An Introduction to Analog Data Assimilation

Pierre Tandeo and Yicun Zhen

IMT Atlantique, Lab-STICC, UBL, Brest, France

October 14, 2019

European Commission

COPERNICUS
MARINE ENVIRONMENT MONITORING SERVICE
Providing PRODUCTS and SERVICES for all marine applications

Q: What is data assimilation?
A: The art of ultilizing the observed data.

Data Assimilation produces (hopefully):

- better state forecasts/reanalysis:
  use the observed data to correct your state estimtates.
- uncertainty quantification:
  provide a covariance matrix that tells you the level of confidence for the current state estimates.
- model/parameter validation:
  the observed data can tell you if your model/parameter has significant errors.
- observation sensitivity test:
  which observations contribute the most to improve the state analysis?
  design of new observations?

Kalman filter/smoother in a flow chart

Given a dynamical system (with noise) and an observation system (with noise):

$$\begin{cases} \mathbf{x}_t = \mathbf{M}\mathbf{x}_{t-1} + \sqrt{Q}\boldsymbol{\eta}_t \\ \mathbf{y}_t^{\text{obs}} = \mathbf{H}_t^{\text{obs}}\mathbf{x}_t + \sqrt{R}\boldsymbol{\epsilon}_t \end{cases}$$

To compute the analysis $\mathbf{x}^a(t)$(Kalman filter)

$$\boxed{\mathbf{x}_1^f} \xrightarrow{\mathbf{y}_1^{\text{obs}}} \boxed{\mathbf{x}_1^a} \xrightarrow{\text{model}} \boxed{\mathbf{x}_2^f} \dashrightarrow \boxed{\mathbf{x}_T^f} \xrightarrow{\mathbf{y}_T^{\text{obs}}} \boxed{\mathbf{x}_T^a}$$

To compute the reanalysis $\mathbf{x}^s(t)$(Kalman smoother)

$$\boxed{\mathbf{x}_1^s} \dashleftarrow \boxed{\mathbf{x}_{T-2}^s} \xleftarrow{\text{KS}} \boxed{\mathbf{x}_{T-1}^s} \xleftarrow{\text{KS}} \boxed{\mathbf{x}_T^s = \mathbf{x}_T^a}$$

Kalman filter/smoother in mathematical formula

Dynamical and observation system:

$$\begin{cases} \mathbf{x}_t = \mathbf{M}\mathbf{x}_{t-1} + \sqrt{\mathbf{Q}}\boldsymbol{\eta}_t \\ \mathbf{y}_t^{\text{obs}} = \mathbf{H}_t^{\text{obs}}\mathbf{x}_t + \sqrt{\mathbf{R}}\boldsymbol{\epsilon}_t \end{cases}$$

Kalman Filter:

$$\mathbf{K}_t = \mathbf{P}_t^f(\mathbf{H}_t^{\text{obs}})^\top(\mathbf{R} + \mathbf{H}_t^{\text{obs}}\mathbf{P}_t^f(\mathbf{H}_t^{\text{obs}})^\top)^{-1} \text{ Kalman gain matrix}$$

$$\mathbf{x}_t^a = \mathbf{x}_t^f + \mathbf{K}_t(\mathbf{y}_t^{\text{obs}} - \mathbf{H}_t^{\text{obs}}\mathbf{x}_t^f) \text{ state analysis}$$

$$\mathbf{P}_t^a = \mathbf{P}_t^f - \mathbf{K}_t\mathbf{H}_t^{\text{obs}}\mathbf{P}_t^f = Cov(\mathbf{x}_t^a, \mathbf{x}_t^a)$$

$$\mathbf{x}_{t+1}^f = \mathbf{M}\mathbf{x}_t + \sqrt{\mathbf{Q}}\eta_t \text{ state forecast}$$

$$\mathbf{P}_{t+1}^f = \mathbf{M}\mathbf{P}_t^a\mathbf{M}^\top + \mathbf{Q} = Cov(\mathbf{x}_{t+1}^f, \mathbf{x}_{t+1}^f)$$

Kalman Smoother:

$$\mathbf{C}_t = \mathbf{P}_t^a\mathbf{M}^\top = Cov(\mathbf{x}_t^a, \mathbf{x}_{t+1}^f)$$

$$\mathbf{J}_t = \mathbf{C}_t(\mathbf{P}_{t+1}^f)^{-1} \text{The gain matrix for the smoother}$$

$$\mathbf{x}_t^s = \mathbf{x}_t^a + \mathbf{J}_t(\mathbf{x}_{t+1}^s - \mathbf{x}_{t+1}^f) \text{ state reanalysis}$$

$$\mathbf{P}_t^s = \mathbf{P}_t^a + \mathbf{J}_t(\mathbf{P}_{t+1}^s - \mathbf{P}_{t+1}^f)\mathbf{J}_t^\top = Cov(\mathbf{x}_t^s, \mathbf{x}_t^s)$$

The algorithm of ensemble Kalman filter/smoother

Ensemble KF/KS:

Ensemble simulation $\Rightarrow \mathbf{P}_t^f, \mathbf{P}_t^a, \mathbf{P}_t^s, \mathbf{C}_t$ can be directly calculated as the sample covariance.

Algorithm 1 Ensemble Kalman Smoother

$t = 1, ..., T$ and $i = 1, ..., N_e$.
Input: $\mathbf{x}_{i,1}^f$, $\mathbf{H}_t, \mathbf{R}, F(\cdot), \mathbf{y}_t$,
Output: $\hat{\mathbf{x}}_t^s$, $\mathbf{P}_t^s$

The forward ensemble Kalman filter

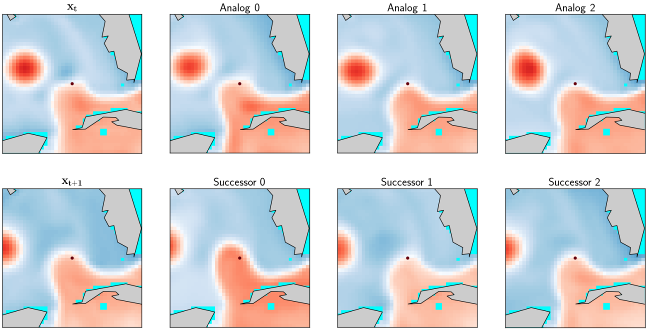
1: **for** $t = 1, 2, ..., T$ **do**:
2: $\quad \bar{\mathbf{x}}_t^f \leftarrow \frac{1}{N_e} \sum_{i=1}^{N_e} \mathbf{x}_{i,t}^f$
3: $\quad \mathbf{P}_t^f \leftarrow \frac{1}{N_e - 1} \sum_{i=1}^{N_e} (\mathbf{x}_{i,t}^f - \bar{\mathbf{x}}_t^f)(\mathbf{x}_{i,t}^f - \bar{\mathbf{x}}_t^f)^\top$
4: $\quad \mathbf{K}_t \leftarrow \mathbf{P}_t^f \mathbf{H}_t^\top (\mathbf{R} + \mathbf{H}_t \mathbf{P}_t^f \mathbf{H}_t^\top)^{-1}$
5: $\quad$ Draw $\varepsilon_{i,t} \sim \mathcal{N}(0, \mathbf{R})$
6: $\quad \mathbf{x}_{i,t}^a \leftarrow \mathbf{x}_{i,t}^f + \mathbf{K}_t (\mathbf{y}_t - \mathbf{H}_t \mathbf{x}_{i,t}^f + \varepsilon_{i,t})$
7: $\quad \mathbf{x}_{i,t+1}^f \leftarrow F(\mathbf{x}_{i,t}^a)$, forecast the state at $t+1$. When EnKS is applied within AnDA, F is replaced by the analog forecast.

The backward ensemble Kalman smoother

8: $\mathbf{x}_{i,T}^s \leftarrow \mathbf{x}_{i,T}^a$
9: **for** t=T-1,T-2,...,1 **do**:
10: $\quad \bar{\mathbf{x}}_t^a \leftarrow \frac{1}{N_e} \sum_{i=1}^{N_e} \mathbf{x}_{i,t}^a$
11: $\quad \mathbf{A}_t \leftarrow \frac{1}{N_e - 1} \sum_{i=1}^{N_e} (\mathbf{x}_{i,t}^a - \bar{\mathbf{x}}_t^a)(\mathbf{x}_{i,t+1}^f - \bar{\mathbf{x}}_{t+1}^f)^\top$
12: $\quad \mathbf{J}_t \leftarrow \mathbf{A}_t (\mathbf{P}_t^f)^{-1}$
13: $\quad \mathbf{x}_{i,t}^s \leftarrow \mathbf{x}_{i,t}^a + \mathbf{J}_t (\mathbf{x}_{i,t+1}^s - \mathbf{x}_{i,t+1}^f)$
14: $\quad \hat{\mathbf{x}}_t^s \leftarrow \frac{1}{N_e} \sum_{i=1}^{N_e} \mathbf{x}_{i,t}^s$
15: $\quad \mathbf{P}_t^s \leftarrow \frac{1}{N_e - 1} \sum_{i=1}^{N_e} (\mathbf{x}_{i,t}^s - \bar{\mathbf{x}}_t^s)(\mathbf{x}_{i,t}^s - \bar{\mathbf{x}}_t^s)^\top$

The analog forecast method

- Have a huge amount of historical data (the catalog);
- For a given initial state $\mathbf{x}_t$, find the similar states (analogs) in the historical database;
- Calculate $\mathbf{x}_{t+1}$ basesd on the analogs and the corresponding successors.



Analog forecast algorithm in detail

$$\mathbf{x}(t) = \mathbf{M}(t-1)\mathbf{x}(t-1) + \sqrt{\mathbf{Q}(t-1)}\boldsymbol{\eta}(t-1)$$

Analog Forecast (locally linear)

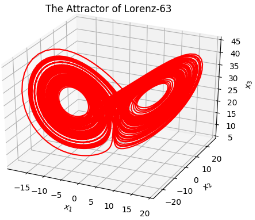
- Given the state estimate $\mathbf{x}_t$
- Find the k (=50, for instance) analogs that are the closest to $\mathbf{x}_t$: $\mathcal{A}_1, ..., \mathcal{A}_k$, and calculate the mean $\bar{\mathcal{A}} = \frac{1}{k}(\mathcal{A}_1 + ... + \mathcal{A}_k)$
- Linearly regress $\mathcal{S}_1, ..., \mathcal{S}_k$ on $\mathcal{A}_1 - \bar{\mathcal{A}}, ..., \mathcal{A}_k - \bar{\mathcal{A}}$: $\mathcal{S} = \mathbf{M}(\mathcal{A} - \bar{\mathcal{A}}) + \mathbf{b}$
- Apply the linear local model on $\mathbf{x}_t$: $\mathbf{x}_{t+1} \leftarrow \mathbf{M}(\mathbf{x}_t - \bar{\mathcal{A}}) + \mathbf{b}$
- Covariance inflation: $\mathbf{x}_{t+1} \leftarrow \mathbf{x}_{t+1} + \mathcal{N}(0, \mathbf{Q}_{t+1})$ where $\mathbf{Q}_{t+1} = cov(\mathcal{S} - (\mathbf{M}(\mathcal{A} - \bar{\mathcal{A}}) + \mathbf{b}))$

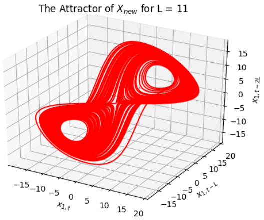
Time-Delayed Analog Forecast

Motivation

Taken's Theorem (1981): under certain conditions, a strange attractor can be resconstructed using lagged partial observations.

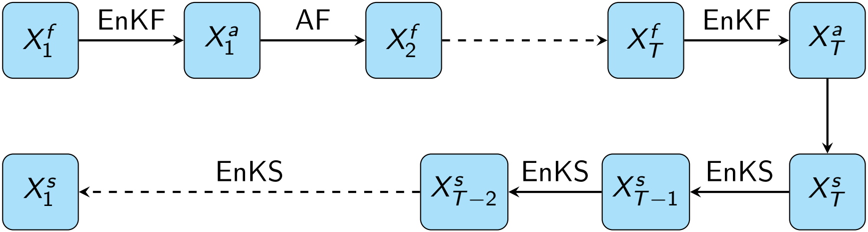
$\Rightarrow$Construct time-delayed analogs $A_t^{new} = (A_t, A_{t-L}, ..., A_{t-kL})$ and similarly for successors and state estimates.





Analog Data Assimilation (AnDA) for Reanalysis

- Ensemble Kalman filter(EnKF) to calculate the state analysis;
- Analog forecast (AF) for state forecast;
- Ensemble Kalman smoother (EnKS) for calculating the state reanalysis.



Objective interpolation (OI)–a widely used model-free method for calculating the reanalysis

Sketch of the algorithm

Let $x_{t,i}$ denote the i-th component of the state vector $\mathbf{x}_t$. Let $\bar{x}_i$ be the temporal mean value of x_i. We want to construct $x^s_{t,i}$ for $t = 1, 2, ..., T$.

- For each pair of (t, i), define a cylinder in space and time. $x^s_{t,i}$ will be calculated only based on the observations in this cylinder. For instance, for each pair of (t, i), we only consider the obs that are within 150(km) from x_i and 10 (days) from t.
- Define a spatial-temporal background covariance matrix $\mathbf{B}$. For instance, $\mathbf{B}(x_{t_1,i}, x_{t_2,j}) = \sqrt{var(x_i)var(x_j)}\exp\{-\frac{d^2_{ij}}{L^2_s} - \frac{d^2_t}{L^2_t}\}$, or $\mathbf{B}(x_{t_1,i}, x_{t_2,j}) = \mathbf{B}^{\text{clim}} \otimes \exp\{-\frac{d^2_t}{L^2_t}\}$, where L_s, L_t are parameters that need to be tuned.
- $x^s_{t,i} = \bar{x}_i + \mathbf{B}_{\text{loc}}\mathbf{H}^\top_{\text{loc}}(\mathbf{R}_{\text{loc}} + \mathbf{H}_{\text{loc}}\mathbf{B}_{\text{loc}}\mathbf{H}^\top_{\text{loc}})^{-1}(\mathbf{y}^{\text{obs}}_{\text{loc}} - \mathbf{H}_{\text{loc}}\bar{\mathbf{x}})$
  $B^s_{t,i} = \{\mathbf{B}_{\text{loc}} - \mathbf{B}_{\text{loc}}\mathbf{H}^\top_{\text{loc}}(\mathbf{R}_{\text{loc}} + \mathbf{H}_{\text{loc}}\mathbf{B}_{\text{loc}}\mathbf{H}^\top_{\text{loc}})^{-1}\mathbf{H}_{\text{loc}}\mathbf{B}_{\text{loc}}\}_{(t,i),(t,i)}$

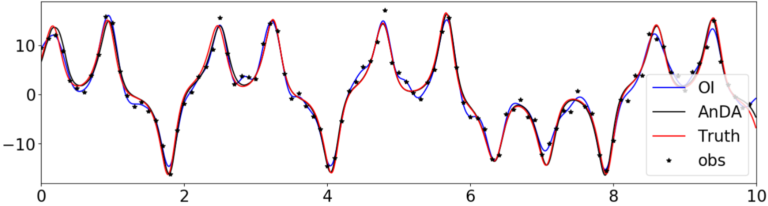
Lorenz model experiments

Lorenz 63

- $\mathrm{dt} = 0.01$, $\mathrm{dt_{obs}} = 0.08$, $y_t^o = x_{1,t} + \xi_t$, $\mathrm{R_{obs}} = 2.0$
- Use time-delayed analogs $X_t^{\text{analog}} = (x_{1,t}, x_{1,t-7dt}, x_{1,t-14dt})$
- 50 ensemble members
- Catalog: a simulation of L63 model for $T = 100$.

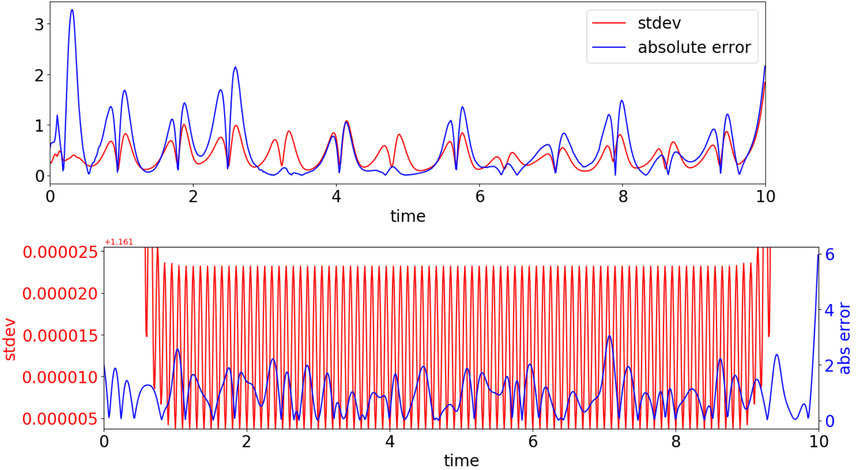
Numerical results for L63

	OI	AnDA	obs
RMSE	1.04	0.68	1.414



- AnDA produces better mean estimates in this experiment.

Numerical results for L63



The standard deviation (stdev) and absolute error of AnDA (top) and OI (bottom) estimates.

- The stdev estimated by AnDA has similar shape as the error.

Reanalysis of sea-surface height (SSH) at Gulf of Mexico

- Dataset: OCCIPUT simulated SSH of 50 members and 20 years;
- Catalog: the time series of the first 100 principal components of OCCIPUT dataset;
- Obs: simulated along-track obs (without error) of SSH from altimeters in 2004.

⇒ Task: Compare the reanalysis results of AnDA and OI.

Numerical results for simulated SSH

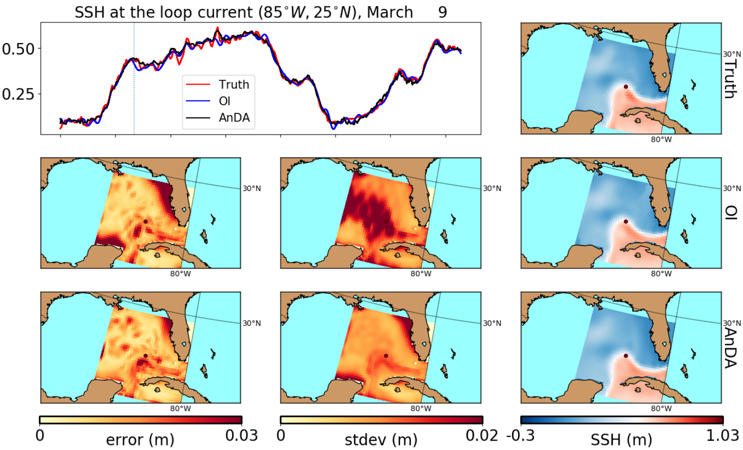
⇒We tuned the parameters for OI and found that the following parameters are optimal:

R	r_x	L_s	L_t
4 (cm^2)	1.5(degrees)	1.5(degrees)	6 (days)

⇒For AnDA (locally-linear), we use $N_e = 1000$ ensemble members and choose $k = 500$, $R = 4(cm^2)$.

	AnDA	OI
RMSE	1.37(cm)	1.76(cm)
RMSE(central region)	1.35(cm)	1.41(cm)

Numerical results for simulated SSH



- AnDA and OI produces similar mean state estimates;
- The stdev estimated by AnDA has similar contour shape as the error; The stdev estimated by OI only relies on the satellite trajectory;

Discussion

Numerical results (summary)

- AnDA outperforms OI in RMSE in L63 model;
- AnDA produces similar RMSE as OI in simulated SSH experiment;
- AnDA provides more informative stdev than OI does.

Reflections of AnDA

- Curse of dimensionality?
- Would AnDA still do well when the catalog and the truth are not from the same source?
- Replace analog forecast by machine learning?

Thank you! Any question?

Lguensat, R., Tandeo, P., Ailliot, P., Pulido, M., and Fablet, R. (2017).
The Analog Data Assimilation.
Monthly Weather Review, 145(10):4093–4107.

Tandeo, P., Ailliot, P., Ruiz, J. J., Hannart, A., Chapron, B., Easton, R., and Fablet, R. (2015).
Combining analog method and ensemble data assimilation: application to the Lorenz-63 chaotic system.
In Machine Learning and Data Mining Approaches to Climate Science, pages 3–12.