A Nonlinear Data Assimilation Scheme for Multi-layer Flow Field with Surface Observations

Zhongrui Wang [1] Nan Chen [1] Di Qi [2]

[1]University of Wisconsin-Madison [2]Purdue University

Background

State estimation of a multi-layer flow field (e.g., ocean) with surface observations is a challenging. One widely used method that linearly connects different layers with a regression model can be inaccurate when the flow is highly turbulent[1].

Multi-layer flow with surface observations: A general form

$$\frac{\mathrm{d}\mathbf{x}_\ell}{\mathrm{d}t} = \mathbf{v}_1(\mathbf{x}_\ell, t) + \Sigma_{\mathbf{x}}\dot{\mathbf{W}}_\ell, \qquad \ell = 1, \ldots, L \tag{1a}$$

$$\frac{\mathrm{d}\mathbf{v}}{\mathrm{d}t} = (\mathsf{L} + \mathsf{D})\mathbf{v} + \mathbf{B}(\mathbf{v}, \mathbf{v}) + \mathbf{F} + \Sigma_{\mathbf{v}}\dot{\mathbf{W}}_{\mathbf{v}}, \tag{1b}$$

where $\mathbf{x}_\ell = (x_\ell, y_\ell)^{\mathrm{T}}$ is the ℓth tracer's displacement. $\mathbf{v} = (\ldots, \mathbf{v}_i, \ldots)^{\mathrm{T}}$ is the planar flow velocities of I layers. L and D are linear dispersion and dissipation. $\mathbf{B}(\mathbf{v}, \mathbf{v})$ is a nonlinear quadratic form. $\mathbf{F}$ is a constant forcing. $\Sigma\dot{\mathbf{W}}$ is the Gaussian white noise $\dot{\mathbf{W}}$ multiplied by the noise strength matrix Σ.

Conditional Gaussian nonlinear system

The conditional Gaussian nonlinear system (CGNS) is very common in geophysical flows:

$$\frac{\mathrm{d}\mathbf{u}_1}{\mathrm{d}t} = \mathbf{A}_0(\mathbf{u}_1, t) + \mathsf{A}_1(\mathbf{u}_1, t)\mathbf{u}_2 + \Sigma_1(\mathbf{u}_1, t)\dot{\mathbf{W}}_1, \tag{2a}$$

$$\frac{\mathrm{d}\mathbf{u}_2}{\mathrm{d}t} = \mathbf{a}_0(\mathbf{u}_1, t) + \mathsf{a}_1(\mathbf{u}_1, t)\mathbf{u}_2 + \Sigma_2(\mathbf{u}_1, t)\dot{\mathbf{W}}_2, \tag{2b}$$

where $\mathbf{u}_1 \in \mathbb{C}^{N_1}$ ares observed variables and $\mathbf{u}_2 \in \mathbb{C}^{N_2}$ are hidden variables.

- **Nonlinear & Non-Gaussian**: $\mathbf{A}_0$, $\mathbf{a}_0$, A_1, a_1, Σ_1, Σ_2 can nonlinearly depend on $\mathbf{u}_1$. Thus, the CGNS can be *highly nonlinear*, the marginal distributions of $\mathbf{u}_1$, $\mathbf{u}_2$ can be *strongly non-Gaussian*.
- **Conditional Gaussian**: Given an observed trajectory of $\mathbf{u}_1$, the posterior of $\mathbf{u}_2$ is Gaussian:

$$\mathbf{u}_2(t)|\mathbf{u}_1(s \le t) \sim \mathcal{N}(\boldsymbol{\mu}_2(t), \mathsf{R}_2(t)), \tag{3}$$

with mean $\boldsymbol{\mu}_2(t)$ and covariance $\mathsf{R}_2(t)$ solvable through the analytic formulae

$$\frac{\mathrm{d}\boldsymbol{\mu}_2}{\mathrm{d}t} = (\mathbf{a}_0 + \mathsf{a}_1\boldsymbol{\mu}_2) + \mathsf{R}_2\mathsf{A}_1^*(\Sigma_1\Sigma_1^*)^{-1}\left(\frac{\mathrm{d}\mathbf{u}_1}{\mathrm{d}t} - (\mathbf{A}_0 + \mathsf{A}_1\boldsymbol{\mu}_2)\right), \tag{4a}$$

$$\frac{\mathrm{d}\mathsf{R}_2}{\mathrm{d}t} = \mathsf{a}_1\mathsf{R}_2 + \mathsf{R}_2\mathsf{a}_1^* + \Sigma_2\Sigma_2^* - (\mathsf{R}_2\mathsf{A}_1^*)(\Sigma_1\Sigma_1^*)^{-1}(\mathsf{A}_1\mathsf{R}_2). \tag{4b}$$

Thanks to this blessing of CGNS, the **conditional Gaussian data assimilation (CGDA)** can solve the posterior mean and covariance *exactly and efficiently* based on (4).

Multi-step DA for multi-layer flow

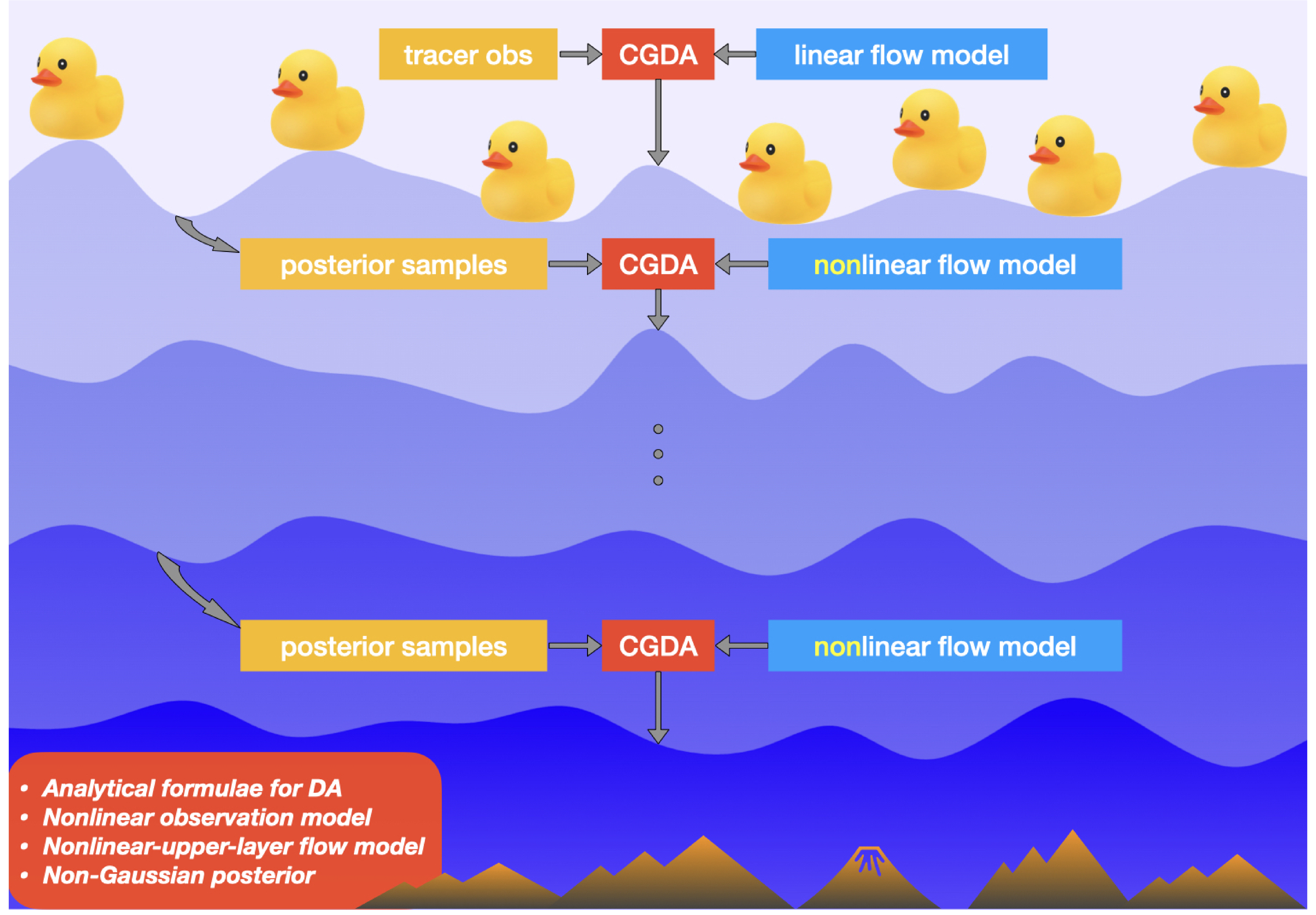
Consider a two-layer flow $\mathbf{v} = (\mathbf{v}_1, \mathbf{v}_2)^{\mathrm{T}}$ with surface tracer observations $\mathbf{x}(s)$. We aim for the posterior

$$P_{\mathbf{V}(t)|\mathbf{X}(s), s \le t}(\mathbf{v}|\{\mathbf{x}(s)\}_{s \le t}), \tag{5}$$

1. The **first DA step** solves the **surface-layer** flow posterior $P(\mathbf{v}_1|\{\mathbf{x}(s)\}_{s \le t})$ given tracer obs.
2. **Sample** from $P(\mathbf{v}_1|\{\mathbf{x}(s)\}_{s \le t})$ to get $\{\mathbf{v}_1^{(n)}\}$, as pseudo-observations of the upper-layer flow.
3. The **second DA step** solves the **lower-layer** flow posterior $P(\mathbf{v}_2|\mathbf{v}_1^{(n)}, \{\mathbf{x}(s)\}_{s \le t})$ for each sample. The ultimate posterior that combines N_s samples is

$$\begin{aligned} P(\mathbf{v}_2|\{\mathbf{x}(s)\}_{s \le t}) &= \int_{\mathbf{v}_1} P(\mathbf{v}_1, \mathbf{v}_2|\{\mathbf{x}(s)\}_{s \le t})d\mathbf{v}_1 \ \textit{(marginal probability)} \\ &= \int_{\mathbf{v}_1} P(\mathbf{v}_2|\mathbf{v}_1, \{\mathbf{x}(s)\}_{s \le t})P(\mathbf{v}_1|\{\mathbf{x}(s)\}_{s \le t})d\mathbf{v}_1 \ \textit{(conditional probability)} \\ &\approx \frac{1}{N_s}\sum_{n}^{N_s} P(\mathbf{v}_2|\mathbf{v}_1^{(n)}, \{\mathbf{x}(s)\}_{s \le t}) \ \textit{(Monte-Carlo estimation)} \\ &\approx \frac{1}{N_s}\sum_{n}^{N_s} P(\mathbf{v}_2|\{\mathbf{v}_1^{(n)}(s)\}_{s \le t}, \{\mathbf{x}(s)\}_{s \le t}) \ \textit{(if conditioned on trajectory)}. \end{aligned} \tag{6}$$

Multi-step CGDA



One-step CGDA

Comparing the flow-observation system (1) to CGNS (2), we can fit (1) into the CGNS framework by dropping the quadratic term $\mathbf{B}(\mathbf{v}, \mathbf{v})$. *The one-step CGDA adopts a linear stochastic flow model and perform CGDA to all layers at once.* It is equivalent to the previous work using linear regression[1].

Multi-step CGDA

1. The one-step CGDA can work as the first step of multi-step CGDA to recover the surface-layer flow.
2. Sample from the upper-layer flow posterior to get pseudo-observations.
3. At the second DA step, we drop nonlinear terms of the *lower-layer flow*, but *preserve nonlinear terms of the upper-layer flow, to fit a CG nonlinear stochastic flow model, and perform CGDA to solve the lower-layer posterior.*
4. Sequentially apply step 2 and 3 top-down till to the bottom.

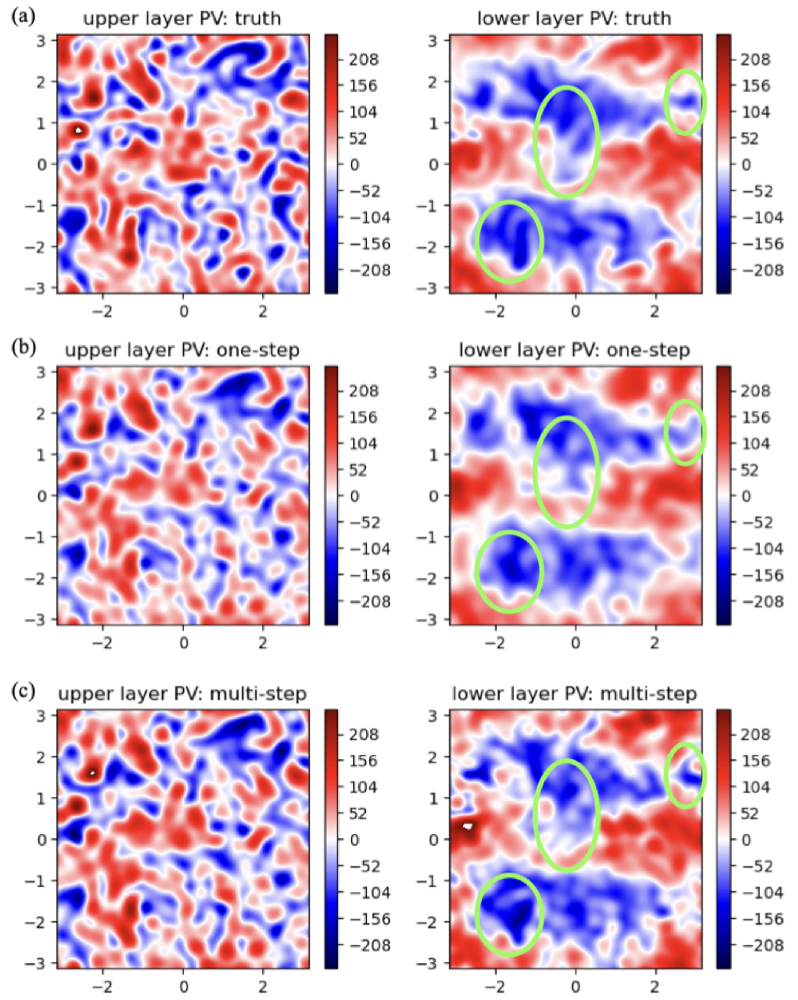
Pros of Multi-step CGDA

The one-step CGDA that completely linearizes the flow model tends to be an aggressive simplification in most real-world applications. As the flow model contributes a major part of the flow-observation system's complexity, and nonlinearities contribute a major part of the flow model's complexity.

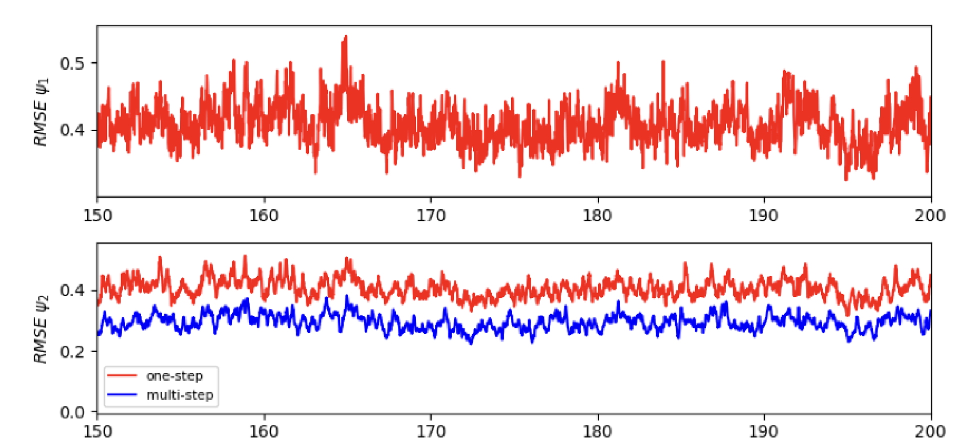
The multi-step CGDA bypasses the oversimplification by adopting **Monte-Carlo estimation**, and **performing CGDA layer by layer**. It allows

- **analytic formulae to solve the posterior mean and covariance**
- **preserves the nonlinearity of the surface observation process** at the first DA step, and the **nonlinearity of the upper-layer flow** at the subsequent DA steps
- **non-Gaussian lower-layer posterior**, which is given by a mixture of Gaussians

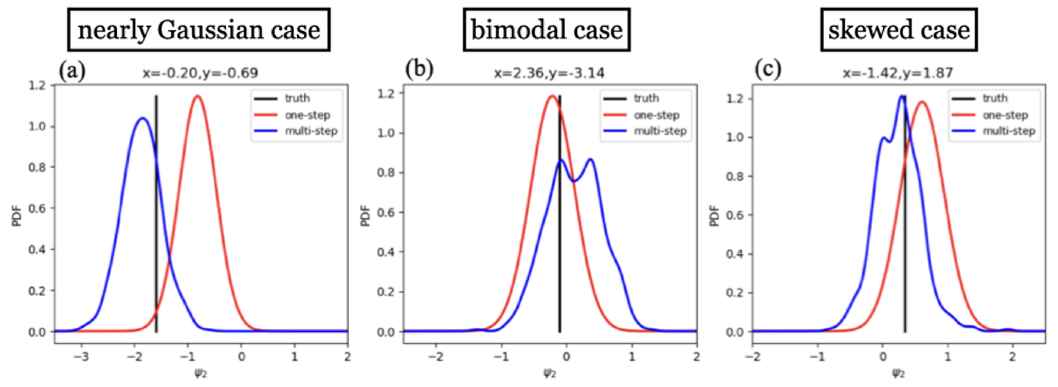
Application to two-layer QG system



Snapshots of the potential vorticity field



Posterior RMSE series



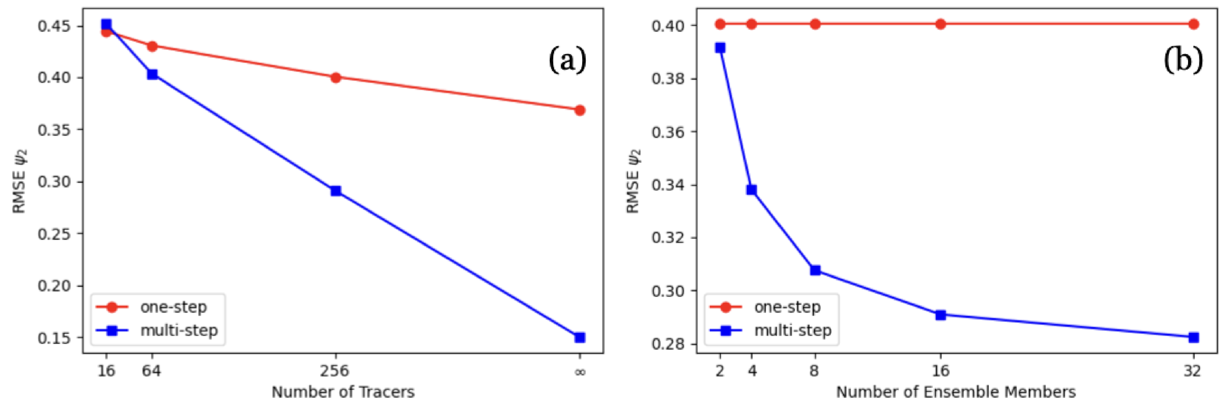
Snapshots of the posterior pdfs at three different displacements

Number of time steps	$N_t = 205000$
Time step	$dt = 0.002$
Domain	$[-\pi, \pi)^2$
Grid points (Fourier modes)	$K_x = K_y = 128$
Spectral truncation radius	$r = 16$
Observation noise strength	$\sigma_x = \sigma_y = 0.1$
Number of tracers	$L = 256$
Ensemble size	$N_s = 16$
Deformation wavenumber	$k_d = 10$
Rossby parameter	$\beta = 22$
Zonal mean flow	$U_0 = 0$
Zonal shear flow	$U = 1$
Ekman damping	$\kappa = 9$
Hyperviscosity	$\nu = 10^{-12}$
Topography	$h = 40(cos(x) + 2cos(2y))$

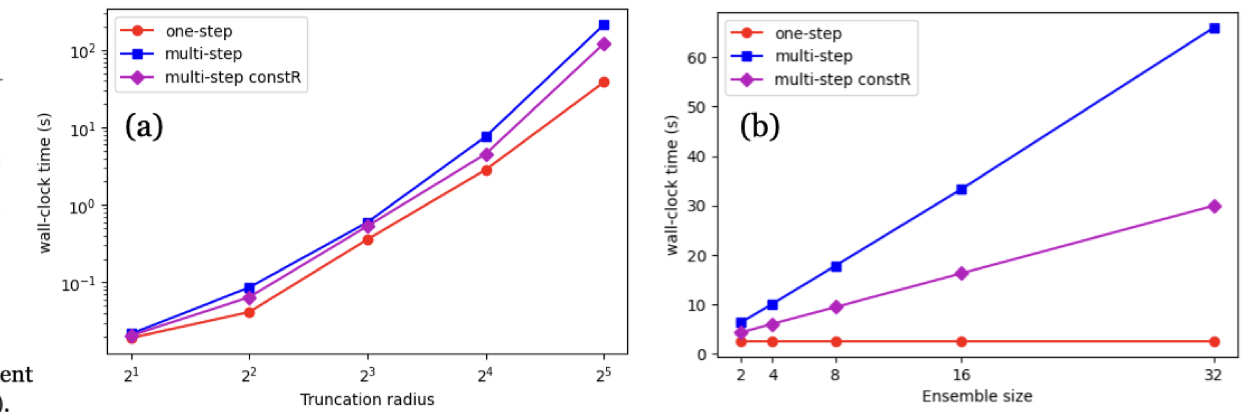
Table 1.Default experiment parameters

Regime	$\beta = 22$	$\beta = 111$
One-step CGDA	0.400	0.137
Multi-step CGDA	0.291	0.113
Multi-step CGDA /w constR	0.303	-

Table 2. Time mean RMSEs of DA methods in turbulent regime ($\beta = 22$) and less turbulent regime ($\beta = 111$). "/w constR" means with constant covariance.



Sensitivities to the (a) number of tracers, and (b) ensemble size



Computational cost varying with (a) spectral truncation radius, and (b) ensemble size

Future work

Parameter estimation: A better state estimation usually also leads to a better parameter estimation. Constant forcing parameters, e.g., topography, can be estimated using the multi-step CG smoother. Applications include **recovering the ocean bottom topography** based on tracers.

Reference

[1] Anne Molcard, Annalisa Griffa, and Tamay M. Özgökmen. Lagrangian Data Assimilation in Multilayer Primitive Equation Ocean Models. January 2005. doi:10.1175/JTECH-1686.1. URL https://journals.ametsoc.org/view/journals/atot/22/1/jtech-1686_1.xml. Section: Journal of Atmospheric and Oceanic Technology.

[2] Zhongrui Wang, Nan Chen, and Di Qi. A Closed-Form Nonlinear Data Assimilation Algorithm for Multi-Layer Flow Fields. 2024. preprint available upon request.