Which filter for data assimilation in water quality models? Focus on oxygen reaeration and heterotrophic bacteria activity

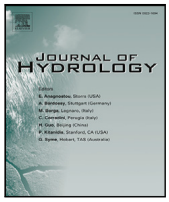
Shuaitao Wang, Nicolas Flipo, Thomas Romary

Journal of Hydrology 620 (2023) 129423

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol



Research papers

Which filter for data assimilation in water quality models? Focus on oxygen reaeration and heterotrophic bacteria activity

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

[a] Sorbonne Université, CNRS, EPHE, UMR Metis, 75005 Paris, France
[b] Mines Paris, PSL University, Center for Geosciences and Geoengineering, 77300 Fontainebleau, France

ARTICLE INFO

This manuscript was handled by Corrado Corradini, Editor-in-Chief, with the assistance of Jonghyun Harry Lee, Associate Editor.

Keywords:
Water quality modeling
Data assimilation
Parameter estimation
Particle filter
Ensemble Kalman filter
ProSe-PA software

ABSTRACT

With the development of sophisticated water quality models and the advances in computational power, data assimilation (DA) techniques, especially ensemble-based methods (the ensemble Kalman filter and particle filter), are attracting considerable attention in water quality modeling for improving the estimation of state variables and parameters in water quality models. The ensemble Kalman filter (EnKF) has become the most popular DA method while the particle filter (PF), which does not rely on Gaussian or quasi-linearity assumptions, is seldom applied in water quality modeling. Here, we present a comparison between the PF and EnKF for the update of model parameters related to river metabolism. The two filters are implemented in ProSe-PA, a hydro-biogeochemical software, and their performance is assessed on two synthetic case studies. The results indicate that PF and EnKF can estimate dissolved oxygen concentrations and the posterior probability distribution function of the associated parameters, either precisely for both filters in the case of a slightly nonlinear system (reaeration at the air–water interface) or more precisely for the PF in the case of a strongly nonlinear system (organic matter degradation) dominated by heterotrophic bacterial activities. Since the PF is more accurate, its usage is recommended for water quality modeling and guidelines are provided for its set-up.

1. Introduction

With the development of sophisticated water quality models (Warn, 1987; Hamrick, 1992; Billen et al., 1994; Whitehead et al., 1997; Even et al., 1998; Pelletier et al., 2006), complete and complex biogeochemical processes of an aquatic system can be simulated to understand its biogeochemical functioning (Flipo et al., 2004; Vilmin et al., 2015a, 2016; Bae and Seo, 2018; Sadeghian et al., 2018; Marescaux et al., 2020). However, a large number of model parameters are incorporated into these models to describe exhaustively the biogeochemical processes. This raises the question of prediction uncertainty (Beven, 1989; Polus et al., 2011; Cho et al., 2020). The model parameters need to be determined experimentally in the laboratory or calibrated by minimizing a loss function, which leads to problems of model validation and extrapolation (Arhonditsis and Brett, 2004; Polus et al., 2011).

To improve model performance, data assimilation techniques have been applied successfully in geosciences (Carrassi et al., 2018) and more specifically in water quality modeling (Wang et al., 2022; Cho et al., 2020). Evensen et al. (2022) also dedicated a book to the problem of state and parameter estimation in data assimilation. The data assimilation method combines ongoing observation data and model forecasts to obtain the optimal estimates of state variables and parameters of water quality models (Wikle and Berliner, 2007). Although numerous data assimilation methods exist in the literature (variational methods, Kalman filter, extended Kalman filter, ensemble Kalman filter, particle filter) and have been widely applied in meteorology and hydrology modeling (Courtier et al., 1994; Kalnay et al., 1996; Gauthier et al., 2007; Moradkhani et al., 2005a; Plaza et al., 2012; Abbaszadeh et al., 2018; Piazzi et al., 2021), few applications of data assimilation can be found in surface water quality modeling (Cho et al., 2020). The first application of data assimilation was published by Beck and Young (1976) using the extended Kalman filter (EKF). The EKF was then used as the main data assimilation technique in water quality modeling until 2009 (Mao et al., 2009). With the advances in computational power, the ensemble Kalman filter (EnKF) has become the most popular data assimilation method in surface water quality modeling (Huang et al., 2013; Kim et al., 2014; Huang and Gao, 2017; Page et al., 2018; Chen et al., 2019; Loos et al., 2020; Park et al., 2020). Most of the cited studies focused on the simulation of harmful algal blooms. The first implementation of a particle filter (PF) for water quality modeling was released very recently by Wang et al. (2019) and offers

* Corresponding author.
E-mail addresses: shuaitao.wang@sorbonne-universite.fr, shuaitaowang@outlook.com (S. Wang).

https://doi.org/10.1016/j.jhydrol.2023.129423
Received 6 September 2022; Received in revised form 15 March 2023; Accepted 17 March 2023
Available online 22 March 2023

promising perspectives that should favor its usage across the freshwater community.

The EnKF is based on the assumption that the forecasts of water quality states and model parameters are normally distributed and it updates them by linear formulas (Evensen, 2003). In nonlinear systems such as the modeling of an aquatic system, the Gaussian assumption cannot hold all the time and therefore EnKF yields biased samples and estimates (Wikle and Berliner, 2007). This assumption has been questioned also in the modeling of hydrologic systems and phytoplankton dynamics (Plaza et al., 2012; Pasetto et al., 2012; Huang et al., 2013). These authors recommended testing the PF, a more advanced method, in order to overcome this problem. Nonetheless, the PF has been applied only by Wang et al. (2022) for assimilating dissolved oxygen (DO) concentrations in the Seine River system, seemingly because the scientific community believes that the PF is more computationally demanding and is not easy to implement. In addition, no comparison between the PF and EnKF can be found in surface water quality modeling, specifically in DO data assimilation and model parameter estimation, which may be one of the reasons that the application of PFs in water quality modeling is so rare.

The aim of this paper is therefore to carry out a comparison of the PF and EnKF for DO data assimilation and for the estimation of model parameters. The performances of the PF and EnKF are evaluated on the basis of synthetic case studies using the hydro-biogeochemical program ProSe-PA (Wang et al., 2019). First, the recovery of DO concentrations by reaeration in a river system is built to model a slightly nonlinear system (Section 2.4.1). The computational time, the simulated DO concentrations, and the posterior distributions of the reaeration coefficient using different ensemble sizes are assessed (Section 3.1). Second, the PF and EnKF are applied in a strongly nonlinear system represented by heterotrophic bacterial activities in the river (Section 2.4.2). The uncertainties in parameter estimation and the PF set-up are finally discussed (Sections 4.2 and 4.3).

2. Material and methods

2.1. ProSe-PA software

The ProSe-PA (ProSe for Parallel computing and data Assimilation – Wang et al. (2019)) software couples the ProSe model (Even et al., 1998, 2004; Flipo et al., 2004; Vilmin et al., 2015a), which is the historical model widely used to investigate the biogeochemical functioning of the Seine River system (Even et al., 1998, 2004, 2007; Flipo et al., 2007; Polus et al., 2011; Raimonet et al., 2015; Vilmin et al., 2015b, 2018), with data assimilation frameworks (Fig. 1).

ProSe-PA is composed of three independent C-libraries: hydrodynamic, transport and biogeochemistry (Fig. 1). The hydrodynamic library calculates water heights and discharges by solving the 1D shallow water equations. Advection and dispersion are modeled using the hydraulic data calculated by the hydrodynamic library. The biogeochemistry library, C-RIVE, is based on the community-centered RIVE model (Billen et al., 1994; Garnier et al., 1995). The RIVE model simulates biogeochemical processes such as the cycles of nutrients, carbon, and DO in the water column and in an unconsolidated sediment layer.

To assimilate high-frequency oxygen concentration and to estimate model parameters, a PF was first implemented in ProSe-PA (Wang et al., 2019, 2022) and more recently an EnKF filter as well.

2.2. Sequential data assimilation frameworks in ProSe-PA: Particle filter and ensemble Kalman filter

Before describing each data assimilation framework, we introduce the state-space model on which each framework depends.

2.2.1. State-space model

A state-space model is a mathematical representation of the evolution of a system over time as a set of input, output, and state variables (Kalman, 1960). In our case, three equations are used to describe the evolution of DO concentrations (state variable in terms of physics). Let $\mathbf{X}$ be the random variable representing the model parameters and let $\mathbf{Y}$ represent the simulated DO concentrations. The random variables $\mathbf{X}$ and $\mathbf{Y}$ are characterized by their probability distribution function (pdf).

$$\mathbf{x_t} = \mathbf{x_{t-1}} + \boldsymbol{\eta_t} \tag{1}$$

$$\mathbf{y_t} = M(\mathbf{y_{t-1}}, \boldsymbol{\mu_t}, \mathbf{x_t}) + \boldsymbol{\nu_t} \tag{2}$$

$$\mathbf{y^{*t}} = \mathbf{H}\mathbf{y_t} + \boldsymbol{\epsilon_t} \tag{3}$$

Lower case $\mathbf{x}$ and $\mathbf{y}$ correspond to realizations of the random variables $\mathbf{X}$ and $\mathbf{Y}$. The observation vector $\mathbf{y}^*$ denotes a realization of the random variable $\mathbf{Y}^*$. The variables ν_t and ϵ_t stand for the unknown model errors and observation errors at time t (Eqs. (2) and (3)). The relative errors of the model are of the order 10^{-5} or 10^{-6}. This is far below observation errors. Therefore, no model errors are considered ($\nu_t = 0$) in our case. The evolution of parameter values is described by a random walk (η_t), which concurs with the prior knowledge of the parameters (Eq. (1)). $\mathbf{H}$ is the linear observation operator that maps simulations to observations.

2.2.2. Particle filter

A PF was implemented in ProSe-PA software and its efficiency was demonstrated in a synthetic case study (Wang et al., 2019) and in a real system (Wang et al., 2022). The PF is a method based on the Bayes theorem (Bayes, 1763) and the Markov property (Markov, 1906). It integrates observations $\mathbf{y}^*$ at each time step into the forward model (ProSe-PA) in order to approximate the posterior pdf $f(\mathbf{x}|\mathbf{y}^*)$ by a set of particles each associated with a weight ($\boldsymbol{\omega}$). For each particle, the weight is calculated using the Bayes theorem and Markov property (Doucet and Johansen, 2011; Wang et al., 2019), as stated by Eq. (4):

$$\omega_t \propto f(\mathbf{y_t^*}|\mathbf{x})\omega_{t-1} \tag{4}$$

where $f(\mathbf{y_t^*}|\mathbf{x})$ is the likelihood, that is, the probability to observe $\mathbf{y_t^*}$ given $\mathbf{x}$ at time t. ω_{t-1} denotes the posterior weight at time $t-1$ that gives the prior knowledge at time t. ω_t is then the posterior weight at time t.

To approximate the filtering distribution $f(\mathbf{x_t}|\mathbf{y^*_{1:t}})$, the normalized weights $\hat{\omega}_t^i = \frac{\omega_t^i}{\sum \omega_t^i}$ are computed. The $f(\mathbf{x_t}|\mathbf{y^*_{1:t}})$ can be approximated by the ensemble particles (Doucet et al., 2001):

$$f(\mathbf{x_t}|\mathbf{y^*_{1:t}}) \approx \sum_{i=1}^{N} \hat{\boldsymbol{\omega}}_t^i \delta_{\mathbf{x}_t^i} \tag{5}$$

where N is the ensemble size and δ the Dirac measure.

Assuming that the observation errors (Eq. (3), ϵ_t) are Gaussian and mutually independent at each observation station, we compute the likelihood $f(\mathbf{y_t^*}|\mathbf{x}^i)$ using the probability density function of the multivariate normal distribution (Eq. (6)).

$$\ln L(\mathbf{y}_t^*|\mathbf{x}_t^i) = -\frac{m}{2}\ln(2\pi) - \frac{1}{2}\ln(|\Sigma|) - \frac{1}{2}(\mathbf{y}_t^* - \mathbf{H}\mathbf{y}_t^i)^{\mathrm{T}}\Sigma^{-1}(\mathbf{y}_t^* - \mathbf{H}\mathbf{y}_t^i) \tag{6}$$

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

where m is the number of monitoring stations and i represents the particle i. The linear observation operator ($\mathbf{H}$) maps the simulated DO concentrations at the monitoring sites ($\mathbf{H}\mathbf{y}_t^i$). Σ corresponds to the error covariance matrix of the observations. Since the set of observation errors is mutually independent, Σ is a diagonal matrix and the diagonal terms correspond to the variance of the measurement errors at each monitoring station.

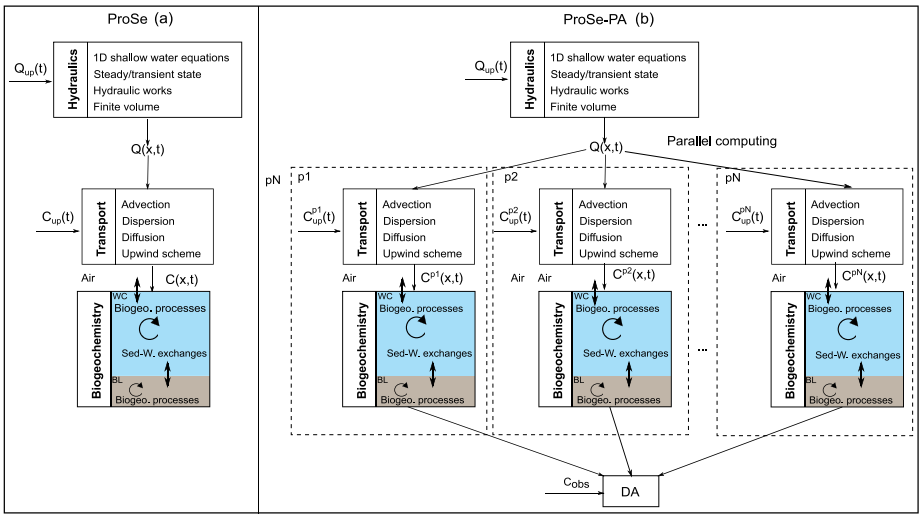


Fig. 1. (a) Schematic description of ProSe with the hydraulic, transport, and biogeochemistry modules. (b) Schematic description of ProSe-PA with a flowchart of the data assimilation framework (particles or ensemble member p1, p2, …, pN).
Source: Modified from Wang et al. (2022).

A common problem when applying the PF is the degeneracy of the particles. This means that almost all the particles get a near-zero weight and only a few particles or none has a high weight after a given simulation time. In such case, the posterior pdf $f(\mathbf{x}_t|\mathbf{y}^*_{1:t})$ cannot be approximated adequately by the ensemble particles (Eq. (5)). To reduce the degeneracy effect, a resampling procedure is used. The resampling procedure roughly duplicates particles with high weights while eliminating particles with near-zero weights. A series of resampling methods were proposed in the literature and reviewed by Li et al. (2015). The systematic resampling technique (Kitagawa, 1996; Moradkhani et al., 2005a; Li et al., 2015) was chosen for the ProSe-PA software.

It is not necessary to resample the particles at each time step, but only when the particles show some signs of degeneracy. The criterion for performing the resampling is based on the variance of the weights, which indicates the degree of degeneracy. Kong et al. (1994) defined the effective sample size (N_{eff}) to monitor the degree of degeneracy. The effective sample size can be approximated as follows, from the normalized weights:

$$\widehat{N_{eff}} = \frac{1}{\sum_{i=1}^{N}(\hat{\omega}_t^i)^2} \tag{7}$$

The particle resampling is carried out once $\widehat{N_{eff}}$ falls below a user-defined threshold ($N_{thres} = \alpha \cdot N$). The $\widehat{N_{eff}}$ itself has a maximum value of N (number of particles) and a minimum value of 1. Since the systematic resampling technique is used in ProSe-PA, the weights are reset to $1/N$ after resampling.

In particle filtering, we consider posterior parameter values at time $t-1$ as prior parameter values at time t, which means that the random variable η_t in Eq. (1) equals 0 ($\mathbf{x_t} = \mathbf{x_{t-1}}$). The posterior distribution $f(\mathbf{x}_t|\mathbf{y}^*_{1:t})$ evolves with the particle weights. However, the particles that have high weights may be duplicated many times during resampling, which results in sample impoverishment. To restore the diversity of the particles, a random perturbation is added to the parameter values after particle resampling (Eq. (8)):

$$\mathbf{x}^i_{t+1} = \mathbf{x}^i_{t,resampling} + \boldsymbol{\eta}^i_t \quad \boldsymbol{\eta}^i_t \sim N(0, (s \cdot \boldsymbol{\Phi})^2) \tag{8}$$

where s is a user-defined parameter and $\boldsymbol{\Phi}$ is the range space of parameters.

2.2.3. Ensemble Kalman filter

The EnKF algorithm (Evensen, 1994, 2003), like the traditional Kalman filter, consists of two sequential steps: forecast and analysis. During the forecast step, the model ensemble is propagated forward in time using the prior model state and parameter values (Eq. (2)). Then the model state and parameter values are updated using the linear Kalman filter analysis formula (Eq. (9)). In our case, only the parameter values are updated during the analysis step to ensure the continuity of mass balance, which is crucial for water quality modeling.

$$\mathbf{x}^a_t = \mathbf{x}^f_t + \mathbf{K}_t(\mathbf{y}^*_t - \mathbf{H}\mathbf{y}_t) \tag{9}$$

$$\mathbf{x}^f_t = \mathbf{x}^a_{t-1} + \eta_t \tag{10}$$

where $\mathbf{x}^a_t \in \mathbb{R}^{p\times N}$ are the analysis values of the parameters (posterior) with p the number of parameters. $\mathbf{x}^f_t \in \mathbb{R}^{p\times N}$ denotes the forecast parameter values (prior) and $\mathbf{K}_t \in \mathbb{R}^{p\times N}$ represents the Kalman gain matrix. The forecast parameter values ($\mathbf{x}^f_t \in \mathbb{R}^{p\times N}$) are obtained by adding the random values to the analyzed parameter values ($\mathbf{x}^a_t \in \mathbb{R}^{p\times N}$, Eq. (10)). The ensemble random values, with ensemble mean equal to 0, are noted as $\eta_t \sim N(0, (s \cdot \boldsymbol{\Phi})^2)$.

The Kalman gain matrix is calculated as follows (Burgers et al., 1998; Evensen et al., 2022):

$$\mathbf{K}_t = \mathbf{P}^f_t\mathbf{H}^{\mathrm{T}}(\mathbf{H}\mathbf{P}^f_t\mathbf{H}^{\mathrm{T}} + \mathbf{R}_t)^{-1} \tag{11}$$

where $\mathbf{P}^f_t$ and $\mathbf{R}_t$ represent the error covariance matrices of forecast and observation at time t.

For parameter estimation, the above form of Kalman gain can be rewritten (Moradkhani et al., 2005b; Huang et al., 2013):

$$\mathbf{K}_t = \boldsymbol{\Sigma}^{\mathbf{xy}}_t(\boldsymbol{\Sigma}^{\mathbf{yy}}_t + \boldsymbol{\Sigma}^{\mathbf{y^*y^*}}_t)^{-1} \tag{12}$$

where $\boldsymbol{\Sigma}^{\mathbf{xy}}_t \in \mathbb{R}^{p\times m}$ is the cross-covariance matrix of the parameter ensemble $\mathbf{x}_t \in \mathbb{R}^{p\times N}$ and the forecast ensemble state at monitoring stations $\mathbf{H}\mathbf{y}_t \in \mathbb{R}^{m\times N}$. $\boldsymbol{\Sigma}^{\mathbf{yy}}_t \in \mathbb{R}^{m\times m}$ denotes the error covariance matrix of the prediction at monitoring stations ($\mathbf{H}\mathbf{y}_t$) and $\boldsymbol{\Sigma}^{\mathbf{y^*y^*}}_t \in \mathbb{R}^{m\times m}$ is the observation error covariance matrix.

In EnKF, the error statistics are represented using an ensemble model state (Evensen, 1994; Burgers et al., 1998; Evensen, 2003; Evensen et al., 2022). The true state is generally unknown and estimated by the ensemble mean. Consequently, the unknown error covariance matrix of the forecast state ($\boldsymbol{\Sigma}^{\mathbf{yy}}_{\mathbf{t}}$) can be estimated as follows:

$$\begin{aligned}\boldsymbol{\Sigma}^{\mathbf{yy}}_t &= \frac{1}{N-1}\sum_{i=1}^{N}(\mathbf{H}y^i_t - \overline{\mathbf{H}\mathbf{y}_t})(\mathbf{H}y^i_t - \overline{\mathbf{H}\mathbf{y}_t})^{\mathbf{T}} \\ &= \frac{1}{N-1}(\mathbf{H}\mathbf{y}_t - \overline{\mathbf{H}\mathbf{y}_t})(\mathbf{H}\mathbf{y}_t - \overline{\mathbf{H}\mathbf{y}_t})^{\mathbf{T}}\end{aligned} \tag{13}$$

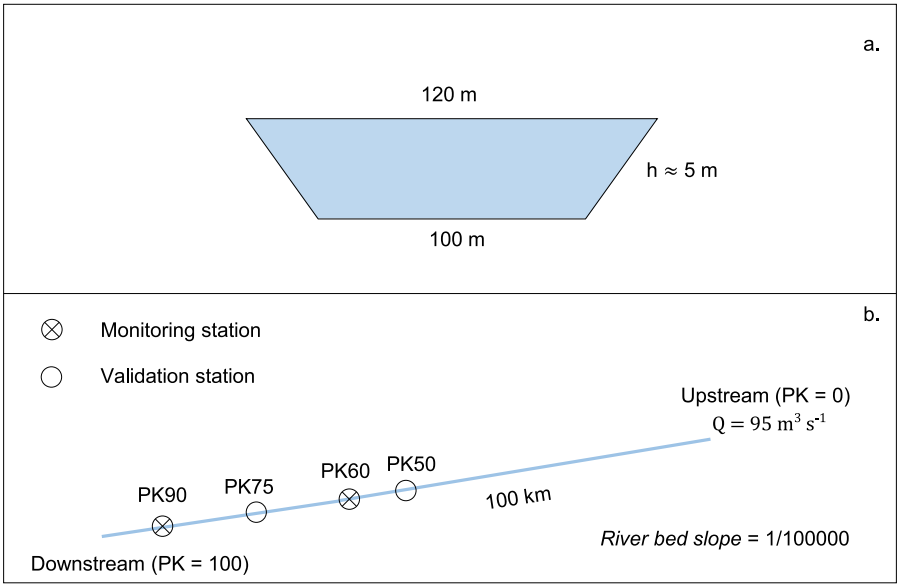


Fig. 2. a. Cross section of the trapezoid-shaped river channel; Light blue area represents water body with a height of about 5 m. b. Location of monitoring and validation stations; PK: kilometric point. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

where $\mathbf{\Sigma_t^{yy}}$ has dimension $m \times m$, where m is the number of monitoring stations. N is the ensemble size and $\mathbf{H}y_t^i \in \mathbb{R}^{m\times 1}$ corresponds to the predicted states of the ensemble member i at m monitoring stations. The ensemble mean is calculated as $\overline{\mathbf{Hy}_t} \in \mathbb{R}^{m\times 1} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{H}y_t^i$.

The observation errors defined in Eq. (3) can be obtained by perturbing the observation values ($y_t^{*,i} = y_t^* + \epsilon_t^i \in \mathbb{R}^{m\times 1}$). The ensemble of perturbations, with ensemble mean equal to 0, can be noted as $\epsilon_t \in \mathbb{R}^{m\times N}$. An estimator of the observation error covariance matrix ($\mathbf{\Sigma_t^{y^*y^*}} \in \mathbb{R}^{m\times m}$) can then be constructed using the ensemble perturbations (Evensen, 2003; Evensen et al., 2022):

$$\mathbf{\Sigma_t^{y^*y^*}} = \frac{\mathbf{1}}{\mathbf{N-1}}\sum_{\mathbf{i=1}}^{\mathbf{N}}\epsilon_t^i(\epsilon_t^i)^{\mathbf{T}} \tag{14}$$

$$= \frac{\mathbf{1}}{\mathbf{N-1}}\boldsymbol{\epsilon}_t\boldsymbol{\epsilon}_t^{\mathbf{T}} \tag{15}$$

Similarly, the cross-covariance matrix of parameter ensembles and forecast state ensembles ($\mathbf{\Sigma_t^{xy}} \in \mathbb{R}^{p\times m}$) is estimated as (Evensen, 2003; Evensen et al., 2022):

$$\mathbf{\Sigma_t^{xy}} = \frac{1}{N-1}\sum_{i=1}^{N}(x_t^i - \overline{\mathbf{x_t}})(\mathbf{H}y_t^i - \overline{\mathbf{Hy_t}})^{\mathbf{T}}$$

$$= \frac{\mathbf{1}}{\mathbf{N-1}}(\mathbf{x_t} - \overline{\mathbf{x_t}})(\mathbf{Hy_t} - \overline{\mathbf{Hy_t}})^{\mathbf{T}} \tag{16}$$

where $\overline{\mathbf{x_t}} \in \mathbb{R}^{p\times 1} = \frac{1}{N}\sum_{i=1}^{N}x_t^i$.

The above three covariance matrix estimates are then plugged into (12) to compute (an approximation of) the Kalman gain.

2.3. Description of the case study: Geometric and hydraulic data

A trapezoid-shaped river channel, which is 100 km long (Fig. 2b), is conceptualized to mimic the Seine River. The trapezoid-shaped cross section has a bottom base of 100 m and top base of 120 m (Fig. 2a). A discharge of 95 $m^3\ s^{-1}$ corresponding to a water velocity of 0.17 m s^{-1} is imposed upstream. The river bed slope and water height corresponding to these values are 10^{-6} and 5 m, respectively. No inflows and dams are considered in the case study. The concept of kilometric point (PK) is used in ProSe-PA to represent the location of a point in the river channel. The downstream and upstream points have PKs of 100 and 0, respectively (Fig. 2b).

2.4. Data assimilation scenarios: from oxygen reaeration to bacterial activities

2.4.1. Recovery of DO by reaeration

The physical oxygen reaeration is modeled as:

$$\frac{d[O_2]}{dt} = \frac{K_{rea}}{h}([O_2]_{sat}(T) - [O_2]) \tag{17}$$

$$K_{rea} = \sqrt{\frac{D_m \times V_w}{h}} + (K_{wind} \times V_{wind}^{2.23} \times (D_m \times 10^4)^{\frac{2}{3}} + K_{navig}) \tag{18}$$

with,

K_{rea}: Reaeration coefficient, [m s^{-1}]

$[O_2]_{sat}(T)$: Saturation concentration of DO at temperature T, [mgO_2 L^{-1}]

$[O_2]$: DO concentration, [mgO_2 L^{-1}]

h: Water height, [m]

D_m: Molecular diffusivity of dissolved oxygen, [m^2 s^{-1}]

V_w and V_{wind}: Water velocity and wind speed, [m s^{-1}]

K_{wind} and K_{navig}: Reaeration coefficients related to wind and navigation, [m s^{-1}]

The oxygen reaeration coefficient (K_{rea}) is composed of three terms: molecular diffusion (D_m), wind turbulence (K_{wind}), and navigation turbulence (K_{navig}). Only K_{navig}, which was identified as the most influential parameter in DO concentrations in winter (Wang et al., 2018), is estimated using PF and EnKF in this case. The wind speed at 10 m elevation is considered null. A K_{navig} value of 0.015 m h^{-1}, as the reference parameter value, is used to generate observation data.

A period of 15 days is simulated with ProSe-PA. To simulate varying oxygen saturation concentrations, the water temperature is set to 10 °C ($[O_2]_{sat}$ = 11.29 mgO_2 L^{-1}) from day 0 to day 5 and then increases to 20 °C ($[O_2]_{sat}$ = 9.11 mgO_2 L^{-1}) on day 15. The initial and upstream oxygen concentrations are set to 3.80 mgO_2 L^{-1} (Table 1), which enables a simulation of the recovery of DO by reaeration.

2.4.2. Bacterial activities

The RIVE model simulates explicitly the heterotrophic bacterial activities: growth, respiration, and mortality. A Monod function (Monod, 1949) is used to describe the growth of heterotrophic bacteria limited by small monomeric substrate concentration ($[SMS]$). The bacterial respiration is expressed as,

$$\frac{d[O_2]}{dt} = -\tau(1-Y)upt \tag{19}$$

Table 1
Initial and boundary conditions of the simulations.

Oxygen reaeration					
Species	Description	C_{ini}	$C_{upstream}$	Unit	Temperature
$[O_2]$	Dissolved oxygen	3.80	3.80	$mgO_2\ L^{-1}$	10 °C – 20 °C
Bacterial activities					
$[O_2]$	Dissolved oxygen	9.11	9.11	$mgO_2\ L^{-1}$	
$[SMS]$	Small monomeric substrate	1.22	1.22	$mgC\ L^{-1}$	20 °C
$[HB]$	Heterotrophic bacteria	0.01	0.01	$mgC\ L^{-1}$	

Table 2
Parameters considered in PF and EnKF.

Oxygen reaeration					
Parameters	Description	Min	Max	Reference	Unit
K_{navig}	Reaeration coefficient due to navigation	0.0	0.05	0.015	$[m\ h^{-1}]$
Bacterial activities					
μ_{max}	Maximum growth rate of bacteria	0.01	0.13	0.04	$[h^{-1}]$
Y	Growth yield of bacteria	0.03	0.5	0.15	[–]

"Reference": Parameter values used to generated virtual observation data.

$$upt = \frac{1}{Y}\mu_{max} e^{\frac{(T-T_{opt})^2}{\sigma^2}} \frac{[SMS]}{[SMS]+K_{SMS}}[HB]$$

with,

τ: $\frac{32}{12}$, when considering the full oxidation of organic matter by the respiration process, $[mgO_2/mgC]$

Y: Bacteria growth yield, [–]

upt: Uptake of small monomeric substrate for bacteria growth, $[mgC\ L^{-1}\ h^{-1}]$

μ_{max}: Maximum growth rate, $[h^{-1}]$

T and T_{opt}: Water temperature and optimal water temperature for bacteria growth, [°C]

K_{SMS}: Half-saturation constant for small monomeric substrate, $[mgC\ L^{-1}]$

$[HB]$: Heterotrophic bacterial biomass, $[mgC\ L^{-1}]$

The DO concentrations are most sensitive to μ_{max} and Y when bacterial activities drive the river metabolism during low-flow periods (Wang et al., 2018). Therefore, only μ_{max} and Y are considered when evaluating the performances of PF and EnKF.

To simulate the strong bacterial activities, the water temperature is the same as the optimal temperature for the growth of bacteria (T_{opt} = 20 °C) during the simulation. The DO concentrations are initially saturated in the system (9.11 $mgO_2\ L^{-1}$). The upstream DO concentrations are constant and saturated during the simulation (Table 1). The initial and upstream conditions of small monomeric substrate ($[SMS]$) and bacteria ($[HB]$) are listed in Table 1. A maximum growth rate of 0.04 h^{-1} and a growth yield of 0.15 are used to generate observation data.

2.4.3. Mimicking oxygen monitoring

As mentioned above, only K_{navig} is considered for oxygen reaeration simulation while μ_{max} and Y are considered for the simulation of bacterial activities. Reference values of those parameters (Table 2) are used to generate reference oxygen data, $[O_2]_{ref}$, every 15 min (Fig. 2b) in all model cells with a forward simulation of ProSe-PA. A random noise is added to those reference data at locations PK60 and PK90, in order to mimic data acquired by monitoring systems, $[O_2]_{obs}$, which are entailed by observational errors. The observational errors are defined in Eq. (20).

$$[O_2]_{obs} = [O_2]_{ref} + \theta, \quad \theta \sim N(0, (0.01 \times [O_2]_{ref})^2) \tag{20}$$

2.5. Statistical criteria for evaluating the performances of PF and EnKF in DO simulation

Two monitoring stations of DO (PK60 and PK90) and two validation stations (PK50 and PK75) are modeled (Fig. 2). The data from the monitoring stations are assimilated by ProSe-PA either by a PF or an EnKF. The performances of PF and EnKF in the simulation of DO concentrations are evaluated through RMSE (root mean square error) and KGE (Kling–Gupta efficiency, (Kling et al., 2012)) at the validation stations (PK50, PK75). For completeness, the values of the criteria are also provided at the monitoring stations. RMSE is the standard deviation of the simulation errors (Eq. (21)). The KGE is based on the decomposition of the Nash–Sutcliffe efficiency, which provides the analysis of the relative importance of its different components (correlation, bias, and variability). KGE ranges from -Inf to 1. The closer to 1, the more accurate the model.

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{k=1}^{N_{obs}} ([O_2]_{sim,k} - [O_2]_{obs,k})^2}{N_{obs}}} \tag{21}$$

$$\mathrm{KGE} = 1 - \sqrt{(r-1)^2 + (\beta-1)^2 + (\gamma-1)^2} \tag{22}$$

where

N_{obs}: Number of observations

$[O_2]_{sim,k}$ and $[O_2]_{obs,k}$: Simulated and observed DO concentrations

r: Correlation coefficient

β: Model bias. $\beta = \frac{\mu_{sim}}{\mu_{obs}}$, with μ the mean of the DO concentrations

γ: Coefficient of variation. $\gamma = \frac{\sigma_{sim}/\mu_{sim}}{\sigma_{obs}/\mu_{obs}}$, with σ the standard deviation of the DO concentrations

2.6. Visualization of the filtering distributions of the considered parameters

To visualize the filtering distributions (Eq. (5)), kernel density estimation is used to determine the results of PF while a normal distribution is assumed for the results of EnKF, which can be calculated using the ensemble mean and standard deviation (Wikle and Berliner, 2007). To compare the results of PF obtained using different random perturbations given the ensemble size, a identical bandwidth is used in the kernel density estimation.

3. Results

3.1. Oxygen reaeration due to navigation (slightly nonlinear system)

3.1.1. Calculation time

Nine simulations with different ensemble sizes (10, 30, 50, 100, 150, 200, 300, 400, 500) are realized to compare the calculation time of PF and EnKF (Fig. 3). Both PF and EnKF simulations are run with 20 threads (Intel(R) Xeon(R) CPU E5-2640 v4 @ 2.40 GHz) and a time step of 15 min for a simulation period of 15 days.

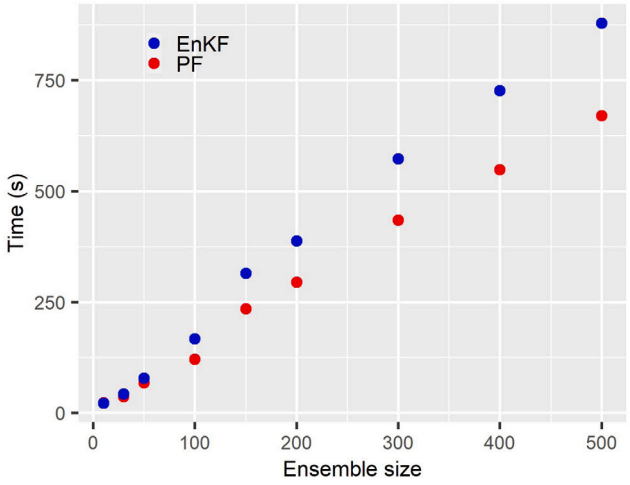


Fig. 3. Calculation time using PF and EnKF. Simulations are run with 20 threads (Intel(R) Xeon(R) CPU E5-2640 v4 @ 2.40 GHz). A time step of 15 min is used for a 15-day period.

No significant differences in the calculation time between PF and EnKF can be observed when the ensemble size is smaller than 100 (Fig. 3). However, the calculation time using PF is lower by 25% compared to the calculation time using EnKF when the ensemble size is greater than 100. When the ensemble size increases, more time is indeed needed to compute the error covariance matrices $\mathbf{\Sigma}_t^{\mathbf{xy}}$ and $\mathbf{\Sigma}_t^{\mathbf{yy}}$ and solve the related linear systems.

3.1.2. Simulated DO concentrations

All simulations with PF can retrieve the DO concentrations very accurately at all stations, even with an ensemble size of 10. Given the remarkable similarity of the results obtained across all stations and ensemble sizes, only the results at station PK75 are shown here (Fig. 5). For PF, a maximum RMSE of 0.02 mgO_2 L^{-1} is estimated at station PK50 (Fig. 4), which is rather small compared to the observation errors (standard deviations between 0.038 mgO_2 L^{-1} and 0.08 mgO_2 L^{-1}, Eq. (20)). All KGEs are over 0.99 for PF.

The RMSE for EnKF first decreases and then increases with the ensemble size. Its minimum value (around 0.10 mgO_2 L^{-1}) is obtained with the ensemble sizes of 200 and 300 (Fig. 4), while being relatively close to the observation errors. With the ensemble size of 300, maximum values of KGE are obtained for EnKF.

Even though the PF is remarkably accurate, the two filters capture the recovery of DO by reaeration and obtain satisfactory results (Fig. 5).

3.1.3. Estimated posterior distributions of K_{navig}

With an ensemble size of 300, both PF and EnKF produce satisfactory posterior distributions of K_{navig} (Fig. 6). The reference value of K_{navig} (0.015 m h^{-1}) is well characterized by the modes of the distributions. However, the distributions estimated by PF are much narrower than those estimated by EnKF, which are relatively broad (Fig. 6).

The simulated oxygen concentrations and posterior distributions of K_{navig} show that PF and EnKF work well for a slightly nonlinear system (oxygen recovery by reaeration) even though PF estimates carry less uncertainty.

3.2. Heterotrophic bacterial activities (strongly nonlinear system)

In this case, a series of simulations are carried out with different ensemble sizes (100, 300, 400, 500, 800 and 1000). A minimum ensemble size of 500 is deemed imperative for obtaining desirable results. Only the results with the ensemble size of 500 are presented in this section. A perturbation of $s = 0.01$ (Eq. (8)) yield satisfactory results for PF while a perturbation of $s = 0.03$ is needed for EnKF. The results show that PF performs much better than EnKF when the bacterial activities control the river metabolism (strongly nonlinear system), both in the identification of parameter values and in the retrieval of DO concentrations.

Table 3
Statistical criteria in the case of bacterial activities.

Criteria	PK50	PK60	PK75	PK90	Unit
RMSE (PF)	0.014	0.014	0.016	0.015	mgO_2 L^{-1}
KGE (PF)	0.991	0.991	0.988	0.992	[–]
RMSE (EnKF)	0.202	0.230	0.241	0.182	mgO_2 L^{-1}
KGE (EnKF)	0.715	0.817	0.885	0.975	[–]

3.2.1. Simulated oxygen concentrations

The simulated oxygen concentrations (ensemble mean) show that PF performs better than EnKF under conditions of strong bacterial activities (Fig. 7). All RMSEs are smaller than 0.02 mgO_2 L^{-1} (Table 3) for PF while all RMSEs are over 0.18 mgO_2 L^{-1} (Table 3), which is higher than the observation errors (around 0.08 mgO_2 L^{-1}). The analysis of the KGEs confirms the fact that the PF outperforms the EnKF. The results indicate that EnKF can capture the depletion of oxygen trend, but it struggles to retrieve the DO concentrations correctly.

3.2.2. Estimated posterior distributions

The estimated posterior distribution of μ_{max} and Y confirms the efficiency of PF for estimating parameters in a water quality model, which is usually a strongly nonlinear system. The parameter values of μ_{max} and Y (0.04 h^{-1} and 0.015, respectively) used to generate observation data are very well identified by the PF (Fig. 8) while the EnKF fails to identify the bacterial yield Y and slightly overestimates the value of μ_{max} (0.04 h^{-1}) with a much broader pdf than the PF (Fig. 8).

4. Discussion

4.1. Recommendations for data assimilation in water quality modeling

Although the ensemble Kalman filter (EnKF) has become the most popular data assimilation method for updates of state variables and model parameters in surface water quality modeling (Cho et al., 2020), the results of this research show that the particle filter (PF) outperforms the EnKF for parameter inference and oxygen concentration estimation, especially in a strongly nonlinear system as for heterotrophic bacteria activity which is one of the main drivers of river metabolism (Odum, 1956; Escoffier et al., 2018). Also in ocean biogeochemical modeling, Gharamti et al. (2017) indicate that the uncertainty associated with the state estimates using two EnKF-based methods increases during the spring blooms (strongly nonlinear behavior). That is because the EnKF assumes Gaussian errors for forecast states and parameters, which is not realistic for the functioning of an aquatic system (lake, river) (Huang et al., 2013; Wang et al., 2022). Evensen et al. (2022, p. 95) also stated that "Commonly-used ensemble data-assimilation methods, like the EnKF ..., only sample the posterior pdf correctly in the Gauss-linear case and typically fail in cases with strong nonlinearity". The PF overcomes the Gaussian assumption of EnKF. The implementation of PF into a complex hydro-biogeochemical program (ProSe-PA) has been proved to be realistic (Wang et al., 2019) and its performance (calculation cost, DO concentration updates, parameter estimation) has been recently evaluated for the Seine River system, France (Wang et al., 2022). Therefore, the authors recommend the use of PF for the update of biogeochemical state variables and model parameters in surface water quality modeling.

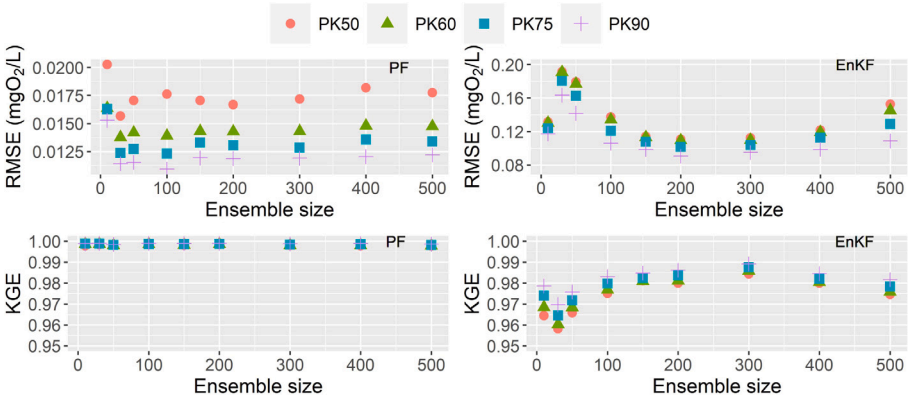


Fig. 4. Performances of PF and EnKF in DO reaeration evaluated by RMSE and KGE.

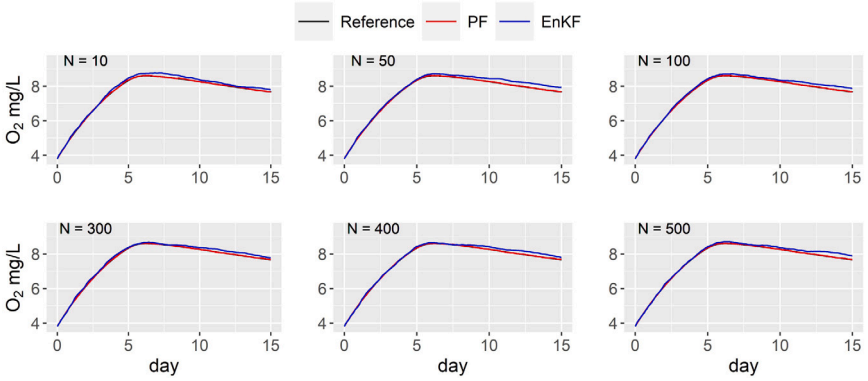


Fig. 5. Simulated DO concentrations (ensemble mean) at station PK75 with different ensemble size. The PF simulation overlaps the reference data.

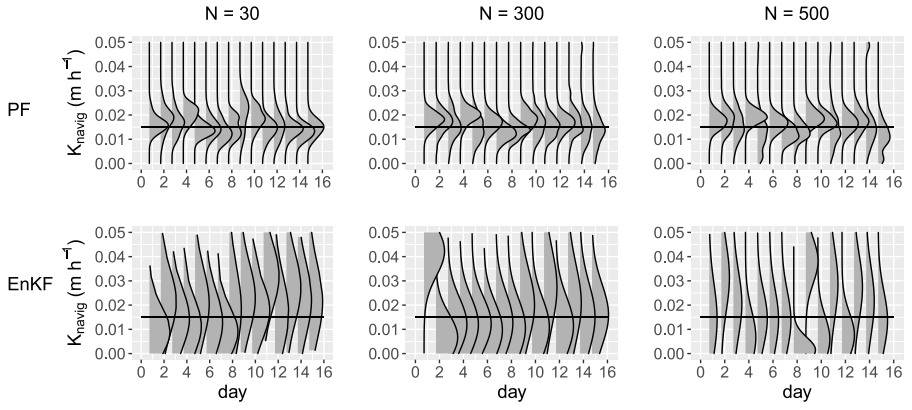


Fig. 6. Estimated posterior distributions of K_{navig} by PF and EnKF. Black line represents the reference value of K_{navig} (0.015 m h^{-1}) that is used to generate observation data.

4.2. Uncertainties in parameter estimation: PF vs. EnKF

Both PF and EnKF are able to quantify the parameter uncertainties in a slightly nonlinear system (oxygen reaeration). These uncertainties are characterized by the posterior pdfs of the parameters. With a random perturbation of $s = 0.10$ (Eqs. (10) and (8)), the distributions estimated by EnKF are broader than those estimated by PF (Fig. 6) and thus depict larger uncertainties. That is because the ensemble members in EnKF are equally weighted and the posterior distribution is assumed to be Gaussian and characterized by its first two moments (mean and variance) (Wikle and Berliner, 2007). Compared with weighted ensemble members in PF (Eq. (5)), the assumption of equally weighted samples in EnKF is generally not valid and results in biased samples (Wikle and Berliner, 2007).

The uncertainties in parameter estimation result from multiple reasons for PF, such as observation error or perturbation after resampling (Eq. (8)). A sensitivity assessment of the perturbation parameter after resampling (s, Eq. (8)) for the case of oxygen recovery is realized. To compare the estimated distributions with different random walks, a fixed bandwidth (0.00011) is used in the kernel density estimation (Fig. 9). The results show that a perturbation of $s = 0.01$ enables a perfect estimation of K_{navig} (Fig. 9). The posterior distributions of K_{navig}

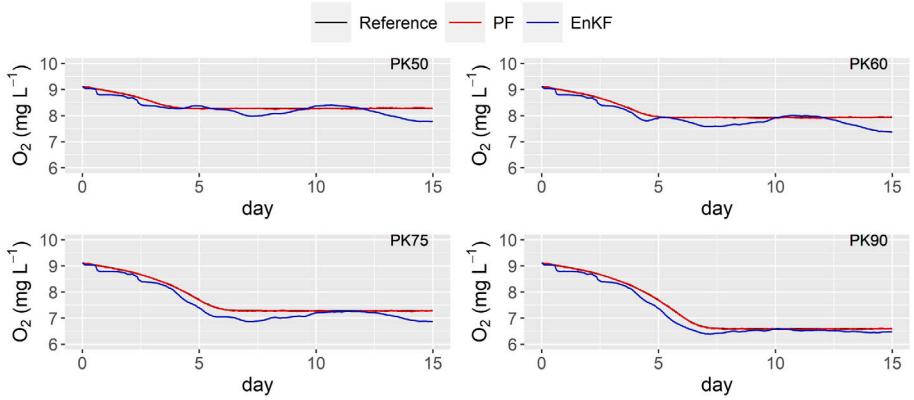


Fig. 7. Simulated DO concentrations (ensemble mean) by PF and EnKF with an ensemble size of 500.

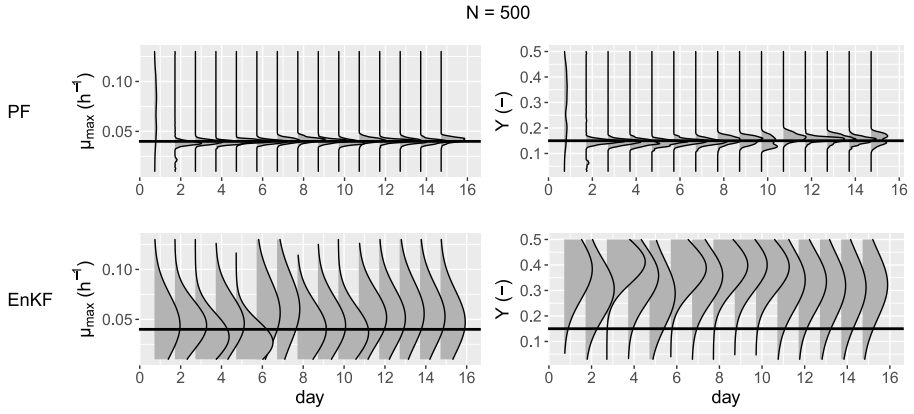


Fig. 8. Daily posterior distributions of μ_{max} and Y estimated by PF and EnKF with an ensemble size of 500.

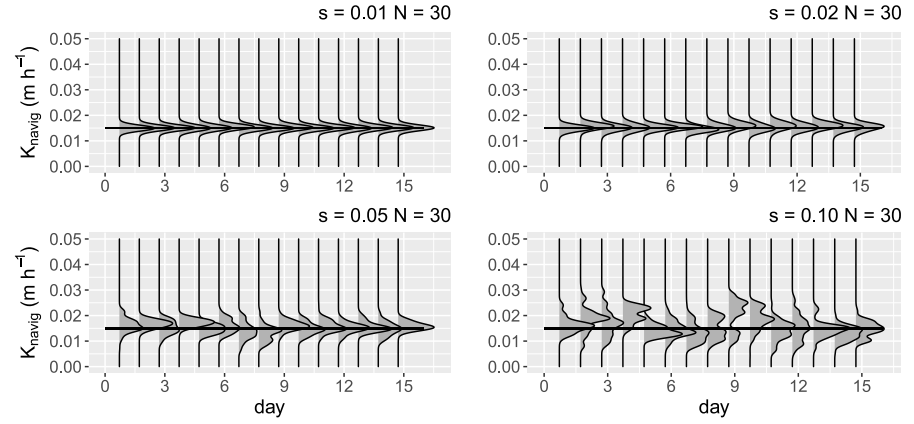


Fig. 9. Daily posterior distributions of K_{navig} estimated by PF with $s = 0.01$, $s = 0.02$, $s = 0.05$, and $s = 0.10$ for an ensemble size of 30 (bandwidth = 0.00011).

are narrow, which depicts small estimation uncertainties. However, when increasing the random walk (s = 0.02, s = 0.05, and s = 0.10), the posterior distributions become increasingly spread out (Fig. 9), which corresponds to an increase in parameter uncertainties with larger perturbations.

4.3. Impact of random walk: precision and capacity of PF

As shown above, a small perturbation (s = 0.01) enables a perfect estimation of oxygen concentration and parameters both in slightly nonlinear systems and strongly nonlinear systems. In other words, the precision of the results is high. That is because the parameter values are stationary during the simulation. Once the posterior distribution (Eq. (5)) is well approximated by the ensemble, it is no longer necessary to explore the parameter space. However, the capacity of the filter with a small random walk perturbation parameter to respond to fast changes in the parameters remains questionable. A change of parameter values was reported in the evolution of microorganism communities, especially in the development of phytoplankton (Mao et al., 2009; Huang et al., 2013; Wang et al., 2022).

To illustrate a time-varying parameter, an extreme scenario mimicking the shift of K_{navig} value from 0.015 m h^{-1} to 0.03 m h^{-1} on day 5 is designed (Fig. 10). With a small random walk (s = 0.01), the filter needs 9 days to capture the change in K_{navig} value, and large discrepancies between simulated and reference DO concentrations

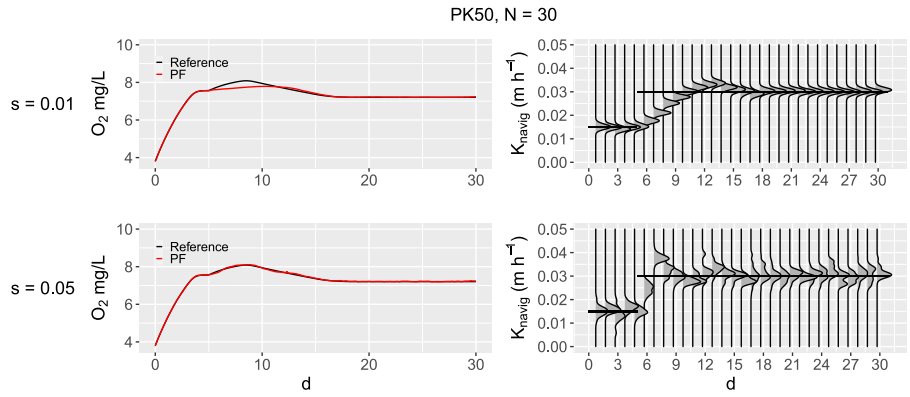


Fig. 10. Simulated DO concentrations (ensemble mean) at station PK50 and daily posterior distributions of K_{navig} with $s = 0.01$ and $s = 0.05$ (bandwidth = 0.00092). PF: particle filter; Reference: DO concentrations with reference parameters; Varying of the K_{navig} value from 0.015 m h^{-1} to 0.03 m h^{-1} (black lines) on day 5.

are obtained during this period. When increasing the random walk ($s = 0.05$), the filter takes 3 days to respond to the fast change in K_{navig} value, and the simulated DO concentrations (ensemble mean) are satisfactory. Once the filter is stabilized, the precision with a small random walk is higher than that with a large random walk.

Nevertheless, the small random walk may stay stuck in a local maximum in the case of multimodal posterior distribution. In this case, the ensemble cannot characterize the posterior distribution adequately and larger random walks are recommended to allow for a maximum search of the posterior distribution (Moradkhani et al., 2012). It should be also noted that a larger random walk after the resampling step could result in an overspread ensemble (biased samples), which cannot account adequately for the prior distribution. Therefore, it is crucial to find a good balance between the precision and the ability of parameter space exploration in the PF.

Further studies can focus on improving the resampling/perturbation procedure using more advanced techniques, such as the auxiliary particle filter (Pitt and Shephard, 1999; Johansen and Doucet, 2008), which performs the resampling at time step $t-1$ using the available measurement at time step t, or the MCMC moves with the metropolis acceptance ratio to determine whether to accept a proposed sample (Metropolis et al., 1953; Hastings, 1970; Gilks and Berzuini, 2001; Doucet and Johansen, 2011; Moradkhani et al., 2012).

5. Conclusions

To compare the performances of PF and EnKF for the updates of water quality states and model parameters, the PF and EnKF implemented in the hydro-biogeochemical program ProSe-PA were assessed via two synthetic case studies. The main conclusions are given below.

- PF is recommended for the updates of DO concentrations and model parameters in surface water quality modeling.
- For quasi-linear oxygen reaeration inference, both PF and EnKF can capture the recovery of DO by reaeration and identify the reaeration coefficient. But the uncertainty associated with oxygen and parameter estimates obtained using EnKF is larger than that obtained using PF.
- The calculation time using PF is lower by 25% compared to EnKF when the ensemble size is more than 100.
- For nonlinear bacterial activities inference, PF shows a high efficiency for both the simulation of DO concentrations and the estimation of bacteria-related parameters, while it is difficult to retrieve the DO concentrations and estimate the parameter values properly with the EnKF.
- The small random walk after the resampling procedure in PF yields high precision of oxygen and parameter estimates, while a larger random walk is necessary to capture efficiently the fast change in the parameters. It is thus important to find a good balance between precision and the ability of the parameter search.

CRediT authorship contribution statement

Shuaitao Wang: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing – original draft, Visualization. **Nicolas Flipo:** Conceptualization, Methodology, Software, Formal analysis, Writing – review & editing, Supervision, Funding acquisition. **Thomas Romary:** Conceptualization, Methodology, Formal analysis, Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgments

This work is a contribution to the PIREN-Seine research program (https://www.piren-seine.fr), part of the french Long-Term Socio-Ecological Research (LTSER) site "Zones Ateliers Seine".

References

Abbaszadeh, P., Moradkhani, H., Yan, H., 2018. Enhancing hydrologic data assimilation by evolutionary particle filter and Markov Chain Monte Carlo. Adv. Water Resour. 111, 192–204. http://dx.doi.org/10.1016/j.advwatres.2017.11.011.

Arhonditsis, G., Brett, M., 2004. Evaluation of the current state of mechanistic aquatic biogeochemical modeling. Mar. Ecol. Prog. Ser. 271, 13–26.

Bae, S., Seo, D., 2018. Analysis and modeling of algal blooms in the Nakdong river, Korea. Ecol. Model. 372, 53–63. http://dx.doi.org/10.1016/j.ecolmodel.2018.01.019.

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

Beck, M., Young, P., 1976. Systematic identification of DO-BOD model structure. J. Env. Eng. Div., Am. Soc. Civ. Eng. 102, 909–927.

Beven, K., 1989. Changing ideas in hydrology. The case of physically-based model. J. Hydrol. 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.

Burgers, G., Jan van Leeuwen, P., Evensen, G., 1998. Analysis scheme in the ensemble Kalman filter. Mon. Weather Rev. 126 (6), 1719–1724. http://dx.doi.org/10.1175/1520-0493(1998)126<1719:ASITEK>2.0.CO;2.

Carrassi, A., Bocquet, M., Bertino, L., Evensen, G., 2018. Data assimilation in the geosciences: An overview of methods, issues, and perspectives. Wiley Interdiscip. Rev. Clim. Change 9 (5), e535. http://dx.doi.org/10.1002/wcc.535.

Chen, C., Huang, J., Chen, Q., Zhang, J., Li, Z., Lin, Y., 2019. Assimilating multi-source data into a three-dimensional hydro-ecological dynamics model using ensemble Kalman filter. Environ. Model. Softw. 117, 188–199. http://dx.doi.org/10.1016/j.envsoft.2019.03.028.

Cho, K.H., Pachepsky, Y., Ligaray, M., Kwon, Y., Kim, K.H., 2020. Data assimilation in surface water quality modeling: A review. Water Res. 186, 116307. http://dx.doi.org/10.1016/j.watres.2020.116307.

Courtier, P., Thépaut, J.N., Hollingsworth, A., 1994. A strategy for operational implementation of 4D-Var, using an incremental approach. Q. J. R. Meteorol. Soc. 120 (519), 1367–1387. http://dx.doi.org/10.1002/qj.49712051912.

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

Doucet, A., Johansen, A.M., 2011. A tutorial on particle filtering and smoothing : fiteen years later. In: Crisan, D., Rozovskii, B. (Eds.), The Oxford Handbook of Nonlinear Filtering. In: Oxford Handbooks in Mathematics, Oxford University Press, Oxford, N.Y., ISBN: 9780199532902, pp. 656–705.

Escoffier, N., Bensoussan, N., Vilmin, L., Flipo, N., Rocher, V., David, A., Métivier, F., Groleau, A., 2018. Estimating ecosystem metabolism from continuous multi-sensor measurements in the Seine river. Environ. Sci. Pollut. Res. 25 (24), 23451–23467. http://dx.doi.org/10.1007/s11356-016-7096-0.

Even, S., Bacq, N., Ruelland, D., Billen, G., Garnier, J., Poulin, M., Théry, S., Blanc, S., 2007. New tools for modelling water quality of hydrosystems: An application in the seine river basin in the frame of the water framework directive. Sci. Total Environ. 375 (1–3), 274–291. http://dx.doi.org/10.1016/j.scitotenv.2006.12.019.

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. J. Geophys. Res.: Oceans 99 (C5), 10143–10162. http://dx.doi.org/10.1029/94JC00572.

Evensen, G., 2003. The ensemble Kalman filter: theoretical formulation and practical implementation. Ocean Dyn. 53 (4), 343–367. http://dx.doi.org/10.1007/s10236-003-0036-9.

Evensen, G., Vossepoel, F.C., Jan van Leeuwen, P., 2022. Data Assimilation Fundamentals - A Unified Formulation of the State and Parameter Estimation Problem. Springer, http://dx.doi.org/10.1007/978-3-030-96709-3.

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.H., Lalande, M., 2007. Primary production in headwater streams of the Seine basin: the Grand Morin case study. Sci. Total Environ. 375, 98–109. http://dx.doi.org/10.1016/j.scitotenv.2006.12.015.

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.

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. Mon. Weather Rev. 135 (6), 2339–2354. http://dx.doi.org/10.1175/MWR3394.1.

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 Model. 112, 65–89. http://dx.doi.org/10.1016/j.ocemod.2017.02.006.

Gilks, W.R., Berzuini, C., 2001. Following a moving target—Monte Carlo inference for dynamic Bayesian models. J. R. Stat. Soc. Ser. B Stat. Methodol. 63 (1), 127–146. http://dx.doi.org/10.1111/1467-9868.00280.

Hamrick, J.M., 1992. A Three-Dimensional Environmental Fluid Dynamics Computer Code : Theoretical and cOmputational Aspects. Special Report in Applied Marine Science and Ocean Engineering; No. 317, Virginia Institute of Marine Science, William & Mary, http://dx.doi.org/10.21220/V5TT6C.

Hastings, W.K., 1970. Monte Carlo sampling methods using Markov Chains and their applications. Biometrika 57 (1), 97–109.

Huang, J., Gao, J., 2017. An improved ensemble Kalman filter for optimizing parameters in a coupled phosphorus model for lowland polders in Lake Taihu Basin, China. Ecol. Model. 357, 14–22. http://dx.doi.org/10.1016/j.ecolmodel.2017.04.019.

Huang, J., Gao, J., Liu, J., Zhang, Y., 2013. State and parameter update of a hydrodynamic-phytoplankton model using ensemble Kalman filter. Ecol. Model. 263, 81–91. http://dx.doi.org/10.1016/j.ecolmodel.2013.04.022.

Johansen, A.M., Doucet, A., 2008. A note on auxiliary particle filters. Statist. Probab. Lett. 78 (12), 1498–1504. http://dx.doi.org/10.1016/j.spl.2008.01.032.

Kalman, R.E., 1960. A new approach to linear filtering and prediction problems. J. Basic Eng. 82 (1), 35–45.

Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., Iredell, M., Saha, S., White, G., Woollen, J., Zhu, Y., Chelliah, M., Ebisuzaki, W., Higgins, W., Janowiak, J., Mo, K.C., Ropelewski, C., Wang, J., Leetmaa, A., Reynolds, R., Jenne, R., Joseph, D., 1996. The NCEP/NCAR 40-year reanalysis project. Bull. Am. Meteorol. Soc. 77 (3), 437–472. http://dx.doi.org/10.1175/1520-0477(1996)077<0437:TNYRP>2.0.CO;2.

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. J. Hydrol. 519, 2810–2821. http://dx.doi.org/10.1016/j.jhydrol.2014.09.073.

Kitagawa, G., 1996. Monte Carlo filter and smoother for non-Gaussian nonlinear state space models. J. Comput. Graph. Statist. 5 (1), 1–25. http://dx.doi.org/10.1080/10618600.1996.10474692.

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

Kong, A., Liu, J.S., Wong, W.H., 1994. Sequential imputations and Bayesian missing data problems. J. Amer. Statist. Assoc. 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. Front. Inf. Technol. Electron. Eng. 16 (11), 969–984. http://dx.doi.org/10.1631/FITEE.1500199.

Loos, S., Shin, C.M., Sumihar, J., Kim, K., Cho, J., Weerts, A.H., 2020. Ensemble data assimilation methods for improving river water quality forecasting accuracy. Water Res. 171, 115343. http://dx.doi.org/10.1016/j.watres.2019.115343.

Mao, J., Lee, J.H., Choi, K., 2009. The extended Kalman filter for forecast of algal bloom dynamics. Water Res. 43 (17), 4214–4224. http://dx.doi.org/10.1016/j.watres.2009.06.012.

Marescaux, A., Thieu, V., Gypens, N., Silvestre, M., Garnier, J., 2020. Modeling inorganic carbon dynamics in the seine river continuum in France. Hydrol. Earth Syst. Sci. 24 (5), 2379–2398. http://dx.doi.org/10.5194/hess-24-2379-2020.

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

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., Teller, E., 1953. Equation of state calculations by fast computing machines. J. Chem. Phys. 21 (6), 1087–1092. http://dx.doi.org/10.1063/1.1699114.

Monod, J., 1949. The growth of bacterial cultures. Annu. Rev. Microbiol. 3 (1), 371–394. http://dx.doi.org/10.1146/annurev.mi.03.100149.002103.

Moradkhani, H., DeChant, C.M., Sorooshian, S., 2012. Evolution of ensemble data assimilation for uncertainty quantification using the particle filter-Markov Chain Monte Carlo method. Water Resour. Res. 48 (12), http://dx.doi.org/10.1029/2012WR012144.

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

Moradkhani, H., Sorooshian, S., Gupta, H.V., Houser, P.R., 2005b. Dual state–parameter estimation of hydrological models using ensemble Kalman filter. Adv. Water Resour. 28 (2), 135–147. http://dx.doi.org/10.1016/j.advwatres.2004.09.002.

Odum, H.T., 1956. Primary production in flowing waters. Limnol. Oceanogr. 1, 795–801.

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 Res. 134, 74–85. http://dx.doi.org/10.1016/j.watres.2018.01.046.

Park, S., Kim, K., Shin, C., Min, J.H., Na, E.H., Park, L.J., 2020. Variable update strategy to improve water quality forecast accuracy in multivariate data assimilation using the ensemble Kalman filter. Water Res. 176, 115711. http://dx.doi.org/10.1016/j.watres.2020.115711.

Pasetto, D., Camporese, M., Putti, M., 2012. Ensemble Kalman filter versus particle filter for a physically-based coupled surface–subsurface model. Adv. Water Resour. 47, 1–13. http://dx.doi.org/10.1016/j.advwatres.2012.06.009.

Pelletier, G., Chapra, S., Tao, H., 2006. QUAL2Kw — A framework for modeling water quality in streams and rivers using a genetic algorithm for calibration. Environ. Modell. Softw. 419–425. http://dx.doi.org/10.1016/j.envsoft.2005.07.002.

Piazzi, G., Thirel, G., Perrin, C., Delaigue, O., 2021. Sequential data assimilation for streamflow forecasting: Assessing the sensitivity to uncertainties and updated variables of a conceptual hydrological model at basin scale. Water Resour. Res. 57 (4), http://dx.doi.org/10.1029/2020WR028390.

Pitt, M.K., Shephard, N., 1999. Filtering via simulation: Auxiliary particle filters. J. Amer. Statist. Assoc. 94 (446), 590–599.

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. Hydrol. Earth Syst. Sci. 16 (2), 375–390. http://dx.doi.org/10.5194/hess-16-375-2012.

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. Hydrol. Process. 25 (2), 217–233. http://dx.doi.org/10.1002/hyp.7838.

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 Res. 73, 373–387. http://dx.doi.org/10.1016/j.watres.2015.01.026.

Sadeghian, A., Chapra, S.C., Hudson, J., Wheater, H., Lindenschmidt, K.E., 2018. Improving in-lake water quality modeling using variable chlorophyll a/algal biomass ratios. Environ. Model. Softw. 101, 73–85. http://dx.doi.org/10.1016/j.envsoft.2017.12.009.

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. http://dx.doi.org/10.1007/s10533-014-0038-3.

Vilmin, L., Flipo, N., de Fouquet, C., Poulin, M., 2015b. Pluri-annual sediment budget in a navigated river system: The Seine river (France). Sci. Total Environ. 502, 48–59. http://dx.doi.org/10.1016/j.scitotenv.2014.08.110.

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? Environ. Sci. Pollut. Res. 25 (24), 23485–23501. http://dx.doi.org/10.1007/s11356-016-7109-z.

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. Glob. Biogeochem. Cycles 30 (7), 1086–1104. http://dx.doi.org/10.1002/2015GB005271.

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 Res. 144, 341–355. http://dx.doi.org/10.1016/j.watres.2018.07.033.

Wang, S., Flipo, N., Romary, T., 2019. Oxygen data assimilation for estimating micro-organism communities' parameters in river systems. Water Res. 165, 115021. http://dx.doi.org/10.1016/j.watres.2019.115021.

Wang, S., Flipo, N., Romary, T., Hasanyar, M., 2022. Particle filter for high frequency oxygen data assimilation in river systems. Environ. Model. Softw. 105382. http://dx.doi.org/10.1016/j.envsoft.2022.105382.

Warn, A.E., 1987. SIMCAT-a catchment simulation model for planning inverstment for river quality. In: Beck, M.B. (Ed.), Systems Analysis in Water Quality Management. IAWPRC Pergamon, Oxford, pp. 211–218.

Whitehead, P.G., Williams, R.J., Lewis, D.R., 1997. Quality simulation along river systems (QUASAR): model theory and development. Sci. Total Environ. 194/195, 447–456.

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