



RESEARCH ARTICLE

10.1029/2024MS004259

Special Collection:Data assimilation for Earth
system models**Key Points:**

- An efficient nonlinear Lagrangian-Eulerian multiscale data assimilation framework with parallel computing is developed
- A set of continuum equations under the Eulerian coordinates are rigorously derived from Boltzmann kinetic theory of particle dynamics
- Closed analytic formulas are available for solving nonlinear data assimilation solutions and deriving reduced-order schemes

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:N. Chen,
chenncan@math.wisc.edu**Citation:**

Deng, Q., Chen, N., Stechmann, S. N., & Hu, J. (2025). LEMDA: A Lagrangian-Eulerian multiscale data assimilation framework. *Journal of Advances in Modeling Earth Systems*, 17, e2024MS004259. <https://doi.org/10.1029/2024MS004259>

Received 31 JAN 2024

Accepted 9 JAN 2025

Author Contributions:

Conceptualization: Quanling Deng, Nan Chen, Samuel N. Stechmann
Methodology: Quanling Deng, Nan Chen, Samuel N. Stechmann
Software: Quanling Deng, Nan Chen
Validation: Quanling Deng, Nan Chen
Writing – original draft: Quanling Deng, Nan Chen

LEMDA: A Lagrangian-Eulerian Multiscale Data Assimilation Framework

 Quanling Deng¹ , Nan Chen² , Samuel N. Stechmann^{2,3} , and Jiuhua Hu²

¹School of Computing, Australian National University, Canberra, ACT, Australia, ²Department of Mathematics, University of Wisconsin–Madison, Madison, WI, USA, ³Department of Atmospheric and Oceanic Sciences, University of Wisconsin–Madison, Madison, WI, USA

Abstract Lagrangian trajectories are widely used as observations for recovering the underlying flow field via Lagrangian data assimilation (DA). However, the strong nonlinearity in the observational process and the high dimensionality of the problems often cause challenges in applying standard Lagrangian DA. In this paper, a Lagrangian-Eulerian multiscale DA (LEMDA) framework is developed. It starts with exploiting the Boltzmann kinetic description of the particle dynamics to derive a set of continuum equations, which characterize the statistical quantities of particle motions at fixed grids and serve as Eulerian observations. Despite the nonlinearity in the continuum equations and the processes of Lagrangian observations, the time evolution of the posterior distribution from LEMDA can be written down using closed analytic formulas after applying the stochastic surrogate model to describe the flow field. This offers an exact and efficient way of carrying out DA, which avoids using ensemble approximations and the associated tunings. The analytically solvable properties also facilitate the derivation of an effective reduced-order Lagrangian DA scheme that further enhances computational efficiency. The Lagrangian DA part within the framework has advantages when a moderate number of particles is used, while the Eulerian DA part can effectively save computational costs when the number of particle observations becomes large. The Eulerian DA part is also valuable when particles collide, such as using sea ice floe trajectories as observations. LEMDA naturally applies to multiscale turbulent flow fields, where the Eulerian DA part recovers the large-scale structures, and the Lagrangian DA part efficiently resolves the small-scale features in each grid cell via parallel computing. Numerical experiments demonstrate the skillful results of LEMDA and its two components.

Plain Language Summary Lagrangian tracers are drifters, such as robotic instruments, balloons, sea ice floes, and litter. The observed Lagrangian tracer trajectories can be combined with a model to improve the state estimation of the flow field. This is called Lagrangian data assimilation (DA), where a large number of tracer trajectories are often needed to infer the multiscale features of the flow field accurately. However, this process is usually quite expensive. In this paper, a new multiscale DA scheme that exploits both the Eulerian and Lagrangian observations is developed. The Eulerian observations are induced by a group of Lagrangian tracers. They are described at fixed grids and represent the statistical quantities of these particle motions. The DA with these Eulerian observations provides a rapid and accurate way to recover the large-scale flow features, while the Lagrangian DA part efficiently resolves the small-scale features in each grid cell via parallel computing. Eulerian DA is also extremely valuable when collisions between tracers happen, such as when sea ice floes are used as tracers; in this case, significant errors appear in Lagrangian DA. Notably, with appropriate stochastic surrogate forecast models, closed analytic formulas are available for the proposed DA framework.

1. Introduction

Data assimilation (DA) seeks to optimally integrate different sources of information to improve the state estimation of a complex system (Asch et al., 2016; Kalnay, 2003; Lahoz & Menard, 2010; Law et al., 2015; A. J. Majda & Harlim, 2012). Typically, the output from a numerical model is combined with observations to reduce the bias and the uncertainty in the estimated states, where observations are often noisy and sparse while the model usually contains model errors. DA plays a vital role in improving the state estimation at the initialization stage, facilitating the skillful prediction of chaotic or turbulent signals. DA is also widely used to recover the unobserved states or dynamically interpolate the partially observed time series. It is an essential tool for postprocessing the data in geophysics and climate science (Kalnay et al., 1996; Uppala et al., 2005).

Writing – review & editing:

Quanling Deng, Nan Chen, Samuel
N. Stechmann

Observational data, especially in geophysics and fluids, can be categorized into two groups, depending on how they are measured. Eulerian observations refer to those estimated at fixed locations, while Lagrangian observations are often the drifters that follow a parcel of fluid's movement (Blunden & Arndt, 2019; Castellari et al., 2001; Griffa et al., 2007; Honnorat et al., 2009; Salman et al., 2008). When these observations are combined with models for state estimation, the corresponding approaches are named Eulerian and Lagrangian DA, respectively. Eulerian observations are usually more straightforward to measure and convenient to use in DA, especially since the observational operator is linear in many practical applications. In contrast, Lagrangian DA is preferred in situations where traditional Eulerian observations are unavailable. Lagrangian DA is a natural approach to recover the multiscale ocean velocity field, which is often not directly measured but can be inferred from the observed tracer trajectories (Apte & Jones, 2013; Apte, Jones, & Stuart, 2008; Apte, Jones, Stuart, & Voss, 2008; Ide et al., 2002). Some of the well-known Lagrangian data sets are the Global Drifter Program (Centurioni et al., 2017), which aims to estimate near-surface currents by tracking the surface drifters deployed throughout the global ocean, and the Argo program (Gould et al., 2004) that utilizes a fleet of robotic instruments drifting with the ocean currents for advancing operational ocean DA. Lagrangian DA also provides a powerful tool to facilitate the recovery of ocean eddies in the Arctic regions, where the sea ice floes play the role of the Lagrangian tracers (Chen, Fu, & Manucharyan, 2022; Mu et al., 2018). In addition to the ocean drifters, other types of Lagrangian tracers include trash or oil in the ocean (García-Sánchez et al., 2022; Van Sebille et al., 2012) and balloons collecting atmospheric data (Businger et al., 1996).

Despite the recent development of Lagrangian data sets, several significant challenges exist in applying Lagrangian DA to recover the underlying flow field. First, as the velocity field is a strongly nonlinear function of displacement, the observational process of Lagrangian DA is always highly nonlinear (Apte, Jones, & Stuart, 2008; Chen et al., 2014). The strong nonlinearity in the observational process requires additional manipulations for applying standard ensemble-based DA methods (Apte & Jones, 2013; Apte, Jones, & Stuart, 2008; Apte, Jones, Stuart, & Voss, 2008; Ide et al., 2002; Sun & Penny, 2019). Second, the dimension of the coupled flow and tracer system for Lagrangian DA is usually quite large, which creates computational difficulties. The high dimensionality comes from both the underlying flow model and a large number of observed Lagrangian tracers. The latter is essential for recovering multiscale features of complex flow fields. Increasing the number of tracers by a significant amount is a straightforward way to allow the observations to collect information in the entire domain. However, such a strategy will significantly increase the computational cost of Lagrangian DA. Third, the trajectories of Lagrangian drifters may be affected by other external forcing sources in addition to the underlying flow field (Cundall & Strack, 1979; Manucharyan & Montemuro, 2022; Potyondy & Cundall, 2004). For example, sea ice floe dynamics is complicated by collisions, where the contact forces between floes gradually dominate the drag force from the ocean when the number of floes or, equivalently, the concentration increases. Modeling the exact instantaneous contact force is a highly challenging task, which substantially affects the accuracy of the state estimation from Lagrangian DA.

To overcome the mathematical and computational challenges in Lagrangian DA, a Lagrangian-Eulerian multiscale DA (LEMMA) framework is developed in this paper, where the directly available observations consist of only the Lagrangian trajectories. One crucial step in building such a framework is to use the Boltzmann equation as a kinetic-theory description of the particle dynamics to derive a set of continuum equations (Deng et al., 2024), which characterize the statistical quantities of particle motions. It has been shown that using observed statistics for DA can resolve some challenging issues in applying standard trajectory observations (Bach et al., 2023; Farchi et al., 2021). Notably, this leads to an alternative consideration of the problem under the Eulerian framework, where the state variables describing the statistical features of the particles, such as the mass density, the momentums, the angular momentums, etc., are defined at fixed grids. Since acquiring the observed Eulerian quantities requires a statistical average of the particle properties, these Eulerian observations are often calculated at coarse grids. Therefore, the focus on the Eulerian DA part is to efficiently recover the large-scale features of the underlying flow field. In contrast, the observed individual particle trajectory allows Lagrangian DA to infer a more refined flow field structure based on finer grids or assuming a higher spatial resolution for solutions. One salient property of LEMMA is that, despite the nonlinearity in the continuum equations and the processes of Lagrangian observations, the time evolutions of the posterior distribution from both Eulerian and Lagrangian DA can be written down using closed analytic formulas. This offers an exact and efficient way of carrying out DA, which avoids using ensemble methods and the associated tuning procedures.

In addition to significantly improving the computational efficiency by exploiting the closed analytic solution of the posterior distribution, the two components of LEMDA, namely the Eulerian and Lagrangian parts, respectively, have the following unique features. On the one hand, the new Eulerian DA can overcome at least two shortcomings in the standard Lagrangian DA. First, the computational cost of Lagrangian DA increases significantly as the number of observed trajectories increases. In contrast, the computational efficiency of Eulerian DA relies only on the mesh size. With an increasing number of observed Lagrangian tracers, the averaged quantities of the Eulerian observations are estimated more accurately while the computational cost of the DA remains the same. Second, the statistical average in constructing the Eulerian observations automatically smooths out the interfering effect in recovering the underlying flow field due to particle collisions. In other words, the dominant forcing that drives the time evolution of the observed Eulerian statistical quantities comes solely from the flow field in such a situation. This unique feature provides an effective way for LEMDA to recover the large-scale flow structures in the presence of particle collisions, which is usually a challenging task for standard Lagrangian DA. On the other hand, rigorous mathematical theory can be applied to improve the computational efficiency of the nonlinear Lagrangian DA in the LEMDA framework under appropriate circumstances. The structure of the Lagrangian DA part allows us to derive approximate schemes that remain analytically solvable for the DA solutions but further accelerate the computations. This includes predetermining the posterior covariance matrix for nearly incompressible flows when the number of tracers becomes large. Finally, the two DA approaches in the LEMDA framework can naturally be combined to develop a multiscale DA scheme that improves the state estimation of complex flow fields. The Eulerian DA part aims to recover the large-scale features while the observed individual particle trajectory advances the Lagrangian DA part to infer a more refined flow field structure within each grid cell that cannot be captured by the Eulerian DA part. Notably, parallel computing can be implemented for applying Lagrangian DA in different grid cells to significantly enhance computational efficiency.

The rest of the paper is organized as follows. Section 2 describes the LEMDA framework. It includes the derivation of the continuum equation from the particle system and the approximate stochastic forecast models of the flow field. The section also contains the crucial analytic solution of the posterior distribution from DA. Section 3 presents the time series and information metrics for assessing the DA skill. Section 4 analyzes the performance of Lagrangian DA and derives a reduced-order approximate Lagrangian DA scheme to significantly reduce the computational cost. Section 5 studies Eulerian DA and illustrates its advantages over Lagrangian DA in the presence of particle collisions. Numerical experiments of LEMDA are presented in Section 6. The paper is concluded in Section 7.

2. The LEMDA Framework

Consider a two-dimensional, doubly periodic domain $[0, 2\pi)^2$. Assume there are L Lagrangian tracers, whose displacements are denoted by $\mathbf{x}_l = (x_l, y_l)^T$ with $l = 1, \dots, L$. Denote by $\mathbf{v}_l = (u_l, v_l)^T$ the corresponding tracer velocity. Further denote by $\mathbf{u} = (\tilde{u}, \tilde{v})^T$ the underlying flow field that drives the motion of the Lagrangian tracers. The target state variables to recover in DA are $\mathbf{u}$ and $\mathbf{v}_l$ with $l = 1, \dots, L$ while $\mathbf{x}_l$ with $l = 1, \dots, L$ are observations.

LEMDA is a DA framework exploiting both Lagrangian and Eulerian DAs. The Lagrangian DA part uses the directly observed trajectories of the tracers. The Eulerian DA part is based on a set of deduced continuum models via the Boltzmann description and the associated observed state variables. These quantities characterize the statistical properties of the particles described at fixed grids. Continuous-in-time observations are considered in this work and therefore the focus below is on continuous DA, which has been widely used in many theoretical studies (Azouani et al., 2014; Bergemann & Reich, 2012; Bröcker, 2010; Carlson & Larios, 2021; Farhat et al., 2020; Yang et al., 2013) and practical problems (Chen, Li, & Liu, 2022; Desamsetti et al., 2019; Lean et al., 2021; Rebholz & Zervas, 2021). The framework can be extended to its discrete-in-time DA analog.

Figure 1 includes an overview of the LEMDA framework. The details of each component will be presented in the remaining subsections. In addition to the general framework, the tests in the remaining paper will also include (a) applying Eulerian DA to the cases where the observed tracers are the ice floes with collisions (see Section 5.2), (b) the development of cheap stochastic surrogate models for the underlying flow field, a crucial part of allowing for

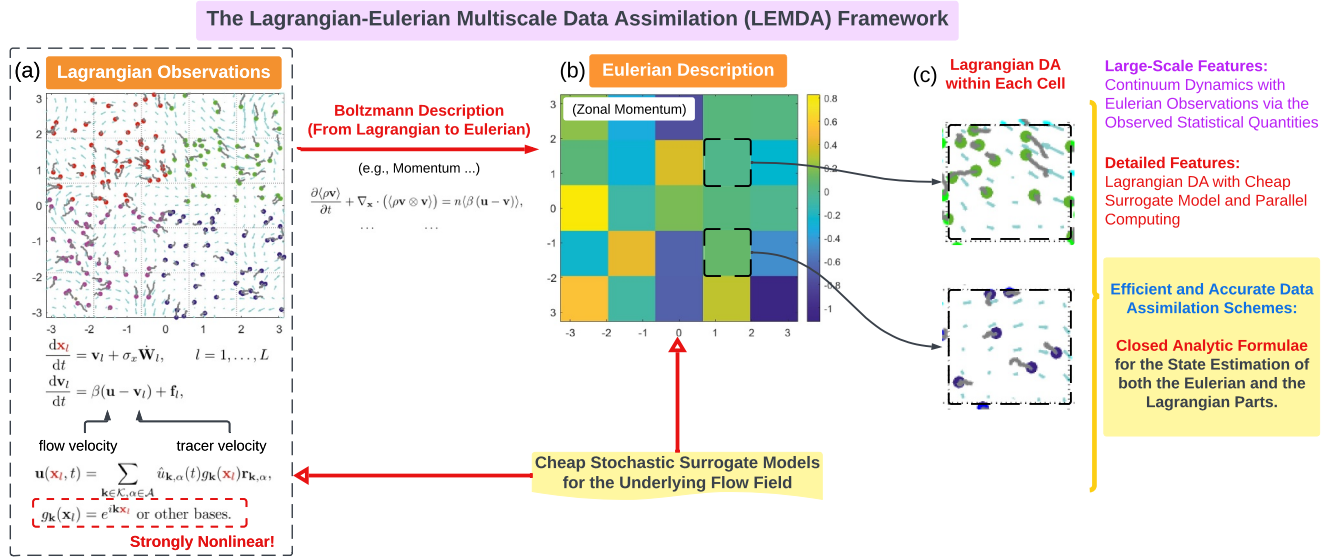


Figure 1. Overview of the LEMMA framework. Panel (a): the Lagrangian observations and Lagrangian DA (see Section 2.1). Note the strong nonlinearity in the observational process of Lagrangian DA. Panel (b): the Eulerian description of particle statistics, which can be derived from the Boltzmann description of the particle system (see Section 2.2). Panel (c): applying Lagrangian DA to each grid cell to recover refined features that are missed by the large-scale Eulerian DA in the multiscale DA framework.

the analytic solvers of both Eulerian and Lagrangian DAs (see Section 2.3), and (c) the rigorous derivation of reduced-order DA schemes for Lagrangian DA to further reduce the computational cost (see Section 4.2).

2.1. The Lagrangian Particle Model

The Lagrangian particle model describes the motion of each tracer (Damsgaard et al., 2018; Manucharyan & Montemuro, 2022).

$$\frac{dx_l}{dt} = v_l + \sigma_x \dot{W}_l, \quad (1a)$$

$$\frac{dv_l}{dt} = \beta(u - v_l) + f_l, \quad (1b)$$

where $\dot{W}_l$ in Equation 1a is a standard Gaussian white noise (zero mean and identity covariance) and σ_x is the noise coefficient. This noise is optional when generating the true signal. Newton's law does not require such a noise. Nevertheless, if it is imposed as a non-zero value, then it represents the unresolved scales of the flow field. On the other hand, when (Equation 1) is utilized as a forecast model in DA, then the noise can account for uncertainty in the observational process (with a scaling). In Equation 1b, β is the drag coefficient, where the drag force provides the basic mechanism for the underlying flow field to drive the tracer velocity. For a finite value of β , the response of the tracer velocity to the ocean has a delayed effect. The drag coefficient is usually proportional to the inverse of the cross-section area of the particle. In many applications, for example, when sea ice floes represent the particles, the drag coefficient is around order $O(1)$ when the equations are written in the standard non-dimensional form, as is used in this work (Chen et al., 2021). Only a linear drag is used in (Equation 1) for simplicity. The equation can be easily modified by utilizing other types of drag force (Damsgaard et al., 2018). The forcing term f_l represents the other external forces, such as the one due to the particle collision (e.g., when the trajectories of the ice floes are used as Lagrangian observations). The forcing f_l is typically a nonlinear function of x_l and can be a function of v_l and u as well (Chen et al., 2021; Cundall & Strack, 1979). If $f_l = 0$, then in the limit of $\beta \rightarrow \infty$, the flow velocity u and the tracer velocity v_l are completely synchronized, which has also been used in many theoretical analysis and idealized tests (Apte & Jones, 2013; Apte, Jones, & Stuart, 2008; Apte, Jones, Stuart, & Voss, 2008; Ide et al., 2002).

2.2. The Continuum Model for Describing Eulerian Quantities

The Boltzmann equation was originally proposed to describe the statistical behavior of a thermodynamic system (Cercignani & Cercignani, 1988). The Lagrangian tracers are analogs to the molecules in thermodynamics. Mathematically, by assuming the tracer number goes to infinity, the Boltzmann-type equations can be naturally derived.

The phase space of the Boltzmann equation is given by the set of all possible positions $\mathbf{x}$ and velocity $\mathbf{v}$. The differential volume element is written as

$$dE = d\mathbf{x} d\mathbf{v}. \quad (2)$$

In the limit $L \rightarrow \infty$, the general Boltzmann equation describing the statistical behavior of the particle motion is given by (Harris, 2004)

$$\frac{\partial g}{\partial t} + \nabla_{\mathbf{x}} \cdot \left(g \frac{d\mathbf{x}}{dt} \right) + \nabla_{\mathbf{v}} \cdot \left(g \frac{d\mathbf{v}}{dt} \right) = C(g), \quad (3)$$

where g is a density function of the space $(t, \mathbf{x}, \mathbf{v})$. The term $C(g) = \left(\frac{\partial g}{\partial t} \right)_{\text{coll}}$ represents the particle interaction, such as the collision in the ice floe case, which satisfies

$$\int C(g) dE = 0. \quad (4)$$

It is worth noting that extra variables can be added to the phase space. For example, angular displacement and angular velocity are additional crucial state variables when the shape of particles, such as the ice floes, is considered. The kinetic equations for such angular/rotating/oriented situations have been investigated in the context of bacteria, cells, molecules, colloidal particles, granular flows, etc., where the angular motion, direction of alignment of particles/rods, and torque from contact interactions are accounted for in Smoluchowski-type/Boltzmann-type equations (Baskaran & Marchetti, 2010; Curtiss, 1956, 1967; Deng et al., 2024; Doi & Edwards, 1978, 1988; Ezhilan et al., 2013; Onsager, 1949; Saintillan & Shelley, 2015; Taxman, 1958). This literature also includes discussions of non-dilute concentration scenarios (Nott et al., 2011; Shaqfeh & Fredrickson, 1990; Wolgemuth, 2008). For the case of the Boltzmann equation in Equation 3, note that it already has five dimensions. The dimension will be further increased with additional variables. As such, it is numerically challenging to solve the Boltzmann equation with additional variables.

To allow practical numerical simulation of the particle statistics, the equations of the so-called statistical moments are derived to reduce the dimension of the system while maintaining the main large-scale dynamical features. To this end, integrating the Boltzmann Equation 3 over the velocity space leads to the continuity equation for the mass density ρ :

$$\frac{\partial \rho}{\partial t} + \nabla_{\mathbf{x}} \cdot (\rho \langle \mathbf{v} \rangle) = 0, \quad (5)$$

where $\langle \mathbf{v} \rangle$ is the particle velocities at fixed locations. Practically, $\langle \mathbf{v} \rangle$ can be regarded as the averaged velocity of the particles appearing in a grid cell. Similarly, the momentum equations can be derived

$$\frac{\partial \langle \rho \mathbf{v} \rangle}{\partial t} + \nabla_{\mathbf{x}} \cdot (\langle \rho \mathbf{v} \otimes \mathbf{v} \rangle) = n \langle \beta (\mathbf{u} - \mathbf{v}) \rangle, \quad (6)$$

where $\otimes$ is the outer product and n is the particle number density. Herein, as the mass of a particle m is assumed to be one in Equation 1b, $\rho = nm = n$.

The flow velocity $\mathbf{u}$ appears on the right-hand side of Equation 6, allowing the equation to be used as an observational process for Eulerian DA. The observation is the momentum $\langle \rho \mathbf{v} \rangle$, which can be computed by

exploiting the observed Lagrangian tracers. Depending on the applications, other statistical moment equations, such as those describing the angular momentums, can be derived (Deng et al., 2024).

In this paper, the two components of the momentum in Equation 6 will be used as observations for the Eulerian DA part.

2.3. Stochastic Approximate Forecast Model for the Flow Field

DA involves running the forecast model to obtain the so-called prior distribution (Evensen & Van Leeuwen, 2000). The forecast model of the underlying flow field is typically complicated and expensive to run. Furthermore, adopting an ensemble of forecast trajectories in DA will bring additional computational challenges in practice.

In this subsection, a general framework for approximating the forecast model that characterizes the underlying turbulent flow field is developed. With appropriate stochastic approximations, the forecast prior distribution remains similar to that using the original model, but the computational cost is reduced significantly. Notably, the structure of the forecast model allows for the use of closed analytic formulas to solve the DA solution, which avoids using expensive ensemble methods and will be discussed in the subsequent sections.

Consider the following spectral representation of the underlying flow velocity field (Chen et al., 2015; A. Majda, 2003),

$$\mathbf{u}(\mathbf{x}, t) = \sum_{\mathbf{k} \in \mathcal{K}, \alpha \in \mathcal{A}} \hat{u}_{\mathbf{k}, \alpha}(t) e^{i\mathbf{k}\mathbf{x}} \mathbf{r}_{\mathbf{k}, \alpha}, \quad (7)$$

where $\mathbf{x} = (x, y)^T$ is the two-dimensional coordinate and the flow field is given by a finite summation of Fourier modes. The index $\mathbf{k} = (k_1, k_2)^T$ is the wavenumber, and the index α represents different types of waves, including, for example, the gravity modes and the geophysically balanced modes in the study of many geophysical flows. The set $\mathcal{K}$ usually consists of all the wavenumbers that satisfy $-K_{\max} \leq k_1, k_2 \leq K_{\max}$ with $K_{\max}$ being an integer that is pre-determined. The vector $\mathbf{r}_{\mathbf{k}, \alpha}$ is the eigenvector, which links the two components of velocity fields, namely $\tilde{u}$ and $\tilde{v}$. As the left hand side of Equation 7 is evaluated in physical space and is real-valued, the coefficients $\hat{u}_{\mathbf{k}, \alpha}$ and $\hat{u}_{-\mathbf{k}, \alpha}$ for all $\mathbf{k}$ are complex conjugates. In addition, $\mathbf{r}_{\mathbf{k}, \alpha}$ and $\mathbf{r}_{-\mathbf{k}, \alpha}$ are also complex conjugates. Note that the Fourier basis functions are adopted here to simplify the description of the framework. Different basis functions and boundary conditions can be utilized in Equation 7 for various applications in practice.

Stochastic models are used to forecast the time evolution of each Fourier coefficient $\hat{u}_{\mathbf{k}, \alpha}$ in Equation 7 in the forecast stage of DA, which is a much more computationally efficient way to mimic the observed turbulent flows generated from a complicated partial differential equation (PDE) system. Although stochastic models will not reproduce the exact single forecast trajectory from the original system, the forecast prior distribution from appropriate stochastic models can be similar to that from the more complicated model. To this end, the stochastic model is often calibrated by matching several key statistics in the observed time series of $\hat{u}_{\mathbf{k}, \alpha}$. Among different stochastic models, the linear stochastic model, namely the complex Ornstein-Uhlenbeck (OU) process (Gardiner, 1985), is a widely used choice:

$$\frac{d\hat{u}_{\mathbf{k}, \alpha}}{dt} = (-d_{\mathbf{k}, \alpha} + i\omega_{\mathbf{k}, \alpha}) \hat{u}_{\mathbf{k}, \alpha} + f_{\mathbf{k}, \alpha}(t) + \sigma_{\mathbf{k}, \alpha} \dot{W}_{\mathbf{k}, \alpha}, \quad (8)$$

where $d_{\mathbf{k}, \alpha}$, $\omega_{\mathbf{k}, \alpha}$ and $f_{\mathbf{k}, \alpha}(t)$ are damping, phase and deterministic forcing, $\sigma_{\mathbf{k}, \alpha}$ is the noise coefficient and $\dot{W}_{\mathbf{k}, \alpha}$ is a white noise. The constants $d_{\mathbf{k}, \alpha}$, $\omega_{\mathbf{k}, \alpha}$ and $\sigma_{\mathbf{k}, \alpha}$ are real-valued while the forcings are complex. The stochastic noise in the linear stochastic model is utilized to effectively parameterize the nonlinear deterministic time evolution of chaotic or turbulent dynamics (Berner et al., 2017; Branicki et al., 2018; Farrell & Ioannou, 1993; Harlim & Majda, 2008; Kang & Harlim, 2012; Li & Stechmann, 2020; A. J. Majda, 2016; A. J. Majda & Chen, 2018). The four parameters $f_{\mathbf{k}, \alpha}$, $d_{\mathbf{k}, \alpha}$, $\omega_{\mathbf{k}, \alpha}$ and $\sigma_{\mathbf{k}, \alpha}$ in the complex-valued linear stochastic model Equation 8 can be uniquely determined by utilizing the four statistics of $\hat{u}_{\mathbf{k}, \alpha}$: the mean $m_{\mathbf{k}, \alpha}$, the variance $E_{\mathbf{k}, \alpha}$, and the real and imaginary parts of the decorrelation time $T_{\mathbf{k}, \alpha} - i\theta_{\mathbf{k}, \alpha}$ (A. J. Majda & Harlim, 2012),

$$f_{k,\alpha} = \frac{(T_{k,\alpha} - i\theta_{k,\alpha})m_{k,\alpha}}{T_{k,\alpha}^2 + \theta_{k,\alpha}^2}, \quad d_{k,\alpha} = \frac{T_{k,\alpha}}{T_{k,\alpha}^2 + \theta_{k,\alpha}^2}, \quad \omega_{k,\alpha} = \frac{\theta_{k,\alpha}}{T_{k,\alpha}^2 + \theta_{k,\alpha}^2}, \quad \sigma_{k,\alpha} = \sqrt{\frac{2E_{k,\alpha}T_{k,\alpha}}{T_{k,\alpha}^2 + \theta_{k,\alpha}^2}}. \quad (9)$$

There are two scenarios in practice. First, if the true forecast model is known, then the above four statistics can be calculated based on a long simulation from the model. With these statistics in hand, the relationships in Equation 9 are utilized to determine the parameters in the linear stochastic model Equation 8 associated with each spectral mode. Second, if the true forecast model is unknown, then an iterative method can be adopted to alternate between estimating the parameters in the linear stochastic model and the DA. This is achieved by starting with a random guess of the parameters to recover the flow field driven by the set of linear stochastic models. It is followed by updating the parameter values from the recovered time series of each spectral mode. Repeat this procedure until the convergence. See (Chen & Fu, 2023) for details. Typically, the estimation of the large-scale modes is relatively accurate since the dynamics are more deterministic, and observations are more effective in reflecting large-scale motions. In contrast, the small-scale modes are more turbulent and contain high frequencies. Therefore, the estimation usually contains more uncertainty, resulting in relatively large noise coefficients in the stochastic forecast model.

The mathematical framework of modeling the random flow field with linear stochastic models characterizing the time series of spectral modes in Equations 7–8 has been widely used to describe various turbulent flow fields, including the rotating shallow water equation (Chen et al., 2015) and the quasi-geostrophic equation (Chen, 2023; Covington et al., 2022). It has been adopted as an effective surrogate forecast model in DA to approximate the Navier-Stokes equations (Branicki et al., 2018), moisture-coupled tropical waves (Harlim & Majda, 2013) and a nonlinear topographic barotropic model (Chen & Fu, 2023). Quantifying the uncertainty using the linear stochastic model as a surrogate model in the statistical forecast and filtering can be found in (Branicki et al., 2013; Chen & Fu, 2023; Chen & Majda, 2016).

In this paper, the linear stochastic models Equation 8 will be used to generate the true signal. They will also be served the forecast model of the flow field for the DA. In practice, the true system can be highly nonlinear while the calibrated linear stochastic models Equation 8 can remain as the forecast model, as discussed above. Note that the observational processes in both Eulerian and Lagrangian DA are nevertheless highly nonlinear. Therefore, the overall DA remains nonlinear.

The strong nonlinearity in DA can be seen as follows. Eulerian DA consists of the observational process of the statistical moments, that is, the momentum Equation 6 in this work, and the underlying flow field driven by a set of stochastic Equation 8. They are linked by the spectral representation of the flow field Equation 8. The coupled DA system is nonlinear due to the terms on the left-hand side of Equation 6. Similarly, the Lagrangian observations are driven by the particle system Equation 1. Notably, once the flow velocity $\mathbf{u}$ in Equation 1 is written explicitly by Equation 7, the state variable $\mathbf{x}$ of the Lagrangian observations appear in the exponential function on the right-hand side of the governing equation. Therefore, both the Eulerian and Lagrangian DA parts are strongly nonlinear.

As a final remark, it is worth highlighting that the stochastic surrogate models are primarily utilized to accelerate the DA. Nevertheless, developing these models can be considered separate from building the hybrid Lagrangian and Eulerian DA framework. The spectral representation in Equation 7 and the subsequent stochastic models for the spectral modes Equation 8 are mostly appropriate when the domain has periodic boundary conditions and the forecast statistics of the original governing and the surrogate model are similar in the spectral space. For problems with complicated geometry or model structures, ensemble approximations, and other methods can be utilized to obtain the forecast statistics. Although the closed analytic formulas for finding the DA solution may not be applicable, the general framework of using the hybrid Lagrangian and Eulerian DA schemes to handle the multiscale properties of nature still holds. See Section 7 for additional discussions.

2.4. The LEMDA Framework With Closed Analytic Solution

The focus of this paper is on developing a continuous-in-time DA scheme that fits into the LEMDA framework. Despite the strong nonlinearity, the posterior solutions of both Eulerian and Lagrangian DAs, after applying the stochastic surrogate model Equations 7–8 to describe the flow field, can be written down using closed analytic

formulas. This crucial feature can be seen by noticing that the two DA systems can both be written in the following abstract form.

$$\frac{d\mathbf{Z}(t)}{dt} = \mathbf{F}_Z(\mathbf{Z}, t) + \mathbf{A}(\mathbf{Z}, t)\mathbf{U}(t) + \boldsymbol{\Sigma}_Z \dot{\mathbf{W}}_Z(t), \quad (10a)$$

$$\frac{d\mathbf{U}(t)}{dt} = \mathbf{F}_U(\mathbf{Z}, t) + \boldsymbol{\Lambda}\mathbf{U}(t) + \boldsymbol{\Sigma}_U \dot{\mathbf{W}}_U(t), \quad (10b)$$

where $\mathbf{F}_Z(\mathbf{Z}, t)$, $\mathbf{F}_U(\mathbf{Z}, t)$ and $\mathbf{A}(\mathbf{Z}, t)$ are nonlinear functions of $\mathbf{Z}$, $\boldsymbol{\Sigma}_Z$ and $\boldsymbol{\Sigma}_U$ are noise coefficients, and $\dot{\mathbf{W}}_Z(t)$ and $\dot{\mathbf{W}}_U(t)$ are standard Gaussian white noises (zero mean and identity covariance). These vectors are defined as the resulting vectors from the Lagrangian or Eulerian DA systems. In Lagrangian DA, $\mathbf{Z}$ is the collection of the L Lagrangian tracer trajectories. Therefore Equation 10a contains the $2L$ equations in the form of Equation 1a, describing x_l and y_l for $l = 1, \dots, L$. The state variable $\mathbf{U}$ is the collection of the spectral modes $\hat{u}_{k,\alpha}$ of the flow field driven by Equation 8 and the velocity v_l of different tracers Equation 1b. In the Eulerian DA part, $\mathbf{Z}$ becomes the observed statistical moments, for example, the momentums $\langle \rho v \rangle$, of the tracers. In other words, the governing equation of $\mathbf{Z}$ is the discrete form of Equation 6 at discrete mesh grids, where N_x and N_y grids are utilized in the x and the y directions, respectively. The term $\mathbf{F}_Z(\mathbf{Z}, t)$ contains the second term on the left-hand side of Equation 6 after discretization. The term $\mathbf{A}(\mathbf{Z}, t)\mathbf{U}(t)$ corresponds to the right-hand side of Equation 6. Additional random noise is added to the resulting equations to compensate for the discretization error and observational error. The latter is mainly due to computing these statistics based on a finite number of particles. See Supporting Information S1 for details. The state variable $\mathbf{U}$ contains the underlying flow velocity driven by Equation 8 and is the variable for state estimation. The variable $\mathbf{Z}$ is the observation, which is not updated.

The DA problems in Equation 10 are highly nonlinear in terms of the joint variables $(\mathbf{Z}, \mathbf{U})$. Nevertheless, conditioned on the observational variable $\mathbf{Z}$, the equations of the unobserved states $\mathbf{U}$ that contain the underlying flow field are linear. Notably, the associated conditional distribution of $\mathbf{U}$ is precisely the DA solution. Due to the conditional linearity, given one realization of the observed variable $\mathbf{Z}(s \leq t)$, the filtering posterior distribution $p(\mathbf{U}(t)|\mathbf{Z}(s \leq t))$ of the DA problem in Equation 10 is conditionally Gaussian, where the time evolutions of the conditional mean $\boldsymbol{\mu}$ and the conditional covariance $\mathbf{R}$ have the following closed analytic formulas (Chen & Majda, 2018; Liptser & Shiryaev, 2013)

$$\frac{d\boldsymbol{\mu}}{dt} = (\mathbf{F}_U + \boldsymbol{\Lambda}\boldsymbol{\mu}) + \mathbf{R}\mathbf{A}^*(\boldsymbol{\Sigma}_Z\boldsymbol{\Sigma}_Z^*)^{-1}\left(\frac{d\mathbf{Z}}{dt} - (\mathbf{F}_Z + \mathbf{A}\boldsymbol{\mu})\right), \quad (11a)$$

$$\frac{d\mathbf{R}}{dt} = \boldsymbol{\Lambda}\mathbf{R} + \mathbf{R}\boldsymbol{\Lambda}^* + \boldsymbol{\Sigma}_U\boldsymbol{\Sigma}_U^* - \mathbf{R}\mathbf{A}^*(\boldsymbol{\Sigma}_Z\boldsymbol{\Sigma}_Z^*)^{-1}\mathbf{A}\mathbf{R}, \quad (11b)$$

with $*$ being the complex conjugate transpose. The derivation from Equations 10 and 11 can be found in (Liptser & Shiryaev, 2013). The continuous filtering solution in Equation 11 can be regarded as a generalized nonlinear version of the Kalman-Bucy filter (Bucy & Joseph, 2005; Jazwinski, 2007; Kalman & Bucy, 1961), which updates the variable $\mathbf{U}$ using Equation 11 by observing the time series $\mathbf{Z}$. Nevertheless, the nonlinear nature of the coupled system leads to a random Riccati equation for solving the covariance matrix. In addition, the gain matrix $\mathbf{R}\mathbf{A}^*(\boldsymbol{\Sigma}_Z\boldsymbol{\Sigma}_Z^*)^{-1}$ is time-dependent due to the nonlinearity in the observational process. These properties distinguish the formulas in Equation 11 from the standard Kalman-Bucy filter and facilitate the filtering solution to better capture turbulent features triggered by strong nonlinearity. It is worth highlighting the following. Both $\mathbf{Z}$ and $\mathbf{U}$ are the state variables in the coupled model Equation 10 regardless of under the Eulerian or the Lagrangian descriptions. However, when applying DA, the state variable for update is only $\mathbf{U}$. The variable $\mathbf{Z}$ can be regarded as the observational variable, which is no longer a variable for state estimation but a nonlinear transformation of the actual state variables $\mathbf{U}$ for estimation. Again, this aligns with the Kalman-Bucy filter, even though the framework here extends to the nonlinear situation.

Besides filtering, analytic formulas are available for the posterior distribution of the nonlinear smoother (Chen, 2020), which is widely used to dynamically interpolate the states and obtain the reanalysis data

(Einicke, 2012; Särkkä & Svensson, 2023). This paper focuses on the filtering, but LEMDA can be easily extended to seek the nonlinear smoother solution.

With these analytic formulas for the posterior estimates, ensemble methods are no longer needed in solving the DA problems. Therefore, localization, covariance inflation, and other tuning procedures can be avoided. This is one of the major computational advantages if the coupled system can be written into the above form, which facilitates exact, accurate, and robust solutions. In addition, with the analytic formulas, computational efficiency is significantly improved.

2.5. Summary of the LEMDA Algorithm

This subsection summarize the LEMDA algorithm with a list of steps.

- Step 1 (Lagrangian observations). Obtain a set of Lagrangian trajectories as the direct observations. In practice, these observations are from the satellite images, buoys, or other measurements. In the test below using synthetic data, the Lagrangian trajectories are generated from the coupled tracer-flow model using Equation 1 and a stochastic model representing the underlying flow field.
- Step 2 (Eulerian mesh grids). Determine the Eulerian mesh grids.
- Step 3 (Eulerian observations). Within each Eulerian grid cell, compute Eulerian quantities, such as $\langle \rho v \rangle$ in Equation 6. This can be done by taking the statistical average of the quantities of interest based on the Lagrangian trajectories.
- Step 4 (Forecast models). Develop a surrogate forecast model for DA. The given governing equations of the flow field can be used directly as the forecast model if it satisfies the structure required in the next step. Otherwise, a special surrogate model, such as those in Equation 7–8, is preferred to accelerate computational efficiency.
- Step 5 (Coupled observation-flow system). Write down the coupled observation-flow system for both Eulerian and Lagrangian DAs in the form of Equation 10.
- Step 6 (DA) Apply the closed analytic formulas in Equation 11 for estimating the state of U given the observations Z from Step 1 and Step 3.

As a remark, the analytic formulas in Equation 11, which requires the model to have a special structure Equation 10, aims to accelerate finding the DA solution. The choice of such a special DA scheme with analytic formulas for the DA solution facilitates an understanding of the hybrid strategy in the LEMDA framework. These analytic solutions avoid using ensemble methods, eliminating the sampling error in the study. Nevertheless, if the forecast model has a general nonlinearity, ensemble-based DA methods can still be applied for the corresponding hybrid Eulerian and Lagrangian DA (namely, using ensemble methods to estimate the prior and posterior mean and covariance instead of exploiting the analytic formulas in Equation 11). The overall structure from Step 1 to Step 4 remains valid. Therefore, the hybrid Eulerian and Lagrangian DA framework is not limited to the special DA scheme used in this paper. Notably, LEMDA is not intended to compete with the ensemble-based DA. Instead, the ensemble-based DA can be incorporated into the LEMDA framework to find the posterior distribution. Section 7 will include more discussions on such a topic.

Finally, Table 1 includes a summary of the symbols used in the LEMDA and the associated models.

3. Assessing the Skill of DA

3.1. Time Series Measurements for Assessing the Skill in the Posterior Mean Estimate

Two time series measurements are utilized to quantify the skill of DA. They are the root-mean-square error (RMSE) and the pattern correlation (Corr) (Hyndman & Koehler, 2006). These two measurements compare the posterior mean and the truth. Since the truth and the posterior mean are time series, the measurements are called “time series measurements.” Given two vectors $\mathbf{a} = (a_1, a_2, \dots, a_I)$ (posterior mean) and $\mathbf{b} = (b_1, b_2, \dots, b_I)$ (truth), they are defined as.

$$\text{RMSE} = \left(\sqrt{\frac{\sum_{i=1}^I (a_i - b_i)^2}{I}} \right) / \text{std}(\mathbf{b}), \quad (12a)$$

Table 1
Symbols Used in This Work and Their Meanings

Symbol	Meaning
$\mathbf{x}_l$	Displacement of the l -th tracer
$\mathbf{v}_l$	Velocity of the l -th tracer
$\mathbf{f}_l$	External forces to the l -th tracer
$\mathbf{u}$	Velocity of the underlying flow field
$\mathbf{u}_f$	Velocity of the underlying flow field corresponding to fine-scale features
$\mathbf{u}_c$	Velocity of the underlying flow field corresponding to coarse-scale features
$\hat{\mathbf{u}}_{\mathbf{k}}$	The $\mathbf{k}$ -th spectral mode of the underlying flow field
β	Drag coefficient
σ_x	Uncertainty in the tracer displacement
$\sigma_{\mathbf{k}}$	Uncertainty in the $\mathbf{k}$ -th spectral mode of the flow field
ρ	Mass density
$\langle \rho \mathbf{v} \rangle$	Momentum of a set of tracers
n	Particle number density
t	Time
$\mathbf{Z}$	Collection of observed variables
$\mathbf{U}$	Collection of the unobserved variables
μ	Posterior mean in DA
$\mathbf{R}$	Posterior covariance in DA

$$\text{Corr} = \frac{\sum_{i=1}^I ((a_i - \text{mean}(\mathbf{a})) (b_i - \text{mean}(\mathbf{b})))}{\sqrt{\sum_{i=1}^I (a_i - \text{mean}(\mathbf{a}))^2} \sqrt{\sum_{i=1}^I (b_i - \text{mean}(\mathbf{b}))^2}}, \quad (12b)$$

where $\text{mean}(\mathbf{a})$ and $\text{mean}(\mathbf{b})$ are the mean value of the vectors $\mathbf{a}$ and $\mathbf{b}$, respectively, and $\text{std}(\mathbf{b})$ is the standard deviation of $\mathbf{b}$. Note that the two vectors can be two time series or the rescaled vectors from matrices describing the solution at different spatial points based on snapshots. The RMSE is non-negative. A more skillful posterior mean estimate corresponds to a smaller RMSE. Note that the RMSE defined in Equation 12a has already been normalized by the standard deviation of the truth. Therefore, when RMSE is above 1, the error in the posterior mean estimate is larger than the mean of the truth. The pattern correlation gives a number between -1 and 1 . A larger value of the pattern correlation corresponds to a more similar profile between the two vectors. Usually, a skillful estimate has a pattern correlation larger than 0.5 . In this paper, the DA solution in spectral space is first projected to the two-dimensional physical space. The two-dimensional data field is reshaped to be stored in a vector. Then, these skill scores are applied to the two vectors of the truth and the posterior mean estimate at each time instant. Either the time series of the skill scores or the time average value will be shown in the following sections.

3.2. An Information Measurement for Assessing the Skill in the Posterior Distribution

Another useful measurement aims to quantify the uncertainty reduction in DA, which uses the information from the entire posterior distribution. The total uncertainty in DA can be quantified by computing the difference between the posterior and the prior distributions. Here, the prior distribution is the statistical equilibrium (namely the background) state of the forecast model, which does not involve additional information from observations. The relative entropy, which is an information criterion, can be utilized to assess the uncertainty reduction (or information gain) in the posterior distribution $p(t)$ compared with the prior one p_{eq} (e.g., the statistical equilibrium or the background distribution) (Branicki et al., 2013; Chen et al., 2014; DelSole, 2005; Giannakis & Majda, 2012; Kleeman, 2002, 2011; A. Majda et al., 2005).

The relative entropy is defined as follows:

$$\mathcal{P}(p(t), p_{eq}) = \int p(t) \log \left(\frac{p(t)}{p_{eq}} \right), \quad (13)$$

which is also known as Kullback-Leibler divergence or information divergence (Kullback, 1987; Kullback & Leibler, 1951). One practical setup in many applications arises when both the distributions involve only the mean and covariance so that $p(t) \sim \mathcal{N}(\bar{\mathbf{u}}, \mathbf{R})$ and $p_{eq} \sim \mathcal{N}(\bar{\mathbf{u}}_{eq}, \mathbf{R}_{eq})$ are Gaussian distributions. In such a case, $\mathcal{P}(p(t), p_{eq})$ has an explicit formula

$$\mathcal{P}(p(t), p_{eq}) = \left[\frac{1}{2} (\bar{\mathbf{u}} - \bar{\mathbf{u}}_{eq})^* \mathbf{R}_{eq}^{-1} (\bar{\mathbf{u}} - \bar{\mathbf{u}}_{eq}) \right] + \left[-\frac{1}{2} \log \det(\mathbf{R} \mathbf{R}_{eq}^{-1}) + \frac{1}{2} (\text{tr}(\mathbf{R} \mathbf{R}_{eq}^{-1}) - |\mathcal{K}|) \right], \quad (14)$$

where “det” and “tr” are the determinant and the trace of a matrix and $|\mathcal{K}|$ is the dimension of the system. In Equation 14, the first term in brackets is called “signal,” reflecting the information difference in the mean but weighted by the inverse of the equilibrium variance, $\mathbf{R}_{eq}$, whereas the second term in brackets, called “dispersion,” involves only the information distance regarding the covariance ratio, $\mathbf{R} \mathbf{R}_{eq}^{-1}$.

Despite the lack of symmetry, the relative entropy (Equation 13) has two attractive features. First, $\mathcal{P}(p(t), p_{eq}) \geq 0$ with equality if and only if $p(t) = p_{eq}$. Second, $\mathcal{P}(p(t), p_{eq})$ is invariant under general nonlinear changes of variables (Cai et al., 2002). The signal and dispersion terms in Equation 14 are individually invariant under any (linear) change of variables which maps Gaussian distributions to Gaussians. In the LEMDA framework with linear stochastic forecast models, both the prior and posterior distributions are Gaussian. Thus, the formula in Equation 14 will be used to quantify the uncertainty reduction. Since $\mathcal{P}(p(t), p_{eq})$ is time-dependent, the skill score shown below is its time average.

4. Lagrangian DA

Let us start understanding the Lagrangian component in the LEMDA framework. The results in this Section are purely based on the Lagrangian DA part.

4.1. Performance of Lagrangian DA

The underlying flow model is given by Equations 7–8 with double periodic boundary conditions. The flow field is assumed incompressible, and no mean background flow is included. The incompressibility can be imposed on the eigenvectors in Equation 7 (Chen et al., 2014). The parameters adopted in the following experiments are:

$$d_k = 1, \quad \omega_k = 0, \quad f_k = \mathbf{0} \quad \text{and} \quad \sigma_k = 0.05, \quad (15)$$

for all $\mathbf{k}$ such that the flow field has an equipartition of the energy that generates turbulent features. The range of the spectral modes is $\mathbf{k} \in [-4, 4]^2$. Excluding the background mode $\mathbf{k} = (0, 0)$, there are in total 80 modes. Translating the non-dimensional flow field to the dimensional form, the choice of the parameters corresponds to a flow field in the square domain with size being 1,000 km $\times$ 1,000 km, e-folding time being about 20 days, and the order of velocity being 0.05 m/s (which corresponds to an amplitude of 0.5 in the non-dimensional form as shown in all the figures). This mimics the typical ocean flow field (Taylor & Thompson, 2023).

The parameters in the tracer dynamics (Equation 1) are taken as

$$\sigma_x = 0.001 \quad \text{and} \quad f_l = \mathbf{0}. \quad (16)$$

In such a setup, tracers do not suffer from collision forces. This mimics the standard cases for drifter motions in mid-latitude oceans. Section 5 will include a comparison of the Eulerian and Lagrangian DA parts in the presence of collision forces. The noise level here represents the observational error in identifying the locations of tracers in the satellite image, for example, the sea ice floes (Covington et al., 2022). Converting to the dimensional values, this noise induces an error of about 160 m per day, which is of the order of one-pixel distance (typical 250 m)

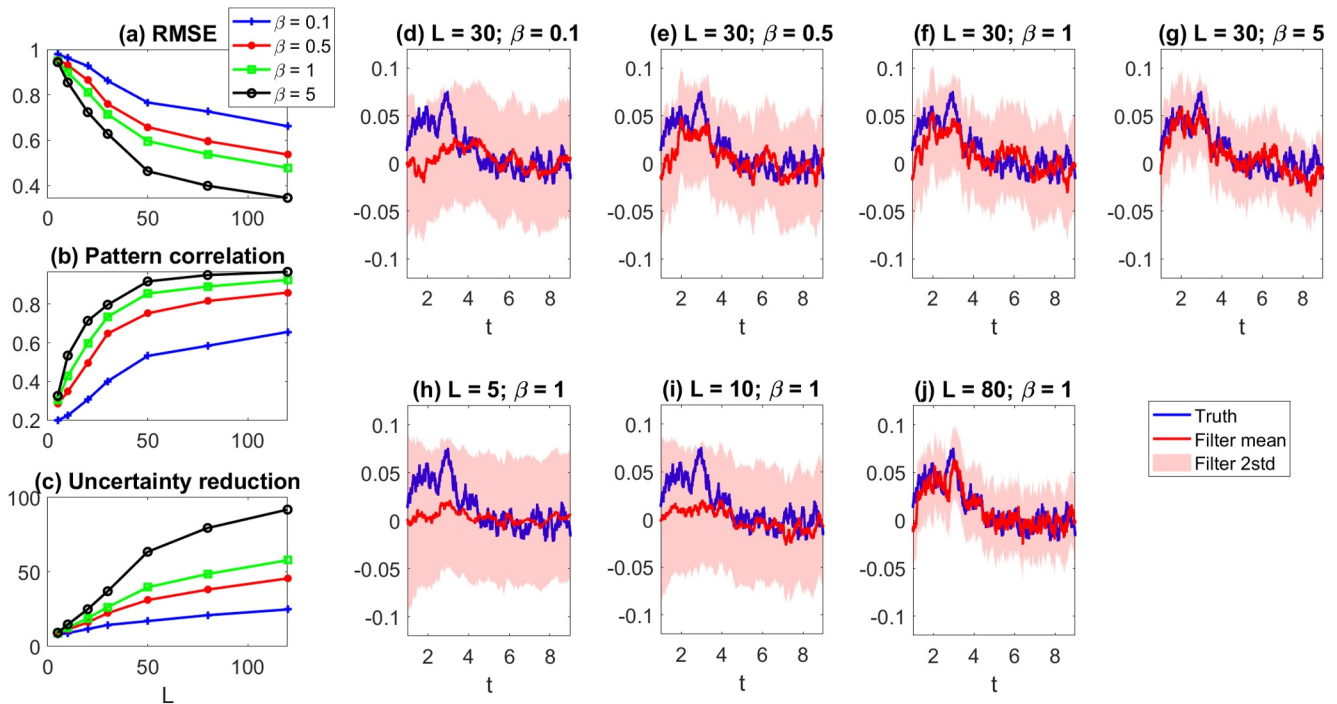


Figure 2. Lagrangian DA with parameters in Equations 15 and 16. Panels (a)–(c) show the RMSE Equation 12a, pattern correlation Equation 12b and uncertainty reduction Equation 14 as a function of the tracer numbers L (x -axis) and drag coefficient β (different curves). Panels (d)–(j) show the real part of the posterior estimate of the mode $(-4, -4)$ with different L and β . Note that in this and the subsequent figures, all flow fields are shown in the non-dimensional units.

(Manucharyan et al., 2022) and is a reasonable choice as the measurement error. The drag coefficient $\beta = 1$ is within the range for typical ice floes in the Arctic region (Damsgaard et al., 2018), and the DA skill of different values of β will be tested in the following numerical simulations. The initial distribution of tracers is uniform. The numerical integration time step is $\Delta t = 0.0001$ (86.4 s). The DA is computed for $t \in [0, 10]$, and the skill scores are calculated based on the solution within the time interval $t \in [1, 10]$ to eliminate the effect from the burn-in period.

Figures 2 and 3 present the Lagrangian DA results. Panels (a)–(c) of Figure 2 show the RMSE (Equation 12a), pattern correlation (Equation 12b) and uncertainty reduction (Equation 14) as a function of the tracer numbers L (x -axis) and drag coefficient β (different curves). There are two main conclusions. First, the error in the posterior mean decreases when the number of observations L increases. Similarly, when L becomes large, the uncertainty reduction becomes more significant, which means the posterior uncertainty shrinks, and the posterior mean estimate is more confident. Second, the DA skill depends on the drag coefficient β . A smaller drag coefficient leads to a decline in the skill in the state estimation of the flow field. This is not surprising since a smaller β implies a smaller signal-to-noise ratio in (Equation 1b). In other words, the observability suffers from the decrease of β when DA applies to recover the flow field. Panels (d)–(j) show the posterior estimate of the mode $(-4, -4)$. Figure 3 shows the snapshots of the flow field in physical space at $t = 8$ with different L and β . The results in Figure 3 support the conclusions drawn from the skill scores in Panels (a)–(c) of Figure 2.

4.2. Reduced-Order Lagrangian DA

Despite the analytic formulas (11), potential computational challenges in Lagrangian DA may appear when the dimension of the flow field becomes large. Recall that the posterior covariance matrix scales as a quadratic function of the dimension of the flow, which is the total number of spectral modes in this paper. Therefore, the major computational burden comes from solving the matrix Riccati Equation 11b for the time evolution of the posterior covariance. Effective approximations in accelerating the estimation of the posterior covariance will significantly facilitate the entire DA procedure.

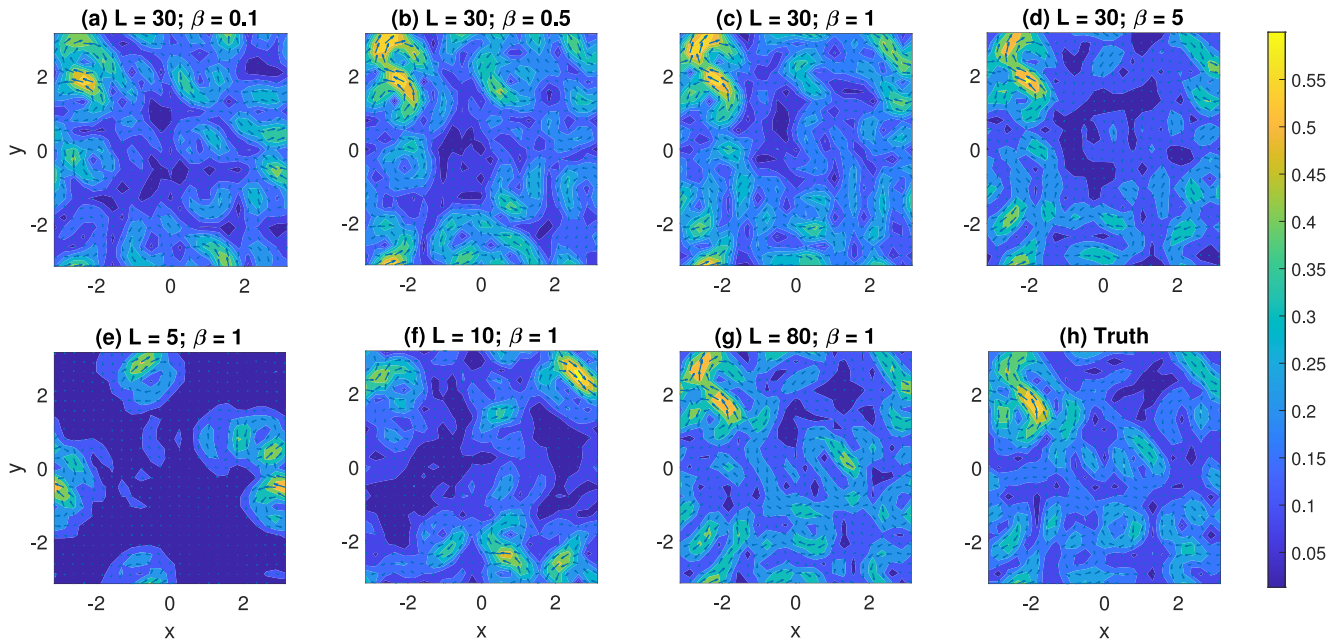


Figure 3. Lagrangian DA with parameters in Equations 15 and 16. Panels (a–g) show the flow field at $t = 8$ with different L and β . Panel (h) includes the truth. The velocity with an amplitude of 0.5 corresponds to 0.05 m/s.

The computational strategy developed below is not specifically designed for LEMDA but applies to other DA methods under the same conditions. If the tracers are nearly uniformly distributed in the domain, then an approximate scheme for reducing the computational cost in solving the posterior covariance matrix can be derived. Note that if the flow field is nearly incompressible and the drag coefficient β is not too small, then the tracers are unlikely to be highly concentrated in a particular area (Chen et al., 2014). Therefore, such a requirement of the tracer distribution is reasonable in many ocean and fluid applications. One essential basis for reaching this approximate scheme is the following. Although different spectral modes are fully coupled in the observational processes via (Equation 7), the linear stochastic models of these modes are by design independent in the forecast model (Equation 10b). This is one of the crucial features of DA in spectral space, where the strong spatial dependence is primarily described by the spatial bases. This conclusion generally holds even when the governing equations of the spectral modes are coupled due to being decomposed from a complicated partial differential equation (A. J. Majda & Harlim, 2012). Therefore, one natural strategy for developing approximate DA is to set the part associated with the flow modes in the posterior covariance (Equation 11b) to be a diagonal matrix. This significantly reduces the computational cost, since only the diagonal entries need to be updated. Furthermore, these diagonal components are near constants in Lagrangian DA and can be predetermined. With this reduced-order strategy, the computational cost scales as a linear function of the state variable, while the cost for full DA is proportional to a quadratic function of the number of state variables due to solving the full covariance matrix.

Denote by R_k one of the diagonal components of the covariance in (11b) associated with one spectral mode of the flow field. By applying the mean-field theory (Chaikin et al., 1995) with the assumption that the tracers are nearly uniformly distributed in the field, its governing equation can be written as

$$\frac{dR_k}{dt} = -2d_k R_k + \sigma_k^2 - \sigma_x^{-2} \xi^2 L R_k^2, \quad (17)$$

where ξ is a parameterization that can be derived. The positive root of R_k in Equation 17 corresponds to the constant approximate solution of the covariance, which is given by

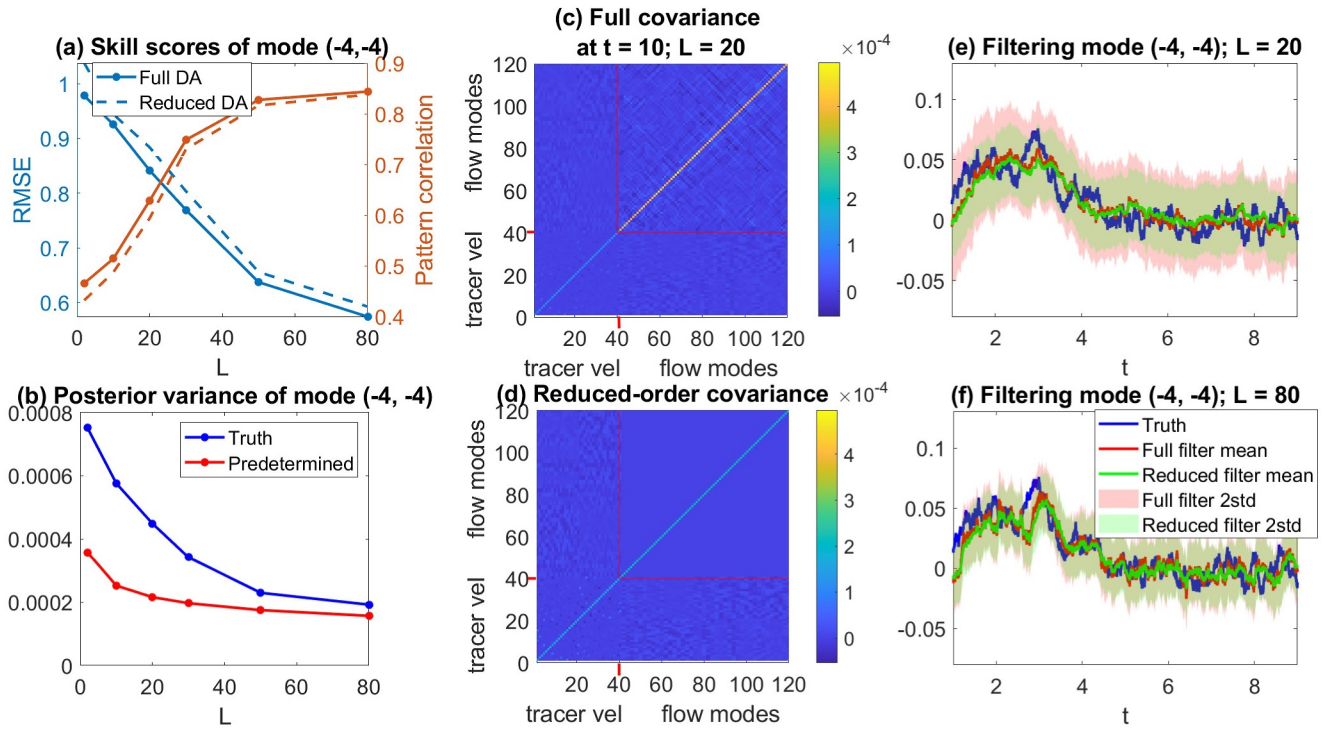


Figure 4. Lagrangian DA and its reduced-order form, where $\beta = 1$ is adopted. Panel (a) compares the skill scores of the full and reduced-order DA solutions in terms of the normalized RMSE and the pattern correlation of mode $(-4, -4)$. Panel (b) shows the diagonal entry of the posterior variance of mode $(-4, -4)$ resulting from the full DA scheme averaged over the time interval $[1, 10]$ with that given by the predetermined value in Equation 18. Panel (c) shows the posterior covariance from the filtering solution at the last time instant $t = 10$ with $L = 20$ tracers, which is diagonal dominant. Panel (d) shows the posterior covariance when the reduced-order DA scheme is applied. Panels (e)–(f) demonstrate the posterior mean time series and the posterior uncertainty for mode $(-4, -4)$. The posterior uncertainty is given by the two standard deviations from the mean value.

$$R_k = \frac{-d_k + \sqrt{d_k^2 + \sigma_x^{-2} \xi^2 L \sigma_k^2}}{\sigma_x^{-2} \xi^2 L}. \quad (18)$$

See Appendix A for details. The implication of the result in Equation 18 is that the covariance of the flow field no longer needs to be solved via solving the matrix Riccati equation, which significantly reduces the computational cost.

Figure 4 shows the results of Lagrangian DA using the exact analytic solution in Section 2.4 and its reduced-order form in Equation 18, where the latter exploits a pre-determined constant diagonal matrix for the posterior covariance of the part associated with the underlying flow field. Panel (a) compares the skill scores of the full and reduced-order DA solutions. The two DA schemes behave similarly, especially when the number of tracers increases. This is consistent with the condition of the mean-field theory appearing in the derivation shown in Appendix A. The result here demonstrates the skillful behavior of the cheap reduced-order DA scheme. Panel (b) shows the diagonal entry of the posterior variance of mode $(-4, -4)$ resulting from the full DA scheme averaged over the time interval $[1, 10]$ with that given by the predetermined value in Equation 18. Note that since the equipartition of the energy is adopted here, other modes behave similarly to this one. The predetermined diagonal entry becomes more consistent with the actual value when L increases. This also provides evidence that the two DA schemes lead to comparable results. Panel (c) shows the posterior covariance from the filtering solution at the last instant $t = 10$ with $L = 20$ tracers. In such a case, the covariance associated with the flow field (the top right part with the boundaries given by the thin red lines) is already diagonally dominant. Therefore, once the reduced-order DA is applied in Panel (d), the results of the other part of the covariance matrix remain similar. Finally, Panels (e)–(f) demonstrate the posterior mean time series and the posterior uncertainty for mode $(-4, -4)$. Even

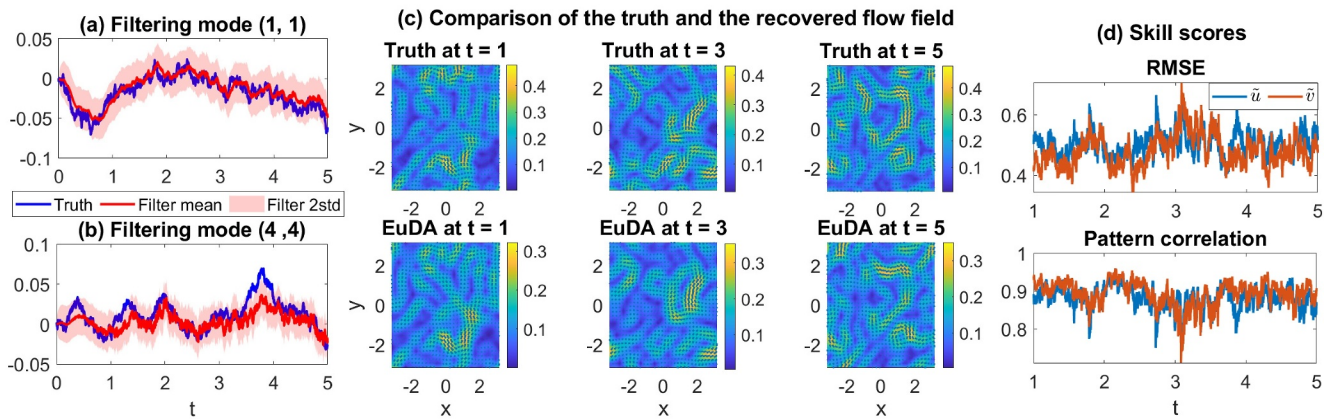


Figure 5. Eulerian DA of the flow field using without particle collisions. Panels (a)–(b) compare the filtering results with the truth for two Fourier modes (1,1) and (4,4). Panel (c) shows the comparison of the truth and the recovered velocity field in physical space at $t = 1, 3$ and 5 . Panel (d) shows the skill scores of the velocity field in physical space.

with $L = 20$, the solutions are close to each other. When L increases to $L = 80$, the two solutions almost overlap and approach the truth.

5. Eulerian DA

This section includes the Eulerian DA results with two different scenarios. In the first case, the collision force is omitted, and the goal is to understand the skill of Eulerian DA using the continuum equations derived in Section 2.2 as the forecast model. In the second scenario, ice floes are utilized as tracers. Therefore, collisions due to the floe-floe contact forces play a vital role in driving the motions of the floes, which significantly affects the direct Lagrangian DA. The skill of Eulerian DA in recovering large-scale features of the flow field will be illustrated.

5.1. Eulerian DA With No Particle Collisions

Let us start with the same case as in Section 4, where the flow and the particle models are equipped with the parameters in Equations 15–16 with 80 spectral modes. In addition, the drag force is $\beta = 1$.

The performance of Eulerian DA with a rather coarse grid of size 9×9 (or equivalently 111 km between two nearby grids) for recovering the velocity of the underlying flow field is studied. A total of $L = 2000$ particles in the flow field are utilized to compute the averaged momentums of particles in each grid cell as observational quantities. On average, each grid cell has about 25 particles. As will be seen later, this number is sufficient to reach a skillful DA result. It is worth highlighting that the computational cost of calculating these averaged momentums is nearly negligible compared with using a large number of particles as the direct observations for Lagrangian DA. Therefore, there should not be any concerns regarding using such a large number of particles (if available) in Eulerian DA from the computational cost aspect. The particle velocity is computed using a finite difference scheme based on the displacement in the two consecutive time instants that differed by Δt . The velocity of each particle suffers from random noise, as seen in Equation 1a. Nevertheless, the statistical average will eliminate the effect of the random noise to a large extent. In Eulerian DA, the observed variables are the momentums in these grid cells. They are denoted by $\mathbf{Z}$ in Equation 10a. For a grid of size 9×9 , the noise in the observational processes $\mathbf{Z}$ is given by $\sigma_z = 0.02$, where σ_z is the diagonal component of the identity matrix Σ_z in Equation 10a. The value σ_z is computed based on the procedure described in Supporting Information S1. As in the Lagrangian DA case, the tracers are uniformly distributed over the domain $[-\pi, \pi]^2$ at the initial time instant.

Figure 5 presents the numerical results. The observations of the momentums are computed based on all the $L = 2000$ observed particle and Eulerian observations are taken at a 9×9 mesh grid. Panels (a)–(b) show the filtering results in recovering two Fourier modes. The posterior mean captures the patterns of the truth quite well with relatively moderate uncertainty. Panel (c) compares the truth and the recovered velocity field in physical space at $t = 1, 3$ and 5 . These results further provide an intuitive understanding that Eulerian DA is skillful in

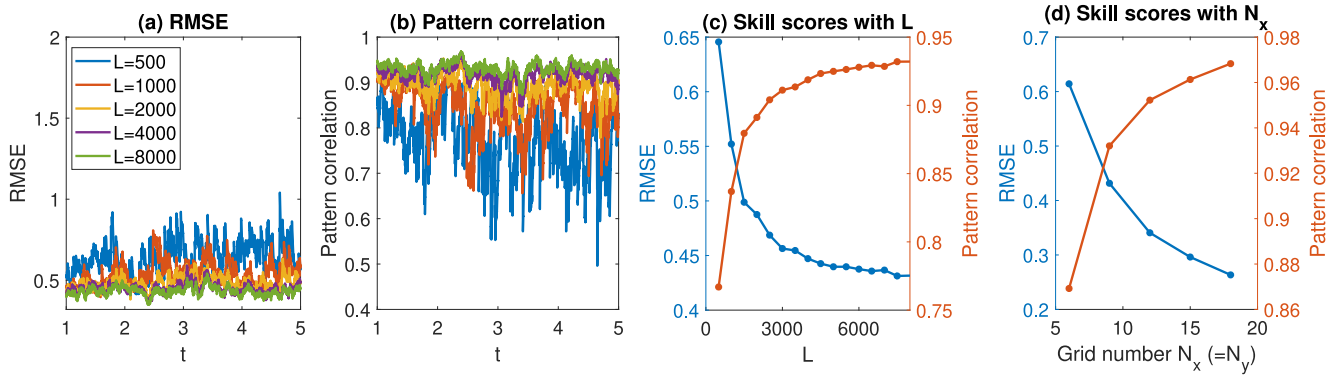


Figure 6. Skill scores of Eulerian DA with different number of particles L and grid size $N_x = N_y$. Panels (a)–(b) show the RMSE and pattern correlation as a function of time using different L , where the mesh size $N_x = N_y = 9$ is fixed. Panel (c) shows the skill scores average over time as a function of L . Panel (d) shows the skill scores as a function of $N_x = N_y$, where $L = 8000$ is used.

capturing the features in the true flow field. Panel (d) shows the skill scores of the velocity field in physical space, which are quantitative indicators that support the findings in the previous panels. Notably, the EuDA, even with such a coarse grid, can perform well with a pattern correlation over 0.85.

Figure 6 provides an understanding of the impacts of the number of particles L and the grid sizes on the performance of Eulerian DA. For simplicity, equal numbers of the grid points in the x and the y directions are taken, namely $N_x = N_y$. Panels (a)–(b) show the time evolutions of the RMSE and the pattern correlation using different L , where the mesh size $N_x = N_y = 9$ is fixed. The time-averaged skill scores are shown in Panel (c). As expected, the skill scores improve as the number of particles used to calculate the observational quantities (i.e., the momentums) at fixed grid cells increases. The least accurate case in these panels corresponds to $L = 500$. Since a 9×9 mesh grid is adopted, each cell has, on average, only six particles. This reduces the accuracy of DA. In particular, the cells contain no particles from time to time. In such a case, the observed values are interpolated using the computed momentum values in the nearby grids. The details are contained in Supporting Information S1. Nevertheless, even using such a small number of particles to calculate the momentums, the pattern correlation is above 0.7, and the RMSE is, on average, below 0.8, which means the DA remains relatively skillful. This indicates the robustness of the method. Next, Panel (d) shows the skill scores as a function of $N_x = N_y$, where $L = 8000$ is used. The skill scores improve as the grid size is refined. Notably, when the grid size is as coarse as 6×6 , the pattern correlation remains above 0.85. This provides a reasonable justification for using Eulerian DA to recover the large-scale features of the underlying flow field.

5.2. Recovery of Large-Scale Flow Feature Using Eulerian DA in the Presence of Particle Collisions

Consider a case where the force in the tracer Equation 1 is nonzero, that is, $f_i \neq 0$. This appears, for example, when ice floes are adopted as tracers in recovering the underlying ocean field in the marginal ice zone (Lopez et al., 2023; Mu et al., 2018). To this end, the particles are also named ice floes in this subsection. For simplicity, the floes are given by disks of uniform size and thickness. Therefore, the force f_i characterizes the particle-particle collision force. Hooke's law of elasticity is naturally used to model this collision force (Damsgaard et al., 2018; Manucharyan & Montemuro, 2022). Decompose the total force into two parts:

$$f_i = \sum_{j \in J} (f_n^{ij} + f_t^{ij}), \quad (19)$$

where f_n^{ij} and f_t^{ij} are the normal and tangential components, respectively, from the j -th floe in the set J . The set J contains all the floes that contact the i -th floe at a certain time instant. A simplification is made. Assume that the floes only have translational momentums, but the angular momentum is omitted. This is sufficient for illustrating the key differences in Eulerian and Lagrangian DAs. In such a case, the contact forces are given by

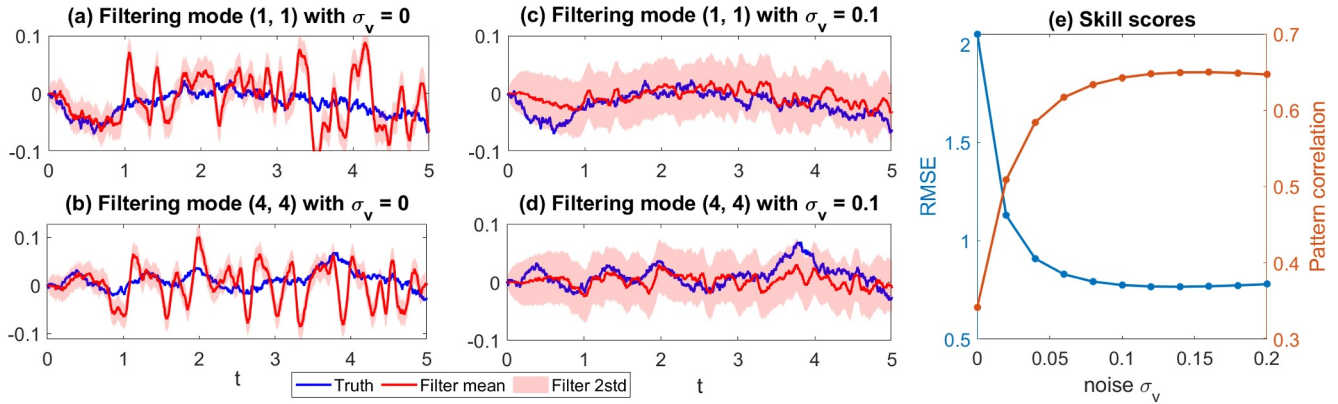


Figure 7. Lagrangian DA by randomly selecting 40 floes from in total 2000 floes in the domain. Here, the full Lagrangian DA rather than the reduced-order version is utilized. Panels (a)–(b) show the filtering results of mode (1,1) and (4,4) using the original forecast model by ignoring the contact force and without additional noise, that is, $\sigma_v = 0$ in Equation 22. Panels (c)–(d) show the results but inflate the noise in the governing equation of the floe velocity by taking $\sigma_v = 0.1$. Panel (e) shows the skill scores as a function of the level of noise inflation.

$$\begin{aligned} f_n^{lj} &= c^{lj} E^{lj} \delta_n^{lj} \mathbf{n}^{lj}, & \delta_n^{lj} &\equiv d^{lj} - 2r, & \text{with } d^{lj} &= |\mathbf{x}^l - \mathbf{x}^j|, \\ f_t^{lj} &= c^{lj} G^{lj} v_t^{lj} \mathbf{t}^{lj}, & v_t^{lj} &= (\mathbf{v}^j - \mathbf{v}^l) \cdot \mathbf{t}^{lj}, \end{aligned} \quad (20)$$

where for simplicity all the floes are assumed to have the same radius $r = 0.01\pi$ (corresponding to the same area $S = \pi r^2 = \pi^3 \times 10^{-4}$). In Equation 20, E^{lj} is the Young's bulk modulus, G^{lj} is the Young's shear modulus, and c^{lj} is the chord length in the transverse direction of the cross-sectional area (Cundall & Strack, 1979). The vectors $\mathbf{n}$ and $\mathbf{t}$ are the unit vectors along the normal and the tangential directions. The Coulomb friction law is further applied to limit the tangential contact force relative to the magnitude of the normal contact force. The restriction $|f_t^{lj}| \leq \mu^{lj} |f_n^{lj}|$ is applied, where μ^{lj} is the coefficient of friction that characterizes the condition of the surfaces of the two floes in contact. In the following simulations, $\mu^{lj} = 0.2$ is adopted. By choosing reasonable values of Young's moduli, the amplitude of the collision forces is comparable to that of the drag forces.

The same setup as that in subsection 5.1 is utilized here except that the collisions are further added in this experiment. The test regime here contains 2000 floes, which corresponds to a concentration of about 0.16. This concentration allows the floes to trigger relatively frequent contact with each other.

5.2.1. Issues in Applying Lagrangian DA

It is worth highlighting that collisions occur essentially instantaneously. Therefore, a tiny numerical integration time step is needed to resolve the collisions in the Lagrangian DA part, affecting computational efficiency. Such an issue can be avoided when Eulerian DA is adopted, which will be presented in Section 5.2.2. In addition, perfectly resolving collisions is inaccessible in practice. A small bias in describing the contact force due to the approximation in geometric properties or the collision law may lead to a significant error in the Lagrangian DA part. Even if the contact force is perfectly characterized, when all the 2000 floes are utilized as the observed trajectories, the Lagrangian DA part becomes quite expensive as the dimension of the state variable increases. To remain an efficient DA, only 40 floes are randomly selected as observations. As a majority of the floes are not included as observational variables, the contact force is hard to accurately characterize in the forecast model. Thus, it is ignored in the forecast model of the floe velocity, which simply becomes

$$\frac{d\mathbf{v}_l}{dt} = \beta(\mathbf{u} - \mathbf{v}_l). \quad (21)$$

Such a setup serves as the benchmark. Panels (a)–(b) of Figure 7 show the filtering results of mode (1,1) and (4,4) using the original forecast model in Equation 21 by ignoring the contact force. Vast oscillations are seen in the posterior mean time series. These large biases happen when collisions occur at these time instants while the

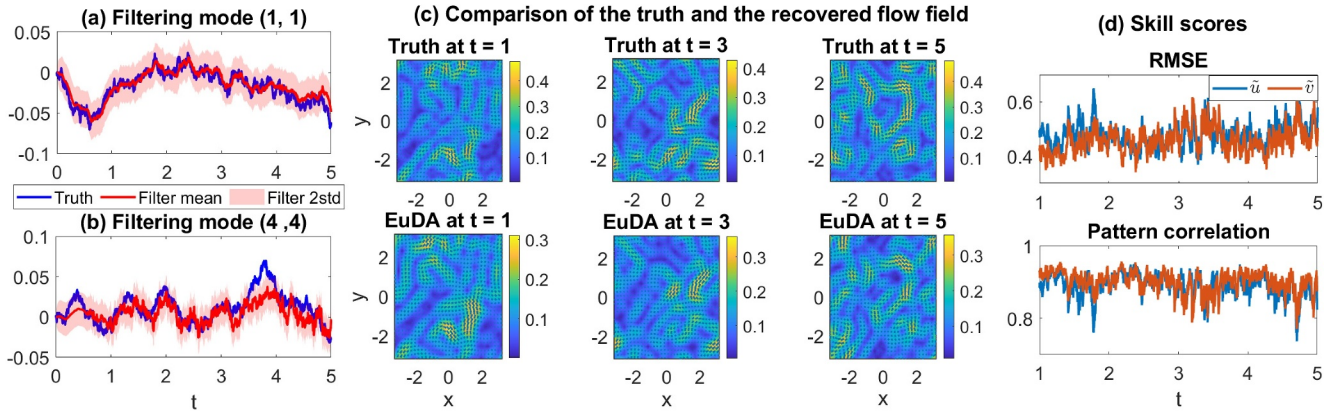


Figure 8. Eulerian DA of the flow field using particle collisions. Panels (a)–(b) compare the filtering results with the truth for two Fourier modes (1,1) and (4,4). Panel (c) shows the comparison of the truth and the recovered velocity field in physical space at $t = 1, 3$ and 5 . Panel (d) shows the skill scores of the velocity field in physical space.

contact force is ignored in the forecast model. The RMSE is twice as much as the standard deviation of the model equilibrium and the pattern correlation is only 0.35, both of which clearly indicate that the Lagrangian DA part loses its skill in recovering the underlying flow field.

One possible way to improve the results of the Lagrangian DA part is to apply the noise inflation (also known as covariance inflation) strategy, which is widely utilized in practice (Hamill & Whitaker, 2005; Silva et al., 2021). To this end, the forecast model of the floe velocity (Equation 1b) is modified by

$$\frac{dv_i}{dt} = \beta(u - v_i) + \sigma_v \dot{W}_i. \quad (22)$$

The additional noise effectively characterizes the uncertainty due to the collisions. It can be regarded as modeling the collision effects in a statistical way in the forecast model. Panels (c)–(d) show the DA results by inflating the noise in the forecast model of the floe velocity by taking $\sigma_v = 0.1$. The posterior estimate is significantly improved. Panels (e)–(f) show the skill scores as a function of the level of noise inflation, which provides a quantitative indication of the improvement of the DA skill. With noise inflation from $\sigma_v = 0.0$ to $\sigma_v = 0.1$, the pattern correlation increases from $\text{Corr} = 0.35$ to $\text{Corr} = 0.65$. Yet, the improved pattern correlation is still below the one associated with the case using a perfect forecast model ($\text{Corr} = 0.80$) shown in Figure 2. These results confirm that noise inflation can help improve the results to a large extent, but a barrier still exists by utilizing such a statistical characterization of the contact force.

5.2.2. Improving the Results Using Eulerian DA

Figure 8 shows the results using Eulerian DA based on the $L = 2000$ floes and a mesh with 9×9 grid cells. The posterior estimates from the Eulerian DA part are much more accurate than those from the Lagrangian DA part in Section 5.2.1. The underlying reason is that the effect of collisions is averaged out when the statistical moments, that is, the momentums, are computed. Notably, Eulerian DA shown in Figure 8 leads to comparable results as the case without particle collision shown in Figure 6.

Figure 9 studies the robustness of Eulerian DA in the presence of particle collisions. Panels (a)–(b) show the skill scores as a function of time with different L where the particle radii are fixed as 0.01π . The time-averaged skill scores are shown in Panel (c). Similar to the case without collision, with an increased number of floes, Eulerian DA has smaller statistical errors and leads to more accurate state estimations. Panel (d) shows the skill scores as a function of the mesh grids $N_x = N_y$, where the total number of floe is $L = 8000$, and the concentration is fixed at $c = 0.63$. The results indicate that a more refined grid helps eliminate the discretization error. Panel (e) shows the skill scores as a function of concentration, where the total number of floes $L = 2000$ is fixed, but the radius varies to reach different concentration values. Note that the maximum concentration is above $\pi/4$ as small overlaps between different floes are allowed when the elastic collision formula is applied. When the concentration

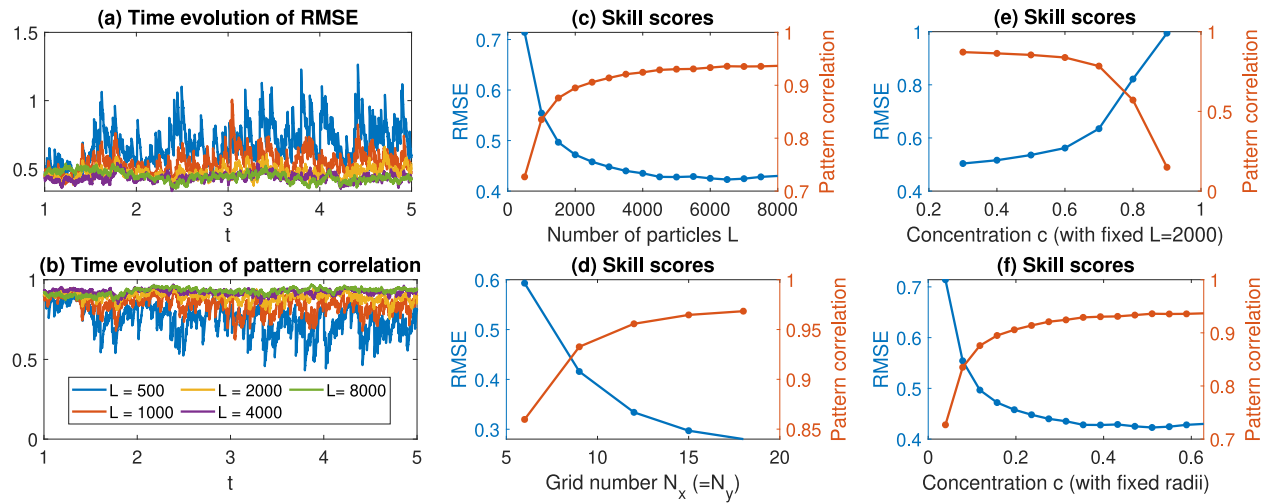


Figure 9. Skill scores of EuDA with particle collisions. Panels (a)–(b) show the skill scores as a function of time with different L . The time-averaged skill scores are shown in Panel (c). Panel (d) shows the skill scores as a function of the mesh grids $N_x = N_y$. Panel (e) shows the skill scores as a function of concentration. Panel (f) shows the skill scores as a function of the concentration, where the radius is fixed and the number of floes L varies.

increases and exceeds a certain threshold of about 0.65, the skill scores decline quickly. The reason is that when the floes are more packed, they become hard to move around. In other words, the contact force becomes more dominant than the drag force. The observability of the ocean is thus significantly affected. In the limit that the domain is fully packed, all the floes are static, which becomes useless in recovering the underlying flow field. Finally, Panel (f) shows the skill scores as a function of the concentration, where the radius is fixed and the number of floes L varies. It is seen that before arriving at the threshold of declining performance, the skill scores remain robust.

6. LEMDA for Recovering Multiscale Turbulent Flows

In this section, the two components will be combined to facilitate the DA of a multiscale turbulent flow field. The general description below assumes the particles have no collisions. In practice, non-interacting floes can naturally be selected to provide observed trajectories for implementing Lagrangian DA.

6.1. The Procedure of Applying LEMDA

LEMMA consists of two steps. The first step is to apply Eulerian DA to recover the large-scale modes in a relatively coarse grid. The second step is to apply Lagrangian DA with a small number of observed particles in each coarse grid cell to recover the small-scale modes, where the large-scale modes are treated as known modes obtained from the first step.

The following crucial features of LEMMA are worth highlighting. First, as the Eulerian DA part is only applied to a relatively coarse grid, there are sufficient numbers of particles to compute the observed statistical quantities, which allows an accurate DA result. Meanwhile, since the grid is coarse, computational efficiency is guaranteed. Second, fundamentally different from the expensive direct implementation of Lagrangian DA to the entire domain, the Lagrangian DA part in the LEMMA framework only applies to each grid cell. Notably, Lagrangian DAs at different grid cells are implemented independently. In other words, parallel computing can naturally be applied for carrying out these Lagrangian DAs, where no message passes across different processors. This parallel DA procedure using Lagrangian observations is highly efficient, especially since only a few observations are used to recover the small-scale flow structures in each grid cell.

6.2. Implementation Details

In the first step, when Eulerian DA is applied to recover the large-scale features of the flow field, the small-scale modes are treated as unresolved in DA. As the spectral method is used to characterize the flow field, the

contribution of the strong signals of the small-scale modes averaged over a coarse grid cell is insignificant overall. This allows Eulerian DA to recover the large-scale features skillfully.

In the second step, when Lagrangian DA is applied in each grid cell, the original system is written into the following form to facilitate DA.

$$\frac{dx_l}{dt} = v_l + \sigma_x \dot{W}_l, \quad (23a)$$

$$\frac{dv_l}{dt} = \beta(u_f + u_c - v_l), \quad u_f(x, t) = \sum_{k \in \mathcal{K}, |k| \geq \eta} \hat{u}_k(t) e^{ikx} r_k, \quad (23b)$$

$$\frac{d\hat{u}_k}{dt} = -d_k \hat{u}_k + f_k(t) + \sigma_k \dot{W}_k, \quad (23c)$$

Where

$$u_c(x, t) = \sum_{k \in \mathcal{K}, |k| < \eta} \hat{u}_k(t) e^{ikx} r_k$$

is the coarse-scale flow components recovered by the Eulerian DA part. In Equation 23, η is pre-defined that separates the large- and small-scale modes. The Lagrangian DA method is applied to each coarse cell to recover the small-scale modes u_f (where the subscript f stands for the fine grids). It is worth clarifying that the recovered spectral modes associated with $u_f(x, t)$ in Equation 23 in each grid cell only describe the small-scale features within that cell. To obtain the spectral modes characterizing the small-scale features in the entire domain, the local spectral modes from each Lagrangian DA are first utilized to reconstruct the flow field in physical space in each cell. Then, the recovered flow field associated with the small-scale structures in different cells is collected to reach the small-scale flow field in physical space. A new spectral decomposition of this field is finally applied to obtain the spectral modes describing the small-scale features in the entire domain.

In principle, one Lagrangian tracer can go from one cell to another as time evolves. This will not affect Lagrangian DA as long as the posterior mean and covariance sizes change adaptively in time according to the number of particle observations in each domain. In the simulation here, a simplified setup is utilized. In each grid cell, 10 particles are preselected (typically, they are chosen near the centers of the grid cells). These 10 particles will stay in the same cell during the DA test period, which allows the posterior mean and covariance to stay in the grid of fixed size.

6.3. Numerical Experiments

The flow field considered in this numerical simulation is characterized by spectral modes $k \in [-8, 8]^2$, and there are in total 288 modes (the (0,0) mode is excluded). An equipartition of the energy spectrum is again adopted that induces a strongly turbulent system. Except that the noise coefficients in characterizing the Fourier modes are reduced to $\sigma_k = 0.01$, all the other parameters remain the same as those in Equations 15–16. The purpose of adopting a reduced noise coefficient level is to guarantee that the order of the velocity in physical space remains around 0.05 m/s, which mimics the realistic situation. There are, in total, 3,136 particles in the entire domain with no collision between different particles.

The coarse grid utilized for the Eulerian DA part has a size of 7×7 , the degree of freedom of which is 49 (or equivalently 143 km between two nearby grids). Thus, it is natural to use the Eulerian DA to estimate the state of the modes with $|k| < 4$ (in total 44 modes), which are referred to as large-scale modes. The remaining 244 modes with $|k| \geq 4$ are called small-scale modes recovered using the Lagrangian DA part, where 10 particle trajectories in each coarse grid cell are utilized as observations.

Figure 10 includes the results of LEMDA and the associated Eulerian and Lagrangian DA components in recovering. Panel (a) compares the truth and the recovered flow fields at $t = 1, 3$, and 5. The first row demonstrates the true flow field. As the equipartition of the energy is utilized, the true flow field is highly turbulent. The second row shows the recovered large-scale modes using Eulerian DA. Note from the color bar that the amplitude

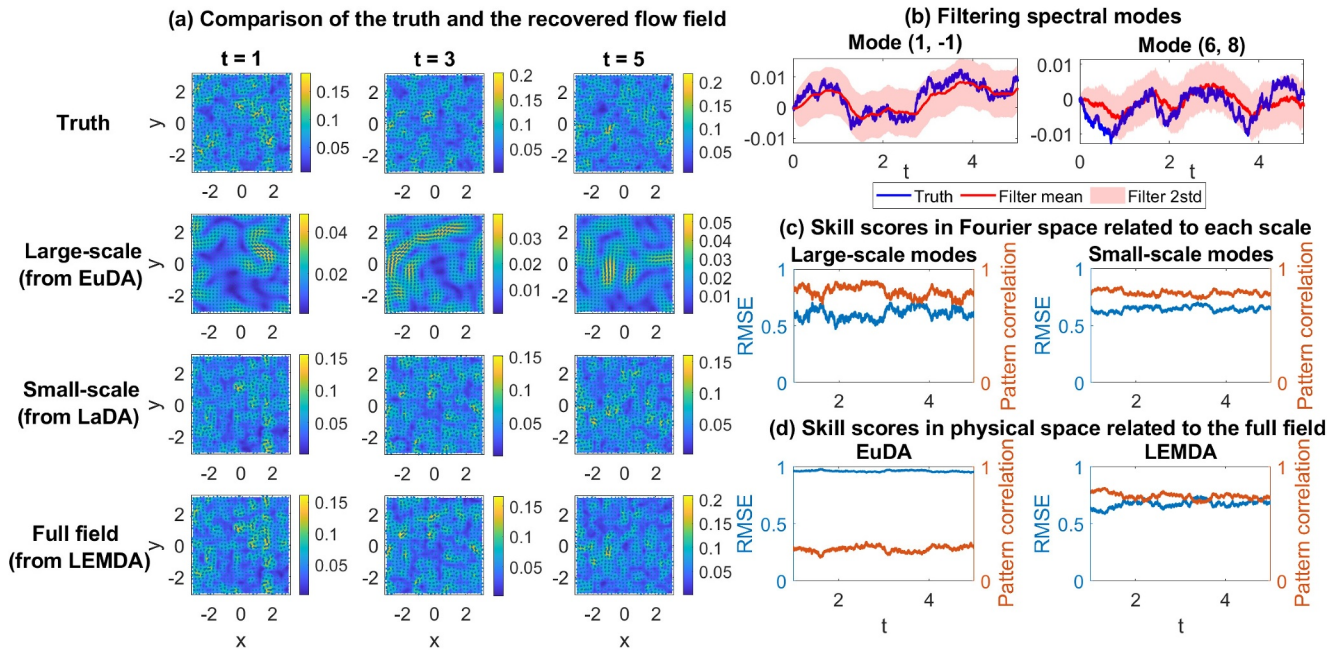


Figure 10. Recovery of multiscale turbulent flow field utilizing LEMDA. Panel (a) compares the truth and the recovered flow fields. Panel (b) shows the filtering results of one large-scale mode (1, -1) and one small-scale mode (6,8). Panel (c) shows the skill score in Fourier space averaged over all the large-scale modes (left) and small-scale modes (right), related to the true large- and small-scale modes, respectively. Panel (d) shows the skill score in physical space associated with the entire flow field for the recovered large-scale features using Eulerian DA (left) and in the recovered entire flow field using LEMDA (right).

(and thus the energy) of the large-scale modes is only a portion of the total flow field. The filtering solution of one of the large-scale Fourier modes (1, -1) is shown in the left column of Panel (b), which indicates the significant skill of the Eulerian DA part. The third row in Panel (a) shows the flow structures associated with the small-scale modes recovered from the parallel Lagrangian DA part. The skill of the Lagrangian DA part in recovering these refined features can be quantitatively seen in the right column of Panel (b). Summing up the two components of the flow field in the second and third rows in Panel (a) leads to the results in the fourth panel, which is the recovered entire flow field. Comparing the first and the fourth rows, the recovered flow field reproduces the multiscale features of the truth. Next, Panel (c) provides a qualitative understanding of the skill scores in recovering the large- and small-scale features of the flows. Both components are recovered accurately, with pattern correlations above 0.75. The large-scale features are recovered slightly more accurately than the small-scale components in this case, although the skill of the latter can be enhanced when the number of the observed Lagrangian trajectories increases. The right column of Panel (d) studies the performance of the entire flow field, which is the superposition of the large- and small-scale structures. Again, the pattern correlation is above 0.75, and the RMSE is below 0.7, consistent with the intuitions obtained from the first and the fourth rows in Panel (a). Note that the small-scale modes account for a large portion of the total signal. If they are omitted, then the skill score drops significantly, as seen in the left column of Panel (d), which shows the error in the flow field reconstructed using only the recovered large-scale modes related to the entire flow field. As a final remark, the results using only the Lagrangian DA part based on all the trajectories of the observed particles are not shown here. The main issue of carrying out Lagrangian DA with a large number of particles is the computational cost. When the number of tracer trajectories increases, the dimension of $\mathbf{Z}$ in Equation 10 becomes large. The row number of the matrix $\mathbf{A}$ also increases. As a result, the computational efficiency of solving Equation 11 is significantly impacted, preventing a fast calculation of the posterior distribution.

7. Conclusion

In this paper, a Lagrangian-Eulerian multiscale data assimilation (LEMMA) framework is developed to facilitate the state estimation of multiscale flow fields. Despite the nonlinearity, the DA solutions can be written down using closed analytic formulas, which allows for efficiently reaching accurate solutions and avoiding empirical tunings. The Lagrangian and Eulerian components have their significant features. Lagrangian DA can be directly applied

when particle trajectories are used as observations. Effective reduced-order filters can be derived to further enhance computational efficiency. Eulerian DA scheme is systematically derived from the Boltzmann equation to advance a helpful and efficient method that recovers large-scale features. Particularly, using a large number of Lagrangian tracers as observations can be computationally expensive while using the induced Eulerian quantities accelerates the calculations. Eulerian DA is also particularly valuable when particles have collisions, such as using sea ice floes as particles. In such a case, the direct Lagrangian DA is not the most appropriate choice for state estimation. Eulerian and Lagrangian DAs are combined for multiscale DA. Parallel computing can be naturally incorporated to significantly accelerate computational efficiency. It is worth mentioning that the LEMDA is designed to handle the resolved scales by Eulerian mesh and the sub-grid scales by the tracer trajectories. There exist many other multiscale DA methods focusing on using different grid resolutions (de Moraes et al., 2020; Wang et al., 2021; Ying, 2020), which have a variety of applications.

When the forecast model can be well approximated by the stochastic surrogate models Equations 7–8, analytic formulas 11 apply to solving the posterior distributions. This facilitates the DA computations. Nevertheless, it is worth remarking that for problems with more complicated domain geometry or when the spectral decomposition is not an appropriate choice in describing the forecast model, the ensemble Kalman (or ensemble Kalman-Bucy) filter or another particle-based DA scheme can be used to solve both Eulerian and Lagrangian DAs. The general framework of LEMDA by creating Eulerian and Lagrangian DA for recovering the large- and small-scale features still holds. LEMDA is not intended to compete with the ensemble-based DA. Instead, the ensemble-based DA can be incorporated into the LEMDA framework to find the posterior distribution. The LEMDA framework also applies to multiscale flow fields when discrete-in-time observations are available, which appear in many applications in fluids and geophysical problems.

There are several other natural extensions of the current study, which remain as future work. First, only the momentum equations are used in Eulerian DA. When the particles are represented by ice floes and other floats with geometric structures, angular momentum is another widely used measurement that helps state estimation. The governing equation of the angular momentum can be derived following the procedure in Section 2.2. Then, such an additional observational quantity can be incorporated into the DA framework to potentially help enhance the DA skill. Second, most of the numerical experiments in this paper consider a perfect model setup. It is essential to study the DA skill when model errors appear in the forecast model. Some crucial sources of model errors are statistical approximate errors using the linear stochastic model to approximate highly nonlinear true underlying systems and using linear drags to approximate the quadratic drag law. Another interesting topic is the optimal selection of a few Lagrangian trajectories to supplement the Eulerian observations that maximize the uncertainty reduction in the estimated flow field from the hybrid Eulerian-Lagrangian observations. Information criteria (Chen et al., 2023) are useful cost functions to resolve such an optimization problem. Furthermore, this work assumes the particles appear almost everywhere in the flow field. In practice, clouds may intervene in the observations (e.g., from satellite images) and bring about missing observations in a certain region. Therefore, studying the skill of LEMDA with partial observations will be a practically meaningful topic.

Appendix A: Derivation of the Reduced-Order Lagrangian DA

Recall the governing equation of Lagrangian DA.

$$\frac{dx_l}{dt} = v_l + \sigma_x \dot{W}_l, \quad (\text{A1a})$$

$$\frac{dv_l}{dt} = \beta(u - v_l), \quad u(x, t) = \sum_{k \in \mathcal{K}} \hat{u}_k(t) e^{ikx} r_k, \quad (\text{A1b})$$

$$\frac{d\hat{u}_k}{dt} = -d_k \hat{u}_k + f_k(t) + \sigma_k \dot{W}_k, \quad (\text{A1c})$$

where the subscript α is omitted for notation simplicity. The flow field can be regarded as having only one type of mode, such as the geophysically balanced mode that corresponds to the incompressible flow. This is also a natural choice to satisfy the condition that the tracers are (nearly) uniformly distributed. For this reason, the phase term $i\omega_k$ is also omitted for simplicity. A similar process can be carried out for the derivation when ω_k is nonzero. The

forcing f_l in Equation A1b, describing the contact force between different tracers, is also dropped to allow the flow field to dominate the tracer motion. Writing the Equation A1 in the general form of Equation 11 in the main text, the observed state $\mathbf{Z}$ contains the displacements of all the tracers $\mathbf{x}_l$, for $l = 1, \dots, 2L$. The unobserved state $\mathbf{U}$ involves the velocity components $\mathbf{v}_l$ of the $2L$ tracers and the $|\mathcal{K}|$ spectral modes of the underlying flow field $\hat{\mathbf{u}}_k$, where $|\mathcal{K}|$ is the total number of the spectral modes in the set $\mathcal{K}$. The observed and unobserved variables have dimensions $2L \times 1$ and $(2L + |\mathcal{K}|) \times 1$, respectively.

Recall the posterior covariance equation of the filtering solution:

$$\frac{d\mathbf{R}}{dt} = \mathbf{\Lambda}\mathbf{R} + \mathbf{R}\mathbf{\Lambda}^* + \mathbf{\Sigma}_U\mathbf{\Sigma}_U^* - \mathbf{R}\mathbf{A}^*(\mathbf{\Sigma}_Z\mathbf{\Sigma}_Z^*)^{-1}\mathbf{A}\mathbf{R}. \quad (\text{A2})$$

The reduced filter aims to simplify the evolution equation of the covariance $\mathbf{R}$ in Equation A2. Notably, the matrix $\mathbf{R}$ can be written in the block form:

$$\mathbf{R} = \begin{pmatrix} \mathbf{R}_{11} & \mathbf{R}_{12} \\ \mathbf{R}_{21} & \mathbf{R}_{22} \end{pmatrix} \quad (\text{A3})$$

where $\mathbf{R}_{11}$ corresponds to the variance of the tracer velocity and $\mathbf{R}_{22}$ contains the variance of the spectral modes of the underlying flow field. The cross-covariance block $\mathbf{R}_{12} = \mathbf{R}_{21}^*$ contains the coupling between the tracer velocity of the underlying flow field. The dimensions of $\mathbf{R}_{11}$, $\mathbf{R}_{22}$ and $\mathbf{R}_{12}$ are $2L \times 2L$, $|\mathcal{K}| \times |\mathcal{K}|$, and $2L \times |\mathcal{K}|$, respectively. The matrix $\mathbf{R}_{11}$ is full because the tracers nearby are highly correlated. The matrix $\mathbf{R}_{12}$ is also full. The off-diagonal components pass the information from the inferred tracer velocity based on their displacement to recover the underlying flow field in DA. If such a matrix becomes diagonal, then the underlying system loses observability to estimate the flow field from observations. The main reduction in the covariance lies in the block $\mathbf{R}_{22}$. Note that all the spectral modes are coupled to form the flow velocity, and therefore, $\mathbf{R}_{22}$ from DA is a full matrix. Nevertheless, since different spectral modes are, by design, independent of each other, it is expected that, with $L \rightarrow \infty$, the posterior covariance block $\mathbf{R}_{22}$ also becomes a diagonal matrix. This has been shown when the velocities of the tracers and the underlying flow field are synchronized (Chen et al., 2014), assuming the flow field is incompressible. Furthermore, the block $\mathbf{R}_{22}$ will asymptotically become a constant matrix as $L \rightarrow \infty$ and can be predetermined, which avoids running any governing equation. Deriving such a constant diagonal matrix is the primary goal in this subsection. The only assumption utilized here is that the tracers are distributed uniformly in the domain. This assumption is roughly satisfied when the underlying flow field is (nearly) incompressible and the drag coefficient β is not too small.

The matrices in Equation A2 have the following form:

$$\begin{aligned} \mathbf{\Lambda} &= \begin{pmatrix} -\beta\mathbf{I}_{2L} & \mathbf{Q} \\ \mathbf{0} & -\mathbf{D} \end{pmatrix}, & \mathbf{A} &= (\mathbf{I}_{2L} \quad \mathbf{0}), \\ \mathbf{\Sigma}_U\mathbf{\Sigma}_U^* &= \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma}_k^2 \end{pmatrix}, & \mathbf{\Sigma}_Z\mathbf{\Sigma}_Z^* &= \sigma_x^2\mathbf{I}_{2L}, \end{aligned} \quad (\text{A4})$$

where $\mathbf{I}_{2L}$ is an identity matrix of size $2L \times 2L$ and $\mathbf{D}$ is a diagonal matrix of size $|\mathcal{K}| \times |\mathcal{K}|$ with the diagonal entries being d_k . The block matrix $\mathbf{Q}$ is of size $2L \times |\mathcal{K}|$ with its entries being the two components of $e^{ikx_l}\mathbf{r}_k$. The matrix $\mathbf{\Sigma}_k^2$ is diagonal with size $|\mathcal{K}| \times |\mathcal{K}|$, where each diagonal component is σ_k^2 .

In addition to $\mathbf{\Sigma}_U\mathbf{\Sigma}_U^*$, the remaining terms on the right-hand side of Equation A2 can be computed:

$$\begin{aligned} \mathbf{R}\mathbf{A}^*(\Sigma_Z\Sigma_Z^*)^{-1}\mathbf{A}\mathbf{R} &= \begin{pmatrix} \mathbf{R}_{11} & \mathbf{R}_{12} \\ \mathbf{R}_{21} & \mathbf{R}_{22} \end{pmatrix} \begin{pmatrix} \mathbf{I}_{2L} \\ \mathbf{0} \end{pmatrix} \sigma_x^{-2} \mathbf{I}_{2L} (\mathbf{I}_{2L} \quad \mathbf{0}) \begin{pmatrix} \mathbf{R}_{11} & \mathbf{R}_{12} \\ \mathbf{R}_{21} & \mathbf{R}_{22} \end{pmatrix} \\ &= \sigma_x^{-2} \begin{pmatrix} \mathbf{R}_{11}\mathbf{R}_{11} & \mathbf{R}_{11}\mathbf{R}_{12} \\ \mathbf{R}_{21}\mathbf{R}_{11} & \mathbf{R}_{21}\mathbf{R}_{12} \end{pmatrix}, \end{aligned} \quad (\text{A5})$$

and

$$\begin{aligned} \Lambda\mathbf{R} + \mathbf{R}\Lambda^* &= \begin{pmatrix} -\beta\mathbf{I}_{2L} & \mathbf{Q} \\ \mathbf{0} & -\mathbf{D} \end{pmatrix} \begin{pmatrix} \mathbf{R}_{11} & \mathbf{R}_{12} \\ \mathbf{R}_{21} & \mathbf{R}_{22} \end{pmatrix} + \begin{pmatrix} \mathbf{R}_{11} & \mathbf{R}_{12} \\ \mathbf{R}_{21} & \mathbf{R}_{22} \end{pmatrix} \begin{pmatrix} -\beta\mathbf{I}_{2L} & \mathbf{0} \\ \mathbf{Q}^* & -\mathbf{D} \end{pmatrix} \\ &= \begin{pmatrix} -\beta\mathbf{R}_{11} + \mathbf{Q}\mathbf{R}_{21} & -\beta\mathbf{R}_{12} + \mathbf{Q}\mathbf{R}_{22} \\ -\mathbf{D}\mathbf{R}_{21} & -\mathbf{D}\mathbf{R}_{22} \end{pmatrix} + \begin{pmatrix} -\beta\mathbf{R}_{11} + \mathbf{R}_{12}\mathbf{Q}^* - \mathbf{R}_{12}\mathbf{D} \\ -\beta\mathbf{R}_{21} + \mathbf{R}_{22}\mathbf{Q}^* - \mathbf{R}_{22}\mathbf{D} \end{pmatrix}. \end{aligned} \quad (\text{A6})$$

In light of Equations A4, A5 and A6, the time evolution of the three blocks in the posterior covariance $\mathbf{R}$ can be written as.

$$\frac{d\mathbf{R}_{11}}{dt} = -2\beta\mathbf{R}_{11} + \mathbf{Q}\mathbf{R}_{21} + \mathbf{R}_{12}\mathbf{Q}^* - \sigma_x^{-2}\mathbf{R}_{11}\mathbf{R}_{11}, \quad (\text{A7a})$$

$$\frac{d\mathbf{R}_{21}}{dt} = -\mathbf{D}\mathbf{R}_{21} - \beta\mathbf{R}_{21} + \mathbf{R}_{22}\mathbf{Q}^* - \sigma_x^{-2}\mathbf{R}_{21}\mathbf{R}_{11}, \quad (\text{A7b})$$

$$\frac{d\mathbf{R}_{22}}{dt} = -2\mathbf{D}\mathbf{R}_{22} + \Sigma_k^2 - \sigma_x^{-2}\mathbf{R}_{21}\mathbf{R}_{12}. \quad (\text{A7c})$$

As time evolves, the distribution of the tracers will roughly have a steady behavior. Although the exact distribution still varies in time due to the randomness in the flow field, the system can be regarded as reaching a quasi-equilibrium state. Under such a condition, the goal is to look for a solution that approximately gives

$$\frac{d\mathbf{R}}{dt} = 0. \quad (\text{A8})$$

with Equation A8 in hand, it can be deduced from Equation A7b that

$$\mathbf{R}_{21} \approx \mathcal{A}\mathbf{R}_{22}\mathbf{Q}^*, \quad (\text{A9})$$

where $\mathcal{A}$ is a diagonal matrix with its i -th diagonal entry corresponding to the k -th spectral mode being

$$\xi = \frac{1}{d_k + \beta + \sigma_x^{-2}\gamma}. \quad (\text{A10})$$

here, γ is used to parameterize the effect of $\mathbf{R}_{11}$, when it is multiplied to $\mathbf{R}_{21}$. Note that similar to d_k , the two scalars ξ and γ depend on the mode k . But for notation simplicity, the subscript k is omitted for these two scalars. Accordingly, the i -th diagonal component of $\mathbf{R}_{22}$ in Equation A7c becomes

$$\frac{dR_k}{dt} = -2d_k R_k + \sigma_k^2 - \sigma_x^{-2}\xi^2 L R_k^2, \quad (\text{A11})$$

where Equation A9 is used to replace $\mathbf{R}_{21}$ in Equation A7c such that Equation A11 becomes a closed equation of R_k , one of the diagonal components of $\mathbf{R}_{22}$, itself. From Equation A9–A11, the mean-field theory is applied such that $\mathbf{Q}^*\mathbf{Q} \approx L \cdot \mathbf{I}_{|K|}$. This can be seen by noting that the m and n entry of $\mathbf{Q}^*\mathbf{Q}$ can be explicitly written as

$$\mathbf{Q}^* \mathbf{Q}_{m,n} = \sum_{l=1}^L \exp(i(\mathbf{m} - \mathbf{n})\mathbf{x}_l) \mathbf{r}_n^* \mathbf{r}_m. \quad (\text{A12})$$

If the tracers are uniformly distributed, then the entries of $\mathbf{P}_{m,n}$ with $\mathbf{m} \neq \mathbf{n}$ become zero in the asymptotic limit according to the mean-field theory while those with $\mathbf{m} = \mathbf{n}$ equal to L .

The positive root of R_k in Equation A11 is given by

$$R_k = \frac{-d_k + \sqrt{d_k^2 + \sigma_x^{-2} \xi^2 L \sigma_k^2}}{\sigma_x^{-2} \xi^2 L}, \quad (\text{A13})$$

which is the constant posterior variance associated with the component $\hat{u}_k$.

What remains is to determine ξ . Multiplying $\mathbf{Q}$ by both sides of Equation A9 yields

$$\mathbf{Q} \mathbf{R}_{21} \approx \mathbf{A} \mathbf{Q} \mathbf{R}_{22} \mathbf{Q}^*. \quad (\text{A14})$$

Thus each diagonal component of $\mathbf{Q} \mathbf{R}_{21}$ is approximated by $\xi L R_k$. Note that $\mathbf{R}_{11}$ is not a diagonal matrix but only the diagonal entries are used here to determine ξ . Plugging the above result into Equation A7a leads to the following equation for each diagonal component of $\mathbf{R}_{11}$, namely the parameterized scalar γ ,

$$-2\beta\gamma + 2\xi L R_k - \sigma_x^{-2} \gamma^2 = 0, \quad (\text{A15})$$

the solution of which is given by

$$\gamma = \frac{-\beta + \sqrt{\beta^2 + 2\sigma_x^{-2} \xi L R_k}}{\sigma_x^{-2}}. \quad (\text{A16})$$

The diagonal entry R_k , namely the posterior variance of mode $\hat{u}_k$, is then obtained by solving the three-dimensional coupled algebraic Equations A10, A13, and A16. These three equations have three unknowns R_k , ξ and γ . A standard Newton's iterative method is utilized to find the solution.

Once all the diagonal entries R_k of $\mathbf{R}_{22}$ are pre-determined, only the Equations A7a and A7b need to be solved. Then finally, the constant diagonal covariance block for the flow modes can be predetermined by the smoother solution. The computational time and storage for the smoothing solutions can similarly be reduced.

Data Availability Statement

The MATLAB codes developed for this work are available in Zenodo repository (Deng & Chen, 2024) (10.5281/zenodo.13308883 Link: <https://zenodo.org/record/13308883>).

References

- Apte, A., & Jones, C. (2013). The impact of nonlinearity in Lagrangian data assimilation. *Nonlinear Processes in Geophysics*, 20(3), 329–341. <https://doi.org/10.5194/npg-20-329-2013>
- Apte, A., Jones, C. K., & Stuart, A. (2008a). A Bayesian approach to Lagrangian data assimilation. *Tellus A: Dynamic Meteorology and Oceanography*, 60(2), 336–347. <https://doi.org/10.1111/j.1600-0870.2007.00295.x>
- Apte, A., Jones, C. K., Stuart, A., & Voss, J. (2008b). Data assimilation: Mathematical and statistical perspectives. *International Journal for Numerical Methods in Fluids*, 56(8), 1033–1046. <https://doi.org/10.1002/fld.1698>
- Asch, M., Bocquet, M., & Nodet, M. (2016). *Data assimilation: Methods, algorithms, and applications*. SIAM.
- Azouani, A., Olson, E., & Titi, E. S. (2014). Continuous data assimilation using general interpolant observables. *Journal of Nonlinear Science*, 24(2), 277–304. <https://doi.org/10.1007/s00332-013-9189-y>
- Bach, E., Colinius, T., & Stuart, A. (2023). Filtering dynamical systems using observations of statistics. *arXiv preprint arXiv:2308.05484*.
- Baskaran, A., & Marchetti, M. C. (2010). Nonequilibrium statistical mechanics of self-propelled hard rods. *Journal of Statistical Mechanics: Theory and Experiment*, 2010(04), P04019. <https://doi.org/10.1088/1742-5468/2010/04/p04019>
- Bergemann, K., & Reich, S. (2012). An ensemble Kalman-Bucy filter for continuous data assimilation. *Meteorologische Zeitschrift*, 21(3), 213–219. <https://doi.org/10.1127/0941-2948/2012/0307>

Acknowledgments

The authors wish to thank the five anonymous reviewers for providing valuable suggestions. The research of Q.D. is partially supported by the Australian National Computing Infrastructure (NCI) national facility under Grant zv32 and the HMI Computing for Social Good Seed Research Grant. The research of S.N.S and N.C. was partially funded by the Office of Naval Research (ONR) Multidisciplinary University Research Initiative (MURI) award N00014-19-1-2421. J.H. is supported as a postdoc research associated under this grant.

- Berner, J., Achatz, U., Batte, L., Bengtsson, L., De La Camara, A., Christensen, H. M., et al. (2017). Stochastic parameterization: Toward a new view of weather and climate models. *Bulletin of the American Meteorological Society*, 98(3), 565–588. <https://doi.org/10.1175/bams-d-15-00268.1>
- Blunden, J., & Arndt, D. (2019). A look at 2018: Takeaway points from the state of the climate supplement. *Bulletin of the American Meteorological Society*, 100(9), 1625–1636. <https://doi.org/10.1175/bams-d-19-0193.1>
- Branicki, M., Chen, N., & Majda, A. J. (2013). Non-Gaussian test models for prediction and state estimation with model errors. *Chinese Annals of Mathematics, Series B*, 34(1), 29–64. <https://doi.org/10.1007/s11401-012-0759-3>
- Branicki, M., Majda, A. J., & Law, K. J. (2018). Accuracy of some approximate Gaussian filters for the Navier–Stokes equation in the presence of model error. *Multiscale Modeling and Simulation*, 16(4), 1756–1794. <https://doi.org/10.1137/17m1146865>
- Bröcker, J. (2010). On variational data assimilation in continuous time. *Quarterly Journal of the Royal Meteorological Society*, 136(652), 1906–1919. <https://doi.org/10.1002/qj.695>
- Bucy, R. S., & Joseph, P. D. (2005). Filtering for stochastic processes with applications to guidance. *American Mathematical Society*, 326.
- Businger, S., Chiswell, S. R., Ulmer, W. C., & Johnson, R. (1996). Balloons as a Lagrangian measurement platform for atmospheric research. *Journal of Geophysical Research*, 101(D2), 4363–4376. <https://doi.org/10.1029/95jd00559>
- Cai, D., Kleeman, R., & Majda, A. (2002). A mathematical framework for quantifying predictability through relative entropy. *Methods and Applications of Analysis*, 9(3), 425–444. <https://doi.org/10.4310/maa.2002.v9.n3.a8>
- Carlson, E., & Larios, A. (2021). Sensitivity analysis for the 2d Navier–Stokes equations with applications to continuous data assimilation. *Journal of Nonlinear Science*, 31(5), 84. <https://doi.org/10.1007/s00332-021-09739-9>
- Castellari, S., Griffa, A., Özgökmen, T. M., & Poulain, P.-M. (2001). Prediction of particle trajectories in the Adriatic sea using Lagrangian data assimilation. *Journal of Marine Systems*, 29(1–4), 33–50. [https://doi.org/10.1016/s0924-7963\(01\)00008-2](https://doi.org/10.1016/s0924-7963(01)00008-2)
- Centurioni, L., Horányi, A., Cardinali, C., Charpentier, E., & Lumpkin, R. (2017). A global ocean observing system for measuring sea level atmospheric pressure: Effects and impacts on numerical weather prediction. *Bulletin of the American Meteorological Society*, 98(2), 231–238. <https://doi.org/10.1175/bams-d-15-00080.1>
- Cercignani, C., & Cercignani, C. (1988). *The Boltzmann equation*. Springer.
- Chaikin, P. M., Lubensky, T. C., & Witten, T. A. (1995). *Principles of condensed matter physics* (Vol. 10). Cambridge university press Cambridge.
- Chen, N. (2020). Learning nonlinear turbulent dynamics from partial observations via analytically solvable conditional statistics. *Journal of Computational Physics*, 418, 109635. <https://doi.org/10.1016/j.jcp.2020.109635>
- Chen, N. (2023). Stochastic methods for modeling and predicting complex dynamical systems: Uncertainty quantification, state estimation, and reduced-order models. *Springer Nature*.
- Chen, N., & Fu, S. (2023). Uncertainty quantification of nonlinear Lagrangian data assimilation using linear stochastic forecast models. *Physica D: Nonlinear Phenomena*, 452, 133784. <https://doi.org/10.1016/j.physd.2023.133784>
- Chen, N., Fu, S., & Manucharyan, G. (2021). Lagrangian data assimilation and parameter estimation of an idealized sea ice discrete element model. *Journal of Advances in Modeling Earth Systems*, 13(10), e2021MS002513. <https://doi.org/10.1029/2021ms002513>
- Chen, N., Fu, S., & Manucharyan, G. E. (2022a). An efficient and statistically accurate Lagrangian data assimilation algorithm with applications to discrete element sea ice models. *Journal of Computational Physics*, 455, 111000. <https://doi.org/10.1016/j.jcp.2022.111000>
- Chen, N., Li, Y., & Liu, H. (2022b). Conditional Gaussian nonlinear system: A fast preconditioner and a cheap surrogate model for complex nonlinear systems. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 32(5), 053122. <https://doi.org/10.1063/5.0081668>
- Chen, N., Lunasin, E., & Wiggins, S. (2023). Launching drifter observations in the presence of uncertainty. *arXiv preprint arXiv:2307.12779*.
- Chen, N., & Majda, A. J. (2016). Model error in filtering random compressible flows utilizing noisy Lagrangian tracers. *Monthly Weather Review*, 144(11), 4037–4061. <https://doi.org/10.1175/mwr-d-15-0438.1>
- Chen, N., & Majda, A. J. (2018). Conditional Gaussian systems for multiscale nonlinear stochastic systems: Prediction, state estimation and uncertainty quantification. *Entropy*, 20(7), 509. <https://doi.org/10.3390/e20070509>
- Chen, N., Majda, A. J., & Tong, X. T. (2014). Information barriers for noisy Lagrangian tracers in filtering random incompressible flows. *Nonlinearity*, 27(9), 2133–2163. <https://doi.org/10.1088/0951-7715/27/9/2133>
- Chen, N., Majda, A. J., & Tong, X. T. (2015). Noisy Lagrangian tracers for filtering random rotating compressible flows. *Journal of Nonlinear Science*, 25(3), 451–488. <https://doi.org/10.1007/s00332-014-9226-5>
- Covington, J., Chen, N., & Wilhelmus, M. M. (2022). Bridging gaps in the climate observation network: A physics-based nonlinear dynamical interpolation of Lagrangian ice floe measurements via data-driven stochastic models. *Journal of Advances in Modeling Earth Systems*, 14(9), 2022MS003218. <https://doi.org/10.1029/2022ms003218>
- Cundall, P. A., & Strack, O. D. (1979). A discrete numerical model for granular assemblies. *Géotechnique*, 29(1), 47–65. <https://doi.org/10.1680/geot.1979.29.1.47>
- Curtiss, C. (1956). Kinetic theory of nonspherical molecules. *Journal of Chemical Physics*, 24(2), 225–241. <https://doi.org/10.1063/1.1742459>
- Curtiss, C. (1967). Transport phenomena in gases. *Annual Review of Physical Chemistry*, 18(1), 125–151. <https://doi.org/10.1146/annurev.pc.18.100167.001013>
- Damsgaard, A., Adcroft, A., & Sergienko, O. (2018). Application of discrete element methods to approximate sea ice dynamics. *Journal of Advances in Modeling Earth Systems*, 10(9), 2228–2244. <https://doi.org/10.1029/2018ms001299>
- DelSole, T. (2005). Predictability and information theory. Part II: Imperfect forecasts. *Journal of the Atmospheric Sciences*, 62(9), 3368–3381. <https://doi.org/10.1175/jas3522.1>
- de Moraes, R. J., Hajibeygi, H., & Jansen, J. D. (2020). A multiscale method for data assimilation. *Computational Geosciences*, 24(2), 425–442. <https://doi.org/10.1007/s10596-019-09839-2>
- Deng, Q., & Chen, N. (2024). MATLAB codes for LEMDA: Lagrangian-Eulerian multiscale data assimilation. *Zenodo*. <https://doi.org/10.5281/zenodo.13308883>
- Deng, Q., Stechmann, S. N., & Chen, N. (2024). Particle-continuum multiscale modeling of sea ice floes. *Multiscale Modeling and Simulation*, 22(1), 230–255. <https://doi.org/10.1137/23M155904X>
- Desamsetti, S., Dasari, H. P., Langodan, S., Titi, E. S., Knio, O., & Hoteit, I. (2019). Efficient dynamical downscaling of general circulation models using continuous data assimilation. *Quarterly Journal of the Royal Meteorological Society*, 145(724), 3175–3194. <https://doi.org/10.1002/qj.3612>
- Doi, M., & Edwards, S. (1978). Dynamics of rod-like macromolecules in concentrated solution. Part 1. *Journal of the Chemical Society, Faraday Transactions 2: Molecular and Chemical Physics*, 74, 560–570. <https://doi.org/10.1039/f29787400560>
- Doi, M., & Edwards, S. F. (1988). *The theory of polymer dynamics* (Vol. 73). Oxford University Press.
- Einicke, G. (2012). *Smoothing, filtering and prediction: Estimating the past, present and future*. BoD—Books on Demand.

- Evensen, G., & Van Leeuwen, P. J. (2000). An ensemble Kalman smoother for nonlinear dynamics. *Monthly Weather Review*, 128(6), 1852–1867. [https://doi.org/10.1175/1520-0493\(2000\)128<1852:acksfn>2.0.co;2](https://doi.org/10.1175/1520-0493(2000)128<1852:acksfn>2.0.co;2)
- Ezhilan, B., Shelley, M. J., & Saintillan, D. (2013). Instabilities and nonlinear dynamics of concentrated active suspensions. *Physics of Fluids*, 25(7), 070607. <https://doi.org/10.1063/1.4812822>
- Farchi, A., Laloyaux, P., Bonavita, M., & Bocquet, M. (2021). Using machine learning to correct model error in data assimilation and forecast applications. *Quarterly Journal of the Royal Meteorological Society*, 147(739), 3067–3084. <https://doi.org/10.1002/qj.4116>
- Farhat, A., Glat-Holtz, N. E., Martinez, V. R., McQuarrie, S. A., & Whitehead, J. P. (2020). Data assimilation in large Prandtl Rayleigh–Benard convection from thermal measurements. *SIAM Journal on Applied Dynamical Systems*, 19(1), 510–540. <https://doi.org/10.1137/19m1248327>
- Farrell, B. F., & Ioannou, P. J. (1993). Stochastic forcing of the linearized Navier–Stokes equations. *Physics of Fluids A: Fluid Dynamics*, 5(11), 2600–2609. <https://doi.org/10.1063/1.858894>
- García-Sánchez, G., Mancho, A. M., Ramos, A. G., Coca, J., & Wiggins, S. (2022). Structured pathways in the turbulence organizing recent oil spill events in the eastern mediterranean. *Scientific Reports*, 12(1), 3662. <https://doi.org/10.1038/s41598-022-07350-w>
- Gardiner, C. W. (1985). *Handbook of stochastic methods* (Vol. 3). Springer.
- Giannakis, D., & Majda, A. J. (2012). Quantifying the predictive skill in long-range forecasting. Part I: Coarse-grained predictions in a simple ocean model. *Journal of Climate*, 25(6), 1793–1813. <https://doi.org/10.1175/2011jcli4143.1>
- Gould, J., Roemmich, D., Wijffels, S., Freeland, H., Ignaszewsky, M., Jianping, X., et al. (2004). Argo profiling floats bring new era of in situ ocean observations. *Eos. Transactions - American Geophysical Union*, 85(19), 185–191. <https://doi.org/10.1029/2004eo190002>
- Griffa, A., Kirwan, A., Jr., Mariano, A. J., Özgökmen, T., & Rossby, H. T. (2007). *Lagrangian analysis and prediction of coastal and ocean dynamics*. Cambridge University Press.
- Hamill, T. M., & Whitaker, J. S. (2005). Accounting for the error due to unresolved scales in ensemble data assimilation: A comparison of different approaches. *Monthly Weather Review*, 133(11), 3132–3147. <https://doi.org/10.1175/mwr3020.1>
- Harlim, J., & Majda, A. (2008). Filtering nonlinear dynamical systems with linear stochastic models. *Nonlinearity*, 21(6), 1281–1306. <https://doi.org/10.1088/0951-7715/21/6/008>
- Harlim, J., & Majda, A. J. (2013). Test models for filtering and prediction of moisture-coupled tropical waves. *Quarterly Journal of the Royal Meteorological Society*, 139(670), 119–136. <https://doi.org/10.1002/qj.1956>
- Harris, S. (2004). An introduction to the theory of the Boltzmann equation. *Courier Corporation*.
- Honnorat, M., Monnier, J., & Le Dimet, F.-X. (2009). Lagrangian data assimilation for river hydraulics simulations. *Computing and Visualization in Science*, 12(5), 235–246. <https://doi.org/10.1007/s00791-008-0089-x>
- Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4), 679–688. <https://doi.org/10.1016/j.ijforecast.2006.03.001>
- Ide, K., Kuznetsov, L., & Jones, C. K. (2002). Lagrangian data assimilation for point vortex systems. *Journal of Turbulence*, 3(1), 053.
- Jazwinski, A. H. (2007). Stochastic processes and filtering theory. *Courier Corporation*.
- Kalman, R. E., & Bucy, R. S. (1961). New results in linear filtering and prediction theory. *Journal of Fluids Engineering*, 83(1), 95–108. <https://doi.org/10.1115/1.3658902>
- Kalnay, E. (2003). *Atmospheric modeling, data assimilation and predictability*. Cambridge University Press.
- Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., et al. (1996). The NCEP/NCAR 40-year reanalysis project. *Bulletin of the American Meteorological Society*, 77(3), 437–472. [https://doi.org/10.1175/1520-0477\(1996\)077<0437:tnyrp>2.0.co;2](https://doi.org/10.1175/1520-0477(1996)077<0437:tnyrp>2.0.co;2)
- Kang, E. L., & Harlim, J. (2012). Filtering nonlinear spatio-temporal chaos with autoregressive linear stochastic models. *Physica D: Nonlinear Phenomena*, 241(12), 1099–1113. <https://doi.org/10.1016/j.physd.2012.03.003>
- Kleeman, R. (2002). Measuring dynamical prediction utility using relative entropy. *Journal of the Atmospheric Sciences*, 59(13), 2057–2072. [https://doi.org/10.1175/1520-0469\(2002\)059<2057:mdpuur>2.0.co;2](https://doi.org/10.1175/1520-0469(2002)059<2057:mdpuur>2.0.co;2)
- Kleeman, R. (2011). Information theory and dynamical system predictability. *Entropy*, 13(3), 612–649. <https://doi.org/10.3390/e13030612>
- Kullback, S. (1987). Letter to the editor: The Kullback-Leibler distance. *AMERICAN STATISTICIAN*.
- Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1), 79–86. <https://doi.org/10.1214/aoms/1177729694>
- Lahoz, B. K. W., & Menard, R. (2010). *Data assimilation*. Springer.
- Law, K., Stuart, A., & Zygalakis, K. (2015). *Data assimilation* (p. 214). Springer.
- Lean, P., Hólm, E., Bonavita, M., Bormann, N., McNally, A., & Järvinen, H. (2021). Continuous data assimilation for global numerical weather prediction. *Quarterly Journal of the Royal Meteorological Society*, 147(734), 273–288. <https://doi.org/10.1002/qj.3917>
- Li, Y., & Stechmann, S. N. (2020). Predictability of tropical rainfall and waves: Estimates from observational data. *Quarterly Journal of the Royal Meteorological Society*, 146(729), 1668–1684. <https://doi.org/10.1002/qj.3759>
- Liptser, R. S., & Shiryaev, A. N. (2013). *Statistics of random processes II: Applications* (Vol. 6). Springer Science & Business Media.
- Lopez, R. M., Gupta, M., Thompson, A., & Andrade, J. (2023). Level set discrete element method for modeling sea ice floes. *Computer Methods in Applied Mechanics and Engineering*, 406, 115891. <https://doi.org/10.1016/j.cma.2023.115891>
- Majda, A. (2003). Introduction to PDES and waves for the atmosphere and ocean. *American Mathematical Society*, 9.
- Majda, A., Abramov, R. V., & Grote, M. J. (2005). Information theory and stochasticity for multiscale nonlinear systems. *American Mathematical Society*, 25.
- Majda, A. J. (2016). *Introduction to turbulent dynamical systems in complex systems*. Springer.
- Majda, A. J., & Chen, N. (2018). Model error, information barriers, state estimation and prediction in complex multiscale systems. *Entropy*, 20(9), 644. <https://doi.org/10.3390/e20090644>
- Majda, A. J., & Harlim, J. (2012). *Filtering complex turbulent systems*. Cambridge University Press.
- Manucharyan, G. E., Lopez-Acosta, R., & Wilhelmus, M. M. (2022). Spinning ice floes reveal intensification of mesoscale eddies in the western Arctic Ocean. *Scientific Reports*, 12(1), 7070. <https://doi.org/10.1038/s41598-022-10712-z>
- Manucharyan, G. E., & Montemuro, B. P. (2022). SubZero: A sea ice model with an explicit representation of the floe life cycle. *Journal of Advances in Modeling Earth Systems*, 14(12), e2022MS003247. <https://doi.org/10.1029/2022ms003247>
- Mu, L., Losch, M., Yang, Q., Ricker, R., Losa, S. N., & Nerger, L. (2018). Arctic-wide sea ice thickness estimates from combining satellite remote sensing data and a dynamic ice-ocean model with data assimilation during the CryoSat-2 period. *Journal of Geophysical Research: Oceans*, 123(11), 7763–7780. <https://doi.org/10.1029/2018jc014316>
- Nott, P. R., Guazzelli, E., & Pouliquen, O. (2011). The suspension balance model revisited. *Physics of Fluids*, 23(4), 043304. <https://doi.org/10.1063/1.3570921>
- Onsager, L. (1949). The effects of shape on the interaction of colloidal particles. *Annals of the New York Academy of Sciences*, 51(4), 627–659. <https://doi.org/10.1111/j.1749-6632.1949.tb27296.x>

- Potyondy, D. O., & Cundall, P. (2004). A bonded-particle model for rock. *International Journal of Rock Mechanics and Mining Sciences*, 41(8), 1329–1364. <https://doi.org/10.1016/j.ijrmms.2004.09.011>
- Rebholz, L. G., & Zervas, C. (2021). Simple and efficient continuous data assimilation of evolution equations via algebraic nudging. *Numerical Methods for Partial Differential Equations*, 37(3), 2588–2612. <https://doi.org/10.1002/num.22751>
- Saintillan, D., & Shelley, M. J. (2015). Theory of active suspensions. In S. E. Spagnolie (Ed.), *Complex fluids in biological systems: Experiment, theory, and computation* (pp. 319–355). Springer New York. https://doi.org/10.1007/978-1-4939-2065-5_9
- Salman, H., Ide, K., & Jones, C. K. (2008). Using flow geometry for drifter deployment in Lagrangian data assimilation. *Tellus A: Dynamic Meteorology and Oceanography*, 60(2), 321–335. <https://doi.org/10.1111/j.1600-0870.2007.00292.x>
- Särkkä, S., & Svensson, L. (2023). *Bayesian filtering and smoothing* (Vol. 17). Cambridge University Press. <https://doi.org/10.1017/9781108917407>
- Shaqfeh, E. S., & Fredrickson, G. H. (1990). The hydrodynamic stress in a suspension of rods. *Physics of Fluids A: Fluid Dynamics*, 2(1), 7–24. <https://doi.org/10.1063/1.857683>
- Silva, T. M., Pesco, S., Barreto, A., Jr., & Onur, M. (2021). A new procedure for generating data covariance inflation factors for ensemble smoother with multiple data assimilation. *Computers & Geosciences*, 150, 104722. <https://doi.org/10.1016/j.cageo.2021.104722>
- Sun, L., & Penny, S. G. (2019). Lagrangian data assimilation of surface drifters in a double-gyre ocean model using the local ensemble transform Kalman filter. *Monthly Weather Review*, 147(12), 4533–4551. <https://doi.org/10.1175/mwr-d-18-0406.1>
- Taxman, N. (1958). Classical theory of transport phenomena in dilute polyatomic gases. *Physical Review*, 110(6), 1235–1239. <https://doi.org/10.1103/physrev.110.1235>
- Taylor, J. R., & Thompson, A. F. (2023). Submesoscale dynamics in the upper ocean. *Annual Review of Fluid Mechanics*, 55(1), 103–127. <https://doi.org/10.1146/annurev-fluid-031422-095147>
- Uppala, S. M., Källberg, P., Simmons, A. J., Andrae, U., Bechtold, V. D. C., Fiorino, M., et al. (2005). The ERA-40 re-analysis. *Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography*, 131(612), 2961–3012. <https://doi.org/10.1256/qj.04.176>
- Van Sebille, E., England, M. H., & Froyland, G. (2012). Origin, dynamics and evolution of ocean garbage patches from observed surface drifters. *Environmental Research Letters*, 7(4), 044040. <https://doi.org/10.1088/1748-9326/7/4/044040>
- Wang, X., Chipilski, H. G., Bishop, C. H., Satterfield, E., Baker, N., & Whitaker, J. S. (2021). A multiscale local gain form ensemble transform Kalman filter (MLGETKF). *Monthly Weather Review*, 149(3), 605–622. <https://doi.org/10.1175/mwr-d-20-0290.1>
- Wolgemuth, C. W. (2008). Collective swimming and the dynamics of bacterial turbulence. *Biophysical Journal*, 95(4), 1564–1574. <https://doi.org/10.1529/biophysj.107.118257>
- Yang, T., Mehta, P. G., & Meyn, S. P. (2013). Feedback particle filter. *IEEE Transactions on Automatic Control*, 58(10), 2465–2480. <https://doi.org/10.1109/tac.2013.2258825>
- Ying, Y. (2020). Assimilating observations with spatially correlated errors using a serial ensemble filter with a multiscale approach. *Monthly Weather Review*, 148(8), 3397–3412. <https://doi.org/10.1175/mwr-d-19-0387.1>