Ensemble Data Assimilation for Meshless Methods

Marius Duvillard[1,2], Loïc Giraldi[1], and Olivier Le Maître[3]

[1]CEA, DES, IRESNE, DEC, SESC, LMCP, Cadarache, F-13108 Saint-Paul-Lez-Durance, France
[3]CNRS, Inria, Centre de Mathématiques Appliquées, Ecole Polytechnique, IPP, Route de Saclay, 91128, Palaiseau Cedex, France
[2]Centre de Mathématiques Appliquées, Ecole Polytechnique, IPP, Route de Saclay, 91128, Palaiseau Cedex, France

Abstract

An ~~adapted~~ method is introduced to apply data assimilation methods for meshless ~~or particle-based~~ simulation with the ensemble Kalman filter (EnKF). EnKF, with meshless methods, allows an efficient estimation for numerous kinds of problems based on a Lagrangian discretization of a solution. Two specific methodologies have been introduced to either complete the analysis on a new grid of particles or the forecast particle discretization of each member. The two variants, ~~called Remesh-EnKF and Particles-EnKF,~~ are tested first on the one-dimensional ~~Gaussian~~ advection-diffusion problem with periodic boundary and then applied to a two-dimensional inviscid flow problem solved with the Vortex-In-Cell method (VIC). In the 1D case, the filters were compared to an Eulerian filter and exhibited similar results to a grid-based assimilation method while highlighting the limitations of the Particles-EnKF. In the second case, the applicability of these methods to more complex scenarios ~~was demonstrated, allowing for a more qualitative assessment.~~

Keywords: meshless methods, Particle-based method, data assimilation, EnKF, Ensemble Methods, Vortex Methods.

Contents

1 Introduction

Numerical simulation enables ~~predictions~~ of complex real-world systems, for instance, to facilitate the optimization of complex systems and perform risk analysis, all while reducing experimental costs. Thanks to the increasing computational resources, they are a tool to understand and design processes, particularly in the mechanical field.

The conventional approach to numerical analysis has historically leaned ~~on~~ finite element, finite volume, and finite difference methods. These techniques necessitate the use of structured meshes. The shift towards meshless methods offers significant promises for complex physics or large deformations (moving interfaces, material disintegration, or distortion) or avoids computing complex meshes.

Meshless methods, specifically particle-based methods, describe geometry as a collection of particles that move with the deformation flow in a Lagrangian fashion. Each particle transports material properties and historical variables.

On one hand, particles represent a discrete medium. Particles are individual entities with kinematic properties that interact locally and balance multi-body equilibrium. The Discrete Element Method (DEM), first introduced by Condall and Strack [9], has gained much popularity in modeling granular materials. On the other hand, particles discretize a continuum medium and are associated with shape functions to reconstruct continuous fields and differential operators. The Smoothed Particle Hydrodynamic (SPH), independently introduced by Gingold, Monaghan, and Lucy [14, 18], is one of the first continuum particle-based methods. It associates a kernel to each particle to approximate the continuous fields and the derivative operator to solve the strong form of the equilibrium equation. It has first been applied to stellar models but also to fluid dynamics. ~~n.~~ The Material Point Method (MPM) introduced by Sulsky [24] is another particle-based that has been part of the Particle-In-Cell family, like Fluid Implicit Particle (FLIP) introduced an auxiliary grid

to project, solve, and interpolate back the solution on the particles.

Like any numerical method, the solution involves errors that must be understood, quantified, and reduced. Uncertainty is a fundamental aspect of scientific inquiry and modeling. It often arises when our knowledge is limited or incomplete. This uncertainty can manifest in various forms, such as the ambiguity surrounding the value of a model parameter, the vagueness regarding initial conditions, or the uncertainty in setting boundary conditions or external forces. Moreover, if numerical models usually bring essential physical principles, they involve some simplifications. The numerical error appears due to the algorithm and discretization. Besides, it extends to the uncertainty associated with forthcoming experimental measurements calibrating numerical models.

Data assimilation is a method used to reduce uncertainty. It involves combining different sources of information to obtain a better estimate of the system's state, thereby reducing bias and uncertainties. It is commonly used in meteorology, oceanography, hydrology, and geosciences [4]. The sources of information typically come from observations, such as acquisition data and a numerical model. The combination is carried out dynamically, making the data assimilation problem an inverse problem over time. The dynamic inference of the state can be done using either a sequential or a variational formulation. In the variational formulation, the data assimilation problem minimizes a loss function that combines the observation and model information. The most commonly used formulations are 3D-VAR, 4D-VAR, and incremental 4D-VAR. On the other hand, sequential approaches are based on recursive formulations of probability laws to estimate, through a Bayesian framework, the state at each new measurement. In this approach, a filter is used to estimate the state based on previous and current observations. Based on Bayes' and marginalization rules, the assimilation process is performed through a forecast and analysis step. These two forms are mathematically equivalent in the sense that they aim to solve the same underlying problem, which is to estimate the best possible state of a system given observational data and model equations [13]. The Kalman filter [16] is an example of a sequential formulation considering a linear model and Gaussian distribution assumptions. However, more advanced filters have been introduced to be adapted to nonlinear and arbitrary distributions. One of the most popular Bayesian filters is undoubtedly the Ensemble Kalman Filter introduced by Evensen [12] mainly due to its adaptability to high dimensional problems with any evolution model. It consists, like the Particle Filter (PF), to approximate the probability distribution of a state thanks to an ensemble of simulations called particles or members. Members are independently forecasted with the numerical model. Then, by approximating covariance matrices, members are updated in a Kalman fashion.

Although EnKF has been extensively used in Eulerian discretization, Lagrangian method implementation becomes challenging due to the presence of particles involved in the discretization process. The main challenges consist of defining properly a common state for all the members and updating it during the analysis step. In the general context of the

Discrete Element Method (DEM), Data Assimilation poses challenges in the update phase. For the classic soft-sphere approach introduced by [9], the interaction is dependent on the geometry of the particle, and moving one particle to another leads to a complex high-dimensional nonlinear optimization problem. In [6], an EnKF algorithm has been applied to a DEM simulation to study the Sea Ice flow. Over, some simplifications have been introduced, like a new parametrization to reduce the number of particles, and changing particle positions have mild stability implications. Moreover, it is a case where all members share the same number of particles. In the particle-based methods for continuum discretization, the particles are punctual entities, meaning that resampling or moving particles do not lead to interpenetration issues. Agglomerating, splitting, or resampling are widespread tasks used to update the particle configuration [25, 8] mainly to reduce distortion, high deformation component, or lower numbers of particles. However, the discretization is still different between every member. The first solution is, like in the previous example, to consider a reference discretization for all members. Darakananda proposed in [10] this formulation for the Vortex Method. The simulation's state comprises each particle's positions and associated intensities. Complex filters have been developed to estimate correctly the posterior discretization based on a nonlinear observation model or a deficient number of pressure sensors [17]. Those methods account hardly for the case of very different particle discretization or where the model flow is highly divergent. The case of state defined on different grids with assimilation has already been addressed in fixed-grid methods with multiple resolutions (MRA) [22] and in the case of moving meshes [5]. In these scenarios, the approach involves performing projection and interpolation to establish the state on a reference grid for computing the updated state. The choice of the reference grid and the updated grid offers a variety of implementations. Moreover, [22] shows that the correction of the EnKF filter is only dependent on the predictions and observation and, consequently, is independent of the state definition.

The goal of this paper is to consider general methods to apply ensemble Data Assimilation filtering to particle-based simulation that discretize a continuum domain. The hypothesis considers several members of different particle distributions. In this general case, using a particle state using all particles for the update is unfeasible. Indeed, the update implies a linear combination of all members, leading to an exponential increase of particles. On the other hand, the state could be associated with the spatial field defined in a functional space. The updated fields could be evaluated on the entire domain. Finally, using approximation or regression, a new particle discretization could be approximated. These types of methods have already been introduced in the Vortex Method to better approximate the vorticity field by changing the particle intensities. Regroup under the label Meshless Rezoning Methods in [19]. It mainly involved iterative methods [3], triangulation [21] or Radial Basis Function (RBF) interpolation [1, 23]. The last ones offer to easily introduce new particles or introduce penalization to regularize optimization problems.

Based on those different approaches, we propose two types of adaptation of the EnKF filter. First, the Remesh-EnKF uses a new reference particle discretization. This way, the

state could be updated, and the number of particles is controlled. This first method is based on the regridding of the particle discretization as described by [8] on which the classical EnKF analysis could be performed. Then, in a case where the particle discretization would be preserved, the Particle-EnKF is introduced. In this case, the analyzed field is approximated with the previous particle discretization. The particle's positions are unchanged; only the strengths are modified by regression. In the next part, background on sequential filtering and EnKF algorithm will be introduced 2.1, then on particle-based methods 2.2. Then, the two categories of method will be described in section 3. Afterward, those filters will be compared with a grid-based filter in a 1D Advection-Diffusion problem in section **??**, and an incompressible viscous flow is solved using a Vortex Method **??** where the filters are quantitatively analyses.

2 Background

In this introductory section, we will delve into two key aspects. Firstly, in section 2.1, we will explore sequential data assimilation methods, focusing on the Ensemble Kalman Filter and the intricacies of calibration. Subsequently, in section 2.2, our attention will shift to particle methods for addressing problems in continuous media. We will not only introduce the concept of particle discretization but also elucidate a range of techniques essential for filter development in the ~~method~~ section 3.

2.1 Data assimilation

One type of data assimilation is called *statistic* Data Assimilation. This family of methods introduced a probabilistic framework in order to rigorously deal with measurement and model error in order to not only deduce an estimate of the real state but also associate uncertainty. Thus, state and observation are modeled as random variables. A filtering approach is then applied to estimate the current state based on past observations sequentially.

The goal is to establish the recurrence in probability distributions that, through Bayesian estimation, will enable us to estimate the current state and predict the future state, including future observations.

2.1.1 Data assimilation setting

This recurrence is simplified by the use of a hidden Markov chain model. The forecast and observation are introduced such as $\forall k \geq 1$

$$\begin{cases} \boldsymbol{x}^{k+1} = \mathcal{M}^k(\boldsymbol{x}^k) + \boldsymbol{\eta}^k, \\ \boldsymbol{y}^k = \mathcal{H}^k(\boldsymbol{x}^k) + \boldsymbol{\varepsilon}^k, \end{cases}$$

where $\mathcal{M}^k$ is the model operator describing the time evolution of the state from time k to time $k+1$ and $\mathcal{H}^k$ is the observation operator. The term $\boldsymbol{x}^k \in \mathbb{R}^n$ is the vector state at time k and $\boldsymbol{y}^k \in \mathbb{R}^m$ the observation vector, $\boldsymbol{\eta}^k$ is the model error that accounts for error in the numerical model and the errors due to discretization, and $\boldsymbol{\varepsilon}^k$ is the observation error which combine measurement error and representativeness error. We assume that $\boldsymbol{\eta}^k$, $\boldsymbol{\varepsilon}^k$ are random variables following Gaussian distributions with zero mean and covariance matrices $\boldsymbol{Q}^k$ and $\boldsymbol{R}^k$ respectively. Finally, we assume that the observation and the model errors are independent though the time: $\langle \boldsymbol{\eta}^k, \boldsymbol{\eta}^j \rangle = \langle \boldsymbol{\varepsilon}^k, \boldsymbol{\varepsilon}^j \rangle = 0$ if $k \neq j$, and that initial error on $\boldsymbol{x}^0$, $\boldsymbol{\varepsilon}^k$ and $\boldsymbol{\eta}^k$ are mutually independent. These assumptions simplify the recurrence of probability distributions by observing that $p_{\boldsymbol{x}^{k+1}|\boldsymbol{x}^1,\dots\boldsymbol{x}^k-1} = p_{\boldsymbol{x}^{k+1}|,x^k-1}$ and $p_{\boldsymbol{y}^1,\dots,\boldsymbol{y}^n|\boldsymbol{x}^1,\dots,\boldsymbol{x}^k} = \prod_{i=1}^{k} p_{\boldsymbol{y}^i|\boldsymbol{x}^i}$ We remove the time subscript k in the rest of the section for simplicity and present the *forecast* and *analysis* step for one time increment.

2.1.2 Ensemble Kalman Filter

The Kalman filter [16] is the Bayesian filter that use, moreover the previous hypothesis, that $\mathcal{M}^k$ and $\mathcal{H}^k$ are linear operators. In this case, the posterior distribution of the state is still Gaussian, so only the mean and the variance are transmit. The Kalman estimator is thus a recursive version of the Minimum Mean Square Error applied to the Gaussian Linear model.

The ensemble Kalman Filter (EnKF) is a data assimilation method adapted to high dimensional non-linear problems introduced by Evensen [12]. The formulation uses an ensemble of discrete samples based on the assumptions of a multivariate Gaussian distribution, as for the Kalman filter. EnKF can be seen as a hybrid method between the Kalman filter and the particle filter. The forecast is performed in the same way as in the particle filter. Still, the analysis step is computed in Kalman's fashion with sample model error covariance. We present the stochastic EnKF, where the observations are perturbed to account for observation errors and to introduce stochasticity into the assimilation process, allowing for a more realistic representation of uncertainties and avoiding filter divergence issues.

Assuming we have an ensemble of N states, we could forecast the ensemble by propagating each state with the dynamic model and obtain a forecast ensemble. The two first moments of the error are given by

$$\begin{aligned} \overline{\boldsymbol{x}}_f &= \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x}_i^f, \\ \boldsymbol{C} &= \frac{1}{N-1}\sum_{i=1}^{N_e} (\boldsymbol{x}_i^f - \overline{\boldsymbol{x}}^f)(\boldsymbol{x}_i^f - \overline{\boldsymbol{x}}^f)^T, \end{aligned}$$

where $\overline{\boldsymbol{x}}_f$ and $\boldsymbol{C}$ are the empirical estimates of the mean and covariance matrix of the state distribution obtained from the ensemble members.

We define the matrix of states and the matrix of anomalies $\boldsymbol{X}_f$, $\bar{\boldsymbol{X}}_f$ whose columns are the member states and the normalized anomalies.

$$[\bar{\boldsymbol{X}}_f]_i = \frac{\boldsymbol{x}_i^f - \overline{\boldsymbol{x}}^f}{\sqrt{N-1}}, \quad \bar{\boldsymbol{X}}_f = \boldsymbol{X}_f \boldsymbol{\Pi},$$

where $\boldsymbol{\Pi} = (\boldsymbol{I} - \frac{1}{N}\mathbf{1}\mathbf{1}^T)/\sqrt{N-1}$ with $\boldsymbol{I}$ is the identity matrix and $\mathbf{1} \in \mathbb{R}^N$ is a vector of one.

Respectively the matrix of observation and observation anomalies are $\boldsymbol{Y}_f$, $\bar{\boldsymbol{Y}}_f$ where columns are

$$[\bar{\boldsymbol{Y}}_f]_i = \frac{\mathcal{H}(\boldsymbol{x}_i^f) - \overline{\boldsymbol{y}}^f}{\sqrt{N-1}} = \boldsymbol{Y}\boldsymbol{\Pi} \quad \text{with} \quad \overline{\boldsymbol{y}}^f = \frac{1}{N}\sum_{j=1}^{N} \mathcal{H}(\boldsymbol{x}_j^f).$$

The ensemble defines the covariance between states and observations $\boldsymbol{C}\boldsymbol{H}^T$, the covariance between observations $\boldsymbol{C}\boldsymbol{H}^T$, and $\tilde{\boldsymbol{K}}$

$$\begin{aligned}
\boldsymbol{C}\boldsymbol{H}^T &= \frac{1}{N-1}\sum_{i=1}^{N} (\boldsymbol{x}_i^f - \overline{\boldsymbol{x}}^f)^T \left[\mathcal{H}^k(\boldsymbol{x}_i^f) - \overline{\boldsymbol{y}}^f\right]^T = \bar{\boldsymbol{X}}_f \bar{\boldsymbol{Y}}_f^T, \\
\boldsymbol{H}\boldsymbol{C}\boldsymbol{H}^T &= \frac{1}{N-1}\sum_{i=1}^{N} \left[\mathcal{H}^k(\boldsymbol{x}_i^f) - \overline{\boldsymbol{y}}^f\right] \left[\mathcal{H}^k(\boldsymbol{x}_i^f) - \overline{\boldsymbol{y}}^f\right]^T = \bar{\boldsymbol{Y}}_f \bar{\boldsymbol{Y}}_f^T, \\
\tilde{\boldsymbol{K}} &= \boldsymbol{C}\boldsymbol{H}^T(\boldsymbol{H}\boldsymbol{C}\boldsymbol{H}^T + \boldsymbol{R})^{-1} = \bar{\boldsymbol{X}}_f \bar{\boldsymbol{Y}}_f^T (\bar{\boldsymbol{Y}}_f \bar{\boldsymbol{Y}}_f^T + \boldsymbol{R})^{-1}.
\end{aligned}$$

This observation matrix-free implementation rely on the *secant method* approximation $\mathcal{H}(\boldsymbol{x}_i^f - \overline{\boldsymbol{x}}^f) \approx \mathcal{H}(\boldsymbol{x}_i^f) - \overline{\boldsymbol{y}}^f$. The forecast is then update to a posterior ensemble $[\boldsymbol{x}_i^a]_{i=1}^N$ such as

$$\boldsymbol{X}^a = \boldsymbol{X}^f + \tilde{\boldsymbol{K}}(\boldsymbol{D} - \boldsymbol{Y}), \tag{1}$$

where $[\boldsymbol{D}]_i = \boldsymbol{y} + \boldsymbol{\varepsilon}_i$ is the perturbed observation with $\boldsymbol{\varepsilon}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{R})$, $\tilde{\boldsymbol{K}}$ the ensemble Kalman gain matrix and $(\boldsymbol{D} - \boldsymbol{Y})$ the *innovation* term. The forecast step is then applied to the analyzed ensemble until the next observation. Based on this formulation, we can deduce a correction formula only based on the member's predictions and observations.

We can rewrite the classical update formula using the previous anomaly matrices.

$$\boldsymbol{X}_a = \boldsymbol{X}_f + \bar{\boldsymbol{X}}_f \bar{\boldsymbol{Y}}_f^T (\bar{\boldsymbol{Y}}_f \bar{\boldsymbol{Y}}_f^T + \boldsymbol{R})^{-1}(\boldsymbol{D} - \boldsymbol{Y}) = \boldsymbol{X}_a.$$

We reformulate the correction term by remarking that $\frac{1}{N}\mathbf{1}\mathbf{1}^T\bar{\boldsymbol{Y}}_f^T = \mathbf{0}$. We define $\boldsymbol{F}$, the correction matrix that gives the update in terms of linear combinations of the forward states

$$\boldsymbol{X}_a = \boldsymbol{X}_f + \boldsymbol{X}_f \boldsymbol{F}, \quad \boldsymbol{F} = \bar{\boldsymbol{Y}}_f^T (\bar{\boldsymbol{Y}}_f \bar{\boldsymbol{Y}}_f^T + \boldsymbol{R})^{-1}(\boldsymbol{D} - \boldsymbol{Y}). \tag{2}$$

where the matrix $\boldsymbol{F}$ only depends on the ensemble members through the predicted observations ensemble $\boldsymbol{Y}_f$ and the observation.

Consequently, the analysis could be entirely defined thanks to the ensemble predictions and the observation. Section 10.2 of the book [13] introduced various forms of the EnKF update where equation 10.2 is equivalent to 2.

2.1.3 Expanded State Model

The Bayesian calibration of model parameters is possible by defining an *expanded state*. Suppose we parametrize the dynamic operator $\mathcal{M}(\boldsymbol{x};\boldsymbol{\theta})$. The Parameter vector $\boldsymbol{\theta}$ is then appended to the model state vector, and the model state is forecast with the parameter. The calibration is thus performed online, allowing a more accurate forecast prediction and saving time on calibration. The parameter is supposed to be constant in this article, meaning that the evolution model is simply identity, such as

$$\boldsymbol{\theta}^{k+1} = \boldsymbol{\theta}^k.$$

The expanded state system model is then

$$\begin{cases} \hat{\boldsymbol{x}}^{k+1} & = \hat{\mathcal{M}}^k(\hat{\boldsymbol{x}}^k) + \hat{\boldsymbol{\eta}}^k \quad , \\ \boldsymbol{y}^k & = \hat{\mathcal{H}}(\hat{\boldsymbol{x}}^k) + \boldsymbol{\varepsilon}^k \quad , \end{cases}$$

where

$$\hat{\boldsymbol{x}}^k = \begin{pmatrix} \boldsymbol{x}^k \\ \boldsymbol{\theta}^k \end{pmatrix} \in \mathbf{R}^{n+q},$$

is the expanded state vector, with q the dimension of the parameter vector, and

$$\hat{\mathcal{M}}^k = \begin{pmatrix} \mathcal{M}^k(\boldsymbol{x}^k;\boldsymbol{\theta}^k) \\ \boldsymbol{\theta}^k \end{pmatrix}$$

Finally, the observation equation still gives the same result, such as

$$\hat{\mathcal{H}}^k(\hat{\boldsymbol{x}}^k) = \mathcal{H}^k(\boldsymbol{x}^k).$$

The analysis step is unchanged but applies to $\hat{\boldsymbol{x}}^k$.

2.2 Particle-based Methods

One class of particle-based method are dedicate to solve continuous problems in fluid or solid mechanics. ~~Those methods differ from mesh methods in the way to discretize the solution. The mesh methods are Eulerian and decompose the domain on a mesh that supports a set of functions generally defined on each node.~~ The Lagrangian methods decompose the domain on a set $\mathcal{P}$ of particles that follow the dynamic of the problem. Each particle of the set '$\mathcal{P}$ brings ~~vectorial~~ quantities $\boldsymbol{Q}_p \in \mathbb{R}^{n_q}$ and the spatial coordinates $\boldsymbol{z}_p$.

~~The velocity field $\boldsymbol{v}$ is used to update the position of the particle such as $\boldsymbol{z}_p(t+dt) = \boldsymbol{z}_p(t) + f(z_p, \boldsymbol{v}\boldsymbol{z}_p)$ with f depending on the time-integration scheme.~~

~~The computation of the velocity field and the solving of the equation of mechanics depend on the class of method.~~ We will focus our work on methods that discretize a solution on a continuous domain that can be defined with field $\boldsymbol{u} : \Omega \in \mathbb{R}^d \to \mathbb{R}^n$ with Ω the spatial domain, d is the space dimension and n the dimension of the solution. This includes methods like Smoothed particle hydrodynamics (SPH) [14, 18] and the Vortex Method (VM) [7] and is extended to other methods like the Material Point Method (MPM) [24].

2.2.1 Particle discretization

Let $\Omega \in \mathbb{R}^d$ be our domain, where d is the space dimension. Any smooth field $\boldsymbol{u}$ on Ω could be written thanks to the *shifting property*

$$\boldsymbol{u}(\boldsymbol{z}) = \int_\Omega \boldsymbol{u}(\boldsymbol{z}')\delta(\boldsymbol{z}' - \boldsymbol{z})d\boldsymbol{z}',$$

with δ the Dirac delta distribution.

A kernel function ϕ_ε, where ε is the smoothing length, is introduced to obtain, by substitution, an average estimate $\langle \boldsymbol{u} \rangle$ of $\boldsymbol{u}$ such that

$$\langle \boldsymbol{u}(\boldsymbol{z}) \rangle = \int_\Omega \boldsymbol{u}(\boldsymbol{z}')\phi_\varepsilon(\boldsymbol{z} - \boldsymbol{z}')d\boldsymbol{z}.$$

The smooth kernel should at least respect the following properties

$$\begin{aligned} &\int_\Omega \phi_\varepsilon(\boldsymbol{z})d\boldsymbol{z} = 1, \\ &\phi_\varepsilon(\boldsymbol{z}) \to \delta(z), \quad \varepsilon \to 0, \\ &\phi_\varepsilon(\boldsymbol{z}) \in C_0^k, \quad k \geq 1, \end{aligned}$$

where the two first properties are remanent properties of the Dirac distribution and the last is a differentiability requirement.

The average function $\langle \boldsymbol{u} \rangle$ is then used to substitute the original function.

Finally, the original domain Ω is subdivided with N_p subdomain Ω_p associated with a lagrangian particle in the location $z_p \in \Omega_p$. We denote by v_p the volume of Ω_p. This discretization and ~~the mean value theorem~~ are then used to approximate the average function such that

$$\begin{aligned}\langle \boldsymbol{u}(\boldsymbol{z})\rangle &= \sum_p \int_{\Omega_p} \boldsymbol{u}(\boldsymbol{z}')\phi_\varepsilon(\boldsymbol{z}-\boldsymbol{z}')d\boldsymbol{z}' \\ &\approx \sum_p \boldsymbol{u}(\boldsymbol{z}_p)v_p\phi_\varepsilon(\boldsymbol{z}-\boldsymbol{z}_p) \\ &\approx \sum_p \boldsymbol{U}_p\phi_\varepsilon(\boldsymbol{z}-\boldsymbol{z}_p).\end{aligned}$$

Thus, any function defined on a particle discretization is defined by an ensemble of particle location $\boldsymbol{z}_p$ associated with a particle value $\boldsymbol{U}_p = \boldsymbol{u}(z_p)v_p$ and a smooth kernel ϕ_ε. ~~For simplicity, we will denote $\phi_\varepsilon(\boldsymbol{z}-\boldsymbol{z}_i)$ by $\varphi_i(\boldsymbol{z})$ and $\phi_\varepsilon(\boldsymbol{z}_i-\boldsymbol{z}_j)$ by $\varphi_{i,j}$, for all $\boldsymbol{z}_i, \boldsymbol{z}_j \in \Omega$.~~

Based on this discretization, the differential operator could be derived through this formulation.

2.2.2 Exemple of kernel functions

Several kernels have been used depending on the method. The original formulation of MPM did not use a substitute kernel and wrote the density such as

$$\boldsymbol{u}(\boldsymbol{z}) = \sum_p \boldsymbol{F}_p\phi_\varepsilon(\boldsymbol{z}-\boldsymbol{z}_p)$$

And the resolution is based on a projection on a background grid associated with some shape function [24].

The GIMP method is a different formulation that uses the Heaviside function [2] and thus associates a volume around each particle

$$M_1(r) = \begin{cases} 1; & r \leq 1 \\ 0; & \text{otherwise} \end{cases}$$

where $r = \|\boldsymbol{z}\|_2$.

This method have been introduced to avoid the cell-crossing issue when a particle moves from one cell to another through the background grid.

In SPH associate, as this name suggests, a smooth kernel to approximate the solution. Theoretically, it could be the Gaussian kernel function

$$\phi_g(r) = \frac{1}{(\pi h^2)^{d/2}}\exp(-r^2/h^2)$$

where h is called the smoothing length.

This kernel is infinitely differentiable but defined on non-compact support. In practice, we use a cut-off to remove negligible value for large distance from a particle.

Other kernels, based on B-Spline functions to work on a compact support. Those functions are also positive which is a requirement for some field like the density.

For instance, the quadratic B-spline, we called M_3 defined with

$$M_3(r) = \frac{\alpha}{h^d} \begin{cases} \frac{3}{4} - |q|^2 & 0 \leq |q| < \frac{1}{2} \\ \frac{1}{2}\left(\frac{3}{2} - |q|\right)^2 & \frac{1}{2} \leq |q| < \frac{3}{2} \\ 0 & \frac{3}{2} \leq |q| \end{cases} \tag{3}$$

with $r = \|z\|_2$ and $q = r/h$ and α the normalization condition and d the spatial dimension.

This kernel ensures the C^1 continuity. The cubic kernel is another B-Spline kernel which is

$$M_4(r) = \frac{\alpha}{h^d} \begin{cases} \frac{1}{6}(-|q|+2)^3 - \frac{4}{6}(-|q|+1)^3 & 0 \leq |q| \leq 1 \\ \frac{1}{6}(-|q|+2)^3 & 1 \leq |q| \leq 2 \\ 0 & 2 \leq |q| \end{cases} \tag{4}$$

In this last case, the normalization factor α is

$$\alpha = \begin{cases} 1; & 1\text{d} \\ 30/14\pi; & 2\text{d} \\ 3/2\pi; & 3\text{d} \end{cases}$$

Note that, those kernel have been define with the radial coordinate r. Another possibility would be to define the multidimensional kernel as the tensor product of the 1d kernel on the normalized coordinate. This is what is used for the following application of the regridding operation 2.3.3 or the transfer define between particle and grid in the MPM scheme [24, 15].

For the periodic boundary problem described in section **??**, we define an equivalent kernel function $\phi^P = \sum_{n=-\infty}^{+\infty} \phi(x - nL, h)$,
where L is the periodic length. All the kernel properties are still verified on a single period.

2.3 Specific operation on the particle discretization

One of the main drawbacks of Lagrangian methods is that they can produce a highly distorted distribution of particles. To conserve the overlapping of particle's shape function, and allow communication between particles, two main techniques have been introduced to reconstruct a new particle discretization, or to adapt the particle's strengths to better fit the current field. The different types of the following operators will be introduced in the assimilation process in order to update particle solution of each members.

2.3.1 Approximator operator

The first category of solutions aims to improve the approximation of the vorticity field by modifying particle strength. For any particle location, $\boldsymbol{z}_p$, the continuous field could be evaluated $\boldsymbol{u}(\boldsymbol{z}_p)$.

A first approximation could be to use the particle approximation to reevaluate the particle intensities based on the Mean Value approximation present in 3 such as

$$\boldsymbol{U}_p = \int_{\Omega_p} \boldsymbol{u}(z) d\boldsymbol{z} = \boldsymbol{u}(\boldsymbol{z}_p) V_p.$$

This approximation is easily computable but do not ensure the conservation of the all the moment of the field.

2.3.2 Regression operator

Based on regression methods, the new intensities of the particles defined $\boldsymbol{U} = [\boldsymbol{U}_1, \dots, \boldsymbol{U}_p]^T$ could be obtain by solving a linear system based on the minimization of the quadratic error. Assume that we have some vector $\boldsymbol{u} = [\boldsymbol{u}_1(z_1), \dots, \boldsymbol{u}_p(\boldsymbol{z}_p)]^T$ of the continuous field evaluations. The particle approximation could be compute on each particle positions $\boldsymbol{z}_p$ given

$$\boldsymbol{u} \simeq \hat{\boldsymbol{u}} = \boldsymbol{\Phi}\boldsymbol{U},$$

where $\boldsymbol{\Phi}_{ij} = \varphi_{i,j}$

Find the new intensities U_p correspond to solving the previous system in the least square sense. It corresponds to the classical problem to find the minimizer of the following quadratic function

$$\boldsymbol{U}_p = \arg\min_{\boldsymbol{U}} \|\boldsymbol{u} - \Phi\boldsymbol{U}\|_2^2 .$$

In this case, the solution is $\boldsymbol{U} = (\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}\boldsymbol{u}$. Because the problem is ill-posed, particularly in the case of large set of non well-distributed particles, the solution is regularized by introducing a Thikhonov penalization term. The Ridge regression we used introduced a penalization on of the form $\lambda^2 \|\boldsymbol{U}\|_2^2$, where λ^2 is a penalization coefficient, such as the new problem is

$$\boldsymbol{U}_{p,\text{ridge}} = \arg\min_{\boldsymbol{U}} \|\boldsymbol{u} - \boldsymbol{\Phi}\boldsymbol{U}\| + \lambda^2 \|\boldsymbol{U}\|_2^2 .$$

given the following solution $\boldsymbol{U}_{p,\text{ridge}} = (\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \lambda\boldsymbol{I})^{-1}\boldsymbol{\Phi}\boldsymbol{u}$.

Note that those methods are completely meshless operators that could be defined on whatever particle discretization.

2.3.3 Remeshing operator

A second type of method is based this time on a complete projection of the solution on a new regular grid of particles. This method introduced by Cottet and Koumoutsakos [7, 8] for the Vortex Method is also a key method in particle in cell method to transfer particle intensities to a grid, like with the MPM method [24, 11, 15].

Those method are based on a redistribution of the old intensities to the new regular distributed particle grid thanks to a kernel of redistribution W.

We first define $\Omega \in \mathbb{R}$, and will generalized for multidimentional space. We denote by z_I and z_p respectively the grid and the old particle location. The new particles are defined on a grid with spacing l_p. We define similarly the particle intensities with $\boldsymbol{U}_I$ and $\boldsymbol{U}_p$. By using the redistribution kernel W through a classical interpolation rule, the new particle intensities are

$$\boldsymbol{U}_p = \sum_I \boldsymbol{U}_I W\left(\frac{z_I - z_p}{l_p}\right).$$

The ~~interpolation~~ kernel W determines the type and quality of interpolation. First, the partition of unity is required

$$\sum_p W\left(\frac{z_p - z_I}{l_p}\right) \equiv 1$$

which leads to the conservation of the first order moment

$$\begin{aligned}\sum_p \boldsymbol{U}_p &= \sum_p \sum_I \boldsymbol{U}_I W\left(\frac{z_I - z_p}{l_p}\right) \\ &= \sum_I \boldsymbol{U}_I \sum_p W\left(\frac{z_I - z_p}{l_p}\right) \\ &= \sum_I \boldsymbol{U}_I\end{aligned}$$

It can be shown moreover that if for $1 \leq |\alpha| \leq m-1$, W satisfies,

$$\sum_I (\boldsymbol{z} - \boldsymbol{z}_I)^\alpha W\left(\frac{\boldsymbol{z} - \boldsymbol{z}_I}{l_p}\right) = 0, \tag{5}$$

The regridding procedure will be of order m. Equivalently, the previous equality lead, for $0 \leq |\alpha| \leq m-1$, to

$$\sum_I \boldsymbol{z}_I^\alpha W\left(\frac{\boldsymbol{z}_p - \boldsymbol{z}_I}{l_p}\right) = \boldsymbol{z}^\alpha$$

obtained by developing $(\boldsymbol{z} - \boldsymbol{z}_q)^\alpha$ and using a recurrence on previous orders. This means that the interpolation is exact for polynomials of degrees less or equal to $m-1$ or that the moment of order $m-1$ is conserved. For instance, if $m = 2$, we obtain that

$$\begin{aligned}\sum_p U_p(z-z_p) &= \sum_p \sum_I U_I W\left(\frac{z-z_I}{l_p}\right)(z_p-z_I)\\ &= \sum_p \sum_I U_I\left[(z-z_I)+(z_I-z_p)\right] W\left(\frac{z-z_I}{l_p}\right)\\ &= \sum_I U_I(z_I-z_p)+\sum_p (z-z_p) W\left(\frac{z-z_I}{l_p}\right)\\ &= \sum_I U_I(z-z_I)\end{aligned}$$

As an redistribution kernel, one may use the piecewise linear interpolation function

$$W(z) = \begin{cases} 1-|z| & 0 \leq |z| < 1 \\ 0 & 1 \leq |z| \end{cases} \tag{6}$$

and ensure the conservation of moment 0.

For higher degrees of freedom, the B-spline function provides a smoothing function for higher order. For instance, the quadratic B-splines defined in 3 ensure moment conservation by respecting the property of equation 5.

However, if higher order B-spline improves the smoothness of the solution, their accuracy is limited to second order and can only interpolate exactly linear functions.

~~To increase the accuracy,~~ Monaghan [20] proposes a systematic way to increase the accuracy and maintain the smoothness based on the extrapolation. The idea is to construct a new kernel based on a cutoff and its radial derivative. For $m = 4$, the cubic B-spline is improved by the following new interpolating kernel

$$M_4'(z) = \begin{cases} 1-\frac{5}{2}z^2+\frac{3}{2}|z|^3 & 0 \leq |z| \leq 1 \\ \frac{1}{2}(2-|z|)^2(1-|z|) & 1 \leq |z| \leq 2 \\ 0 & 2 \leq |z| \end{cases} \tag{7}$$

The drawback of those method is to not be positive.

Finally, for multidimensional space, the redistribution kernel W could be obtain as the product of the one dimensional kernel applied to each coordinate. such as

$$\begin{aligned}\boldsymbol{U}_p &= \sum_I \boldsymbol{U}_I W\left(\boldsymbol{z}_I - \boldsymbol{z}_p, l_p\right)\\ &= \sum_I \boldsymbol{U}_I \prod_{i=1}^{d} W_{1\mathrm{D}}\left(\frac{\boldsymbol{z}_{I,i}-\boldsymbol{z}_{p,i}}{l_p}\right)\end{aligned}$$

3 Methods

This section presents the development of ensemble data assimilation techniques for particle-based simulations.

If the forward step is still the same in our approach as in the classical Ensemble Kalman Filter, the main challenge resides in the update in the analysis step.

Based on the analysis formula 1 and the two families of transformation 2.3 on the particle discretization, we propose two different EnKF-adapted filters.

The first common step is to compute the correction matrix $\boldsymbol{F}$ defined in 2, which defines the linear combination of the members. This step is possible because the analysis is independent of the state discretization thanks to the observation matrix-free implementation. Thus the analysis depends only on the observation $\boldsymbol{y}$, the predictive observation $\{\boldsymbol{h_i}\}_{i=1}^{N_{obs}}$ and the noise sample $\{\boldsymbol{\varepsilon_i}\}_{i=1}^{N_{obs}}$.

The differences come with the state update process and the new state definition. On the one hand, the Remesh-EnKF filter 3.1 uses new particle discretization, which will be the same for all the members. On the other hand, the particle positions are still the same for the Particles-EnKF filter 3.2, but the intensities will be updated.

The Remesh-EnKF uses a regridding process to update the particle quantities directly. The second one is intended to keep as much as possible the forward member discretization based on the approximation operator.

3.1 Remesh-EnKF

The first method consists of redefining the discrete discretization on a new common one to perform the update with the transformation described in 2.3.3. After the remeshing, the previous discretization is lost, and the simulation continues with the new set of particles.

After the forecast step, the particle discretization for a member i is defined by the set $\left\{(\boldsymbol{z}_{ip}, \hat{\boldsymbol{U}}_{ip}^f)\right\}_{p=1}^{N_i}$ with N_i the number of particles. After the remesh transformation 3.1, a new particles discretization on the regular grid is obtained and gives $\left\{(\boldsymbol{z}_q, \boldsymbol{U}_{iq}^f)\right\}_{q=1}^{N_g}$ where $(\boldsymbol{z}_q, \boldsymbol{U}_{iq}^f)$ are the new coordinates and intensities. The state of the forecast member could be directly expressed as the vector of intensities $\boldsymbol{U}_i^f = [\boldsymbol{U}_1^f, ..., \boldsymbol{U}_{N_g}^f]^T$ with the ordered particle intensities. The linear combination is directly performed on particle intensities for each member $i = 0, \ldots, N$ such as

$$\boldsymbol{U}_i^a = \boldsymbol{U}_i^f + \sum_j F_{ij} \boldsymbol{U}_j^f$$

.

3.2 Particles-EnKF

The Particles-EnKF formulation will not change the forward particle discretization. Each member will keep the same particle positions during the analysis step. The particle volumes do not change also. The only change will concern the particle intensities. This way, the Lagrangian representation of the solution at the end of the forward step is kept the same as much as possible.

In this method, the analysis functions u_i^a are approximate. The analysis fields are obtained at any spacial coordinates thanks to the particle approximation of each member field u_i^f such as $\forall z \in \Omega$

$$\boldsymbol{u}_i^a(\boldsymbol{z}) = \boldsymbol{u}_i^f(\boldsymbol{z}) + \sum_j \boldsymbol{F}_{ij}\boldsymbol{u}_j^f(\boldsymbol{z}) \quad i = 1, \dots, N$$

Thus, solutions are also described on a particle discretization $Z_i^a = \cup_k Z_k^f$. Such as

$$\boldsymbol{u}_i^a(\boldsymbol{z}) = \sum_p \boldsymbol{U}_{ip}^f \varphi_{ip}(\boldsymbol{z}) + \sum_j \boldsymbol{F}_{ij} \sum_p \boldsymbol{U}_{jp}^f \varphi_{jp}(\boldsymbol{z}) \quad i = 1, \dots, N \tag{8}$$

One solution is to approximate this solution on the previous discretization such that $Z_i^a = Z_i^f$ to avoid exponential growth of the number of particles.

This would be done thanks to agglomeration to reduce the number of particles like in [25], or with regression operation defined in section 2.3.1.

By this way, the analyzed field is approximated by $\hat{\boldsymbol{u}}_i^a$ such as

$$\hat{\boldsymbol{u}}_i^a(\boldsymbol{z}) \simeq \sum_p \boldsymbol{U}_{ip}^a \varphi_{ip}(\boldsymbol{z})$$

where $\boldsymbol{U}_{ip}^a$ have been determined by approximation or regression.

However, because this regression is only performed on the support, the current forecast discretization Z_i^f additional particles could be introduced at the support border to allow a better estimate.

To do so, a border of null intensities particles is introduced after a remesh process. This introduced a collocation point that better fit the velocity flow during the simulation.

3.3 Complexity

Before evaluating the accuracy of the two filters, we could first evaluate their computational complexity. In the Remesh-EnKF, the complexity is led by the remeshing step for the ensemble of size N. The analysis is just a matrix matrix multiplication.

The remeshing process, described in section 2.3.3, necessitates redistributing the particle strengths to a new grid of particles. A loop over the N_p particles is first performed. If the size of the kernel is finite of size N_s, the redistribution of one particle needs N_s^d kernel evaluations, where d is the space dimension. Thus, the complexity of the Remesh-filter is $\mathcal{O}(NN_pN_s^d)$.

Finally, the updated step needs to compute a matrix multiplication between two matrices of shape N^2, leading to a global complexity of $\mathcal{O}(N^2 + NN_pN_s^d)$.

On the other hand, the Particles-EnKF complexity is led by an evaluation of the analysis fields and the regression or approximation step. First, the N analyses fields have to be evaluated in the particle positions of each member. Due to the linear combination of the forecast field as described in equation 8 such as

$$\boldsymbol{u}_i^a(\boldsymbol{z}_k) = \sum_{p \in V_k} \boldsymbol{U}_{ip}^f \varphi_{ip}(\boldsymbol{z}_k) + \sum_j \boldsymbol{F}_{ij} \sum_{p \in V_k} \boldsymbol{U}_{jp}^f \varphi_{jp}(\boldsymbol{z}_k) \quad i = 1, \ldots, N, \quad k = 1, \ldots, N_{ip}. \tag{9}$$

A grid search strategy could be used instead. For each particle, a grid cell is determined from which it belongs. All the particles are classified in a grid of size N_g. For each position evaluation, the distance is evaluated only on the particles in the next cell of size N_s. Because the distance is symmetric, the evaluation could be divided by two.

A loop is performed over the N_g^d grid cells, and the distance of the particles of one cell is computed with the $N_s^d/2 + 1$ cells. We approximate the mean number of particles in each cell as $N_p ic = fract N N_p / N_g^d$. Thus, the complexity is estimated as $\mathcal{O}(N N_p N_s^d)$.

Then, the particle strengths have to be computed. An approximation, as described in section 2.3.1, does not add complexity. However, a regression method, like in section 2.3.2, needs to solve N linear system of size N_p. If the system is already computed thanks to the computation of the distances, the resolution needs the inversion of a N_p matrix. This step uses a conjugate gradient algorithm, the best solver for a sparse system. In this case, the complexity is about $\mathcal{O}((N_0 + N_p)k)$ where k is the number of iterations and N_0 is the non-zero value of the matrix to inverse, which is about $\mathcal{O}(N_p N_s^d)$. Bringing everything together, the complexity is about $\mathcal{O}(k N N_p N_s^d)$. Together the global complexity is about $\mathcal{O}((k+1) N N_p N_s^d)$.

References

[1] Lorena A. Barba. *Vortex Method for Computing High-Reynolds Number Flows: Increased Accuracy with a Fully Mesh-Less Formulation*. PhD thesis, California Institute of Technology, May 2004.

[2] S. Bardenhagen and Edward Kober. The Generalized Interpolation Material Point Method. *CMES - Computer Modeling in Engineering and Sciences*, 5, June 2004.

[3] J. Thomas Beale. On the accuracy of vortex methods at large times. 1988.

[4] Marc Bocquet. Introduction to the principles and methods of data assimilation in the geosciences. page 89, 2014.

[5] Bertrand Bonan, Nancy Nichols, Michael Baines, and Dale Partridge. Data assimilation for moving mesh methods with an application to ice sheet modelling. *Nonlinear Processes in Geophysics*, 24:515–534, September 2017.

[6] Nan Chen, Quanling Deng, and Samuel N. Stechmann. Superfloe Parameterization with Physics Constraints for Uncertainty Quantification of Sea Ice Floes, March 2022. arXiv:2105.13569 [physics].

[7] G.-H Cottet and Petros Koumoutsakos. *Vortex methods - theory and practice.* March 2000.

[8] G.-H. Cottet, M.-L. Ould Salihi, and M. El Hamroui. Multi-purpose regridding in vortex methods. *ESAIM: Proceedings*, 7:94–103, 1999.

[9] P. A. Cundall and O. D. L. Strack. A discrete numerical model for granular assemblies. *Géotechnique*, 29(1):47–65, March 1979. Publisher: ICE Publishing.

[10] Darwin Darakananda, Andre Fernando De Castro da Silva, Tim Colonius, and Jeff Eldredge. Data-assimilated low-order vortex modeling of separated flows. *Physical Review Fluids*, November 2018.

[11] Alban de Vaucorbeil, Vinh Phu Nguyen, Sina Sinaie, and Jian Ying Wu. Material point method after 25 years: Theory, implementation, and applications. In *Advances in Applied Mechanics*, volume 53, pages 185–398. Elsevier, 2020.

[12] Geir Evensen. 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, 1994.

[13] Geir Evensen, Femke C. Vossepoel, and Peter Jan van Leeuwen. *Data Assimilation Fundamentals: A Unified Formulation of the State and Parameter Estimation Problem.* Springer Textbooks in Earth Sciences, Geography and Environment. Springer International Publishing, Cham, 2022.

[14] R. A. Gingold and J. J. Monaghan. Smoothed particle hydrodynamics: theory and application to non-spherical stars. *Monthly Notices of the Royal Astronomical Society*, 181(3):375–389, 12 1977.

[15] Chenfanfu Jiang, Craig Schroeder, Andrew Selle, Joseph Teran, and Alexey Stomakhin. The affine particle-in-cell method. *ACM Transactions on Graphics*, 34(4):1–10, July 2015.

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

[17] Mathieu Le Provost, Ricardo Baptista, Youssef Marzouk, and Jeff Eldredge. *A low-rank nonlinear ensemble filter for vortex models of aerodynamic flows.* January 2021.

[18] Leon B. Lucy. A numerical approach to the testing of the fission hypothesis. *The Astronomical Journal*, 82:1013–1024, 1977.

[19] Chloé Mimeau and Iraj Mortazavi. A Review of Vortex Methods and Their Applications: From Creation to Recent Advances. *Fluids*, 6(2):68, February 2021.

[20] J. J Monaghan. Extrapolating B splines for interpolation. *Journal of Computational Physics*, 60(2):253–262, September 1985.

[21] Giovanni Russo and John A. Strain. Fast triangulated vortex methods for the 2d euler equations. *Journal of Computational Physics*, 111:291–323, 1994.

[22] A. Siripatana, L. Giraldi, O. P. Le Maître, O. M. Knio, and I. Hoteit. Combining ensemble Kalman filter and multiresolution analysis for efficient assimilation into adaptive mesh models. *Computational Geosciences*, 23(6):1259–1276, December 2019.

[23] Pietro Sperotto, Sandra Pieraccini, and Miguel Alfonso Mendez. A meshless method to compute pressure fields from image velocimetry. *Measurement Science and Technology*, 33, 2021.

[24] D Sulsky, Z Chenb, and H L Schreyer. A particle method for history-dependent materials. page 18, 1994.

[25] Yonghao Yue, Breannan Smith, Christopher Batty, Changxi Zheng, and Eitan Grinspun. Continuum Foam: A Material Point Method for Shear-Dependent Flows. *ACM Transactions on Graphics*, 34(5):1–20, November 2015.