Comparaison méthodologique pour l'assimilation de données appliquée à la simulation de la qualité des eaux de surface

Shuaitao Wang[1*], Nicolas Flipo[2], Thomas Romary[2]

[1] Sorbonne Université, CNRS, EPHE, UMR Metis, 75005 Paris, France

[2] Mines Paris - PSL, Centre de Géosciences, 77300 Fontainebleau, France

[*] shuaitao.wang@sorbonne-universite.fr

Résumé

Avec le développement de modèles de qualité de l'eau sophistiqués et l'amélioration de la puissance de calcul, les techniques d'assimilation de données, en particulier les méthodes d'ensemble (le filtre de Kalman d'ensemble et le filtre particulaire), attirent une attention considérable dans la modélisation de la qualité de l'eau pour améliorer l'estimation des variables d'état et la caractérisation des paramètres dans les modèles de qualité de l'eau. Le filtre de Kalman d'ensemble (EnKF) est devenu la méthode la plus populaire, tandis que le filtre particulaire (PF), qui est une méthode plus avancée, est rarement appliqué à la modélisation de la qualité de l'eau. Nous présentons ici une comparaison entre le PF et l'EnKF pour caractériser les paramètres du modèle liés au métabolisme des rivières. Les deux filtres sont implémentés dans le logiciel hydro-biogéochimique PROSE-PA et leurs performances sont évaluées via deux cas d'études synthétiques. Les résultats indiquent que le PF et l'EnKF peuvent tous les deux estimer les concentrations en oxygène dissous et la fonction de densité de probabilité a posteriori des paramètres associés, soit avec une même précision pour les deux filtres dans le cas d'un système légèrement non linéaire (réaération à l'interface air-eau), soit plus précisément pour le PF dans le cas d'un système fortement non linéaire (dégradation de la matière organique) dominé par les activités des bactéries hétérotrophes. Il est ainsi recommandé de favoriser le PF, plus précis, pour la modélisation de la qualité de l'eau.

Points clés

- *Le filtre particulaire est plus performant que le filtre de Kalman d'ensemble pour traiter la dégradation de la matière organique en rivière*
- *Les deux filtres sont satisfaisants pour traiter la réoxygénation de la colonne d'eau par la navigation*
- *Les estimations des paramètres sont plus précises avec le filtre particulaire*

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 the improvement of 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), as a more advanced method, 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 the hydro-biogeochemical software ProSe-PA and their performances are assessed via two synthetic case studies. The results indicate that both the PF and EnKF can estimate dissolved oxygen concentrations and the posterior probability density function of 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.

Key points

- In a strongly non-linear system (degradation of organic matter), the PF outperforms the EnKF
- In a slightly non-linear system (reaeration through navigation), the results of EnKF and PF are both satisfactory
- Parameter estimates are more precise with PF

Introduction

Avec le développement de modèles de qualité de l'eau sophistiqués (Warn, 1987; Hamrick, 1992; Billen et al., 1994; Whitehead et al., 1997; Even et al., 1998; Pelletier et al., 2006), les processus biogéochimiques d'un système aquatique peuvent être simulés afin de comprendre son fonctionnement biogéochimique (Flipo et al., 2004; Vilmin et al., 2015, 2016; Bae and Seo, 2018; Sadeghian et al., 2018; Marescaux et al., 2020). Cependant, un grand nombre de paramètres sont utilisés dans ces modèles pour décrire de manière exhaustive les processus biogéochimiques. Cela soulève la question de l'incertitude des prédictions (Beven, 1989; Polus et al., 2011; Cho et al., 2020). Les valeurs des paramètres doivent être déterminées en laboratoire ou calibrées en minimisant une fonction objectif, ce qui entraîne des problèmes de validation et d'extrapolation du modèle (Arhonditsis and Brett, 2004; Polus et al., 2011).

Pour améliorer les performances des modèles, les techniques d'assimilation de données ont été appliquées avec succès dans le domaine des géosciences (Carrassi et al., 2018) et plus particulièrement dans la modélisation de la qualité de l'eau (Wang et al., 2022; Cho et al., 2020). Evensen et al. (2022) ont également rédigé un livre sur le problème de l'estimation de l'état et des paramètres dans l'assimilation de données. La méthode d'assimilation de données combine les observations et les prévisions du modèle pour obtenir les estimations optimales des variables d'état et des paramètres des modèles de qualité de l'eau (Wikle and Berliner, 2007). Bien que de nombreuses méthodes d'assimilation de données existent dans la littérature (notamment les méthodes variationnelles, le filtre de Kalman, le filtre de Kalman étendu, le filtre de Kalman d'ensemble et le filtre particulaire) et ont été largement appliquées à la modélisation de la météorologie et de l'hydrologie (Courtier

et al., 1994; Kalnay et al., 1996; Gauthier et al., 2007; Moradkhani et al., 2005; Plaza et al., 2012; Abbaszadeh et al., 2018; Piazzi et al., 2021), peu d'applications de l'assimilation de données peuvent être trouvées sur la modélisation de la qualité des eaux de surface (Cho et al., 2020). La première application de l'assimilation de données a été publiée par Beck and Young (1976) en utilisant le filtre de Kalman étendu (EKF). L'EKF a ensuite été utilisé comme principale technique d'assimilation de données dans la modélisation de la qualité de l'eau jusqu'en 2009 (Mao et al., 2009). Avec les avancées de la puissance de calcul, le filtre de Kalman d'ensemble (EnKF) est devenu la méthode la plus populaire dans la modélisation de la qualité des eaux de surface (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). La plupart des études citées se sont concentrées sur la simulation des efflorescences algales. L'implémentation d'un filtre particulaire (PF) pour la première fois en modélisation de la qualité de l'eau a été publiée très récemment par Wang et al. (2019) et offre des perspectives prometteuses qui devraient favoriser son utilisation dans la communauté scientifique.

L'EnKF est basé sur l'hypothèse que les prévisions des états de la qualité de l'eau et des paramètres du modèle sont normalement distribuées et il les analyse par les formules linéaires (Evensen, 2003). Or, dans les systèmes non linéaires tels que la modélisation d'un système aquatique, l'hypothèse gaussienne ne peut pas être toujours maintenue et des estimations biaisées peuvent alors être produites (Wikle and Berliner, 2007). Cette hypothèse a également été remise en question dans la modélisation des systèmes hydrologiques et de la dynamique du phytoplancton (Plaza et al., 2012; Pasetto et al., 2012; Huang et al., 2013). Ces auteurs ont donc recommandé de tester un filtre particulaire (PF), une méthode plus avancée, afin de surmonter l'hypothèse gaussienne. Cependant, la communauté scientifique n'a pas encore adopté le PF, qu'elle perçoit plus exigeant en termes de calcul et pas simple à implémenter numériquement. Le PF n'a ainsi été utilisé que par Wang et al. (2022) pour assimiler avec succès les concentrations d'oxygène dissous dans la Seine. Afin de contribuer aux développements méthodologiques de demain en termes de mise en synergie des données et des modèles, le présent rapport s'attache à comparer les performances du PF et de l'EnKF pour l'assimilation de données d'oxygène dissous et l'estimation des paramètres *ad hoc*. Les performances des deux filtres sont évaluées dans le cadre d'un cas synthétique, représentatif de la Seine à la traversée de l'agglomération parisienne, s'appuyant sur le programme hydro-biogéochimique ProSe-PA (Wang, 2019; Wang et al., 2019).

1 Matériel et méthodes

1.1 Le logiciel ProSe-PA

Le logiciel ProSe-PA (ProSe en Parallèle et Assimilation de données – Wang (2019); Wang et al. (2019)) couple le modèle ProSe (Even et al., 1998; Flipo et al., 2004; Even et al., 2007; Flipo et al., 2007; Vilmin et al., 2015), qui est le modèle historique largement utilisé pour étudier le fonctionnement biogéochimique de la Seine depuis l'amont de l'agglomération parisienne jusqu'au barrage de Poses en amont de son estuaire (Even et al., 1998, 2004, 2007; Polus et al., 2011; Raimonet et al., 2015; Vilmin et al., 2015, 2016, 2018), avec une méthode d'assimilation de données (Fig. 1).

ProSe-PA est notamment composé de trois librairies fonctionnelles indépendantes en ANSI C : **libhyd** – hydrodynamique fluviale – , **libttc** – transport advectif et dispersif –, et **librive** – fonctionnement biogéochimique de la colonne d'eau, des sédiments et du périphyton – (Fig. 1). La librairie hydrodynamique calcule les hauteurs d'eau et les débits en résolvant les équations de Saint-Venant 1D intégrant l'impact des ouvrages hydrauliques sur les écoulements. L'advection et la dispersion sont modélisées en résolvant une équation de transport d'espèces dissoutes 1D forcée par les données hydrauliques calculées par libhyd. Librive est basée sur le modèle RIVE centré sur les communautés de micro-organismes (Billen et al., 1994; Garnier et al., 1995). Elle simule les processus biogéochimiques tels que les cycles des nutriments, du carbone et de l'oxygène dans la colonne d'eau et dans une couche de sédiments non consolidée.

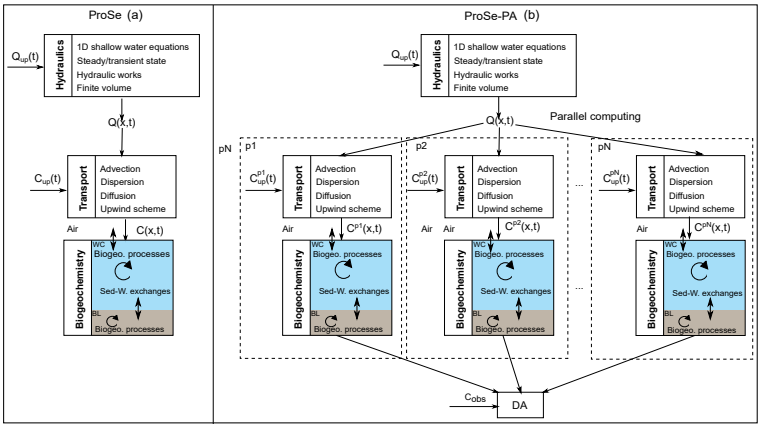

Figure 1. (a) Description schématique de ProSe avec les modules hydraulique, transport et biogéochimie. (b) Description schématique de ProSe-PA avec l'assimilation de données (particules ou membre de l'ensemble p1, p2, ... , pN). Modifié à partir de Wang et al. (2022).

Afin d'assimiler les concentrations en oxygène dissous haute fréquence et estimer les paramètres du modèle, le PF a d'abord été implémenté dans le logiciel ProSe-PA (Wang, 2019; Wang et al., 2019, 2022), puis l'EnKF a récemment été ajouté afin d'en évaluer les performances.

1.2 Assimilation de données séquentielles dans ProSe-PA : le filtre particulaire et le filtre de Kalman d'ensemble

1.2.1 Le filtre particulaire – PF

Un PF a été implémenté dans le logiciel ProSe-PA et son efficacité a été démontrée pour un cas synthétique (Wang et al., 2019) et pour l'étude de la qualité de l'eau de la Seine durant l'année 2011 (Wang et al., 2022). Le PF est une méthode basée sur le théorème de Bayes (Bayes, 1763) et la propriété de Markov (Markov, 1906). Elle intègre les observations à chaque pas de temps dans le modèle direct (ProSe-PA) afin de caractériser les lois de distribution des paramètres a posteriori par un ensemble de particules associées chacune à un poids. Pour chaque particule, le poids est calculé en utilisant le théorème de Bayes et la propriété de Markov. Pour faciliter la lecture du rapport, les formalismes ne sont pas présentés ici. L'implémentation du filtre particulaire est détaillée dans (Doucet and Johansen, 2011; Wang, 2019; Wang et al., 2019; Evensen et al., 2022).

1.2.2 Le filtre de Kalman d'ensemble – EnKF

L'EnKF (Evensen, 1994, 2003), comme le filtre de Kalman traditionnel, se compose de deux étapes séquentielles : la prévision et l'analyse. Au cours de l'étape de prévision, l'ensemble est propagé dans le temps à l'aide des variables d'état et des valeurs des paramètres antérieurs. Ensuite, les variables d'état et les valeurs des paramètres sont mises à jour en utilisant la formule d'analyse du filtre de Kalman linéaire (Kalman, 1960). En revanche, seules les valeurs des paramètres sont mises à jour pendant l'étape d'analyse afin d'assurer la continuité du bilan massique dans notre cas, ce qui est crucial pour la modélisation de la qualité de l'eau. Plus de détails sur ce filtre sont consultables dans Evensen et al. (2022).

1.3 Description des cas synthétiques

1.3.1 Données géométriques et hydrauliques

Un tronçon de rivière de forme trapézoïdale et de 100 km de long (Fig. 2b) est conceptualisé pour imiter la Seine. La section transversale de forme trapézoïdale présente une base inférieure de 100 m et une base supérieure de 120 m (Fig. 2a). Un débit de 95 m^3 s^{-1} correspondant à une vitesse de l'écoulement de 0.17 m s^{-1} est imposé en amont. Aucun débit entrant ni barrage n'est considéré dans ce cas. Le concept de point kilométrique (PK) est utilisé dans ProSe-PA pour représenter l'emplacement d'un point dans la rivière. Les points amont et aval sont respectivement situés au PK0 et au PK100 (Fig. 2b).

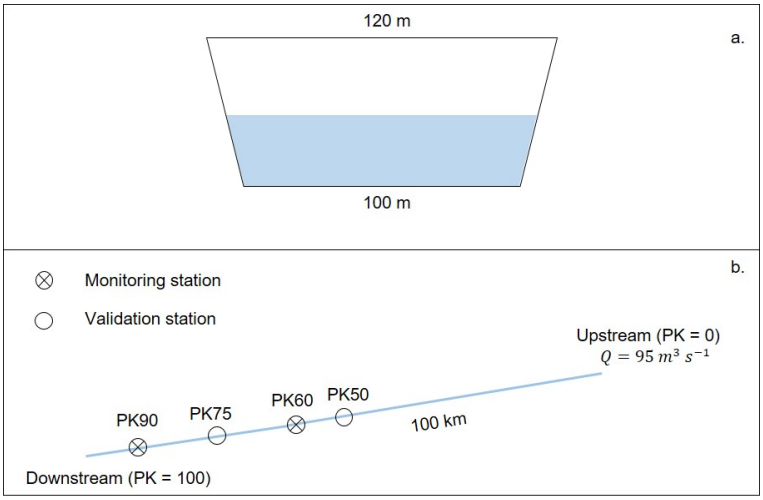

Figure 2. Cas d'étude synthétique imitant la Seine à l'étiage à la traversée de l'agglomération parisienne. Positionnement des stations de monitoring sur lesquelles se base l'assimilation de données et des stations de validation des calculs issus de ProSe-PA

1.3.2 Stations d'observation et validation

Deux stations d'observation en oxygène dissous sont positionnées aux PK60 et PK90 (Fig. 2). Ces deux stations servent de support à l'assimilation de données dans le logiciel. Les observations sont générées en ajoutant une erreur aléatoire θ (Eq. (1)) aux données de référence qui sont obtenues à l'aide d'une simulation directe réalisée avec ProSe avec des valeurs de paramètres dites de référence.

$$[O_2]_{obs} = [O_2]_{ref} + \theta, \quad \theta \sim N(0, (0.01 \times [O_2]_{ref})^2) \tag{1}$$

Pour évaluer les performances des deux filtres, PF et EnKF, les concentrations simulées par ProSe-PA à deux stations de validation (PK50 et PK75) sont comparées aux données d'oxygène de référence. Seules les données issues des deux stations d'observation sont assimilées par ProSe-PA. Les stations de validation sont utilisées uniquement pour évaluer les performances du logiciel.

1.4 Scénario d'assimilation de données

1.4.1 Réaeration à l'interface air-eau

La réaération de l'oxygène est modélisée comme suit :

$$\frac{d[O_2]}{dt} = \frac{K_{rea}}{h}([O_2]_{sat}(T) - [O_2]) \tag{2}$$

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

avec,

K_{rea} : Coefficient de réaération, [m s^{-1}]

$[O_2]_{sat}(T)$: Concentration d'oxygène dissous à saturation à la température T, [mgO_2 L^{-1}]

$[O_2]$: Concentration d'oxygène dissous, [mgO_2 L^{-1}]

h : Hauteur d'eau, [m]

D_m : Diffusion moléculaire de l'oxygène dissous, [m^2 s^{-1}]

V_w et V_{wind} : Vitesse de l'écoulement et vitesse du vent, [m s^{-1}]

K_{wind} et K_{navig} : Coefficients de réaération liés au vent et à la navigation, [m s^{-1}].

Le coefficient de réaération (K_{rea}) est composé de trois termes : la diffusion moléculaire (D_m), la turbulence liée au vent (K_{wind}) et la turbulence liée à la navigation (K_{navig}). Seul K_{navig}, qui a été identifié comme le paramètre le plus influent sur les concentrations en oxygène dissous en hiver (Wang et al., 2018), est estimé à l'aide de PF et EnKF dans cette étude. La vitesse du vent à 10 m d'altitude est considérée comme nulle. Une valeur de K_{navig} de 0.015 m h^{-1} est utilisée pour générer les observations toutes les 15 min aux stations PK60 et PK90. Les observations sont assimilées par PF et EnKF afin de reconstruire les concentrations aux stations PK50 et PK75 et de caractériser la distribution de K_{navig}.

Une période de 15 jours est simulée avec ProSe-PA. Pour inclure l'effet de la température, la température de l'eau est fixée à 10 °C ($[O_2]_{sat}$ = 11.29 mgO_2 L^{-1}) du jour 0 au jour 5, puis augmente linéairement à 20 °C ($[O_2]_{sat}$ = 9.11 mgO_2 L^{-1}) au jour 15. Les concentrations initiales et amont en oxygène sont fixées à 3.80 mgO_2 L^{-1} (Tab. 1), ce qui permet de simuler une réoxygénation conséquente par la réaération à l'interface air-eau.

Tableau 1. Conditions initiales et conditions aux limites des deux problèmes simulés : réoxygénation à l'interface air-eau et minéralisation de la matière organique par les bactéries hétérotrophes

Réaeration				
Variables	C_{init}	C_{amont}	Unité	Température
$[O_2]$	3.80	3.80	mgO_2 L^{-1}	10 °C - 20 °C
Minéralisation de la matière organique				
$[O_2]$	9.11	9.11	mgO_2 L^{-1}	
$[S]$	1.22	1.22	mgC L^{-1}	20 °C
$[HB]$	0.01	0.01	mgC L^{-1}	

1.4.2 Minéralisation de la matière organique

Le modèle RIVE simule explicitement les activités des bactéries hétérotrophes : la croissance, la respiration et la mortalité. Une fonction de Monod (Monod, 1949) est utilisée pour décrire la croissance des bactéries

hétérotrophes en fonction de la concentration en substrat organique ($[S]$). La respiration bactérienne est exprimée par :

$$\frac{d[O_2]}{dt} = -\tau(1-Y)uptakeS \tag{4}$$
$$uptakeS = \frac{1}{Y}\mu_{max} e^{\frac{(T-T_{opt})^2}{\sigma^2}} \frac{[S]}{[S]+K_S}[HB]$$

avec,

τ : $\frac{32}{12}$, en considérant l'oxydation complète de la matière organique par respiration, [mgO_2/mgC]

Y : Rendement de croissance, [-]

$uptakeS$: Absorption de substrat pour la croissance des bactéries, [mgC L^{-1} h^{-1}]

μ_{max} : Taux de croissance maximal, [h^{-1}]

T et T_{opt} : Température de l'eau et température optimale de l'eau pour la croissance des bactéries, [°C]

K_S : Constante de demi-saturation du substrat, [mgC L^{-1}]

$[HB]$: Biomasse bactérienne, [mgC L^{-1}]

Les concentrations en oxygène dissous sont plus sensibles à μ_{max} et Y lorsque les activités bactériennes dominent le système pendant les périodes de basses eaux (Wang et al., 2018; Hasanyar et al., 2022). Par conséquent, seuls μ_{max} et Y sont pris en compte pour évaluer les performances de PF et EnKF.

Pour simuler les fortes activités bactériennes, la température de l'eau est la même que la température optimale pour la croissance des bactéries (T_{opt} = 20 °C) lors de la simulation . Les concentrations en oxygène dissous sont initialisées à la saturation dans le système (9.11 mgO_2 L^{-1}). Les concentrations en amont sont constantes et également saturées (Tab. 1). Les conditions initiales et amont du substrat ($[S]$) et des bactéries ($[HB]$) sont listées dans le tableau 1.

1.5 Paramètres considérés pour l'assimilation de données

Comme mentionné ci-dessus, seul le paramètre K_{navig} est pris en compte pour la simulation de la réaération de l'oxygène, tandis que les paramètres μ_{max} et Y sont considérés pour la simulation de la minéralisation de la matière organique via les activités bactériennes. Les gammes de variation des paramètres et les valeurs des paramètres utilisées pour générer les observations sont indiquées dans le tableau 2.

Tableau 2. Paramètres considérés dans PF et EnKF. Valeurs extraites de Wang et al. (2019)

Réaeration à l'interface air-eau					
Paramètres	Description	Min	Max	Référence	Unité
K_{navig}	Coefficient de réaeration lié à la navigation	0.0	0.05	0.015	[m h^{-1}]
Activités bactériennes					
μ_{max}	Taux de croissance maximal	0.01	0.13	0.04	[h^{-1}]
Y	Rendement de croissance	0.03	0.5	0.15	[-]

"Référence" : Valeurs des paramètres utilisés pour générer les observations virtuelles

1.6 Critères statistiques pour évaluer la performance de PF et EnKF sur la simulation d'oxygène dissous

Les performances des deux filtres, PF et EnKF, sont évaluées par RMSE (*root mean square error* – eq. (5) –) et KGE (*Kling–Gupta efficiency*, (Kling et al., 2012) – eq. (6) –) aux stations de validation (PK50 et PK75). Le KGE est basé sur la décomposition du critère de Nash-Sutcliffe, qui permet d'analyser l'importance relative de ses différentes composantes (corrélation, biais et variabilité). Le KGE varie de -Inf à 1 : plus le KGE est proche de 1, plus le modèle est précis.

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{k=1}^{N_{obs}} ([O_2]_{sim,k} - [O_2]_{obs,k})^2}{N_{obs}}} \tag{5}$$

$$\mathrm{KGE} = 1 - \sqrt{(r-1)^2 + (\beta-1)^2 + (\gamma-1)^2} \tag{6}$$

où

N_{obs} : Nombre d'observations

$[O_2]_{sim,k}$ et $[O_2]_{obs,k}$: Concentrations simulées et observées

r : Coefficient de corrélation

β : Biais du modèle. $\beta = \frac{\mu_{sim}}{\mu_{obs}}$, avec μ la moyenne des concentrations

γ : Coefficient de variation. $\gamma = \frac{\sigma_{sim}/\mu_{sim}}{\sigma_{obs}/\mu_{obs}}$, avec σ l'écart type des concentrations

2 Résultats

2.1 Réaération liée à la navigation (système légèrement non-linéaire)

2.1.1 Temps de calcul

Neuf simulations avec différentes tailles d'ensembles (10, 30, 50, 100, 150, 200, 300, 400, 500) sont réalisées pour comparer le temps de calcul de PF et EnKF (Fig. 3). Les simulations PF et EnKF sont exécutées avec 20 *threads* (Intel(R) Xeon(R) CPU E5-2640 v4 @ 2.40GHz) et un pas de temps de 15 min pour une période de simulation de 15 jours.

Aucune différence significative entre le temps de calcul de PF et celui de EnKF ne peut être observée lorsque la taille de l'ensemble est inférieure à 100 (Fig. 3). Cependant, le temps de calcul avec PF est inférieur de 25% au temps de calcul avec EnKF lorsque la taille de l'ensemble est supérieure à 100. Lorsque la taille de l'ensemble augmente, il faut en effet plus de temps pour calculer les matrices de covariance des erreurs.

2.1.2 Concentrations en oxygène dissous simulées

Toutes les simulations avec PF estiment précisément les concentrations en oxygène dissous à toutes les stations, même avec une taille d'ensemble de 10 (colonne de gauche de la figure 4). Quelle que soit la taille de l'ensemble, la qualité des résultats du PF est très stable. Une RMSE maximale de 0.02 $\mathrm{mgO_2\ L^{-1}}$ est estimée à la station PK50, ce qui est relativement faible par rapport aux erreurs d'observation (écarts types entre 0.038 $\mathrm{mgO_2\ L^{-1}}$ et 0.08 $\mathrm{mgO_2\ L^{-1}}$). Tous les KGE sont supérieurs à 0.99 pour PF.

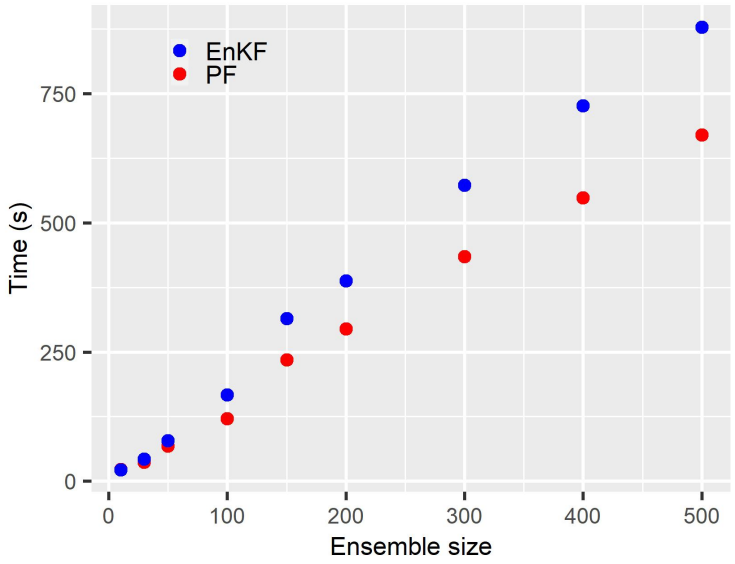

Figure 3. Temps de calcul pour PF et EnKF. Simulations sont réalisées avec 20 threads (Intel(R) Xeon(R) CPU E5-2640 v4 @ 2.40GHz). Un pas de temps de 15 min est utilisé pour une période de 15 jours.

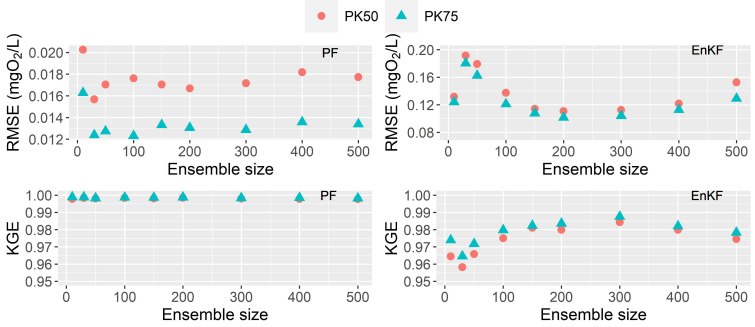

Figure 4. Critères statistiques (RMSE, en haut, et KGE, en bas) calculés aux stations de validations pour PF à gauche, et EnKF à droite

EnKF présente des résultats légèrement dégradés par rapport à ceux du PF (colonne de droite de la figure 4). En effet, la RMSE pour EnKF diminue d'abord et augmente ensuite avec la taille de l'ensemble. Sa valeur minimale (environ 0.10 $mgO_2\ L^{-1}$) est obtenue avec les tailles d'ensemble de 200 et 300, tout en étant relativement bien contenue dans le faisceau d'erreurs d'observation (écarts types entre 0.038 $mgO_2\ L^{-1}$ et 0.08 $mgO_2\ L^{-1}$). Pour EnKF, les valeurs maximales de KGE sont obtenues avec la taille d'ensemble de 300.

2.1.3 Distributions *a posteriori* de K_{navig}

Les concentrations en oxygène dissous simulées et les distributions *a posteriori* de K_{navig} montrent que PF et EnKF fonctionnent bien pour un système légèrement nonlinéaire (réaération), avec cependant un avantage pour le PF qui est plus précis (Fig. 5). Pour une taille d'ensemble de 300, tant PF que EnKF produisent des distributions postérieures satisfaisantes de K_{navig}. La valeur de référence de K_{navig} (0.015 m h^{-1}) est bien caractérisée par les modes des distributions. Cependant, les distributions estimées par PF sont beaucoup plus précises que celles estimées par EnKF, qui sont relativement larges (Fig. 5).

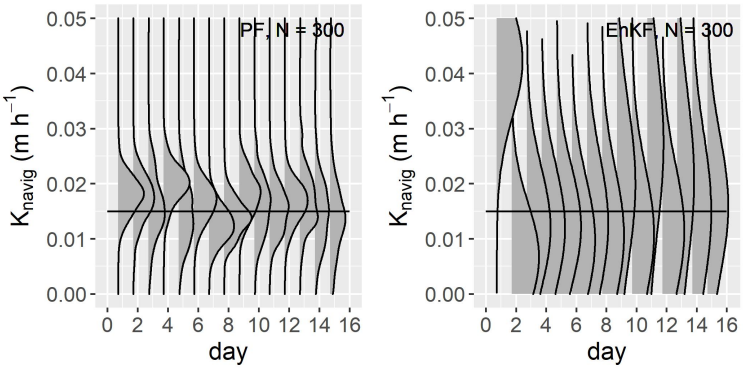

Figure 5. Distributions a posteriori de K_{navig} estimées par PF à gauche et EnKF à droite avec la taille d'ensemble de 300. La ligne noire représente la valeur de référence de K_{navig} (0.015 m h^{-1}) qui est utilisée pour générer les observations virtuelles.

2.2 Minéralisation de la matière organique et activités des bactéries hétérotrophes (système fortement non-linéaire)

Une série de simulations est effectuée avec différentes tailles d'ensemble (100, 300, 400 et 500). La taille d'ensemble de 500 donne des résultats satisfaisants pour les deux filtres PF et EnKF. Par conséquent, seuls les résultats avec la taille d'ensemble de 500 sont présentés dans cette section. Les résultats montrent que PF est beaucoup plus performant que EnKF lorsque les activités bactériennes contrôlent le métabolisme de la rivière (système fortement non-linéaire), tant pour l'identification des valeurs des paramètres que pour l'estimation des concentrations en oxygène dissous.

2.2.1 Concentrations en oxygène simulées

Les concentrations en oxygène simulées (moyenne d'ensemble) montrent que PF est plus performant que EnKF sous les conditions de fortes activités bactériennes (Fig. 6). En effet, toutes les RMSE sont inférieures à 0.02 mgO_2 L^{-1} pour PF alors que toutes les RMSE sont d'environ 0.2 mgO_2 L^{-1} pour EnKF (Tab. 3), ce qui, pour EnKF, est significativement plus élevé que les erreurs d'observation (environ 0.08 mgO_2 L^{-1}). Les KGEs montrent également une meilleure estimation pour PF que pour EnKF. Les résultats indiquent que EnKF peut capturer la tendance à la baisse de l'oxygène, mais qu'il a du mal à retrouver correctement les concentrations en oxygène dissous.

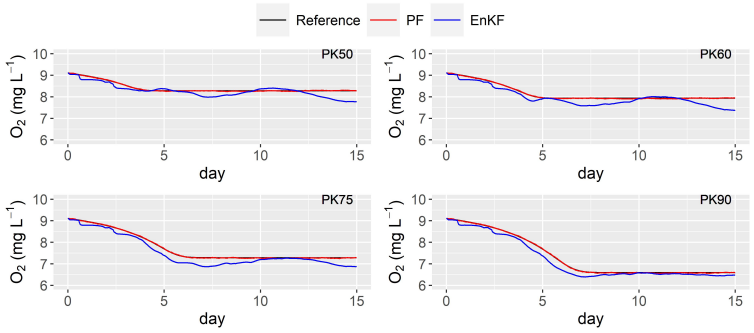

Figure 6. Concentrations en oxygène simulées par PF et EnKF avec la taille d'ensemble de 500. La courbe PF (rouge) se superpose à la courbe de référence (noire).

Tableau 3. Critères statistiques sur l'oxygène dissous aux stations de validation pour la minéralisation de la matière organique par les bactéries hétérotrophes

Criteria	PK50	PK75	Unit
RMSE (PF)	0.014	0.016	$mgO_2\ L^{-1}$
KGE (PF)	0.991	0.988	[-]
RMSE (EnKF)	0.202	0.241	$mgO_2\ L^{-1}$
KGE (EnKF)	0.715	0.885	[-]

2.2.2 Distributions a posteriori de μ_{max} et Y

Les distributions a posteriori estimées de μ_{max} et Y (Fig. 7) confirment l'efficacité de PF pour estimer les paramètres d'un modèle de qualité de l'eau qui simule souvent un système fortement non-linéaire. Ainsi, les valeurs des paramètres de μ_{max} et Y (0.04 h^{-1} et 0.015, respectivement) utilisées pour générer les observations sont très bien identifiées par PF tandis que EnKF ne parvient pas à identifier le rendement de croissance Y (les estimations atteignant le double de la valeur ciblée) et surestime légèrement la valeur de μ_{max} (Fig. 7).

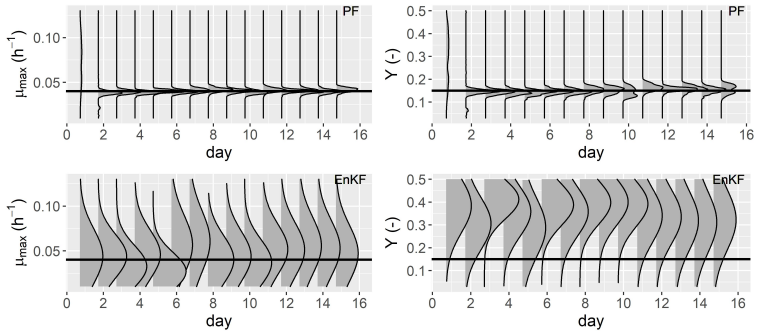

Figure 7. Distributions a posteriori journalières du taux de croissance bactérienne maximale, μ_{max} à gauche, et du rendement de croissance associé, Y à droite, estimées par PF, en haut, et EnKF, en bas, avec une taille d'ensemble de 500.

3 Recommandations pour l'assimilation de données appliquée à la simulation de la qualité de l'eau

Bien que le filtre de Kalman d'ensemble (EnKF) soit devenu la méthode d'assimilation de données la plus populaire dans la modélisation de la qualité des eaux de surface (Cho et al., 2020), les résultats de ce rapport montrent que le filtre particulaire (PF) est beaucoup plus performant, aussi bien pour l'estimation des paramètres du modèle que pour l'estimation de la concentration en oxygène dissous dans un système fortement non-linéaire (processus biogéochimiques complexes). Dans le cadre de la modélisation biogéochimique océanique, Gharamti et al. (2017) a également montré que l'incertitude associée aux estimations de l'état à l'aide de deux méthodes basées sur EnKF augmente pendant les efflorescences printanières (système fortement non-linéaire). Cela s'explique par le fait que l'EnKF suppose des erreurs gaussiennes pour les états et les paramètres de prévision, ce qui n'est pas réaliste pour le fonctionnement d'un système aquatique (lac ou rivière) (Huang et al., 2013; Wang et al., 2022). Ces résultats sont confirmés par Evensen et al. (2022, p. 95) qui mentionnent le fait que "*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.*" En revanche, le PF ne formule pas d'hypothèse sur la forme des distributions de probabilités, ce qui lui permet de dépasser ces

contraintes déjà mises en lumière et confirmées par notre étude. D'ailleurs, l'implémentation du PF dans le logiciel hydro-biogéochimique complexe ProSe-PA s'est avérée réaliste (Wang et al., 2019) et pertinente via l'évaluation de ses performances (coût de calcul, estimation des paramètres et des concentrations en oxygène dissous) à l'échelle de 200 km de Seine à la traversée de l'agglomération parisienne (Wang et al., 2022). Par conséquent, nous recommandons l'utilisation de PF pour l'assimilation de données appliquée à la simulation de la qualité des eaux de surface.

Conclusion

Afin de comparer les performances des filtres particulaire et de Kalman d'ensemble pour l'assimilation de données appliquée à la simulation de la qualité des eaux de surface, ils ont été implémentés numériquement dans le logiciel hydro-biogéochimique ProSe-PA. Les deux filtres ont été évalués via deux cas d'étude spécifiques, représentant une complexité et des processus distincts allant d'un système quasi-linéaire pour le cas de la réaération à l'interface air-eau à un système très fortement non-linéaire pour la dégradation de la matière organique par les bactéries hétérotrophes. Les principales conclusions se résument ainsi :

- Le filtre particulaire est recommandé pour l'assimilation de données appliquée à la simulation de la qualité des eaux de surface ;
- Pour un système fortement non-linéaire (activités des bactéries hétérotrophes), le PF est plus efficace à la fois pour la simulation des concentrations en oxygène dissous et pour l'estimation des paramètres ; ces estimations entachées d'erreurs significatives avec EnKF ;
- Pour un système légèrement non-linéaire (réaération), les deux filtres capturent l'évolution de l'oxygène dissous et identifient le coefficient de réaération, avec toutefois une meilleure précision pour le filtre particulaire ;
- L'algorithme du filtre particulaire est plus véloce, avec un temps de calcul, pour les échantillons de plus de 100 particules, 25% plus court par rapport au EnKF.

Ce rapport est un extrait traduit d'un article en cours de relecture dans la revue internationale *Journal of Hydrology*, soumis par les auteurs du présent rapport.

Bibliographie

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

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

Bae, S. and D. Seo (2018). Analysis and modeling of algal blooms in the nakdong river, korea. *Ecological Modelling 372*, 53–63.

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. and P. Young (1976, 10). Systematic identification of do-bod model structure. *Jnl. of Env. Eng. Division , American Society of Civil Engineers 102*, 909–927.

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

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

Carrassi, A., M. Bocquet, L. Bertino, and G. Evensen (2018). Data assimilation in the geosciences : An overview of methods, issues, and perspectives. *Wiley Interdisciplinary Reviews : Climate Change 9*(5), e535.

Chen, C., J. Huang, Q. Chen, J. Zhang, Z. Li, and Y. Lin (2019). Assimilating multi-source data into a three-dimensional hydro-ecological dynamics model using ensemble kalman filter. *Environmental Modelling & Software 117*, 188 – 199.

Cho, K. H., Y. Pachepsky, M. Ligaray, Y. Kwon, and K. H. Kim (2020). Data assimilation in surface water quality modeling : A review. *Water Research 186*, 116307.

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

Doucet, A. and A. M. Johansen (2011). A tutorial on particle filtering and smoothing : fiteen years later. In D. Crisan and B. Rozovskii (Eds.), *The Oxford handbook of nonlinear filtering*, Oxford handbooks in mathematics, pp. 656–705. Oxford ; N.Y. : Oxford University Press. ISBN 9780199532902.

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

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

Even, S., M. Poulin, J. M. Mouchel, M. Seidl, and P. Servais (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. (2003, Nov). The Ensemble Kalman Filter : theoretical formulation and practical implementation. *Ocean Dynamics 53*(4), 343–367.

Evensen, G. ., F. C. Vossepoel, and P. Jan van Leeuwen (2022). *Data Assimilation Fundamentals - A Unified Formulation of the State and Parameter Estimation Problem*. Springer.

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

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

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

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

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.

Hasanyar, M., T. Romary, S. Wang, and N. Flipo (2022). How much do bacterial growth properties and biodegradable dissolved organic matter control water quality at low flow ? *Biogeosciences Discussions 2022*, 1–33.

Huang, J. and J. Gao (2017). An improved ensemble kalman filter for optimizing parameters in a coupled phosphorus model for lowland polders in lake taihu basin, china. *Ecological Modelling 357*, 14 – 22.

Huang, J., J. Gao, J. Liu, and Y. Zhang (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.

Kalnay, E., M. Kanamitsu, R. Kistler, W. Collins, D. Deaven, L. Gandin, M. Iredell, S. Saha, G. White, J. Woollen, Y. Zhu, M. Chelliah, W. Ebisuzaki, W. Higgins, J. Janowiak, K. C. Mo, C. Ropelewski, J. Wang, A. Leetmaa, R. Reynolds, R. Jenne, and D. Joseph (1996). The ncep/ncar 40-year reanalysis project. *Bulletin of the American Meteorological Society 77*(3), 437–472.

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

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

Loos, S., C. M. Shin, J. Sumihar, K. Kim, J. Cho, and A. H. Weerts (2020). Ensemble data assimilation methods for improving river water quality forecasting accuracy. *Water Research 171*, 115343.

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

Marescaux, A., V. Thieu, N. Gypens, M. Silvestre, and J. Garnier (2020). Modeling inorganic carbon dynamics in the seine river continuum in france. *Hydrology and Earth System Sciences 24*(5), 2379–2398.

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.

Monod, J. (1949). The growth of bacterial cultures. *Annual Review of Microbiology 3*(1), 371–394.

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

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

Park, S., K. Kim, C. Shin, J.-H. Min, E. H. Na, and L. J. Park (2020). Variable update strategy to improve water quality forecast accuracy in multivariate data assimilation using the ensemble kalman filter. *Water Research 176*, 115711.

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

Pelletier, G., S. Chapra, and H. Tao (2006, 03). QUAL2Kw — a framework for modeling water quality in streams and rivers using a genetic algorithm for calibration. *Environmental Modelling and Software*, 419–425.

Piazzi, G., G. Thirel, C. Perrin, and O. Delaigue (2021). Sequential data assimilation for streamflow forecasting : Assessing the sensitivity to uncertainties and updated variables of a conceptual hydrological model at basin scale. *Water Resources Research 57*(4).

Plaza, D. A., R. De Keyser, G. J. M. De Lannoy, L. Giustarini, P. Matgen, and V. R. N. Pauwels (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., N. Flipo, C. de Fouquet, and M. Poulin (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., L. Vilmin, N. Flipo, V. Rocher, and A. Laverman (2015). Modelling the fate of nitrite in an urbanized river using experimentally obtained nitrifier growth parameters. *Water Research 73*, 373–387.

Sadeghian, A., S. C. Chapra, J. Hudson, H. Wheater, and K.-E. Lindenschmidt (2018). Improving in-lake water quality modeling using variable chlorophyll a/algal biomass ratios. *Environmental Modelling & Software 101*, 73–85.

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

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

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

Wang, S. (2019, Septembre). *Simulation du métabolisme de la Seine par assimilation de données en continu*. Theses, Mines Paris – PSL.

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

Wang, S., N. Flipo, and T. Romary (2019). Oxygen data assimilation for estimating micro-organism communities' parameters in river systems. *Water Research 165*, 115021.

Wang, S., N. Flipo, T. Romary, and M. Hasanyar (2022). Particle filter for high frequency oxygen data assimilation in river systems. *Environmental Modelling & Software*, 105382.

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

Whitehead, P. G., R. J. Williams, and D. R. Lewis (1997). Quality simulation along river systems (quasar) : model theory and development. *Sciences of Total Environment 194/195*, 447–456.

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