Introduction to Data Assimilation Lecture Notes
Lecture 1 on Oct. 16, 2025

Quanling Deng*

Abstract

This file prepares and documents the lecture 1 notes for the 2025 Fall YMSC public course: Introduction to Data Assimilation.

Keywords Data assimilation, forward Euler, RMSE, PCC, stability

1 Course overview

Data assimilation is a powerful framework for combining observational data with mathematical models to improve predictions and understanding of complex systems. Widely used in geosciences and many areas of applied science, it provides essential tools for weather forecasting, sea ice, ocean, and climate modeling, as well as a growing range of industrial applications. This course offers an introduction to the fundamental ideas and practical techniques of data assimilation. We will cover key concepts such as state estimation and filtering techniques (Kalman filters, EAKF, ETKF, etc), and explain how these methods integrate theory, computation, and data. We will explore case studies drawn from atmospheric and oceanic sciences. If time permits, we will also discuss advanced topics, such as the Lagrangian–Eulerian Multiscale Data Assimilation (LEMDA) method and nonlinear filtering strategies such as the Yau–Yau filter.

Prerequisite: Basic knowledge of calculus, linear algebra, and statistics is recommended. Motivated first-year students are welcome after a brief discussion with the instructor.

Reference:

- Majda, Andrew J., and John Harlim. Filtering complex turbulent systems. Cambridge University Press, 2012.
- Law, Kody, Andrew Stuart, and Kostas Zygalakis. "Data assimilation." Cham, Switzerland: Springer 214 (2015): 52.
- Asch, Mark, Marc Bocquet, and Maëlle Nodet. Data assimilation: methods, algorithms, and applications. Society for Industrial and Applied Mathematics, 2016.

*Corresponding author. Yau Mathematical Science Center, Tsinghua University, Beijing, China 100048; School of Computing, Australian National University, Canberra, ACT 2601, Australia. E-mail addresses: qldeng@tsinghua.edu.cn; Quanling.Deng@anu.edu.au

2 Definition and main idea of DA

Definition 2.1 (Data Assimilation)**.** Data assimilation (DA) is a mathematical framework for combining observational data with numerical models to obtain the best possible estimate of the state of a physical system.

Essentially, it is Model + Data = Data Assimilation. Precisely, the model has errors, and data has measuring errors:

$$\text{Model with bias and uncertainty} + \text{Data with noise} = \text{Data Assimilation} \tag{2.1}$$

- Also known as filtering or state estimation
- Applications: weather forecasting, climate modeling, etc.
- Goal: obtain the best statistical estimate of a natural system from partial observations of the true signal from nature

3 Two-step procedure

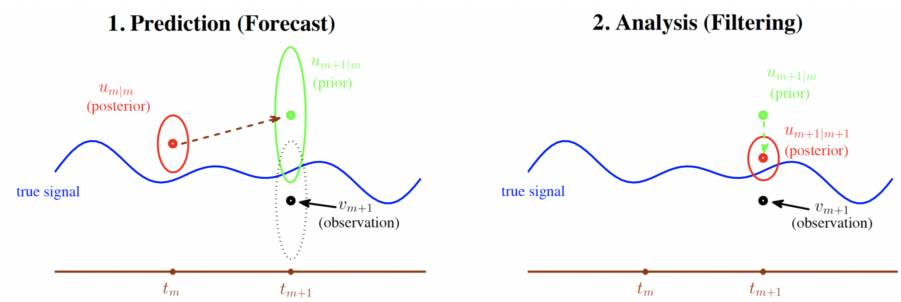

Figure 1: Two-step procedure for DA.

At each time $t_m = m\Delta t$:

1. **Model prediction:** Estimate statistical prediction of a probability distribution $u_{m+1|m}$ (prior) from $u_{m|m}$ using the dynamical model.
2. **Analysis:** Update $u_{m+1|m}$ using the statistical input of noisy observation v_{m+1} to obtain $u_{m+1|m+1}$ (posterior).

4 A deterministic example: accuracy with DA

We consider a first-order ordinary differential equation (ODE) for the unknown function $y = y(t)$ given by

$$y' = \frac{dy}{dt} = f(t, y), \qquad y(0) = y_0, \tag{4.1}$$

where $f(t,y)$ is a prescribed forcing function and y_0 is the given initial value.
Example. Consider the specific case where

$$f(t,y) = 2t, \qquad y_0 = 0, \qquad t \in [0,T].$$

The corresponding initial value problem becomes

$$\frac{dy}{dt} = 2t, \qquad y(0) = 0. \tag{4.2}$$

The analytical solution is obtained by direct integration:

$$y(t) = t^2.$$

Fix a time step $\Delta t > 0$ and define $t_n := n\,\Delta t$ for $n = 0,1,\dots,N$, where $N := T/\Delta t \in \mathbb{N}$.

4.1 Step 1 Prediction/Forecast: Forward (explicit) Euler scheme

Let $Y^n \approx y(t_n)$ with $Y^0 = y(0) = 0$. The forward Euler update reads

$$Y^{n+1} = Y^n + \Delta t\, f(t_n, Y^n) = Y^n + \Delta t\,(2t_n) = Y^n + 2\,t_n\,\Delta t, \qquad n = 0,1,\dots,N-1.$$

First few steps. Since $t_n = n\,\Delta t$ and $Y^0 = 0$:

$$\begin{aligned}
Y^1 &= Y^0 + 2\,t_0\,\Delta t = 0 + 2\cdot 0\cdot \Delta t = 0,\\
Y^2 &= Y^1 + 2\,t_1\,\Delta t = 0 + 2\,(\Delta t)\,\Delta t = 2\,(\Delta t)^2,\\
Y^3 &= Y^2 + 2\,t_2\,\Delta t = 2(\Delta t)^2 + 2\,(2\Delta t)\,\Delta t = 6\,(\Delta t)^2,\\
Y^4 &= Y^3 + 2\,t_3\,\Delta t = 6(\Delta t)^2 + 2\,(3\Delta t)\,\Delta t = 12\,(\Delta t)^2.
\end{aligned}$$

Closed form for Y^n. By summation,

$$Y^n = \sum_{k=0}^{n-1} 2\,t_k\,\Delta t = 2\,\Delta t \sum_{k=0}^{n-1} k\,\Delta t = 2\,(\Delta t)^2 \sum_{k=0}^{n-1} k = 2\,(\Delta t)^2 \cdot \frac{n(n-1)}{2} = n(n-1)\,(\Delta t)^2.$$

Error. The exact solution is $y(t) = t^2$, hence $y(t_n) = (n\Delta t)^2 = n^2(\Delta t)^2$. The nodal error

$$e^n := y(t_n) - Y^n = n^2(\Delta t)^2 - n(n-1)(\Delta t)^2 = n\,(\Delta t)^2.$$

In particular, at time $T = t_N$ the error is

$$e^N = N(\Delta t)^2 = T\,\Delta t,$$

so the global error scales like $\mathcal{O}(\Delta t)$ for fixed T.

4.2 Root mean square error and pattern correlation coefficient

Given two vectors $\mathbf{a} = (a_1, \dots, a_N)$ and $\mathbf{b} = (b_1, \dots, b_N)$,

$$\mathrm{RMSE}(\mathbf{a}, \mathbf{b}) := \sqrt{\frac{1}{N} \sum_{i=1}^{N} (a_i - b_i)^2}.$$

Let $\overline{a} = \frac{1}{N} \sum_{i=1}^{N} a_i$ and $\overline{b} = \frac{1}{N} \sum_{i=1}^{N} b_i$. The *pattern correlation coefficient* (PCC, i.e. Pearson correlation of anomalies) is

$$\mathrm{PCC}(\mathbf{a}, \mathbf{b}) := \frac{\sum_{i=1}^{N} (a_i - \overline{a})(b_i - \overline{b})}{\sqrt{\sum_{i=1}^{N} (a_i - \overline{a})^2} \sqrt{\sum_{i=1}^{N} (b_i - \overline{b})^2}}.$$

Setup for the example $y'(t) = 2t$, $y(0) = 0$. Let $\Delta t > 0$, $t_n := n\Delta t$ for $n = 0, 1, \dots, N$ with $T = N\Delta t$. The exact trajectory at grid points is

$$y(t_n) = t_n^2 = n^2 (\Delta t)^2.$$

The forward–Euler approximation from the previous derivation is

$$Y^n = n(n-1)\,(\Delta t)^2, \qquad n = 0, 1, \dots, N.$$

Denote the two length-$(N{+}1)$ vectors by

$$\mathbf{y} = (y(t_0), \dots, y(t_N)), \qquad \mathbf{Y} = (Y^0, \dots, Y^N).$$

RMSE for this solution. The pointwise error at nodes is $e^n := y(t_n) - Y^n = n(\Delta t)^2$, hence

$$\mathrm{RMSE}(\mathbf{Y}, \mathbf{y}) = \sqrt{\frac{1}{N+1} \sum_{n=0}^{N} \left(e^n\right)^2} = \sqrt{\frac{1}{N+1} \sum_{n=0}^{N} n^2 (\Delta t)^4} = (\Delta t)^2 \sqrt{\frac{N(2N+1)}{6}}.$$

Equivalently, in terms of $T = N\Delta t$,

$$\mathrm{RMSE}(\mathbf{Y}, \mathbf{y}) = \Delta t \sqrt{\frac{2T^2 + T\Delta t}{6}} = \frac{T}{\sqrt{3}} \Delta t + \mathcal{O}(\Delta t^{3/2}).$$

PCC for this solution. With means

$$\overline{y} = \frac{1}{N+1} \sum_{n=0}^{N} n^2 (\Delta t)^2 = (\Delta t)^2 \frac{N(2N+1)}{6}, \qquad \overline{Y} = \frac{1}{N+1} \sum_{n=0}^{N} n(n{-}1)(\Delta t)^2 = (\Delta t)^2 \frac{N(N-1)}{3},$$

a direct computation yields the closed form

$$\mathrm{PCC}(\mathbf{Y}, \mathbf{y}) = \frac{(16N+3)\sqrt{N-1}}{2\sqrt{(2N+1)(4N-3)(8N-3)}} \xrightarrow[N\to\infty]{} 1,$$

i.e. the pattern correlation approaches 1 as the grid is refined.

L^2 error over $[0,T]$ with linear interpolation of Y^n. Let $Y_h(t)$ be the piecewise-linear interpolant of $\{(t_n, Y^n)\}_{n=0}^N$. On $[t_n, t_{n+1}]$, write $t = t_n + s$ with $s \in [0, \Delta t]$. Using $Y^{n+1} - Y^n = 2n(\Delta t)^2$, the linear segment is

$$Y_h(t) = Y^n + \frac{s}{\Delta t}\left(Y^{n+1} - Y^n\right) = n(n-1)(\Delta t)^2 + 2n\,\Delta t\,s.$$

Since $y(t) = (t_n + s)^2 = n^2(\Delta t)^2 + 2n\,\Delta t\,s + s^2$, the interval error is

$$e(t) := y(t) - Y_h(t) = n(\Delta t)^2 + s^2.$$

Hence

$$\int_{t_n}^{t_{n+1}} e(t)^2\,dt = \int_0^{\Delta t} \left(n(\Delta t)^2 + s^2\right)^2 ds = \Delta t^5\left(n^2 + \frac{2}{3}n + \frac{1}{5}\right).$$

Summing $n = 0, \dots, N-1$ gives the exact squared L^2 error

$$\|y - Y_h\|^2_{L^2(0,T)} = \Delta t^5\left(\frac{N(N-1)(2N+1)}{6} + \frac{N}{5}\right),$$

and therefore

$$\|y - Y_h\|_{L^2(0,T)} = \Delta t^{5/2}\sqrt{\frac{N(N-1)(2N+1)}{6} + \frac{N}{5}} = \sqrt{\frac{T^3}{3}}\,\Delta t \;+\; \mathcal{O}(\Delta t^{3/2}),$$

showing first-order convergence in Δt for fixed T.

4.3 Step 2 analysis/filtering: model + data

Setting. Let $\Delta t > 0$, $t_n := n\,\Delta t$ for $n = 0, 1, \dots, N$ with $T = N\Delta t$. The exact samples are

$$y(t_n) = n^2(\Delta t)^2.$$

From the forward Euler forecast for $y'(t) = 2t$, $y(0) = 0$, we have

$$Y^n = n(n-1)\,(\Delta t)^2, \qquad n = 0, 1, \dots, N.$$

Assume point observations

$$Z^n \;=\; n^2(\Delta t)^2 + \beta\,(\Delta t)^2$$

with a (constant) measurement-bias parameter $\beta \in \mathbb{R}$. Define the (time-independent) linear analysis/filter

$$U^n \;=\; \gamma\,Y^n + (1-\gamma)\,Z^n, \qquad \gamma \in \mathbb{R}.$$

Errors at grid points. With $e^n_{(\cdot)} := y(t_n) - (\cdot)^n$, we have

$$e^n_Y = n(\Delta t)^2, \qquad e^n_Z = -\beta\,(\Delta t)^2, \qquad e^n_U = \gamma\,e^n_Y + (1-\gamma)\,e^n_Z = (\Delta t)^2(\gamma n - (1-\gamma)\beta)\,.$$

Remark 4.1. Herein, the filtering process is in one step. It shall be done iteratively for all steps.

Model and exact samples. Consider

$$y'(t) = 2t, \qquad y(0) = 0, \qquad t \in [0, T],$$

with exact solution $y(t) = t^2$. On the grid $t_n = n\,\Delta t$ for $n = 0, 1, \ldots, N$ (so $T = N\Delta t$),

$$y(t_n) = n^2(\Delta t)^2.$$

4.4 Step 1 prediction/forecast (forward Euler, cycled; using Y^n)

Let U^n denote the *analysis* at time t_n (with $U^0 = 0$). The model forecast variable is Y^{n+1}, advanced from the current analysis:

$$\boxed{Y^{n+1} \;=\; U^n + \Delta t \cdot f(t_n, U^n) \;=\; U^n + 2\,t_n\,\Delta t \;=\; U^n + 2n\,(\Delta t)^2\,.}$$

4.5 Step 2 analysis/filtering (cycled)

Assume observations with a constant additive error β:

$$Z^n \;=\; y(t_n) + \beta \;=\; n^2(\Delta t)^2 + \beta.$$

The analysis at t_{n+1} blends the forecast and the observation:

$$\boxed{U^{n+1} \;=\; \gamma\,Y^{n+1} + (1-\gamma)\,Z^{n+1} \;=\; \gamma\big(U^n + 2n(\Delta t)^2\big) + (1-\gamma)\big((n{+}1)^2(\Delta t)^2 + \beta\big),}$$

with a constant blending weight $\gamma \in \mathbb{R}$. This U^{n+1} is then used for the *next* forecast step.

Closed-form error recursion and solution. Let the analysis error be $e^n := y(t_n) - U^n$. Using the two boxed updates,

$$\begin{aligned}
e^{n+1} &= (n{+}1)^2(\Delta t)^2 - \Big[\gamma\,Y^{n+1} + (1-\gamma)\big((n{+}1)^2(\Delta t)^2 + \beta\big)\Big] \\
&= (n{+}1)^2(\Delta t)^2 - \gamma\big(U^n + 2n(\Delta t)^2\big) - (1-\gamma)\big((n{+}1)^2(\Delta t)^2 + \beta\big) \\
&= \gamma\Big((n{+}1)^2(\Delta t)^2 - U^n - 2n(\Delta t)^2\Big) - (1-\gamma)\beta = \gamma\big(e^n + (\Delta t)^2\big) - (1-\gamma)\beta.
\end{aligned}$$

With $e^0 = 0$, this linear forced recurrence ($\sum_{k=0}^{N} a\,r^k = a\,\frac{1-r^{N+1}}{1-r}, r \neq 1.$) solves to

$$\boxed{e^n \;=\; A\big(1-\gamma^n\big), \qquad A \;=\; \frac{\gamma}{1-\gamma}\,(\Delta t)^2 \;-\; \beta,}$$

for $\gamma \neq 1$ (the case $\gamma = 1$ follows by continuity).

Final-time error.

$$\boxed{e^N \;=\; y(t_N) - U^N \;=\; A\big(1-\gamma^N\big).}$$

RMSE and PCC on $\{0,\dots,N\}$. Since $e^n = A(1-\gamma^n)$,

$$\mathrm{RMSE}(\mathbf{U},\mathbf{y})^2 = \frac{1}{N+1}\sum_{n=0}^{N}\big(e^n\big)^2 = \frac{A^2}{N+1}\Bigg[(N+1) \;-\; 2\sum_{n=0}^{N}\gamma^n \;+\; \sum_{n=0}^{N}\gamma^{2n}\Bigg],$$

with the geometric sums (for $\gamma \neq \pm 1$)

$$\sum_{n=0}^{N}\gamma^n = \frac{1-\gamma^{N+1}}{1-\gamma}, \qquad \sum_{n=0}^{N}\gamma^{2n} = \frac{1-\gamma^{2(N+1)}}{1-\gamma^2}.$$

For PCC, write

$$U^n \;=\; y(t_n) - e^n \;=\; n^2(\Delta t)^2 \;-\; A \;+\; A\gamma^n.$$

Adding/subtracting a constant does not affect correlation, hence

$$\mathrm{PCC}(\mathbf{U},\mathbf{y}) = \mathrm{PCC}(\mathbf{V},\mathbf{y}), \qquad V^n := n^2(\Delta t)^2 + A\gamma^n.$$

Let

$$\mu_y := \frac{1}{N+1}\sum_{n=0}^{N} n^2(\Delta t)^2 = (\Delta t)^2\frac{N(2N+1)}{6}, \quad \mu_\gamma := \frac{1}{N+1}\sum_{n=0}^{N}\gamma^n,$$

$$S_{y^2} := \sum_{n=0}^{N} n^4(\Delta t)^4 = (\Delta t)^4\frac{N(N+1)(2N+1)(3N^2+3N-1)}{30}, \quad S_{\gamma^2} := \sum_{n=0}^{N}\gamma^{2n}, \quad S_{y\gamma} := (\Delta t)^2\sum_{n=0}^{N} n^2\gamma^n.$$

Then

$$\mathrm{PCC}(\mathbf{U},\mathbf{y}) = \frac{\big[S_{y^2} - (N+1)\mu_y^2\big] + A\big[S_{y\gamma} - (N+1)\mu_y\mu_\gamma\big]}{\sqrt{S_{y^2} - (N+1)\mu_y^2}\;\sqrt{S_{y^2} - (N+1)\mu_y^2 \;+\; 2A\big[S_{y\gamma} - (N+1)\mu_y\mu_\gamma\big] \;+\; A^2\big[S_{\gamma^2} - (N+1)\mu_\gamma^2\big]}}.$$

(Closed forms for $\sum n^2\gamma^n$ are standard but omitted for brevity.)

Optimizing γ

(i) Minimize the absolute final-time error. Since $e^N = A(1-\gamma^N)$, a sufficient and γ-only condition for $e^N = 0$ (for any N) is to cancel the forcing,

$$A = 0 \quad\Longleftrightarrow\quad \frac{\gamma}{1-\gamma}(\Delta t)^2 = \beta \quad\Longleftrightarrow\quad \boxed{\gamma^\star_{\text{final}} = \frac{\beta}{\beta + (\Delta t)^2}}.$$

With this choice, the recurrence has zero forcing and, with $e^0 = 0$, produces $e^n \equiv 0$ for all n (perfect tracking).

(ii) Minimize RMSE on $\{0,\dots,N\}$. Define

$$K(\gamma) := \frac{1}{N+1}\left[(N+1) - 2\sum_{n=0}^{N}\gamma^n + \sum_{n=0}^{N}\gamma^{2n}\right], \qquad A(\gamma) = \frac{\gamma}{1-\gamma}(\Delta t)^2 - \beta,$$

so that $\mathrm{RMSE}^2 = A(\gamma)^2 K(\gamma)$. Similar arguments yield:

$$\boxed{\gamma^\star_{\mathrm{RMSE}} \;=\; \gamma^\star_{\mathrm{final}} = \frac{\beta}{\beta + (\Delta t)^2}.}$$

5 A deterministic example: stability with DA

Model and exact solutions. Consider

$$y'(t) = \lambda\, y(t), \qquad y(0) = 1, \qquad t \in [0,T],$$

with exact continuous solution $y(t) = e^{\lambda t}$. On the grid $t_n = n\,\Delta t$ for $n = 0,1,\dots,N$ (so $T = N\Delta t$), define

$$a := e^{\lambda \Delta t}, \qquad y(t_n) = a^n.$$

5.1 Step 1 prediction/forecast: forward Euler scheme and RMSE/PCC

Forward Euler (model-only baseline).

$$Y^{n+1} = (1+\lambda\Delta t)\, Y^n, \qquad Y^0 = 1, \quad \Rightarrow \quad Y^n = b^n, \qquad b := 1 + \lambda \Delta t.$$

Final-time error (baseline).

$$e_Y^N := y(t_N) - Y^N \;=\; a^N - b^N \;=\; e^{\lambda T} - (1+\lambda\Delta t)^N.$$

RMSE and PCC on $\{0,\dots,N\}$ (baseline). Let the geometric sums

$$\mu_a := \frac{1}{N+1}\sum_{n=0}^{N} a^n, \quad \mu_b := \frac{1}{N+1}\sum_{n=0}^{N} b^n, \quad S_{a^2} := \sum_{n=0}^{N} a^{2n}, \quad S_{b^2} := \sum_{n=0}^{N} b^{2n}, \quad S_{ab} := \sum_{n=0}^{N} (ab)^n,$$

with $\sum_{n=0}^{N} q^n = \frac{1-q^{N+1}}{1-q}$ and $\sum_{n=0}^{N} q^{2n} = \frac{1-q^{2(N+1)}}{1-q^2}$ for $q \neq 1$. Then

$$\mathrm{RMSE}(\mathbf{Y},\mathbf{y}) = \sqrt{\frac{1}{N+1}\Big(S_{a^2} - 2S_{ab} + S_{b^2}\Big)}, \qquad \mathrm{PCC}(\mathbf{Y},\mathbf{y}) = \frac{S_{ab} - (N+1)\mu_a\mu_b}{\sqrt{S_{a^2} - (N+1)\mu_a^2}\,\sqrt{S_{b^2} - (N+1)\mu_b^2}}.$$

5.2 Step 2 analysis/filtering: model + data (cycled DA)

Observation and cycled DA step. Assume point observations with constant additive error β:

$$Z^n = y(t_n) + \beta = a^n + \beta.$$

Cycled DA: the analysis at step n is

$$U^n = \gamma Y^n + (1-\gamma)Z^n,$$

and the *next* forecast uses the current analysis as its initial value:

$$\widehat{Y}^{n+1} = b\,U^n, \qquad b = 1 + \lambda\Delta t,$$

followed by the next analysis

$$U^{n+1} = \gamma\widehat{Y}^{n+1} + (1-\gamma)Z^{n+1} = \gamma b\,U^n + (1-\gamma)\,a^{n+1} + (1-\gamma)\beta,$$

with $U^0 = 1$ (exact initial condition).

Closed-form error dynamics. Define the analysis error $e^n := a^n - U^n$. Then

$$e^{n+1} = \gamma b\,e^n + \gamma(a-b)\,a^n - (1-\gamma)\beta, \qquad e^0 = 0.$$

This linear, forced recurrence admits the explicit solution

$$\boxed{e^n = p\,a^n + q - (p+q)\,(\gamma b)^n, \quad p = \frac{\gamma\,(a-b)}{a-\gamma b}, \quad q = -\frac{(1-\gamma)\,\beta}{1-\gamma b},}$$

provided $a \neq \gamma b$ and $\gamma b \neq 1$. (Degenerate cases follow by continuity.)

Final-time error (cycled DA).

$$e_U^N := y(t_N) - U^N \;=\; p\,a^N + q - (p+q)\,(\gamma b)^N.$$

RMSE and PCC on $\{0,\ldots,N\}$ (cycled DA). Since

$$e^n = p\,a^n \;+\; q \;-\; (p+q)(\gamma b)^n,$$

we obtain

$$\mathrm{RMSE}(\mathbf{U},\mathbf{y})^2 = \frac{1}{N+1}\sum_{n=0}^{N}\big(e^n\big)^2 = \frac{1}{N+1}\Big(p^2 S_{a^2} + q^2(N+1) + (p+q)^2 S_{(\gamma b)^2} - 2p(p+q)S_{a\gamma b} + 2pq\,S_a - 2q(p+q)S_{\gamma b}\Big),$$

where the additional geometric sums are

$$S_{\gamma b} := \sum_{n=0}^{N}(\gamma b)^n, \quad S_{(\gamma b)^2} := \sum_{n=0}^{N}(\gamma b)^{2n}, \quad S_{a\gamma b} := \sum_{n=0}^{N}(a\,\gamma b)^n, \quad S_a := \sum_{n=0}^{N} a^n.$$

(Each has the same closed form as above with the appropriate ratio.)

For PCC, note that subtracting the mean removes the constant q, but the two geometric components remain. Let

$$W^n := p\,a^n - (p+q)(\gamma b)^n, \quad \mu_W := \frac{1}{N+1}\sum_{n=0}^{N} W^n, \quad S_{aW} := \sum_{n=0}^{N} a^n W^n.$$

Then

$$\mathrm{PCC}(\mathbf{U},\mathbf{y}) = \frac{S_{aW} - (N+1)\mu_a \mu_W}{\sqrt{S_{a^2} - (N+1)\mu_a^2}\,\sqrt{\sum_{n=0}^{N}(W^n - \mu_W)^2}},$$

with

$$S_{aW} = p\,S_{a^2} - (p+q)\,S_{a\,\gamma b}, \qquad \sum_{n=0}^{N}(W^n - \mu_W)^2 = p^2\big(S_{a^2} - (N+1)\mu_a^2\big) - 2p(p+q)\big(S_{a\,\gamma b} - (N+1)\mu_a \mu_{\gamma b}\big) + (p+q)^2\big(S_{(\gamma b)}$$

and $\mu_{\gamma b} = \frac{1}{N+1} S_{\gamma b}$.

5.3 Stability

Forecast (model-only) stability. The Euler amplification is $b = 1 + \lambda \Delta t$. Asymptotic stability requires

$$|b| < 1 \quad \Longleftrightarrow \quad |1 + \lambda \Delta t| < 1,$$

i.e. the Euler stability disc (center -1, radius 1) in the $\lambda \Delta t$-plane. For real $\lambda < 0$: $0 < \Delta t < -\frac{2}{\lambda}$.

DA-cycled stability of the analysis error. From the error solution,

$$e^n = p\,a^n + q - (p+q)(\gamma b)^n,$$

we see that, for $\mathrm{Re}(\lambda) < 0$ (so $|a| < 1$), the analysis error remains bounded and in fact converges provided

$$|\gamma b| \; < \; 1 \qquad \Big(\text{i.e. } |\gamma(1 + \lambda \Delta t)| < 1\Big).$$

In that case,

$$e^n \xrightarrow[n \to \infty]{} q \; = \; -\frac{(1-\gamma)\,\beta}{1 - \gamma b},$$

a finite bias determined by the observation error and blending weight. Hence, even when the *model-only* forecast is unstable ($|b| > 1$), one can *stabilize the DA-cycled estimate* by choosing

$$0 < \gamma < \frac{1}{|b|},$$

which damps the unstable model mode in the analysis loop. The DA step does not change the classical integrator stability region, but it ensures the *estimate* is stable (and biased only by q) by anchoring forecasts to observations each cycle.

Remark. If $\beta = 0$ and $|\gamma b| < 1$, then $q = 0$ and $e^n \to 0$; the DA-cycled estimate becomes asymptotically exact even when $|b| > 1$, provided the blending is sufficiently small.

6 Practical/fundamental issues/difficulties/challenges [1]

1. (**Complex/Multiscale dynamics.**) True signal from nature: Turbulent dynamical systems with extremely complex noisy spatio-temporal signals. "Turbulent dynamical systems to generate the true signal. The true signal from nature arises from a turbulent nonlinear dynamical system with extremely complex noisy spatio-temporal signals which have significant amplitude over many spatial scales."

2. (**Model errors.**) model itself (incomplete physical understanding) + numerical errors. "A major difficulty in accurate filtering of noisy turbulent signals with many degrees of freedom is model error; the fact that the true signal from nature is processed for filtering and prediction through an imperfect model where by practical necessity, important physical processes are parametrized due to inadequate numerical resolution or incomplete physical understanding. The model errors of inadequate resolution often lead to rough turbulent energy spectra for the truth signal to be filtered on the order of the mesh scale for the dynamical system model used for filtering."

3. (**Curse of ensemble size.**) "For forward models for filtering, the state space dimension is typically large, of order 10^4–10^8, for these turbulent dynamical systems, so generating an ensemble size with such a direct approach of order 50–100 members is typically all that is available for real-time filtering."

4. (**Data limitation.**) Sparse, noisy, spatio-temporal observations for only a partial subset of the variables. In systems with multiple spatio-temporal scales, the sparse observations of the truth signal might automatically couple many spatial scales

Specifically, the true solution to the model is not available.

1. for 1, not much can be done.

2. for 2, better physical understanding; developing better models, higher accuracy numerical solvers, such as FEM, Isogeometric analysis, high-order generalized-alpha method

3. for 3, computational strategies such as parallelization, stochastic parameterization algorithms,

4. for 4, exactly solvable models, new computational algorithms

References

[1] Andrew J Majda and John Harlim, *Filtering complex turbulent systems*, Cambridge University Press, 2012.