathbf{Y}^{\mathrm{T}},\mathbf{X}^{\mathrm{T}}]^{\mathrm{T}}$. $\mathbf{y}^i$ and $\mathbf{x}^i$ signify model state and parameter set of particle i. Q_t and Σ_t correspond to the model error covariance matrix and the observation error covariance matrix respectively. In this work, no model errors are considered, $\boldsymbol{\nu}_t = 0$.

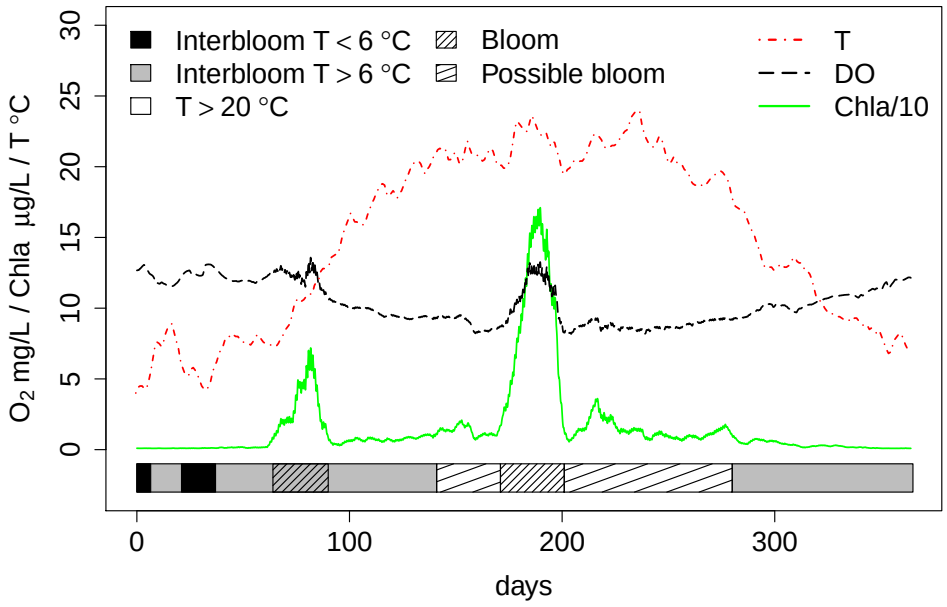


Fig. 3: State classification of synthetic case study over time, concluded by Wang et al. (2018). The red line (dash-dotted) represents water temperature; The black line (dashed) denotes reference DO concentrations at the Bougival station; The green line (solid) shows the evolution of chl*a* concentrations. The polygons at the bottom correspond to the different periods. The water temperature increases from black to white.

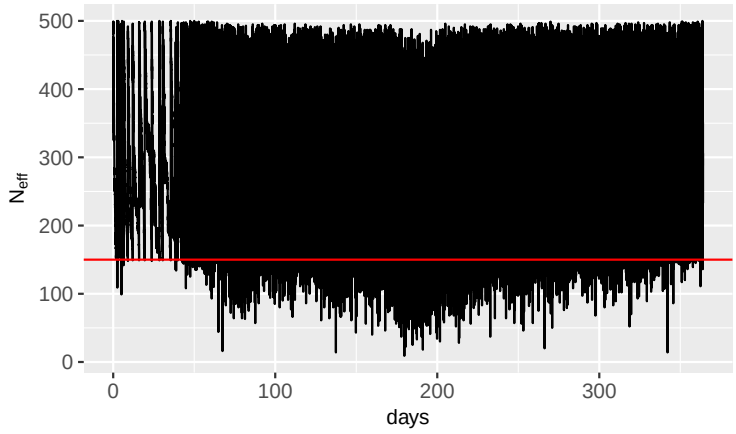


Fig. 4: The evolution of effective sample size during assimilation. Red line corresponds to the minimum effective sample size 150. Frequent resamplings ensure a proper searching of the parameter space.

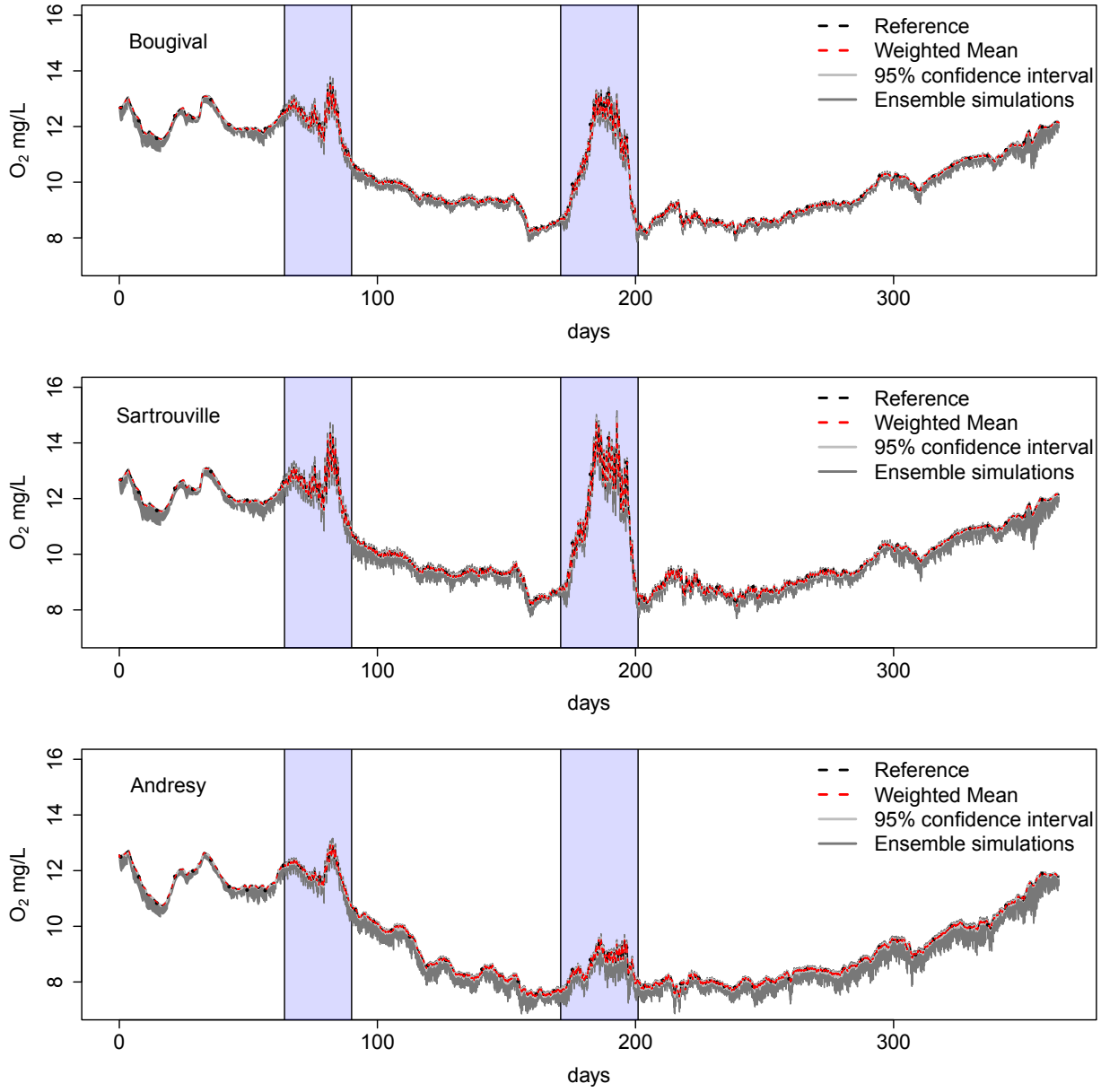


Fig. 5: Simulated DO concentrations at Bougival, Sartrouville and Andresy stations. The dashed red lines denote ensemble weighted means. The reference DO data are represented by dashed black lines. To clarify the illustration, weekly reference DO concentrations are shown by black points. The gray areas correspond to the ensemble simulations (dark gray areas) and the 95 percentile confidence intervals (light gray areas). The two algal bloom periods are shown by blue polygons.

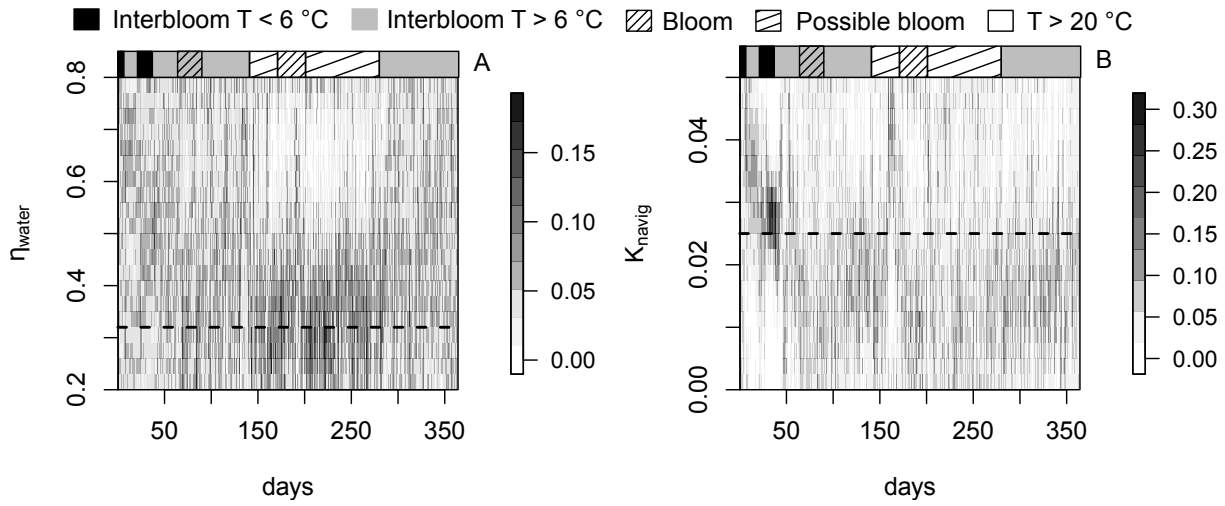


Fig. 6: Daily normalised weights of physical parameters: light extinction coefficient for pure water (η_{water}) and reaeration coefficient due to navigation (K_{navig}). The dashed line represents the predefined reference value. Line shaded polygons correspond to algal bloom periods and water temperature increases from black to white (color gradient). See Fig. 3 for detailed state classification.

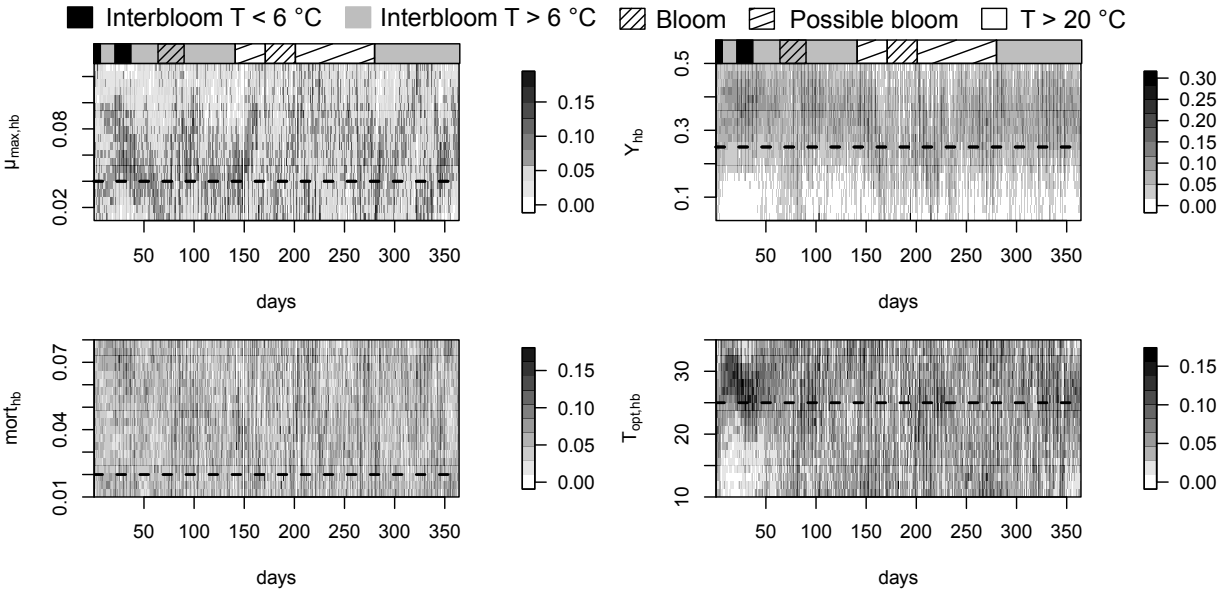


Fig. 7: The normalised importance weights of bacterial parameters. See Tab. 1 for parameter definition. The dashed line represents the predefined reference value. Line shaded polygons correspond to algal bloom periods and water temperature increases from black to white (color gradient). See Fig. 3 for detailed classification.

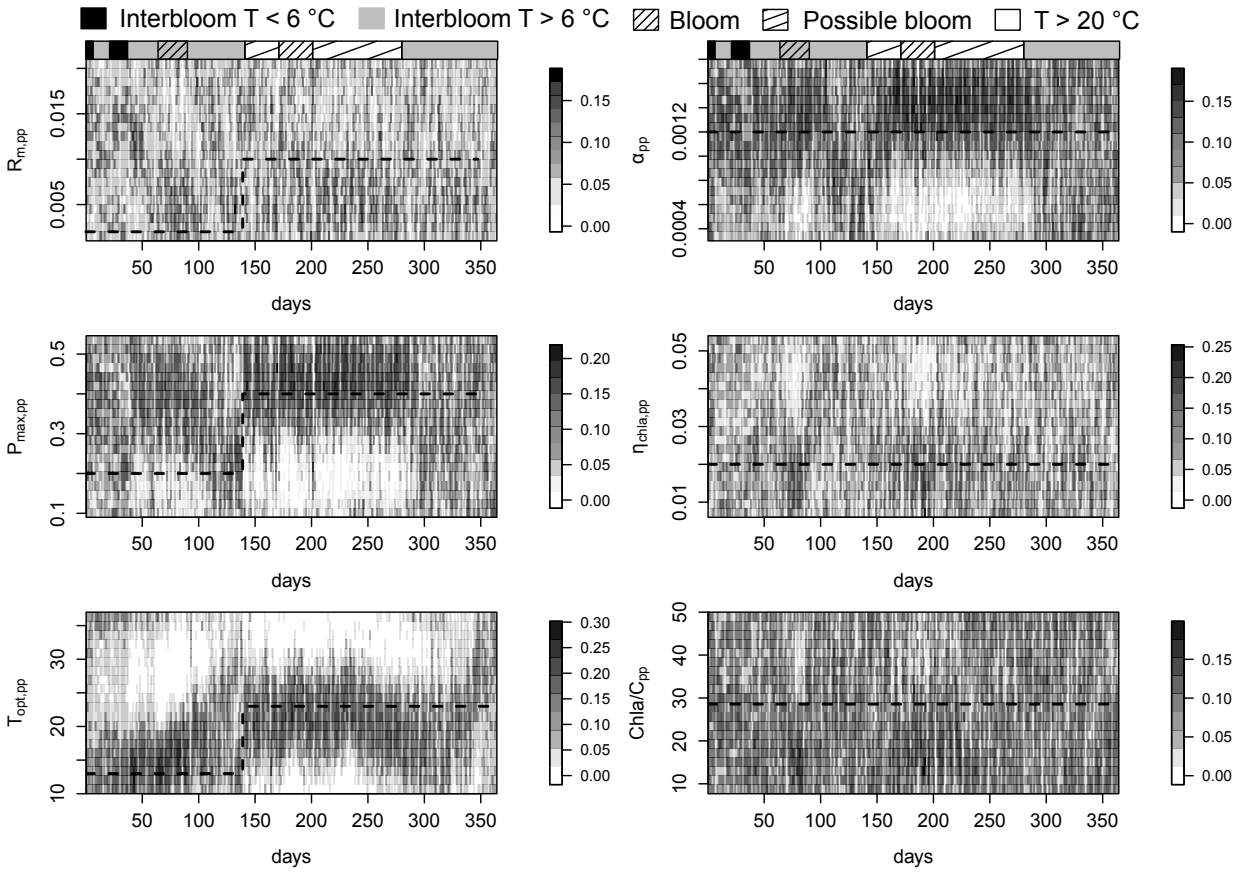


Fig. 8: The normalised importance weights of phytoplanktonic parameters. See Tab. 1 for parameter definition. The dashed line represents the predefined reference value. Line shaded polygons correspond to algal bloom periods and water temperature increases from black to white (color gradient). See Fig. 3 for detailed classification.

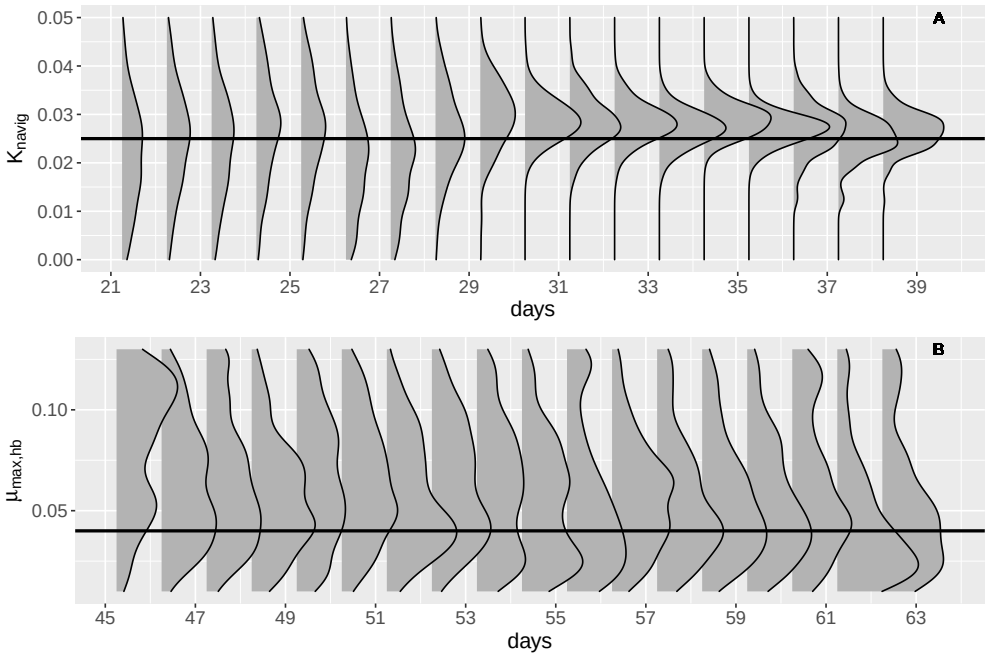


Fig. 9: The posterior pdf of K_{navig} during low temperature period (A) and the posterior pdf of $\mu_{max,hb}$ with a moderate temperature (B); Black lines represent the predefined values.

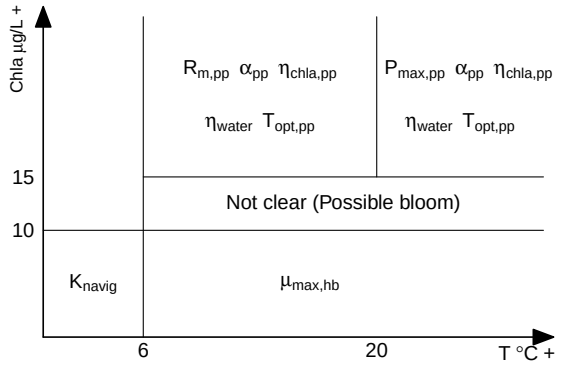


Fig. 10: The identification of parameters in different trophic (C_{chla}) and thermal ($T^{\circ}C$) contexts of river system. See Tab. 1 for parameter definition.

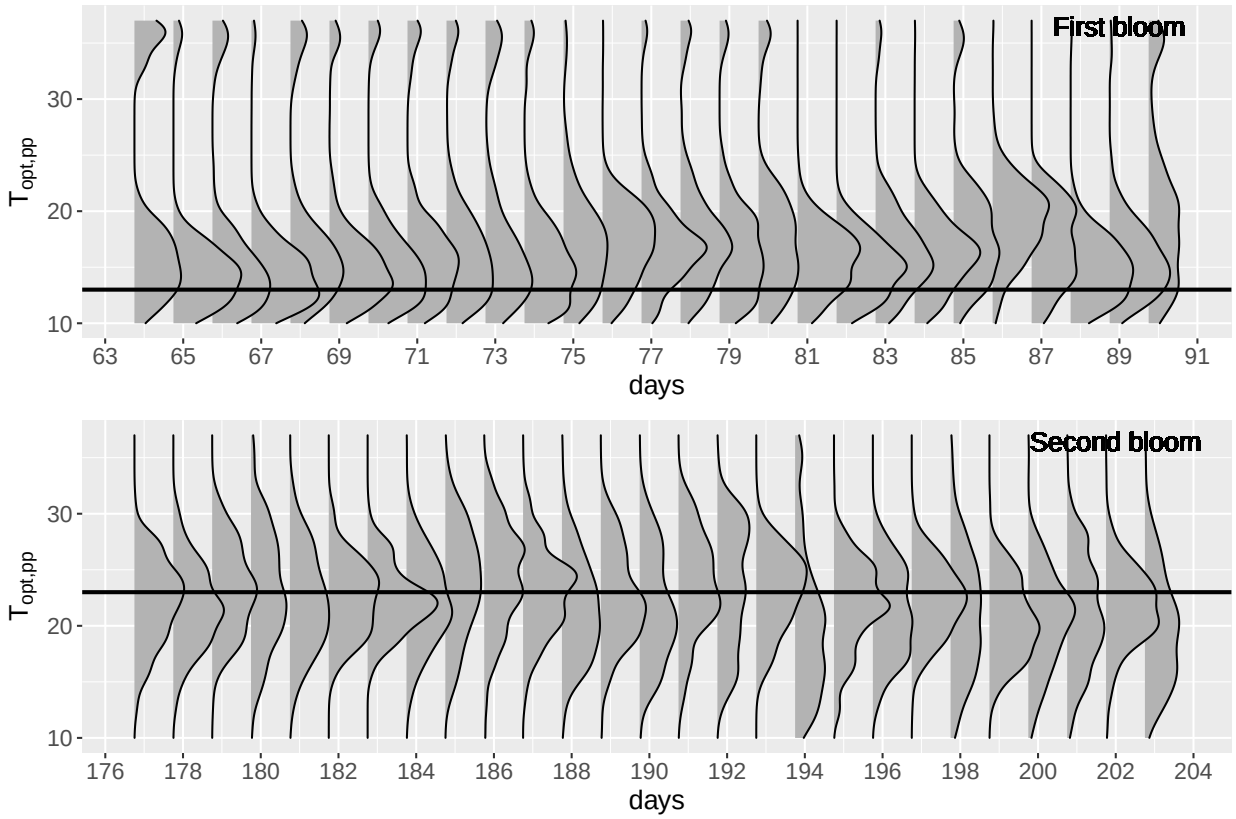


Fig. 11: The posterior distributions of $T_{opt,pp}$ during the first and the second algal bloom at 6 pm. Black line represents the predefined value.

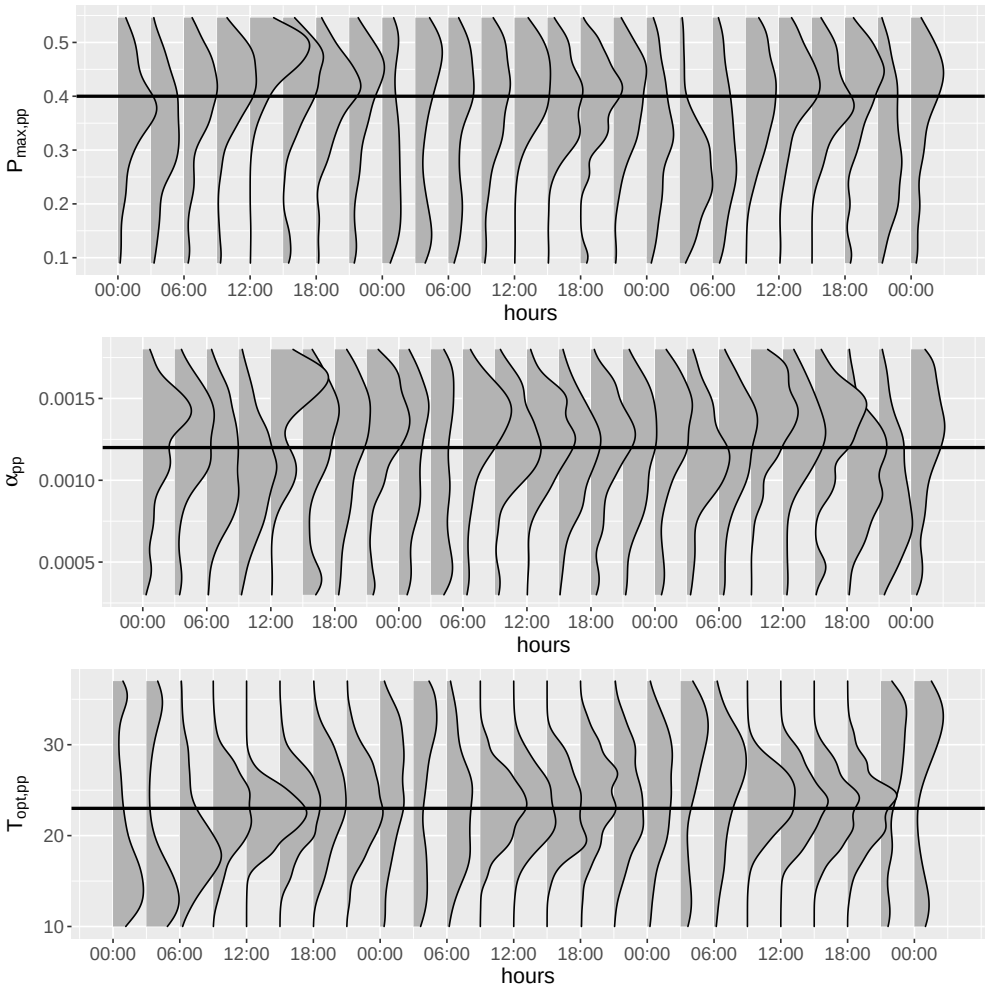


Fig. 12: The posterior distributions of $P_{max,pp}$, α_{pp} and $T_{opt,pp}$ during 184-186 days

List of tables

Table 1: Reference parameters considered in data assimilation (Parameter range from Wang et al. (2018))

Parameters	Description	Range	*Reference March*	*Reference July*	Unit
Physical parameters					
η_{water}	Light extinction coefficient for pure water	[0.2, 0.8]	0.32	0.32	[m^{-1}]
K_{navig}	Reaeration coefficient related to the navigation	[0, 0.05]	0.025	0.025	[$m.h^{-1}$]
Bacterial parameters					
$\mu_{max,hb}$	Maximum growth rate of bacteria	[0.01, 0.13]	0.04	0.04	[h^{-1}]
$mort_{hb}$	Maximum mortality rate of bacteria	[0.01, 0.08]	0.02	0.02	[h^{-1}]
$T_{opt,hb}$	Optimal temperature for bacterial growth	[10, 35]	25	25	[°C]
Y_{hb}	Bacterial growth yield	[0.03, 0.5]	0.25	0.25	[-]
Phytoplanktonic parameters					
α_{pp}	Photosynthetic capacity	[0.0003, 0.0018]	0.0012	0.0012	[$m^2.s.\mu E^{-1}.h^{-1}$]
$\eta_{chla,pp}$	Light extinction coefficient by algal self-shading	[0.006, 0.054]	0.02	0.02	[$L.\mu gchla^{-1}.m^{-1}$]
$Chla/C_{pp}$	Ratio of chlorophyll *a* to carbon	[50, 7.69]	28.57	28.57	[$\mu gchla.mgC^{-1}$]
$P_{max,pp}$	Maximum photosynthesis rate	[0.09, 0.546]	0.2	0.4	[h^{-1}]
$R_{m,pp}$	Respiration of maintenance	[0.001, 0.021]	0.002	0.01	[h^{-1}]
$T_{opt,pp}$	Optimal temperature for growth of phytoplankton	[10, 37]	13	23	[°C]

Table 2: Statistical criteria of the weighted average DO concentrations of 500 particles/simulations at five monitoring sites

Stations	RMSE	NSE	MAE	r
Suresnes	0.0345	0.9995	0.3209	0.9998
Chatou	0.0173	0.9999	0.1347	0.9999
Bougival	0.0192	0.9998	0.1627	0.9999
Sartrouville	0.0201	0.9998	0.1765	0.9999
Andresy	0.0163	0.9998	0.1350	0.9999

RMSE: Root-Mean-Square-Error
NSE: Nash-Sutcliffe Efficiency
MAE: Maximum Absolute Error
r: Correlation coefficient

Table 3: Thresholds of the relative observation error for the parameter identifiability (1%-10%)

Parameters	Thresholds	Parameters	Thresholds	Parameters	Thresholds
$R_{m,pp}$	10% (T < 20 °C)	η_{water}	10%	Y_{hb}	< 7.5%
$P_{max,pp}$	10%	Chla/C_{pp}	Not stable	$mort_{hb}$	Not stable
α_{pp}	10%	$T_{opt,pp}$	10%	$T_{opt,hb}$	< 5%
$\eta_{chla,pp}$	10%	$\mu_{max,hb}$	< 5%	K_{navig}	< 5%

10%: The parameters are always identified for the relative observation errors tested.
See the table 1 for the parameters' definition and units.

Appendix A. Sobol sensitivity analysis of C-RIVE module during algal bloom at different temperatures

The sensitivity analysis of C-RIVE model during algal blooms is realized every two degrees from 2 °C to 26 °C. The results show that the optimal temperature for the growth of phytoplankton ($T_{opt,pp}$) is the most influential parameter at low temperatures (< 6 °C). When the water temperature exceeds 20 °C, the importance of the respiration of maintenance ($R_{m,pp}$) reduces drastically. For more detailed information about the sensitivity analysis of C-RIVE model, one can consult Wang et al. (2018).

Appendix B. Importance sampling and definition of the importance weights in particle filtering

We suppose here that we want to simulate a random variable $\mathbf{U}$ which may have a complex distribution. The importance sampling method uses an instrumental distribution called the importance distribution $\pi(\cdot)$ from which we know how to draw samples (Särkkä, 2013). The importance sampling relies on the following formula:

$$\int f_{\mathbf{U}}(\mathbf{u})d\mathbf{u} = \int \Big[\frac{f_{\mathbf{U}}(\mathbf{u})}{\pi_{\mathbf{U}}(\mathbf{u})}\Big]\pi_{\mathbf{U}}(\mathbf{u})d\mathbf{u} \tag{B.1}$$

where $\pi_{\mathbf{U}}(\mathbf{u})$ is a proposal pdf for the random variable $\mathbf{U}$ from which we can draw samples and $f_{\mathbf{U}}(\mathbf{u})$ is the posterior

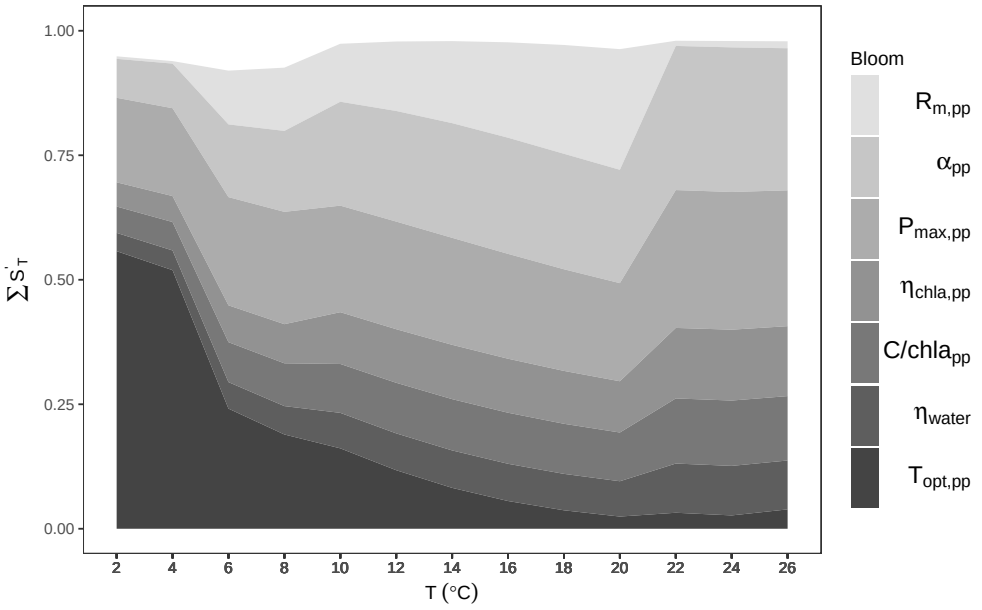


Fig. A.1: The evolution of the total sensitivity indices (normalised to 100%) with temperature for Bloom condition (From bottom to top: $T_{opt,pp}$, η_{water}, $C/Chla_{pp}$, $P_{max,pp}$, α_{pp}, $R_{m,pp}$). See Table 1 for parameter definition. For more detailed information about sensitivity analysis, one can consult Wang et al. (2018).

pdf of given state $\mathbf{U}$ from which we cannot directly or efficiently draw samples. The equation (B.1) transforms the expectation of $\mathbf{U}$ into the expectation of the term $\left[\mathbf{u}\frac{f_{\mathbf{U}}(\mathbf{u})}{\pi_{\mathbf{U}}(\mathbf{u})}\right]$. We can perform Monte Carlo approximation to estimate the expectation of $g(\mathbf{U})$, denoted $\mathbf{E}(g(\mathbf{U}))$, for any function g:

$$\begin{aligned} \mathbf{E}(g(\mathbf{U})) &= \int \left[g(\mathbf{u})\frac{f_{\mathbf{U}}(\mathbf{u})}{\pi_{\mathbf{U}}(\mathbf{u})}\right]\pi_{\mathbf{U}}(\mathbf{u})d\mathbf{u} \\ &\approx \frac{1}{N}\sum_{i=1}^{N} g(\mathbf{u}^i)\frac{f_{\mathbf{U}}(\mathbf{u}^i)}{\pi_{\mathbf{U}}(\mathbf{u}^i)} \\ \omega^i &= \frac{1}{N}\frac{f_{\mathbf{U}}(\mathbf{u}^i)}{\pi_{\mathbf{U}}(\mathbf{u}^i)} \end{aligned} \tag{B.2}$$

where $\mathbf{u}^i$ ($i = 1, \cdots, N$) is ith sample (particle) drawn from the importance distribution $\pi_{\mathbf{U}}(\mathbf{u})$ and ω^i denotes the weight associated with ith sample (particle). With particular choices of g, we can hence retrieve the expectation of $\mathbf{U}$, its variance, quantiles, etc.

We can use importance sampling to address the problem of sampling directly from the posterior pdf $f(\mathbf{z}_{1:t}|\mathbf{y}^*_{1:t})$ of the trajectory $\mathbf{Z}_{1:t}$. Let the importance distribution, $\pi(\mathbf{z}_{1:t}|\mathbf{y}^*_{1:t})$, from equation (7) we can compute importance weights,

$$\omega_t = \frac{f(\mathbf{z}_{1:t}|\mathbf{y}^*_{1:t})}{\pi(\mathbf{z}_{1:t}|\mathbf{y}^*_{1:t})} \propto \frac{f(\mathbf{y}^*_t|\mathbf{z}_t)f(\mathbf{z}_t|\mathbf{z}_{t-1})f(\mathbf{z}_{1:t-1}|y^*_{1:t-1})}{\pi(\mathbf{z}_{1:t}|\mathbf{y}^*_{1:t})} \tag{B.3}$$

Then, assuming the importance distribution $\pi(\mathbf{z}_{1:t}|\mathbf{y}^*_{1:t})$ to be Markovian, that is

$$\begin{aligned} \pi(\mathbf{z}_{1:t}|\mathbf{y}^*_{1:t}) &= \pi(\mathbf{z}_t|\mathbf{z}_{1:t-1}, \mathbf{y}^*_{1:t})\pi(\mathbf{z}_{1:t-1}|\mathbf{y}^*_{1:t-1}) \\ &= \pi(\mathbf{z}_t|\mathbf{z}_{t-1}, \mathbf{y}^*_t)\pi(\mathbf{z}_{1:t-1}|\mathbf{y}^*_{1:t-1}), \end{aligned} \tag{B.4}$$

then equation (B.3) becomes,

$$\begin{aligned} \omega_t &\propto \frac{f(\mathbf{y}^*_t|\mathbf{z}_t)f(\mathbf{z}_t|\mathbf{z}_{t-1})f(\mathbf{z}_{1:t-1}|y^*_{1:t-1})}{\pi(\mathbf{z}_t|\mathbf{z}_{t-1}, \mathbf{y}^*_t)\pi(\mathbf{z}_{1:t-1}|\mathbf{y}^*_{1:t-1})} \\ &\propto \frac{f(\mathbf{y}^*_t|\mathbf{z}_t)f(\mathbf{z}_t|\mathbf{z}_{t-1})}{\pi(\mathbf{z}_t|\mathbf{z}_{t-1}, \mathbf{y}^*_t)}\omega_{t-1} \end{aligned} \tag{B.5}$$

In other words, we derive an update formula for the importance weights in the same form as the update formula for the posterior pdf ((7)). If we can draw N samples (particles) from the importance distribution $\pi(\mathbf{z}_t|\mathbf{y}^*_{1:t})$,

$$\mathbf{z}^i_t \sim \pi(\mathbf{z}_t|\mathbf{y}^*_{1:t}) \quad i = 1, \cdots, N \tag{B.6}$$

a weight update formula for each particle can be written as follows,

$$\begin{aligned} \omega^i_t &\propto \frac{f(\mathbf{y}^*_t|\mathbf{z}^i_t)f(\mathbf{z}^i_t|\mathbf{z}^i_{t-1})}{\pi(\mathbf{z}^i_t|\mathbf{z}^i_{t-1}, \mathbf{y}^*_t)}\omega^i_{t-1} \\ \hat{\omega}^i_t &= \frac{\omega^i_t}{\sum \omega^i_t} \end{aligned} \tag{B.7}$$

where we define the normalized importance weight ($\hat{\omega}^i_t$) associated with the particle i. Normalizing the weights allows to evacuate the problem of the normalizing constant in the Bayesian representation of the posterior ((5)). Equation (B.7) shows that we only need to store the state $\mathbf{z}_t$ at time t. In addition, it is usually not necessary to estimate the full pdf $f(\mathbf{z}_{1:t}|y^*_{1:t})$ in practice. In this work, we are interested in the marginal distribution $f(\mathbf{z}_t|y^*_{1:t})$ at time

t, called filtering distribution. The filtering posterior pdf $f(\mathbf{z}_t|y^*_{1:t})$ at time t can be approximated as,

$$f(\mathbf{z}_t|y^*_{1:t}) \approx \sum_{i=1}^{N} \hat{\omega}_t^i \delta(\mathbf{z}_t - \mathbf{z}_t^i) \tag{B.8}$$

where $\delta(\cdot)$ is a Dirac delta function.

The performance of the above algorithm depends on the quality of the importance distribution $\pi(\cdot)$. Typically, the optimal importance distribution is (Doucet et al., 2001; Särkkä, 2013),

$$\pi(\mathbf{z}_t|\mathbf{z}_{t-1}, \mathbf{y}^*_t) = f(\mathbf{z}_t|\mathbf{z}_{t-1}, \mathbf{y}^*_t) \tag{B.9}$$

In practice, however we generally do not know the distribution $f(\mathbf{z}_t|\mathbf{z}_{t-1}, \mathbf{y}^*_t)$ and we rather propagate $\mathbf{z}_t$ using the evolution equations (1) and (2), which terms to draw samples from $f(\mathbf{z}_t|\mathbf{z}_{t-1})$. In that case, the equations (B.5) and (B.7) simplifies into

$$\boldsymbol{\omega}_t \propto f(\mathbf{y}^*_t|\mathbf{z}_t)\boldsymbol{\omega}_{t-1} \tag{B.10}$$

References

Abbaszadeh, P., Moradkhani, H., Yan, H., 2018. Enhancing hydrologic data assimilation by evolutionary Particle Filter and Markov Chain Monte Carlo. Advances in Water Resources 111, 192–204.

Andreadis, K., Clark, E., Lettermaier, D., Alsdorf, D., 2007. Prospects for river discharge and depth estimation through assimilation of swath-altimetry into a raster-based hydrodynamics model. Geophys. Res. Lett. 34, L10403.

Arulampalam, M. S., Maskell, S., Gordon, N., Clapp, T., 2002. A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. IEEE Transactions on Signal Processing 50 (2), 174–188.

Bayes, T., 1763. An essay towards solving a problem in the doctrine of chances. Phil. Trans. of the Royal Soc. of London 53, 370–418.

Beck, M., 1987. Water quality modelling : a review of the analysis of uncertainty. Water Resourses Research 23 (8), 1393–1442.

Beven, K., 1989. Changing ideas in hydrology. The case of physically-based model. Journal of Hydrology 105, 157–172.

Billen, G., Garnier, J., Hanset, P., 1994. Modelling phytoplankton development in whole drainage networks: the RIVERSTRAHLER Model applied to the Seine river system. Hydrobiologia 289, 119–137.

Cappe, O., Godsill, S. J., Moulines, E., 2007. An Overview of Existing Methods and Recent Advances in Sequential Monte Carlo. Proceedings of the IEEE 95 (5), 899–924.

Courtier, P., Andersson, E., Heckley, W., Vasiljevic, D., Hamrud, M., Hollingsworth, A., Rabier, F., Fisher, M., Pailleux, J., 1998. The ECMWF

implementation of three-dimensional variational assimilation (3D-Var). I: Formulation. Quarterly Journal of the Royal Meteorological Society 124 (550), 1783–1807.

Courtier, P., Thépaut, J. N., Hollingsworth, A., 1994. A strategy for operational implementation of 4D-Var, using an incremental approach. Quarterly Journal of the Royal Meteorological Society 120 (519), 1367–1387.

DeChant, C. M., Moradkhani, H., 2012. Examining the effectiveness and robustness of sequential data assimilation methods for quantification of uncertainty in hydrologic forecasting. Water Resources Research 48 (4).

Descy, J.-P., Leitao, M., Everbecq, E., Smitz, J.-S., Deliège, J.-F., 2012. Phytoplankton of the River Loire, France: a biodiversity and modelling study. J. Plankton. Res. 34 (2), 120–135.

Doucet, A., de Freitas, N., Gordon, N., 2001. Sequential Monte Carlo Methods in Practice. Springer.

Doucet, A., Godsill, S., Andrieu, C., 2000. On sequential monte carlo sampling methods for bayesian filtering. Statistics and Computing 10 (3), 197–208.

Even, S., Bacq, N., Ruelland, D., Billen, G., Garnier, J., Poulin, M., Théry, S., Blanc, S., 2007a. New tools for modelling water quality of hydrosystems: An application in the Seine River basin in the frame of the Water Framework Directive. Sciences of Total Environment 375 (1-3), 274–291.

Even, S., Mouchel, J. M., Servais, P., Flipo, N., Poulin, M., Blanc, S., Chabanel, M., Paffoni, C., 2007b. Modeling the impacts of Combined Sewer Overflows on the river Seine water quality. Sciences of Total Environment 375 (1-3), 140–151.

Even, S., Poulin, M., Garnier, J., Billen, G., Servais, P., Chesterikoff, A., Coste, M., 1998. River ecosystem modelling: Application of the ProSe model to the Seine river (France). Hydrobiologia 373, 27–37.

Even, S., Poulin, M., Mouchel, J. M., Seidl, M., Servais, P., 2004. Modelling oxygen deficits in the Seine river downstream of combined sewer overflows. Ecol. Model. 173, 177–196.

Evensen, G., 1994. Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. Journal of Geophysical Research: Oceans 99 (C5), 10143–10162.

Evensen, G., Nov 2003. The Ensemble Kalman Filter: theoretical formulation and practical implementation. Ocean Dynamics 53 (4), 343–367.

Flipo, N., Even, S., Poulin, M., Tusseau-Vuillemin, M. H., Améziane, T., Dauta, A., 2004. Biogeochemical modelling at the river scale: Plankton and periphyton dynamics - Grand Morin case study, France. Ecol. Model. 176, 333–347.

Flipo, N., Rabouille, C., Poulin, M., Even, S., Tusseau-Vuillemin, M., Lalande, M., 2007. Primary production in headwater streams of the Seine basin: the Grand Morin case study. Sciences of Total Environment 375, 98–109.

Garnier, J., Billen, G., Coste, M., 1995. Seasonal succession of diatoms and chlorophycae in the drainage network of the river Seine: Observations and modelling. Limnol. Oceanogr. 40 (4), 750–765.

Garnier, J., Marescaux, A., Guillon, S., Vilmin, L., Rocher, V., Billen, G., Thieu, V., Silvestre, M., Passy, P., Raimonet, M., Groleau, A., Théry, S., Tallec, G., Flipo, N., 2019. The Seine River Basin. Handbook of Environmental Chemistry. Springer, Ch. Ecological functioning of the Seine River: from long-term modelling approaches to high-frequency data analysis, p. In press.

Gauthier, P., Tanguay, M., Laroche, S., Pellerin, S., Morneau, J., 2007. Extension of 3DVAR to 4DVAR: Implementation of 4DVAR at the Meteorological Service of Canada. Monthly Weather Review 135 (6), 2339–2354.

Gharamti, M., Tjiputra, J., Bethke, I., Samuelsen, A., Skjelvan, I., Bentsen, M., Bertino, L., 2017. Ensemble data assimilation for ocean biogeochemical state and parameter estimation at different sites. Ocean Modelling 112, 65 – 89.

Gupta, H. V., Kling, H., Yilmaz, K. K., Martinez, G. F., 2009. Decomposition of the mean squared error and nse performance criteria: Implications for improving hydrological modelling. Journal of Hydrology 377 (1), 80 – 91.

Huang, J., Gao, J., Liu, J., Zhang, Y., 2013. State and parameter update of a hydrodynamic-phytoplankton model using ensemble Kalman filter.

Ecological Modelling 263, 81–91.

Kalman, R. E., 1960. A New Approach to Linear Filtering and Prediction Problems. Journal of Basic Engineering 82 (1), 35–45.

Kattwinkel, M., Reichert, P., 2017. Bayesian parameter inference for individual-based models using a Particle Markov Chain Monte Carlo method. Environmental Modelling & Software 87, 110–119.

Kim, K., Park, M., Min, J., Ryu, I., Kang, M., Park, L. J., 2014. Simulation of algal bloom dynamics in a river with the ensemble Kalman filter. Journal of Hydrology 519, 2810–2821.

Kitagawa, G., 1996. Monte Carlo Filter and Smoother for Non-Gaussian Nonlinear State Space Models. Journal of Computational and Graphical Statistics 5 (1), 1–25.

Kleist, D. T., Ide, K., 2015. An OSSE-Based Evaluation of Hybrid Variational-Ensemble Data Assimilation for the NCEP GFS. Part II: 4DEnVar and Hybrid Variants. Monthly Weather Review 143 (2), 452–470.

Kling, H., Fuchs, M., Paulin, M., 2012. Runoff conditions in the upper danube basin under an ensemble of climate change scenarios. Journal of Hydrology 424-425, 264 – 277.

Kong, A., Liu, J. S., Wong, W. H., 1994. Sequential imputations and Bayesian missing data problems. Journal of the American statistical association 89 (425), 278–288.

Li, T., Villarrubia, G., Sun, S., Corchado, J. M., Bajo, J., 2015. Resampling methods for particle filtering: identical distribution, a new method, and comparable study. Frontiers of Information Technology & Electronic Engineering 16 (11), 969–984.

Liu, J. S., 2001. Monte Carlo Strategies in Scientific Computing. Springer.

Mao, J., Lee, J. H., Choi, K., 2009. The extended Kalman filter for forecast of algal bloom dynamics. Water Research 43 (17), 4214 – 4224.

Markov, A. A., 1906. Extension of the law of large numbers to dependent quantities (in russian). Izvestiia Fiz.-Matem. Obsch. Kazan Univ.,(2nd Ser.) 15, 135–156.

Moradkhani, H., Hsu, K., Gupta, H., Sorooshian, S., 2005. Uncertainty assessment of hydrologic model states and parameters: Sequential data assimilation using the particle filter. Water Resources Research 41 (5).

Page, T., Smith, P. J., Beven, K. J., Jones, I. D., Elliott, J. A., Maberly, S. C., Mackay, E. B., Ville, M. D., Feuchtmayr, H., 2018. Adaptive forecasting of phytoplankton communities. Water Research 134, 74 – 85.

Pasetto, D., Camporese, M., Putti, M., 2012. Ensemble Kalman filter versus particle filter for a physically-based coupled surfacesubsurface model. Advances in Water Resources 47, 1–13.

Pastres, R., Ciavatta, S., Solidoro, C., 2003. The extended Kalman Filter (EKF) as a tool for the assimilation of high frequency water quality data. Ecological Modelling 170 (2), 227–235.

Pearson, K., 1905. The Problem of the Random Walk. Nature 72 (1865), 294.

Plaza, D. A., De Keyser, R., De Lannoy, G. J. M., Giustarini, L., Matgen, P., Pauwels, V. R. N., 2012. The importance of parameter resampling for soil moisture data assimilation into hydrologic models using the particle filter. Hydrology and Earth System Sciences 16 (2), 375–390.

Polus, E., Flipo, N., de Fouquet, C., Poulin, M., 2011. Geostatistics for assessing the efficiency of distributed physically-based water quality model. Application to nitrates in the Seine River. Hydrological Processes 25 (2), 217–233.

Raimonet, M., Vilmin, L., Flipo, N., Rocher, V., Laverman, A., 2015. Modelling the fate of nitrite in an urbanized river using experimentally obtained nitrifier growth parameters. Water Research 73, 373–387.

Rocher, V., Garcia-Gonzalez, E., Paffoni, C., Thomas, W., 2011. La production de nitrites lors de la dénitrification des eaux usées: un sujet sensible et complexe ! L'Eau, l'Industrie, les Nuisances 344, 80–83.

Salamon, P., Feyen, L., 2009. Assessing parameter, precipitation, and predictive uncertainty in a distributed hydrological model using sequential data assimilation with the particle filter. Journal of Hydrology 376, 428–442.

Särkkä, S., 2013. Bayesian Filtering and Smoothing. Cambridge University Press.

Sasaki, Y., 1955. A Fundamental Study of the Numerical Prediction Based on the Variational Principle. Journal of the Meteorological Society of Japan. Ser. II 33 (6), 262–275.

Sasaki, Y., 1958. An Objective Analysis Based on the Variational Method. Journal of the Meteorological Society of Japan. Ser. II 36 (3), 77–88.

Shi, Y., Davis, K. J., Zhang, F., Duffy, C. J., Yu, X., 2014. Parameter estimation of a physically based land surface hydrologic model using the ensemble Kalman filter: A synthetic experiment. Water Resources Research 50 (1), 706–724.

Simon, E., Bertino, L., 2012. Gaussian anamorphosis extension of the denkf for combined state parameter estimation: Application to a 1d ocean ecosystem model. Journal of Marine Systems 89 (1), 1 – 18.

Simon, E., Samuelsen, A., Bertino, L., Dumont, D., 2012. Estimation of positive sum-to-one constrained zooplankton grazing preferences with the denkf: a twin experiment. Ocean Science 8 (4), 587–602.
URL https://www.ocean-sci.net/8/587/2012/

Sobol, I., 1993. Sensitivity estimates for on linear mathematical models. Mathematical Modelling and Computational Experiments 1.

Vilmin, L., 2014. Modélisation du fonctionnement biogéochimique de la seine de l'agglomération parisienne à l'estuaire à différentes échelles temporelles. Ph.D. thesis, MINES ParisTech.

Vilmin, L., Aissa-Grouz, N., Garnier, J., Billen, G., Mouchel, J. M., Poulin, M., Flipo, N., 2015a. Impact of hydro-sedimentary processes on the dynamics of soluble reactive phosphorus in the Seine River. Biogeochemistry 122, 229–251.

Vilmin, L., Flipo, N., de Fouquet, C., Poulin, M., 2015b. Pluri-annual sediment budget in a navigated river system: The Seine River (France). Sciences of Total Environment 502, 48–59.

Vilmin, L., Flipo, N., Escoffier, N., Groleau, A., 2018. Estimation of the water quality of a large urbanized river as defined by the european WFD: what is the optimal sampling frequency? Environmental Science and Pollution Research 25 (24), 23485–23501.

Vilmin, L., Flipo, N., Escoffier, N., Rocher, V., Groleau, A., 2016. Carbon fate in a large temperate human-impacted river system: Focus on benthic dynamics. Global Biogeochem. Cycles 30 (7), 1086–1104.

Vrugt, J. A., ter Braak, C. J., Diks, C. G., Schoups, G., 2013. Hydrologic data assimilation using particle Markov chain Monte Carlo simulation: Theory, concepts and applications. Advances in Water Resources 51, 457–478.

Wang, S., Flipo, N., Romary, T., 2018. Time-dependent global sensitivity analysis of the C-RIVE biogeochemical model in contrasted hydrological and trophic contexts. Water Research 144, 341–355.

Weerts, A. H., El Serafy, G. Y. H., 2006. Particle filtering and ensemble Kalman filtering for state updating with hydrological conceptual rainfall-runoff models. Water Resources Research 42 (9).

Wikle, C. K., Berliner, L. M., 2007. A Bayesian tutorial for data assimilation. Physica D: Nonlinear Phenomena 230 (1), 1–16.

Xue, P., Chen, C., Beardsley, R. C., 2012. Observing system simulation experiments of dissolved oxygen monitoring in Massachusetts Bay. Journal of Geophysical Research: Oceans 117 (C5).

Yu, L., Fennel, K., Bertino, L., Gharamti, M. E., Thompson, K. R., 2018. Insights on multivariate updates of physical and biogeochemical ocean variables using an ensemble kalman filter and an idealized model of upwelling. Ocean Modelling 126, 13 – 28.

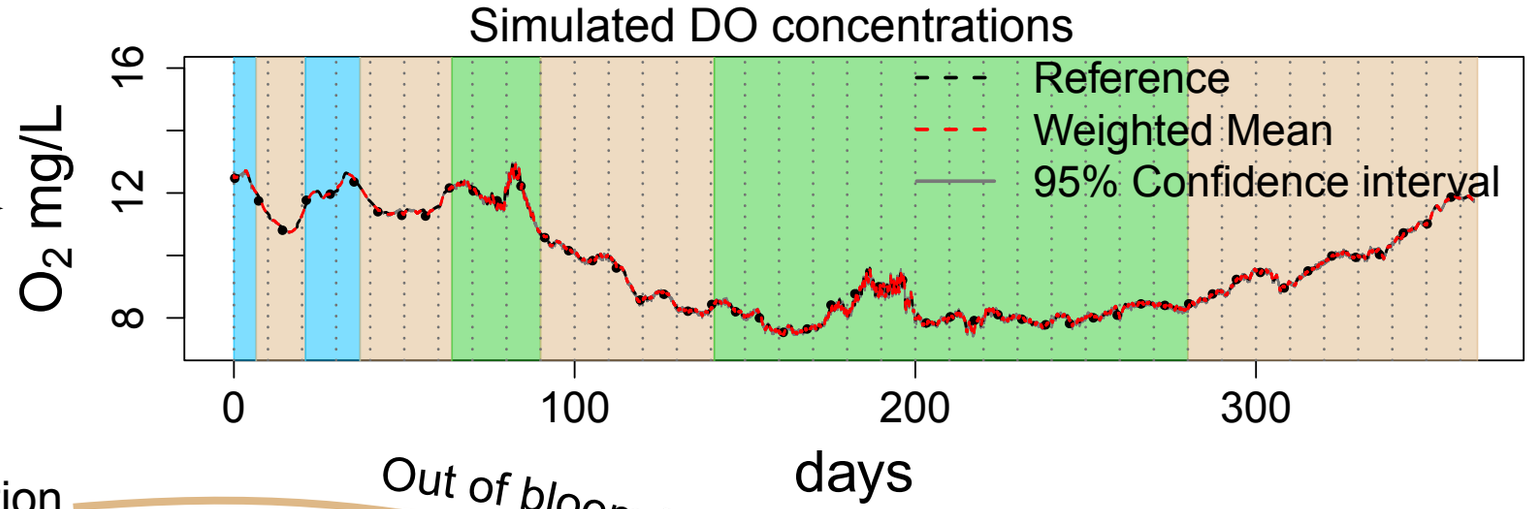
Yucel, I., Onen, A., Yilmaz, K., Gochis, D., 2015. Calibration and evaluation of a flood forecasting system: Utility of numerical weather prediction model, data assimilation and satellite-based rainfall. Journal of Hydrology 523, 49–66.

DO: Dissolved Oxygen

ProSe-PA model

Particle Filter

DO simulation



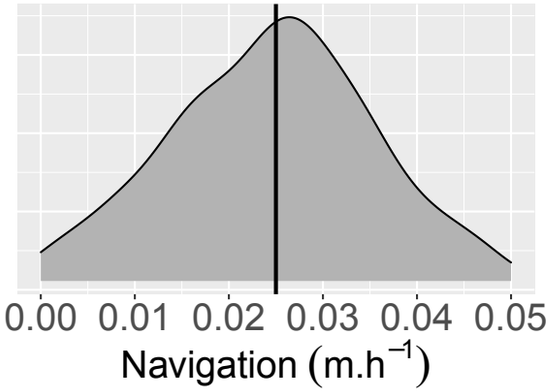
Parameter estimation

Out of bloom T > 8 °C

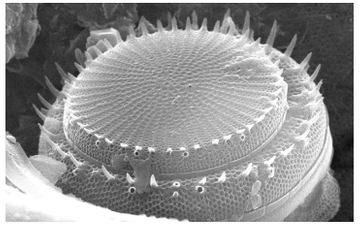
T < 6 °C

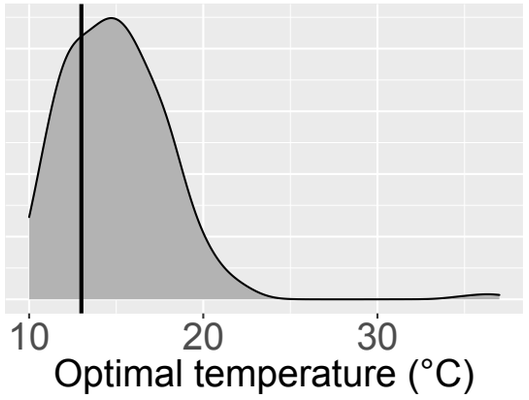
Potential algal bloom

Reaeration



Phytoplanktonic





Bacterial