Ocean data assimilation based on MMD between sets of profiles

Nozomi Sugiura (JAMSTEC, Japan)

15 Nov 2022

Motivation

- In Global Ocean Data Assimilation, we are interested in heat and flow fields in climatological time scale, rather than focusing on phenomena that occur over a short period of time.
- Because we have plenty of profile observations, we would better compare observational and model profiles directly, rather than comparing Temperature and Salinity at each point in a profile.
- If we divide the model domain into spatio-temporal meshes of about 1 degree and 1 month, there are often multiple observed profiles in a mesh, because Argo has about 4000 floats distributed heterogeneously, which are drifting slowly around one location and making observations a few times a month.
- Taking into account these situations, we propose a problem setting of data assimilation based on the comparison of signature averages of observation and model profiles in each mesh.
- A version of `MRI.com` is employed as ocean general circulation model (OGCM).
- We will solve it by 4D-Var.

Signature transform (1/2)

A vertical profile can be regarded as a sequence of 3-d vectors

$$\{X_t \in \mathbb{R}^3 |\ t = t_0, t_1, \cdots, t_L; t_0 = 0, t_L = 1\},$$

where X_t is composed of pressure $X_t^{(1)} = P_t$, salinity $X_t^{(2)} = S_t$, and temperature $X_t^{(3)} = T_t$.

- We interpolate X_t in interval $[t_\ell, t_{\ell+1}] \subset [0, 1]$ as

$$X_t = \frac{t_{\ell+1} - t}{t_{\ell+1} - t_\ell} X_{t_\ell} + \frac{t - t_\ell}{t_{\ell+1} - t_\ell} X_{t_{\ell+1}}.$$

- We then define the first iterated integrals as

$$\mathcal{S}^{(i_1)}(X)_{0,1} = \int_{u=0}^{1} dX_u^{(i_1)} = X_1^{(i_1)} - X_0^{(i_1)}, \quad i_1 = 1, 2, 3,$$

which is just the vector from the starting point to the endpoint.

Signature transform (2/2)

- The second iterated integrals are defined as

$$\mathcal{S}^{(i_1 i_2)}(X)_{0,1} = \int_{u=0}^{1} \mathcal{S}^{(i_1)}(X)_{0,u} dX_u^{(i_2)} = \int_{u=0}^{1} (X_u^{(i_1)} - X_0^{(i_1)}) dX_u^{(i_2)}.$$

- Similarly, the n-th iterated integrals are defined recursively as

$$\mathcal{S}^{(i_1 i_2 \cdots i_n)}(X)_{0,1} = \int_{u=0}^{1} \mathcal{S}^{(i_1, i_2, \cdots, i_{n-1})}(X)_{0,u} dX_u^{(i_n)}.$$

- Using the basis $\{e_1, e_2, e_3\}$ of $\mathbb{R}^3$, we define a formal power series *

$$\mathcal{S}(X) = 1 + \sum_{i_1=1,2,3} \mathcal{S}^{(i_1)}(X) e_{i_1} + \sum_{i_1, i_2=1,2,3} \mathcal{S}^{(i_1 i_2)}(X) e_{i_1} e_{i_2} + \cdots,$$

  which is called the signature of path X (Chevyrev and Kormilitzin, 2016; Lyons et al., 2007).
- In practical use, we truncate the signature up to the n-th iterated integrals:

$$\mathcal{S}_n(X) = 1 + \sum_{i_1=1,2,3} \mathcal{S}^{(i_1)}(X) e_{i_1} + \cdots + \sum_{i_1, \cdots, i_n=1,2,3} \mathcal{S}^{(i_1 \cdots i_n)}(X) e_{i_1} \cdots e_{i_n}.$$

*The subscript $_{0,1}$ in $\mathcal{S}^{(i_1)}(X)_{0,1}$ is omitted.

Lévy Area

Among others, an important feature of a path is represented by the Lévy area:

$$\frac{\mathcal{S}^{(i_1 i_2)}(X) - \mathcal{S}^{(i_2 i_1)}(X)}{2} = \int_{0\le u_1<u_2\le 1} \left(dX^{(i_1)}_{u_1} dX^{(i_2)}_{u_2} - dX^{(i_2)}_{u_1} dX^{(i_1)}_{u_2} \right)/2,$$

where $1 \le i_1 < i_2 \le 3$. As shown in Fig. 1, Lévy area is the area enclosed by the path and the chord.

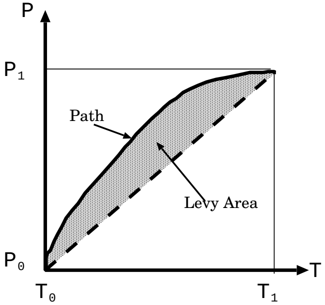


Figure 1: Lévy area for P and T, $\int_{0\le u_1<u_2\le 1} \left(dP_{u_1} dT_{u_2} - dT_{u_1} dP_{u_2}\right)/2$, which is analogous to heat content.

How signature grasps the shape (order-1)

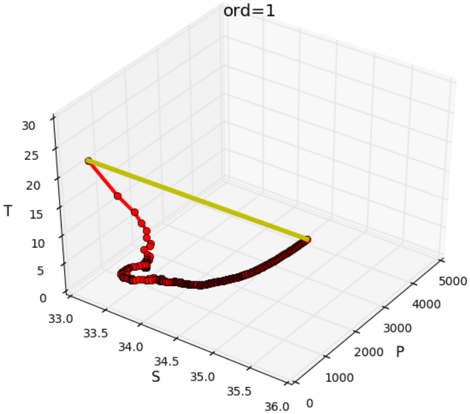


Figure 2: How order-1 signature $\mathcal{S}_1(X)$ grasps the shape of an oceanic profile. T: Water Temperature, S: Salinity, P: Water Pressure.

How signature grasps the shape (order-2)

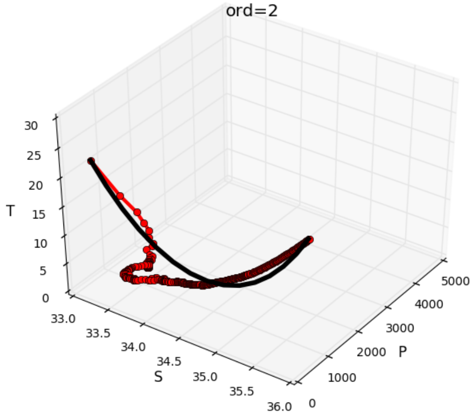


Figure 3: How order-2 signature $\mathcal{S}_2(X)$ grasps the shape of an oceanic profile. T: Water Temperature, S: Salinity, P: Water Pressure.

How signature grasps the shape (order-3)

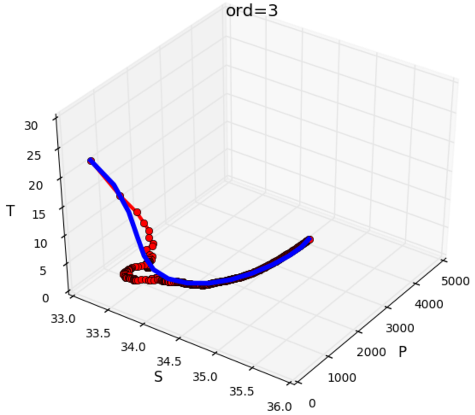


Figure 4: How order-3 signature $\mathcal{S}_3(X)$ grasps the shape of an oceanic profile. T: Water Temperature, S: Salinity, P: Water Pressure.

How signature grasps the shape (order-4)

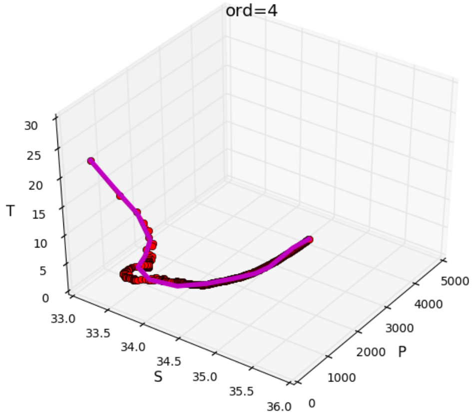


Figure 5: How order-4 signature $\mathcal{S}_4(X)$ grasps the shape of an oceanic profile. T: Water Temperature, S: Salinity, P: Water Pressure.

Data Assimilation Problem

- Suppose we have an inversion problem

$$y = G(u) + \eta, \tag{1}$$

  where G is an ocean general circulation model (OGCM), u is the control variables (initial and boundary conditions), $G(u)$ is the output variables (a set of profiles), y is the observation (a set of Argo profiles), and η is the observational error.
- We set the simulation domain to global ocean from Jan 2012 to Dec 2012. It is divided into spatial meshes of resolution 1×0.5 degree and temporal meshes of monthly resolution. For example, a mesh is defined in the range of 10N to 10.5N, 140E to 141E, February 2012.
- Let π_m be the restriction operator to the m-th spatio-temporal mesh, we define the problem for mesh m as

$$y_m = G_m(u) + \eta_m, \tag{2}$$

  where $G_m := \pi_m \circ G$ is the OGCM that generates profiles in mesh m, u is the control variables (initial condition), $G_m(u) := \pi_m \circ G(u)$ is the set of profiles in mesh m, y_m is the set of Argo profiles in mesh m, and η_m is the observational error for mesh m (assumed to be independent).

Measures and Empirical Measures

- Instead of assuming the probability distribution (measure) of η_m to be Gaussian, we compare the model and observational measures for mesh m:

$$\text{profile } X \in G_m(u) \implies X \sim P_{m,u}, \tag{3}$$
$$\text{profile } \quad Y \in y_m \implies Y \sim Q_m. \tag{4}$$

- These measures, $P_{m,u}$ and Q_m, are approximated by empirical measures:

$$\tilde{P}_{m,u} = \frac{1}{|G_m(u)|} \sum_{X \in G_m(u)} \delta_X, \tag{5}$$
$$\tilde{Q}_m = \frac{1}{|y_m|} \sum_{Y \in y_m} \delta_Y, \tag{6}$$

where $|y_m|$ denotes the number of observational profiles in mesh m, and δ_Y is the delta function.

Maximum Mean Discrepancy

- The distance between two measures can be evaluated by kernel averages, which constitute Maximum Mean Discrepancy (MMD). That has recently been used in estimation problems (Chérief-Abdellatif and Alquier, 2020).
- When paths $X \sim P$ are embedded in the tensor space $\mathcal{T}$ of the signatures by $\mathcal{S} : X \mapsto \mathcal{S}(X) \in \mathcal{T}$, we can define the kernel mean embedding of measure P as $\mu_k(P) := \mathbb{E}_{X\sim P}[\mathcal{S}(X)]$. Then, the MMD between the two measures is defined as

$$\mathrm{MMD}(\tilde{P}_{m,u}, \tilde{Q}_m) = \|\mu_k(\tilde{P}_{m,u}) - \mu_k(\tilde{Q}_m)\|_{\mathcal{T}}. \tag{7}$$

- In our case, Eq. (7) is thus written as

$$\mathrm{MMD}^2(\tilde{P}_{m,u}, \tilde{Q}_m) = \left\| \frac{1}{|G_m(u)|} \sum_{X\in G_m(u)} \mathcal{S}(X) - \frac{1}{|y_m|} \sum_{Y\in y_m} \mathcal{S}(Y) \right\|_{\mathcal{T}}^2 . \tag{8}$$

- This is nothing but the comparison of signature averages for the sets of model profiles in a mesh and those for observation profiles.
- It can be seen as a path-to-path version of moment matching, $(x-y)^2, (x^2-y^2)^2, \cdots$, in the case of point-to-point comparison.

Homogeneous cost function

How we define the norm, $\| \quad \|_{\mathcal{T}}$, in Eq. (8)?

- The signature does not live in a linear space, but in tensor space $\mathcal{T}$.
- If we dilate path X to λX, the k-th iterated integral scales to

$$\mathcal{S}^{(i_1 \cdots i_k)}(\lambda X) = \lambda^{k} \mathcal{S}^{(i_1 \cdots i_k)}(X), \quad k = 1, \cdots, n. \tag{9}$$

- To be consist with this scaling property, we define the MMD (8) as

$$\mathrm{MMD}^2 = \sum_m \sum_{k=1}^{n} \left[\sum_{i_1, \cdots, i_k} \left(\frac{1}{|G_m(u)|} \sum_{X \in G_m(u)} \mathcal{S}^{(i_1 \cdots i_k)}(X) - \frac{1}{|y_m|} \sum_{Y \in y_m} \mathcal{S}^{(i_1 \cdots i_k)}(Y) \right)^2 \right]^{1/k}. \tag{10}$$

- We use this as observational cost function: $J_{\mathrm{obs}}(u) = \frac{1}{2}\mathrm{MMD}^2$. Thereby, if we dilate path X to λX and Y to λY, then $J_{\mathrm{obs}}(u)$ scales to $\lambda^2 J_{\mathrm{obs}}(u)$.
- We apply $n = 4$, so that the 1-st to 4-th iterated integrals are taken into account.
- By minimizing the cost function, we can make closer the probability distributions of the model and the observation, each of these distribution defines how to generate the profiles in a mesh.

4D-Var

- Our control vector u comprises of the initial condition and the air-sea fluxes.
- We define the background cost as

$$J_{\text{bg}}(u) = \frac{1}{2}(u - u_b)^T B^{-1}(u - u_b), \tag{11}$$

where u_b is the firstguess vector and B is the background error covariance.
- By introducing a spatial smoothing operator S, we change variable u into v as

$$u = B^{\frac{1}{2}} S v + u_b. \tag{12}$$

- We finally define the cost function with respect to v as

$$\begin{aligned} \mathcal{J}(v) &:= J_{\text{bg}}(B^{\frac{1}{2}} S v + u_b) + \lambda^2 J_{\text{obs}}(B^{\frac{1}{2}} S v + u_b) \\ &= \frac{1}{2} v^T S^T S v + \lambda^2 J_{\text{obs}}(B^{\frac{1}{2}} S v + u_b). \end{aligned} \tag{13}$$

where λ is a tunable dilation factor.
- Our 4D-Var minimizes the cost function $\mathcal{J}(v)$ by using the BFGS iterations.

Experimental Result

- The south of the Greenland was excluded from observation because of the poor representation there by the model.
- We used dilation factor $\lambda = 10^3$, and 35 iterations were performed.
- The variation of the total and observational costs went as follows.

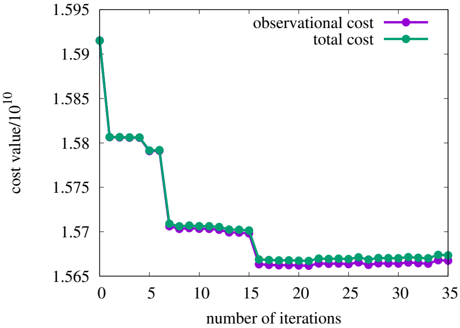


Figure 6: The variation of the cost function. Horizontal axis is the number of iterations and vertical axis is the cost value.

Comparison of the errors for the Iterated Integrals

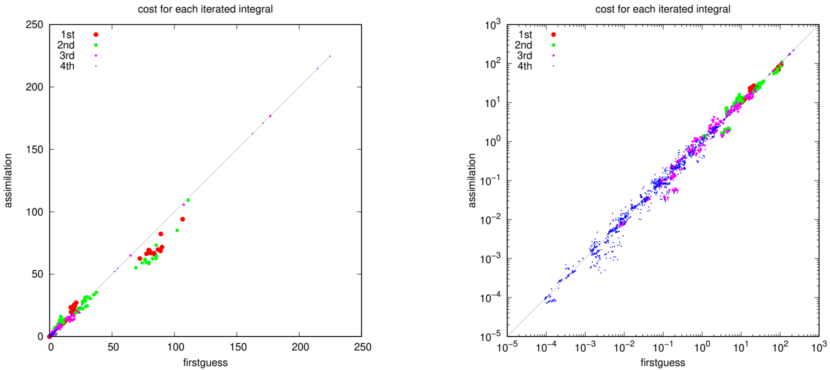


Figure 7: Comparison of the error for the iterated integrals,
$\mathbb{E}_{\text{global}}\left[\left(\overline{\mathcal{S}^{(i_1\cdots i_n)}(X)}-\overline{\mathcal{S}^{(i_1\cdots i_n)}(Y)}\right)^2\right]^{\frac{1}{2}}$, in linear scale (left) and log-log scale (right), where X is from model and Y is from observation. Vertical axis is firstguess, and horizontal axis is assimilation. Dots in lower-right side indicate improvement. The 1-st to 4-th iterated integrals are shown in red, green, magenta, and blue, respectively.

Global Distribution of Lévy area for P and T

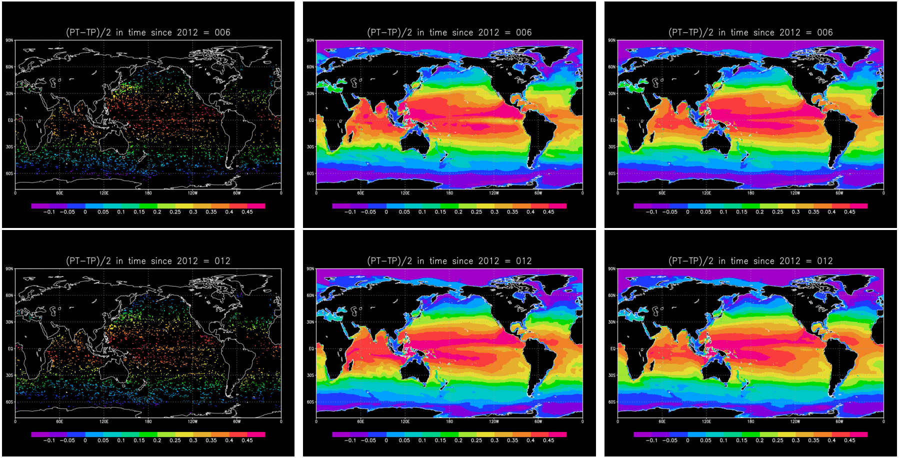


Figure 8: The averages of Lévy area $\int (dPdT - dTdP)/2$ for observation (left), assimilation (center) and firstguess (right) in June (top) and December (bottom) of 2012.

Global Distribution of Lévy area for P and S

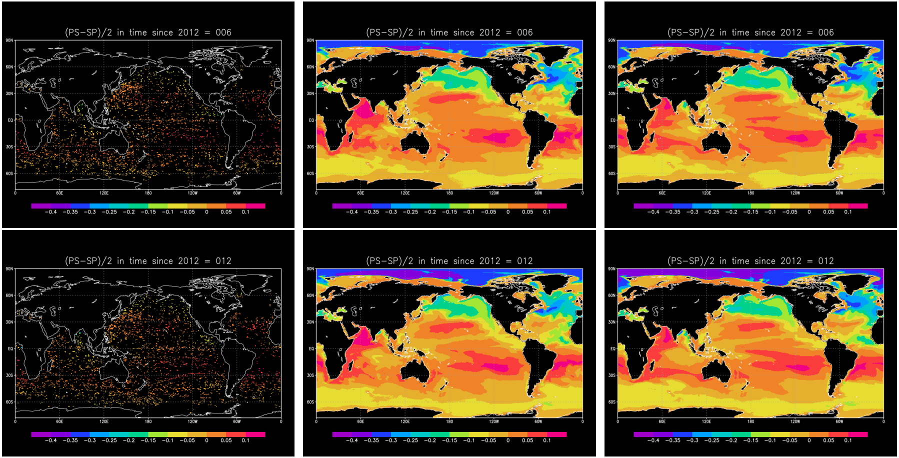


Figure 9: The averages of Lévy area $\int(dPdS - dSdP)/2$ for observation (left), assimilation (center) and firstguess (right) in June (top) and December (bottom) of 2012.

Global Distribution of Lévy area for S and T

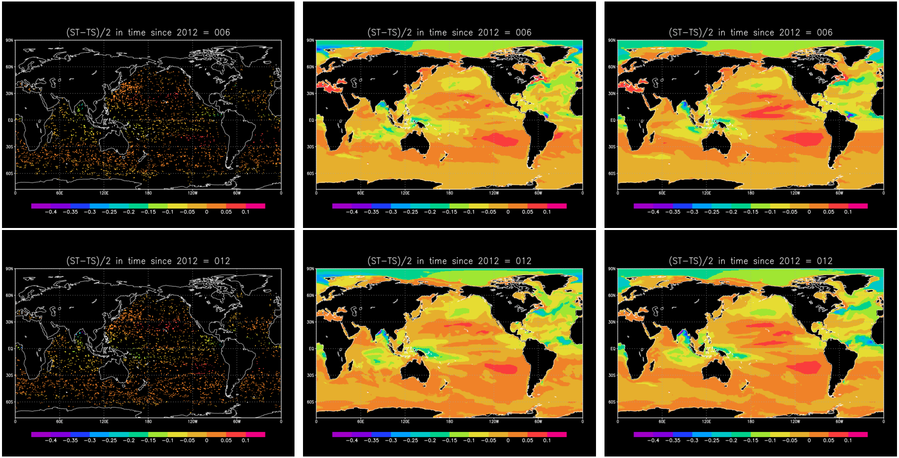
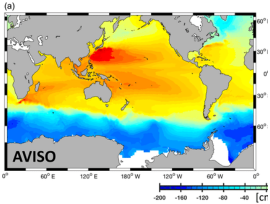


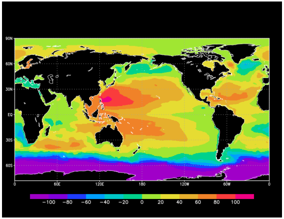
Figure 10: The averages of Lévy area $\int(dSdT - dTdS)/2$ for observation (left), assimilation (center) and firstguess (right) in June (top) and December (bottom) of 2012.

Sea Surface Height

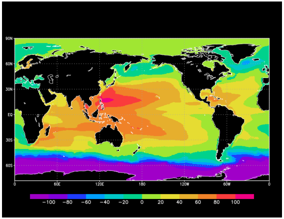
Considerable changes were seen in sea surface height.



obs. (not assimilated)



assimilation



firstguess

Figure 11: Annual mean sea surface height for assimilation (center) and firstguess (right), and a climatology from AVISO (left; Dietze et al. (2020)).

Conclusion

- We designed an ocean data assimilation problem in which the signature-based MMD[†] is employed as the distance between observation and model measures at each mesh.
- Preliminary experiment showed that the problem can be successfully solved by 4D-Var.
- It will be important to examine whether or not this approach can create a good estimation of ocean circulation field.
- There remains to be solved how to determine the dilation factor.
- This approach is novel because it is based on the comparison of profile to profile, rather than of point to point.

[†]MMD: maximum mean discrepancy

References

Chérief-Abdellatif, B.-E. and Alquier, P. (2020). MMD-Bayes: Robust Bayesian estimation via maximum mean discrepancy. In *Symposium on Advances in Approximate Bayesian Inference*, pages 1–21. Proceedings of Machine Learning Research.

Chevyrev, I. and Kormilitzin, A. (2016). A Primer on the Signature Method in Machine Learning. *arXiv preprint arXiv:1603.03788*.

Dietze, H., Löptien, U., and Getzlaff, J. (2020). MOMSO 1.0 – an eddying Southern Ocean model configuration with fairly equilibrated natural carbon. *Geoscientific Model Development*, 13:71–97.

Lyons, T. J., Caruana, M., and Lévy, T. (2007). *Differential Equations Driven by Rough Paths*, volume 1908 of *Lecture Notes in Mathematics*. Springer.

Acknowledgment

This work was supported by JST, AIP Trilateral AI Research, Grant Number JPMJCR20G5, Japan.