A kernel extension of the Ensemble Transform Kalman Filter

Sophie Mauran[‡,*], Sandrine Mouysset[‡], Ehouarn Simon[‡], Laurent Bertino[†]

[‡] IRIT, Université de Toulouse, CNRS, Toulouse INP, France
[†] NERSC, Bergen, Norvège
[*] Now at INRIA, Rennes, France
sophie.mauran@inria.fr

Plan

What is data assimilation

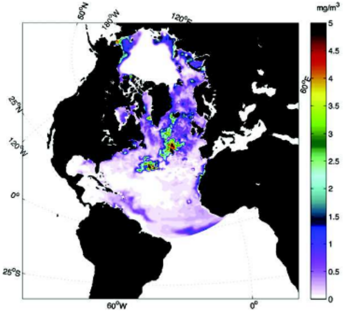
- Aim of data assimilation : estimate the state of a system
  $\Rightarrow$ Estimation of the phytoplankton concentration in the North Atlantic Ocean.

What is data assimilation

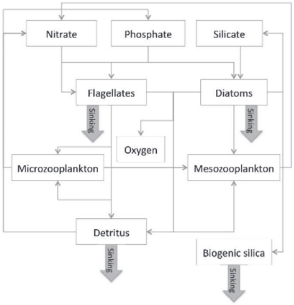
- Aim of data assimilation : estimate the state of a system
  $\Rightarrow$ Estimation of the phytoplankton concentration in the North Atlantic Ocean.
- 3 main components :

What is data assimilation

- Aim of data assimilation : estimate the state of a system
  $\Rightarrow$ Estimation of the phytoplankton concentration in the North Atlantic Ocean.
- 3 main components :
  - Model : mathematical equations



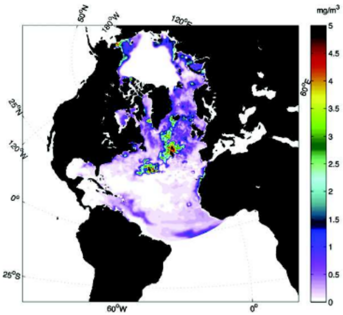
Model forecast



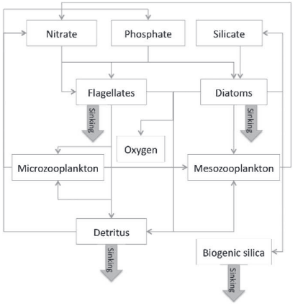
NORWECOM model : ocean ecosystem model

What is data assimilation

- Aim of data assimilation : estimate the state of a system
  $\Rightarrow$ Estimation of the phytoplankton concentration in the North Atlantic Ocean.
- 3 main components :
  - Model : mathematical equations $\Rightarrow$ **IMPERFECT**;



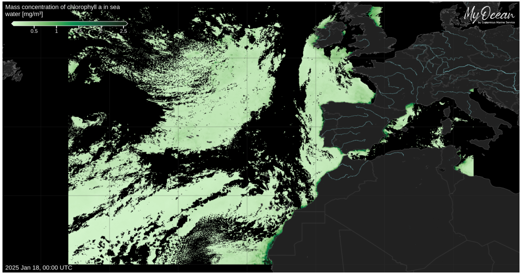
Model forecast



NORWECOM model : ocean ecosystem model

What is data assimilation

- Aim of data assimilation : estimate the state of a system
  $\Rightarrow$ Estimation of the phytoplankton concentration in the North Atlantic Ocean.
- 3 main components :
  - Model : mathematical equations $\Rightarrow$ **IMPERFECT**;
  - Observations : satellites, in situ

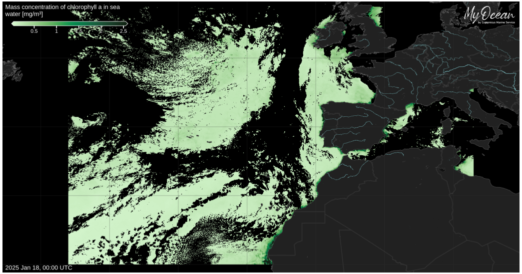


Chlorophyll concentration in the North Atlantic Ocean

Source : Global Ocean Colour (Copernicus-GlobColour)

What is data assimilation

- Aim of data assimilation : estimate the state of a system
  $\Rightarrow$ Estimation of the phytoplankton concentration in the North Atlantic Ocean.
- 3 main components :
  - Model : mathematical equations $\Rightarrow$ **IMPERFECT**;
  - Observations : satellites, in situ $\Rightarrow$ **INCOMPLETE & IMPERFECT**;



Chlorophyll concentration in the North Atlantic Ocean

Source : Global Ocean Colour (Copernicus-GlobColour)

What is data assimilation

- Aim of data assimilation : estimate the state of a system
  $\Rightarrow$ Estimation of the phytoplankton concentration in the North Atlantic Ocean.
- 3 main components :
  - Model : mathematical equations $\Rightarrow$ **IMPERFECT**;
  - Observations : satellites, in situ $\Rightarrow$ **INCOMPLETE & IMPERFECT**;
  - Assimilation algorithm : combination of the model and observations.

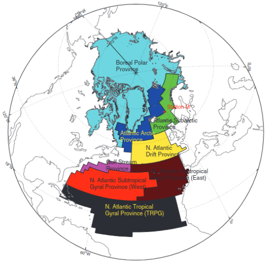
What is data assimilation

- Aim of data assimilation : estimate the state of a system
  $\Rightarrow$ Estimation of the phytoplankton concentration in the North Atlantic Ocean.
- 3 main components :
  - Model : mathematical equations $\Rightarrow$ **IMPERFECT**;
  - Observations : satellites, in situ $\Rightarrow$ **INCOMPLETE & IMPERFECT**;
  - Assimilation algorithm : combination of the model and observations.

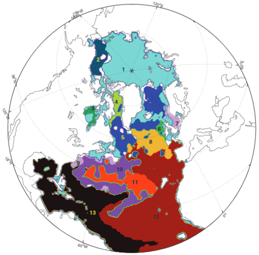
Coarse regions from Longhurst [**Longhurst, 1995**]

What is data assimilation

- Aim of data assimilation : estimate the state of a system
  $\Rightarrow$ Estimation of the phytoplankton concentration in the North Atlantic Ocean.
- 3 main components :
  - Model : mathematical equations $\Rightarrow$ **IMPERFECT**;
  - Observations : satellites, in situ $\Rightarrow$ **INCOMPLETE & IMPERFECT**;
  - Assimilation algorithm : combination of the model and observations.



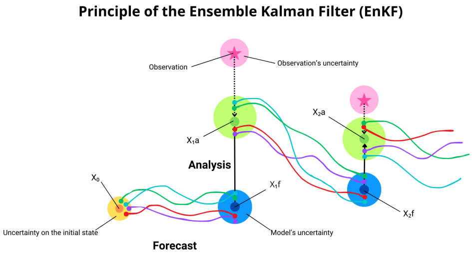
Coarse regions from Longhurst [**Longhurst, 1995**]



Refined regions from assimilation and clustering ([**Simon, Samuelsen, Bertino and Mouysset 2015**])

Assimilation Cycle

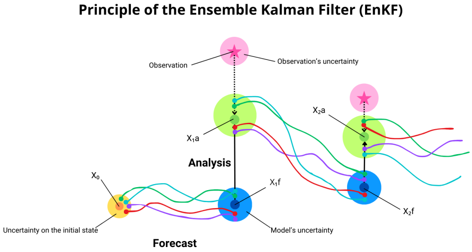
- **Forecast step**: State estimation over time using the **model**.
- **Analysis step**: Incorporation of **observations** to correct the estimation.
  $\Rightarrow$ **Minimization of the cost function**:
  $J(\mathbf{x}) = \frac{1}{2}(\mathbf{x} - \mathbf{x}^f)^T \mathbf{B}^{-1}(\mathbf{x} - \mathbf{x}^f) + \frac{1}{2}(\mathbf{y} - \mathcal{H}(\mathbf{x}))^T \mathbf{R}^{-1}(\mathbf{y} - \mathcal{H}(\mathbf{x}))$



Notations

- $\mathbf{x}^f$: the state estimate after forecast;
- $\mathbf{y}$: the observation to be assimilated;
- $\mathcal{H}$: the observation operator (state to observation space);
- $\mathbf{B}$: the background error covariance matrix;
- $\mathbf{R}$: the observation error covariance matrix.

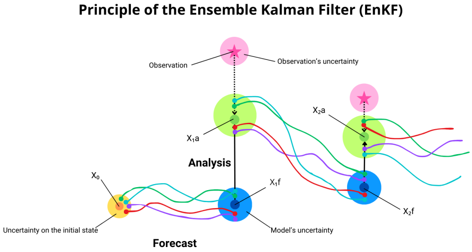
The Ensemble Kalman Filter (Evensen, 1994, 2003)



The EnKF: an ensemble-based method

- Ensemble mean: estimate of the **state**.
- Ensemble covariance matrix $\boldsymbol{P}^f \approx \boldsymbol{B}$: estimate of the **forecast error**.
- **Under-dispersion** $\Rightarrow$ Overconfidence in the estimate $\Rightarrow$ Observations given little weight $\Rightarrow$ **Filter divergence**.

The Ensemble Kalman Filter (Evensen, 1994, 2003)



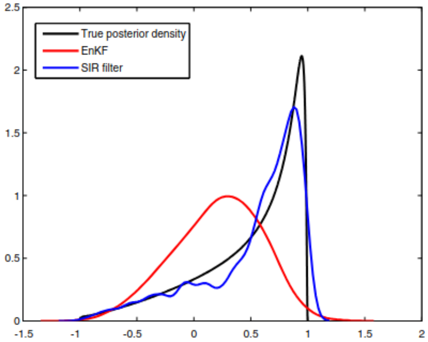
The EnKF: an ensemble-based method

- Ensemble mean: estimate of the **state**.
- Ensemble covariance matrix $\boldsymbol{P}^f \approx \boldsymbol{B}$: estimate of the **forecast error**.
- **Under-dispersion** $\Rightarrow$ Overconfidence in the estimate $\Rightarrow$ Observations given little weight $\Rightarrow$ **Filter divergence**.

Geophysical fluids

- High dimensional problems.
- Nonlinear propagation of the errors.
- **Linear-Gaussian assumptions**: linear models and Gaussian error distributions.

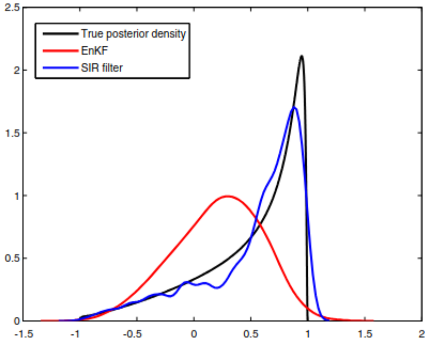
Limitations of Gaussian assumptions



A posteriori density for the Bernoulli model and estimates from EnKF and a particle filter ([**Stordal et al., 2011**]).

- EnKF tends to produce an *a posteriori* density **close to a Gaussian density** (Central Limit Theorem).
- Particle filter: **curse of dimensionality**: N must grow **exponentially** with the log-likelihood of the observations ([**Snyder et al., 2008**]).

Limitations of Gaussian assumptions

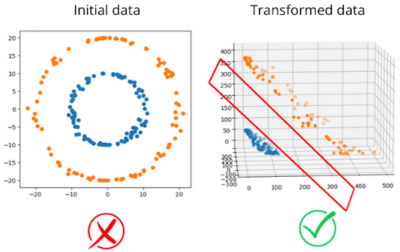


A posteriori density for the Bernoulli model and estimates from EnKF and a particle filter ([**Stordal et al., 2011**]).

- EnKF tends to produce an *a posteriori* density **close to a Gaussian density** (Central Limit Theorem).
- Particle filter: **curse of dimensionality**: N must grow **exponentially** with the log-likelihood of the observations ([**Snyder et al., 2008**]).

How can we (properly) estimate non-Gaussian densities while keeping a reasonable ensemble size?

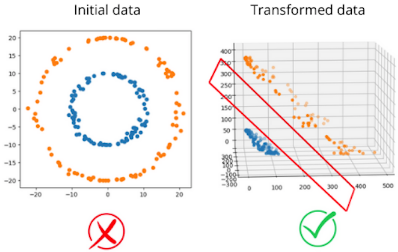
Relevance of kernel methods for nonlinear problems



Transformation function:

$$\forall\ \boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix},\ \varphi(\boldsymbol{x}) = (x_1^2, \sqrt{2}x_1x_2, x_2^2)$$

.



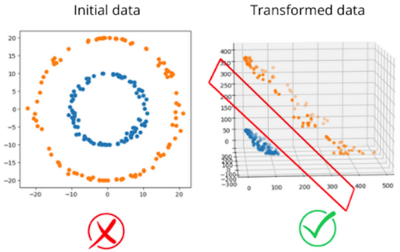
Transformation function:

$$\forall\ \boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix},\ \varphi(\boldsymbol{x}) = (x_1^2, \sqrt{2}x_1x_2, x_2^2)$$

.

$\Rightarrow$ **Problem can be solved in 3D space.**

Relevance of kernel methods for nonlinear problems



Transformation function:

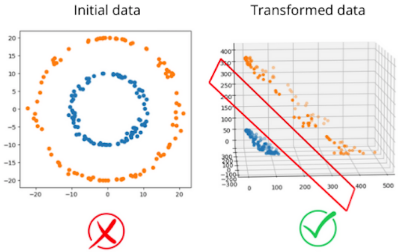
$$\forall\ \boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix},\ \varphi(\boldsymbol{x}) = (x_1^2, \sqrt{2}x_1x_2, x_2^2)$$

.

$\Rightarrow$ **Problem can be solved in 3D space.**

- Solve the original problem on transformed data: **How to deal with dimensionality increase?**

Relevance of kernel methods for nonlinear problems



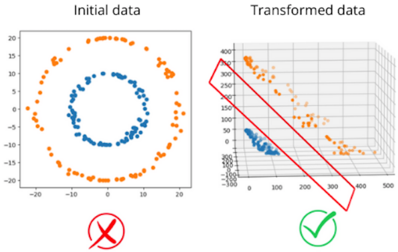
Transformation function:

$$\forall\ \boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix},\ \varphi(\boldsymbol{x}) = (x_1^2, \sqrt{2}x_1x_2, x_2^2)$$

.

$\Rightarrow$ **Problem can be solved in 3D space.**

- Solve the original problem on transformed data: **How to deal with dimensionality increase?**
- Kernel function: $\kappa(\boldsymbol{x}, \boldsymbol{y}) =<\varphi(\boldsymbol{x}), \varphi(\boldsymbol{y})>= (\boldsymbol{x}^\top \boldsymbol{y})^2 \in \mathbb{R}$.



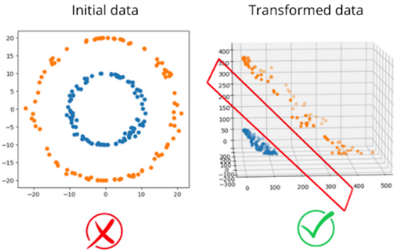
Transformation function:

$$\forall\ \boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix},\ \varphi(\boldsymbol{x}) = (x_1^2, \sqrt{2}x_1x_2, x_2^2)$$

.

$\Rightarrow$ **Problem can be solved in 3D space.**

- Solve the original problem on transformed data: **How to deal with dimensionality increase?**
- Kernel function: $\kappa(\boldsymbol{x}, \boldsymbol{y}) =< \varphi(\boldsymbol{x}), \varphi(\boldsymbol{y}) >= (\boldsymbol{x}^\top \boldsymbol{y})^2 \in \mathbb{R}$.
  $\Rightarrow$ **Implicit use of feature space without explicit mapping or computation.**



Transformation function:

$$\forall\ \boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix},\ \varphi(\boldsymbol{x}) = (x_1^2, \sqrt{2}x_1x_2, x_2^2)$$

$\Rightarrow$ **Problem can be solved in 3D space.**

- Solve the original problem on transformed data: **How to deal with dimensionality increase?**
- Kernel function: $\kappa(\boldsymbol{x}, \boldsymbol{y}) =< \varphi(\boldsymbol{x}), \varphi(\boldsymbol{y}) >= (\boldsymbol{x}^\top \boldsymbol{y})^2 \in \mathbb{R}$.

  $\Rightarrow$ **Implicit use of feature space without explicit mapping or computation.**

This kernel function determines a Reproducing Kernel Hilbert Space (RKHS) associated with this function.

Representer theorem

Let $\mathcal{F}$ be an RKHS with reproducing kernel k. Let C be a cost function and $J : \mathbb{R}^+ \to \mathbb{R}$ a strictly increasing function. Let $(\boldsymbol{x}_i)_{i \in [\![1\ ;\ n]\!]} \in \chi^n$ and $y \in \mathbb{R}$.

We define:

$$\begin{array}{cccc} Q & : & \mathcal{F} & \to & \mathbb{R} \\ & & f & \mapsto & C(y, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_n)) + J(||f||^2_{\mathcal{F}}) \end{array}.$$

If it exists, any minimizer $f^* \in \mathcal{F}$ of Q belongs to the subspace $Vect(k(\cdot, \boldsymbol{x}_i)_{i \in [\![1\ ;\ n]\!]})$.

Mathematical formalism

Representer theorem

Let $\mathcal{F}$ be an RKHS with reproducing kernel k. Let C be a cost function and $J : \mathbb{R}^+ \to \mathbb{R}$ a strictly increasing function. Let $(\boldsymbol{x}_i)_{i \in [\![1\ ;\ n]\!]} \in \chi^n$ and $y \in \mathbb{R}$.
We define:

$$\begin{array}{cccc} Q & : & \mathcal{F} & \to & \mathbb{R} \\ & & f & \mapsto & C(y, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_n)) + J(||f||^2_{\mathcal{F}}) \end{array}.$$

If it exists, any minimizer $f^* \in \mathcal{F}$ of Q belongs to the subspace $Vect(k(\cdot, \boldsymbol{x}_i)_{i \in [\![1\ ;\ n]\!]})$.

$\Rightarrow$ **Optimization problem in a finite-dimensional space of size equal to the data.**

Kernel methods in data assimilation

- Approximation of model errors by a combination of exponential kernels ([**Luo, 2019**]): **specific kernel**.
- Model emulation by a linear combination of hyperbolic tangent kernels ([**Gottwald and Reich, 2021**]): **specific kernel**.

Kernel methods in data assimilation

- Approximation of model errors by a combination of exponential kernels ([**Luo, 2019**]): **specific kernel**.
- Model emulation by a linear combination of hyperbolic tangent kernels ([**Gottwald and Reich, 2021**]): **specific kernel**.
- Definition of a rigorous mathematical framework to estimate realizations/trajectories of the system by linear combinations of predictions, transposed into an **evolving RKHS manifold** characterizing the system dynamics ([**Dufée et al., 2024; Jaouen et al., 2025**]).

Kernel methods in data assimilation

- Approximation of model errors by a combination of exponential kernels ([**Luo, 2019**]): **specific kernel**.
- Model emulation by a linear combination of hyperbolic tangent kernels ([**Gottwald and Reich, 2021**]): **specific kernel**.
- Definition of a rigorous mathematical framework to estimate realizations/trajectories of the system by linear combinations of predictions, transposed into an **evolving RKHS manifold** characterizing the system dynamics ([**Dufée et al., 2024; Jaouen et al., 2025**]).

Thesis idea:

Propose a kernel-based data assimilation algorithm **modular with respect to the kernel**.

Outline

Plan

The formulation of [Bishop et al., 2001]

Assuming the observation operator is linear, the ETKF algorithm can be formalized as:

ETKF formulation

$$\arg\min_{\mathbf{w}\in\mathbb{R}^N} \mathcal{J}(\mathbf{w}) = \frac{1}{2}||\widetilde{\boldsymbol{d}} - \widetilde{\boldsymbol{H}}\boldsymbol{w}||_2^2 + \frac{N-1}{2}||\boldsymbol{w}||_2^2. \quad (1)$$

Notations

- $\widetilde{\boldsymbol{d}} = \boldsymbol{R}^{-1/2}(\boldsymbol{y} - \boldsymbol{H}\bar{\boldsymbol{x}}^f) \in \mathbb{R}^d$ the innovation vector;
- $\widetilde{\boldsymbol{H}} = \boldsymbol{R}^{-1/2}\boldsymbol{H}\boldsymbol{X}^f = \begin{bmatrix} \widetilde{\boldsymbol{h}_1}^\top \\ \vdots \\ \widetilde{\boldsymbol{h}_d}^\top \end{bmatrix}$ the matrix of weighted anomalies in observation space;
- $\bar{\boldsymbol{x}}^f$: the ensemble mean state;
- $\boldsymbol{X}^f \in \mathbb{R}^{m\times N}$: the anomaly matrix of centered states;
- $\boldsymbol{R} \in \mathbb{R}^{d\times d}$: the observation error covariance matrix;
- $\boldsymbol{H}\boldsymbol{X}^f \in \mathbb{R}^{d\times N}$: the anomaly matrix in observation space.

Reformulation of the ETKF with kernel methods [Mauran et al., 2023]

By identifying the linear kernel $\kappa(\boldsymbol{z}, \boldsymbol{t}) = \boldsymbol{z}^\top \boldsymbol{t}$, $\forall\ (\boldsymbol{z}, \boldsymbol{t}) \in \mathbb{R}^N \times \mathbb{R}^N$, (1) is **equivalent** to:

$$\Leftrightarrow \underset{\boldsymbol{w} \in \mathbb{R}^N}{\arg\min}\, \mathcal{J}(\boldsymbol{w}) = \frac{1}{2} \sum_{i=1}^{d} (\kappa(\widetilde{\boldsymbol{h}}_i, \boldsymbol{w}) - y_i)^2 + \frac{1}{2} \kappa(\boldsymbol{w}, \boldsymbol{w}) \tag{2}$$

with $\widetilde{\boldsymbol{h}}_i^\top$ the rows of $\widetilde{\boldsymbol{H}}$ (the matrix of weighted anomalies in observation space).

Reformulation of the ETKF with kernel methods [Mauran et al., 2023]

By identifying the linear kernel $\kappa(\boldsymbol{z}, \boldsymbol{t}) = \boldsymbol{z}^\top \boldsymbol{t}$, $\forall\ (\boldsymbol{z}, \boldsymbol{t}) \in \mathbb{R}^N \times \mathbb{R}^N$, (1) is **equivalent** to:

$$\Leftrightarrow \underset{\boldsymbol{w} \in \mathbb{R}^N}{\arg\min}\, \mathcal{J}(\boldsymbol{w}) = \frac{1}{2} \sum_{i=1}^{d} (\kappa(\widetilde{\boldsymbol{h}}_i, \boldsymbol{w}) - y_i)^2 + \frac{1}{2} \kappa(\boldsymbol{w}, \boldsymbol{w}) \tag{2}$$

with $\widetilde{\boldsymbol{h}}_i^\top$ the rows of $\widetilde{\boldsymbol{H}}$ (the matrix of weighted anomalies in observation space).
Moreover, (2) is **equivalent** to the optimization problem extended in the Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}_\kappa$ with linear reproducing kernel:

Reformulation of the ETKF

$$\underset{f \in \mathcal{H}_\kappa}{\arg\min}\, \widetilde{\mathcal{J}}(f) = \frac{N-1}{2} ||f||^2_{\mathcal{H}_\kappa} + \frac{1}{2} \sum_{i=1}^{d} (f(\widetilde{\mathbf{h}}_\mathbf{i}) - \widetilde{d}_i)^2. \tag{3}$$

Reformulation of the ETKF with kernel methods [Mauran et al., 2023]

By identifying the linear kernel $\kappa(\boldsymbol{z}, \boldsymbol{t}) = \boldsymbol{z}^\top \boldsymbol{t}$, $\forall\ (\boldsymbol{z}, \boldsymbol{t}) \in \mathbb{R}^N \times \mathbb{R}^N$, (1) is **equivalent** to:

$$\Leftrightarrow \underset{\boldsymbol{w} \in \mathbb{R}^N}{\arg\min}\, \mathcal{J}(\boldsymbol{w}) = \frac{1}{2} \sum_{i=1}^{d} (\kappa(\widetilde{\boldsymbol{h}}_i, \boldsymbol{w}) - y_i)^2 + \frac{1}{2} \kappa(\boldsymbol{w}, \boldsymbol{w}) \tag{2}$$

with $\widetilde{\boldsymbol{h}}_i^\top$ the rows of $\widetilde{\boldsymbol{H}}$ (the matrix of weighted anomalies in observation space).
Moreover, (2) is **equivalent** to the optimization problem extended in the Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}_\kappa$ with linear reproducing kernel:

Reformulation of the ETKF

$$\underset{f \in \mathcal{H}_\kappa}{\arg\min}\, \widetilde{\mathcal{J}}(f) = \frac{N-1}{2} ||f||^2_{\mathcal{H}_\kappa} + \frac{1}{2} \sum_{i=1}^{d} (f(\widetilde{\mathbf{h}}_\mathbf{i}) - \widetilde{d}_i)^2. \tag{3}$$

$\Rightarrow$ Generalization.

KETKF Problem

New cost function (representer theorem)

$$\underset{\boldsymbol{\alpha}\in\mathbb{R}^d}{\arg\min}\,\tilde{\tilde{\mathcal{J}}}(\boldsymbol{\alpha}) = \frac{N-1}{2}\boldsymbol{\alpha}^\top \mathbf{K}\boldsymbol{\alpha} + \frac{1}{2}\|\widetilde{\mathbf{d}} - \mathbf{K}\boldsymbol{\alpha}\|_2^2, \tag{4}$$

with $\mathbf{K} = (\kappa(\widetilde{\mathbf{h}}_\mathbf{i}, \widetilde{\mathbf{h}}_\mathbf{j}))_{1\leq i,j\leq d} \in \mathbb{R}^{d\times d}$, the kernel matrix on the **observed data**.

KETKF Problem

New cost function (representer theorem)

$$\arg\min_{\boldsymbol{\alpha}\in\mathbb{R}^d} \widetilde{\widetilde{\mathcal{J}}}(\boldsymbol{\alpha}) = \frac{N-1}{2}\boldsymbol{\alpha}^\top \mathbf{K}\boldsymbol{\alpha} + \frac{1}{2}\|\widetilde{\mathbf{d}} - \mathbf{K}\boldsymbol{\alpha}\|_2^2, \qquad (4)$$

with $\mathbf{K} = (\kappa(\widetilde{\mathbf{h}}_\mathbf{i}, \widetilde{\mathbf{h}}_\mathbf{j}))_{1\le i,j\le d} \in \mathbb{R}^{d\times d}$, the kernel matrix on the **observed data**.

Ensemble augmentation strategy to handle nonlinear observation operators (Evensen, 2006)

$$\begin{bmatrix} \mathbf{E}^a \\ \mathbf{H}\mathbf{E}^a \end{bmatrix} = \begin{bmatrix} \hat{\boldsymbol{\mu}}^f \\ \mathbf{H}\hat{\boldsymbol{\mu}}^f \end{bmatrix} + \begin{bmatrix} \mathbf{X}^f \\ \mathbf{H}\mathbf{X}^f \end{bmatrix} \mathbf{w} + \sqrt{N-1}\mathbf{A}^a, \qquad (5)$$

with $\mathbf{A}^a \in \mathbb{R}^{(m+d)\times N}$ obtained from ${\mathbf{P}^a}^{1/2}$, the covariance matrix of the augmented ensemble.

KETKF Problem (general case)

New cost function (representer theorem)

$$\arg\min_{\boldsymbol{\alpha}\in\mathbb{R}^{m+d}} \tilde{\tilde{\mathcal{J}}}(\boldsymbol{\alpha}) = \frac{N-1}{2}\boldsymbol{\alpha}^\top \mathbf{K}\boldsymbol{\alpha} + \frac{1}{2}\|\widetilde{\mathbf{d}} - \mathbf{\Pi_H}\mathbf{K}\boldsymbol{\alpha}\|_2^2. \quad (6)$$

Notations

- $\mathbf{\Pi_H} = \begin{bmatrix} \mathbf{0_{mm}} & \mathbf{0_{md}} \\ \mathbf{0_{dm}} & \mathbf{I_d} \end{bmatrix} \in \mathbb{R}^{(m+d)\times(m+d)}$ projection onto the observation space;
- $\mathbf{K} = \begin{bmatrix} \mathbf{K_X} & \mathbf{K_{XH}} \\ \mathbf{K_{XH}}^\top & \mathbf{K_H} \end{bmatrix} \in \mathbb{R}^{(m+d)\times(m+d)}$ with:
  - $\mathbf{K_X} = (\kappa(\mathbf{a_i^f}, \mathbf{a_j^f}))_{1\leq i,j\leq m} \in \mathbb{R}^{m\times m}$;
  - $\mathbf{K_{HX}} = (\kappa(\mathbf{a_i^f}, \widetilde{\mathbf{h}}_\mathbf{j}))_{1\leq i\leq m, 1\leq j\leq d} \in \mathbb{R}^{m\times d}$;
  - $\mathbf{K_H} = (\kappa(\widetilde{\mathbf{h}}_\mathbf{i}, \widetilde{\mathbf{h}}_\mathbf{j}))_{1\leq i,j\leq d} \in \mathbb{R}^{d\times d}$;
  - $(\mathbf{a}_i^{f\,\top})_{1\leq i\leq m}$ are the rows of $\boldsymbol{X}^f$.

KETKF: Update of the Mean

New cost function (representer theorem)

$$\underset{\boldsymbol{\alpha}\in\mathbb{R}^{d+m}}{\arg\min}\ \widetilde{\widetilde{\mathcal{J}}}(\boldsymbol{\alpha}) = \frac{N-1}{2}\boldsymbol{\alpha}^\top \mathbf{K}\boldsymbol{\alpha} + \frac{1}{2}\|\widetilde{\mathbf{d}} - \mathbf{\Pi_H}\mathbf{K}\boldsymbol{\alpha}\|_2^2. \quad (7)$$

Solution of (7) (First-order Necessary and Sufficient Condition)

$$\boldsymbol{\alpha}^* = \begin{bmatrix}\boldsymbol{\alpha}_X^* \\ \boldsymbol{\alpha}_H^*\end{bmatrix} = \begin{bmatrix}\mathbf{0_{n1}} \\ [(N-1)\mathbf{I_d} + \mathbf{K_H}]^{-1}\widetilde{\mathbf{d}}\end{bmatrix}. \quad (8)$$

New cost function (representer theorem)

$$\underset{\boldsymbol{\alpha}\in\mathbb{R}^{d+m}}{\arg\min}\ \tilde{\tilde{\mathcal{J}}}(\boldsymbol{\alpha}) = \frac{N-1}{2}\boldsymbol{\alpha}^\top \mathbf{K}\boldsymbol{\alpha} + \frac{1}{2}\|\widetilde{\mathbf{d}} - \mathbf{\Pi_H}\mathbf{K}\boldsymbol{\alpha}\|_2^2. \quad (7)$$

Solution of (7) (First-order Necessary and Sufficient Condition)

$$\boldsymbol{\alpha}^* = \begin{bmatrix} \boldsymbol{\alpha}_X^* \\ \boldsymbol{\alpha}_H^* \end{bmatrix} = \begin{bmatrix} \mathbf{0_{n1}} \\ [(N-1)\mathbf{I_d} + \mathbf{K_H}]^{-1}\widetilde{\mathbf{d}} \end{bmatrix}. \quad (8)$$

Ensemble Mean after Analysis

$$\bar{\mathbf{x}}^{\mathbf{a}} = \bar{\mathbf{x}}^{\mathbf{f}} + \mathbf{K_{XH}}[(N-1)\mathbf{I_d} + \mathbf{K_H}]^{-1}\widetilde{\mathbf{d}}. \quad (9)$$

Computation of the Analysis Error Covariance Matrix $\mathbf{P}^a_X$

- Random variable perspective: $\boldsymbol{\alpha} \sim \mathcal{N}(\boldsymbol{\mu}_\alpha, \mathbf{P}^\alpha)$.

Computation of the Analysis Error Covariance Matrix $\mathbf{P}_X^a$

- Random variable perspective: $\boldsymbol{\alpha} \sim \mathcal{N}(\boldsymbol{\mu}_\alpha, \mathbf{P}^\alpha)$.
- A realization of the state after analysis: $\mathbf{x}^\mathbf{a} = \bar{\mathbf{x}}^\mathbf{f} + \mathbf{\Pi}_\mathbf{X}\mathbf{K}\boldsymbol{\alpha}$.

Computation of the Analysis Error Covariance Matrix $\mathbf{P}_X^a$

- Random variable perspective: $\boldsymbol{\alpha} \sim \mathcal{N}(\boldsymbol{\mu}_\alpha, \mathbf{P}^\alpha)$.
- A realization of the state after analysis: $\mathbf{x}^\mathbf{a} = \bar{\mathbf{x}}^\mathbf{f} + \mathbf{\Pi}_\mathbf{X}\mathbf{K}\alpha$.

Expression of $\mathbf{P}_X^a$ (analysis error covariance matrix)

$$\mathbf{P}_X^a = \mathbf{\Pi}_\mathbf{X}\mathbf{K}\mathbf{P}^\alpha\mathbf{K}\mathbf{\Pi}_\mathbf{X}. \quad (10)$$

Computation of the Analysis Error Covariance Matrix $\mathbf{P}_X^a$

- Random variable perspective: $\boldsymbol{\alpha} \sim \mathcal{N}(\boldsymbol{\mu}_\alpha, \mathbf{P}^\alpha)$.
- A realization of the state after analysis: $\mathbf{x}^\mathbf{a} = \bar{\mathbf{x}}^\mathbf{f} + \boldsymbol{\Pi}_\mathbf{X}\mathbf{K}\boldsymbol{\alpha}$.

Expression of $\mathbf{P}_X^a$ (analysis error covariance matrix)

$$\mathbf{P}_X^a = \boldsymbol{\Pi}_\mathbf{X}\mathbf{K}\mathbf{P}^\alpha\mathbf{K}\boldsymbol{\Pi}_\mathbf{X}. \tag{10}$$

We can also approximate $\mathbf{P}^\alpha$:

Approximation of $\mathbf{P}^\alpha$ [Auroux, 2003]

$$\mathbf{P}^\alpha \approx [\nabla^2 \tilde{\tilde{\mathcal{J}}}(\boldsymbol{\alpha}^*)]^{-1}, \tag{11}$$

with

$$\forall \boldsymbol{\alpha} \in \mathbb{R}^{m+d}, \quad \tilde{\tilde{\mathcal{J}}}(\boldsymbol{\alpha}) = \frac{N-1}{2}\boldsymbol{\alpha}^\top\mathbf{K}\boldsymbol{\alpha} + \frac{1}{2}\|\widetilde{\mathbf{d}} - \boldsymbol{\Pi}_\mathbf{H}\mathbf{K}\boldsymbol{\alpha}\|_2^2.$$

Plan

Experimental framework

Experimental framework

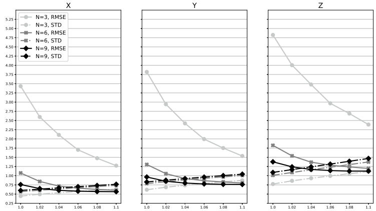
- Lorenz-63 system with the DAPPER package (Raanes, 2024).
- Observations: first two variables, every 25 time steps, $\sigma_o^2 = 2$.
- Analysis RMS error and standard deviation averaged over 10 different seeds.
- Different inflations and ensemble sizes.

Choice of the kernel

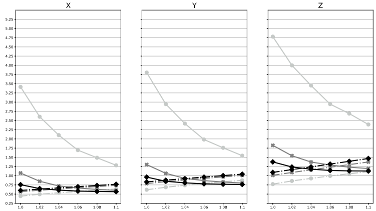
- Linear kernel
  - $\forall (\mathbf{x}, \mathbf{y}) \in \mathbb{R}^N \times \mathbb{R}^N, \quad \kappa_L(\mathbf{x}, \mathbf{y}) = \mathbf{x}^T \mathbf{y}$.
- Polynomial kernel
  - ϕ: data preprocessing function ($\phi(\mathbf{x}) \in [-1, 1]^N$).
  - $\forall (\mathbf{x}, \mathbf{y}) \in \mathbb{R}^N \times \mathbb{R}^N, \quad \kappa_P(\mathbf{x}, \mathbf{y}) = (1 + \phi(\mathbf{x})^T \phi(\mathbf{y}))^2$.
- Weighting the linear kernel
  - $\forall (\mathbf{x}, \mathbf{y}) \in \mathbb{R}^N \times \mathbb{R}^N, \quad \kappa_W(\mathbf{x}, \mathbf{y}) = \kappa_L(\mathbf{x}, \mathbf{y}) e^{-\frac{1-(\mathbf{x}^T\mathbf{y})^2}{\sqrt{N-1}}}$, with κ_L the linear kernel.
  - $\sigma = \dfrac{1-(\mathbf{x}^T\mathbf{y})^2}{\sqrt{N-1}}$: Estimator of the standard error of the Pearson correlation (Soper, 1913)

Analysis: averaged RMSE and standard deviation

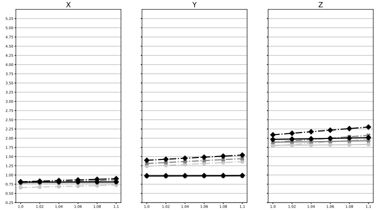
ETKF



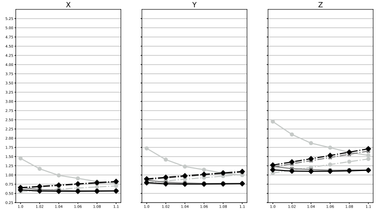
Linear kernel



Polynomial kernel



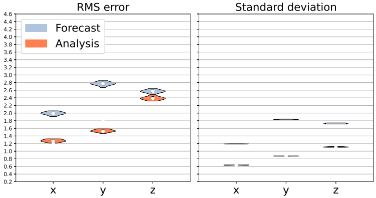
Weighted linear kernel



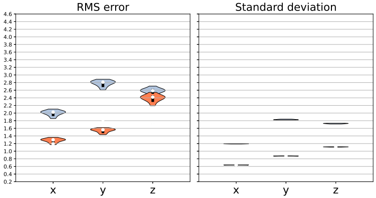
- ETKF and Linear kernel: similar results.
- Weaker impact of the inflation with the Kernel ETKF.
- Lower to similar RMSE for the Weighted linear kernel for a fixed (N, infl).

Results for the 10 experiments ($N = 3$ and $infl = 1.1$)

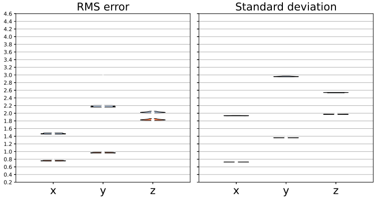
ETKF



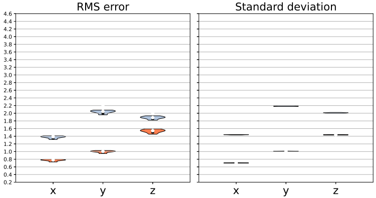
Linear kernel



Polynomial kernel



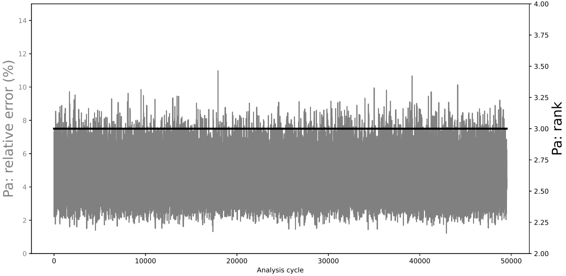
Weighted linear kernel



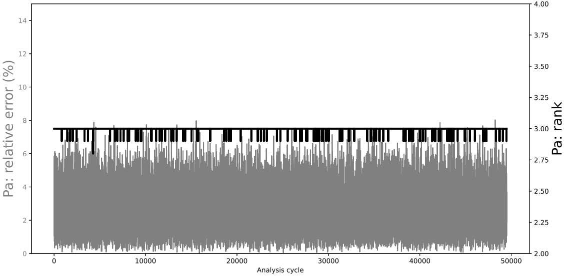
- Larger variability in the error for le linear Kernel compared to the ETKF.
- No variability for the polynomial Kernel.

Averaged rank of $\mathbf{P}_a$ and relative error ($N = 3$ and $infl = 1.1$)

Polynomial kernel



Weighted linear kernel



- $\mathbf{P}_a$: frequently full rank.
- Low approximation error from the anomaly matrix.

Plan

Conclusion and perspectives

Conclusion

- Development of a Kernel extension of the ETKF.
- Experiments:
  - Similar performances for the linear kernel ETKF and classical ETKF .
  - Interest of using other kernels in the presence of nonlinearities.

Perspectives

- More advanced experiments: different kernels and models.
- Dealing with the size of $\mathbf{K} \in \mathbb{R}^{(n+p)\times(n+p)}$: localisation strategy.
- Choice of the kernels and tuning of their hyperparameters.

Bibliographie I

Auroux, D. (2003). *Hessienne et covariances d'erreur*. https://math.unice.fr/~auroux/Work/These/html/node40.html (accessed: 29.08.2021).

Bishop, C. H., Etherton, B. J., & Majumdar, S. J. (2001). Adaptive sampling with the ensemble transform kalman filter. part i: Theoretical aspects. *Monthly Weather Review*, *129*(3), 420–436. https://doi.org/10.1175/1520-0493(2001)129⟨0420:ASWTET⟩2.0.CO;2

Dufée, B., Hug, B., Mémin, É., & Tissot, G. (2024). Ensemble forecasts in reproducing kernel hilbert space family. *Physica D: Nonlinear Phenomena*, *459*, 134044. https://doi.org/https://doi.org/10.1016/j.physd.2023.134044

Evensen, G. (2006, January). Data assimilation. the ensemble kalman filter. Springer. https://doi.org/10.1007/978-3-540-38301-7

Gottwald, G. A., & Reich, S. (2021). Supervised learning from noisy observations: Combining machine-learning techniques with data assimilation. *Physica D: Nonlinear Phenomena*, *423*, 132911. https://doi.org/10.1016/j.physd.2021.132911

Jaouen, M., Dufée, B., Mémin, E., & Tissot, G. (2025). *Ensemble Filter on dynamically driven RKHS: Application to a Multilayer Quasi-Geostrophic ocean model* [working paper or preprint]. https://inria.hal.science/hal-04915412

Bibliographie II

Longhurst, A. (1995). Seasonal cycles of pelagic production and consumption. *Progress in Oceanography, 36*(2), 77–167. https://doi.org/https://doi.org/10.1016/0079-6611(95)00015-1

Luo, X. (2019). Ensemble-based kernel learning for a class of data assimilation problems with imperfect forward simulators. *PLOS ONE, 14*(7), 1–40. https://doi.org/10.1371/journal.pone.0219247

Mauran, S., Mouysset, S., Simon, E., & Bertino, L. (2023). A kernel extension of the ensemble transform kalman filter. *23rd International Conference on Computational Science (ICCS 2023), 10476*(part 4), 438–452. https://doi.org/10.1007/978-3-031-36027-5_35

Snyder, C., Bengtsson, T., Bickel, P., & Anderson, J. (2008). Obstacles to high-dimensional particle filtering. *Monthly Weather Review, 136*(12), 4629–4640. https://doi.org/10.1175/2008MWR2529.1

Stordal, A., Karlsen, H., Naevdal, G., Skaug, H., & Vallès, B. (2011). Bridging the ensemble kalman filter and particle filters: The adaptive gaussian mixture filter. *Computational Geosciences, 15*. https://doi.org/10.1007/s10596-010-9207-1

Equivalence KETKF noyau linéaire / EnKF : mise à jour de la moyenne

- Mise à jour de la moyenne du KETKF :

$$\hat{\mu}^a = \hat{\mu}^f + \boldsymbol{K}_{XH}[(N-1)\boldsymbol{I}_d + \boldsymbol{K}_H]^{-1}\widetilde{\boldsymbol{d}}. \quad (12)$$

- Pour le noyau linéaire :
  - $\boldsymbol{K}_H = \widetilde{\boldsymbol{H}}\widetilde{\boldsymbol{H}}^\top = \boldsymbol{R}^{-1/2}\boldsymbol{H}\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{H}^\top\boldsymbol{R}^{-1/2} = (N-1)\boldsymbol{R}^{-1/2}\boldsymbol{H}\hat{\boldsymbol{P}}^f\boldsymbol{H}^\top\boldsymbol{R}^{-1/2}$
  - $\boldsymbol{K}_{XH} = \boldsymbol{X}\widetilde{\boldsymbol{H}}^\top = \boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{H}^\top\boldsymbol{R}^{-1/2} = (N-1)\hat{\boldsymbol{P}}\boldsymbol{H}^\top\boldsymbol{R}^{-1/2}$

Dans ce cas :

$$\begin{aligned}\hat{\mu}^a &= \hat{\mu}^f + \hat{\boldsymbol{P}}^f\boldsymbol{H}^\top[\boldsymbol{R} + \boldsymbol{H}\hat{\boldsymbol{P}}\boldsymbol{H}^\top]^{-1}(\boldsymbol{y} - \boldsymbol{H}\bar{\boldsymbol{x}}^f) \\ &= \hat{\mu}^f + \hat{\boldsymbol{K}}(\boldsymbol{y} - \boldsymbol{H}\hat{\mu}^f)\end{aligned}$$

Equivalence KETKF noyau linéaire / EnKF : mise à jour de la covariance

- Mise à jour de la covariance du KETKF :

$$\boldsymbol{P}_X^a = \frac{1}{N-1}[\boldsymbol{K}_X + \boldsymbol{K}_{XH}[(N-1)\boldsymbol{I}_d + \boldsymbol{K}_H]^{-1}\boldsymbol{K}_{XH}]. \tag{13}$$

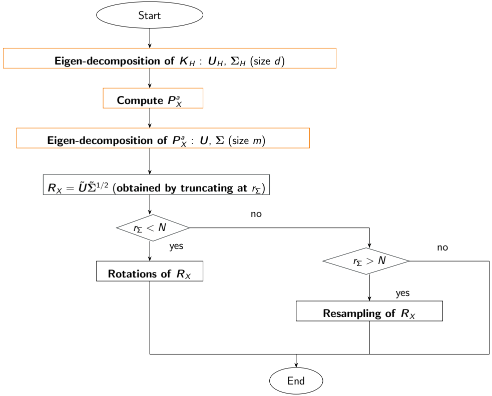
- Pour le noyau linéaire :
  - $\boldsymbol{K}_X = \boldsymbol{X}\boldsymbol{X}^\top = (N-1)\hat{\boldsymbol{P}}^f$
  - $\boldsymbol{K}_H = \widetilde{\boldsymbol{H}}\widetilde{\boldsymbol{H}}^\top = \boldsymbol{R}^{-1/2}\boldsymbol{H}\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{H}^\top\boldsymbol{R}^{-1/2} = (N-1)\boldsymbol{R}^{-1/2}\boldsymbol{H}\hat{\boldsymbol{P}}^f\boldsymbol{H}^\top\boldsymbol{R}^{-1/2}$
  - $\boldsymbol{K}_{XH} = \boldsymbol{X}\widetilde{\boldsymbol{H}}^\top = \boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{H}^\top\boldsymbol{R}^{-1/2} = (N-1)\hat{\boldsymbol{P}}^f\boldsymbol{H}^\top\boldsymbol{R}^{-1/2}$

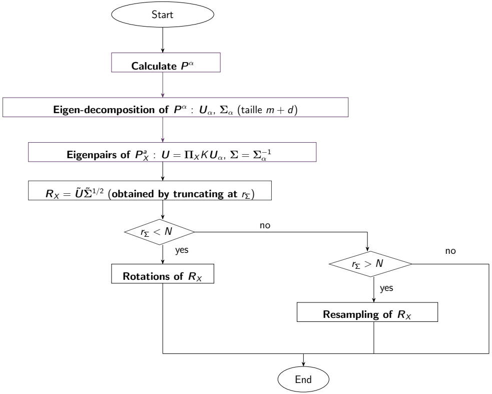
Dans ce cas :

$$\begin{aligned}
\boldsymbol{P}_X^a &= \hat{\boldsymbol{P}}^f - \hat{\boldsymbol{P}}^f\boldsymbol{H}^\top[\boldsymbol{R} + \boldsymbol{H}\hat{\boldsymbol{P}}\boldsymbol{H}^\top]^{-1}\boldsymbol{H}\hat{\boldsymbol{P}}^f \\
&= \hat{\boldsymbol{P}}^f - \hat{\boldsymbol{K}}\boldsymbol{H}\hat{\boldsymbol{P}}^f \\
&= [\boldsymbol{I}_m - \hat{\boldsymbol{K}}\boldsymbol{H}]\hat{\boldsymbol{P}}^f
\end{aligned}$$

KETKF Algorithm

1. Build the kernel matrix $\boldsymbol{K}$.
2. Update the ensemble mean (solve a linear system of size d).
3. Update the ensemble covariance $\boldsymbol{P}_X^a$.
   Two approaches: 1 or 2 diagonalizations ($\mathcal{O}(n^3)$ complexity).





Algorithm Etape de rotation de $\boldsymbol{R}_X$, dérivé directement de [**Farchi2019**]

Require: $1 \leq i \leq N - r_\Sigma$ ▷ *l'indice de la boucle*
Require: $\boldsymbol{R}_X$ ▷ *la matrice que nous allons augmenter d'une colonne*

$\epsilon \leftarrow 1.0$

$q \leftarrow r_\Sigma + i$

$\theta \leftarrow \frac{\sqrt{q}}{\sqrt{q-\epsilon}}$

Calculer $\boldsymbol{Q}_\epsilon \leftarrow \frac{-\theta}{q} \times \begin{bmatrix} \frac{\epsilon}{\sqrt{q}} & \cdots & \cdots & \cdots & \cdots & \frac{\epsilon}{\sqrt{q}} \\ \vdots & 1-\frac{\theta}{q} & \frac{-\theta}{q} & \cdots & \cdots & \frac{-\theta}{q} \\ \vdots & \frac{-\theta}{q} & 1-\frac{\theta}{q} & \frac{-\theta}{q} & \cdots & \frac{-\theta}{q} \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 1-\frac{\theta}{q} & \frac{-\theta}{q} \\ \frac{\epsilon}{\sqrt{q}} & \frac{-\theta}{q} & \cdots & \cdots & \frac{-\theta}{q} & 1-\frac{\theta}{q} \end{bmatrix} \in \mathbb{R}^{q \times q}$

$\boldsymbol{W} \leftarrow \begin{bmatrix} \boldsymbol{0}_m & \boldsymbol{R}_X \end{bmatrix} \in \mathbb{R}^{m \times d}$

Calculer $\boldsymbol{R}_X \leftarrow \boldsymbol{W}\boldsymbol{Q}_\epsilon \ =0$

Rééchantillonnage de R_X

Algorithm Rééchantillonnage de $\boldsymbol{R}_X$

Require: $r_{\text{samp}} \in \mathbb{N}$ ▷ *le facteur de rééchantillonnage*

Require: $\widetilde{\boldsymbol{U}} \in \mathbb{R}^{m \times r_{\boldsymbol{\Sigma}}}$ ▷ *la matrice des vecteurs propres de $\boldsymbol{P}_X^a$ tronquée à la colonne $r_{\boldsymbol{\Sigma}}$, le rang de $\boldsymbol{P}_X^a$*

Require: $\widetilde{\boldsymbol{\Sigma}} \in \mathbb{R}^{r_{\boldsymbol{\Sigma}} \times r_{\boldsymbol{\Sigma}}}$ ▷ *la matrice des valeurs propres de $\boldsymbol{P}_X^a$ tronquée de ses valeurs propres nulles*

$N_{\text{samp}} \leftarrow r_{\text{samp}} \times r_{\boldsymbol{\Sigma}}$

$\boldsymbol{O} \in \mathcal{O}(N_{\text{samp}}, \mathbb{R})$ ▷ matrice orthogonale de taille N_{samp}

$\boldsymbol{X} \leftarrow \left[\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{N_{\text{samp}}}\right] \in \mathbb{R}^{m \times N_{\text{samp}}}$ avec $\boldsymbol{x}_i \sim \mathcal{N}(\mathbf{0}, \widetilde{\boldsymbol{U}}\widetilde{\boldsymbol{\Sigma}}^{1/2}),\ \ i \in [\![1, N_{\text{samp}}]\!]$

$$\boldsymbol{M} \leftarrow \begin{bmatrix} 1 + \frac{-1}{N_{\text{samp}}} & \frac{-1}{N_{\text{samp}}} & \cdots & \frac{-1}{N_{\text{samp}}} \\ \frac{-1}{N_{\text{samp}}} & 1 + \frac{-1}{N_{\text{samp}}} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ \frac{-1}{N_{\text{samp}}} & \frac{-1}{N_{\text{samp}}} & \cdots & 1 + \frac{-1}{N_{\text{samp}}} \end{bmatrix} \in \mathbb{R}^{N_{\text{samp}} \times N_{\text{samp}}}$$

Rééchantillonnage de R_X

Algorithm Rééchantillonnage de $\boldsymbol{R}_X$

$\boldsymbol{X} \leftarrow \widetilde{\boldsymbol{U}}^\top \boldsymbol{X}\boldsymbol{M}$ ▷ *matrice d'anomalies centrée*
for $i \in [\![1, r_{\boldsymbol{\Sigma}}]\!]$ **do**
  $\boldsymbol{x}_i \leftarrow \sqrt{\lambda_i} \times \boldsymbol{x}_i$
end for
$\tilde{\mathbf{1}} \leftarrow \tilde{\boldsymbol{U}}\boldsymbol{X}$
$\boldsymbol{U}_{\text{samp}}\boldsymbol{\Sigma}_{\text{samp}}\boldsymbol{V}_{\text{samp}}^\top \leftarrow SVD(\tilde{\boldsymbol{A}})$ ▷ avec $\boldsymbol{\Sigma}_{\text{samp}} = diag([\lambda_{\text{samp}_i}])$
$\tilde{\boldsymbol{O}} \leftarrow \boldsymbol{O}$ tronquée à sa $N^{ième}$ ligne et à la colonne $r_{\boldsymbol{\Sigma}}$.
$\tilde{\boldsymbol{O}} \leftarrow \tilde{\boldsymbol{O}}^\top$
for $i \in [\![1, r_{\boldsymbol{\Sigma}}]\!]$ **do**
  $\boldsymbol{U}_{\text{samp}_i} \leftarrow \sqrt{\frac{1}{r_{\text{samp}}}} \times \lambda_{\text{samp}_i} \times \boldsymbol{U}_{\text{samp}_i}$ ▷ avec $\boldsymbol{U}_{\text{samp}_i}$ les colonnes de $\boldsymbol{U}_{\text{samp}}$
end for $=0$

Rééchantillonnage de R_X

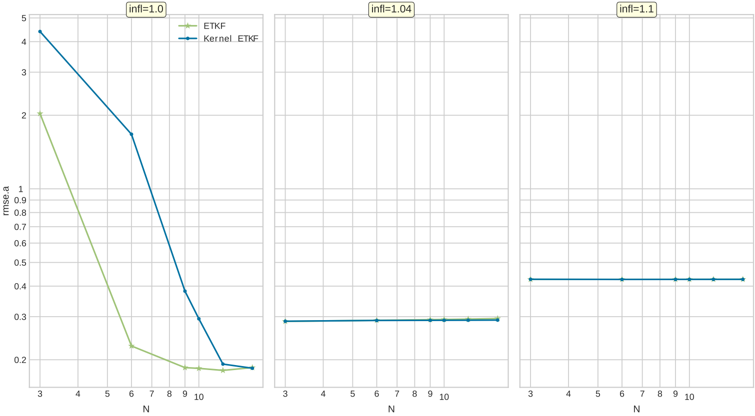
Algorithm Rééchantillonnage de $\boldsymbol{R}_X$

$$\tilde{\boldsymbol{M}} \leftarrow \begin{bmatrix} 1+\frac{-1}{N} & \frac{-1}{N} & \cdots & \frac{-1}{N} \\ \frac{-1}{N} & 1+\frac{-1}{N} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ \frac{-1}{N} & \frac{-1}{N} & \cdots & 1+\frac{-1}{N} \end{bmatrix} \in \mathbb{R}^{N\times N}$$

2: $\boldsymbol{R_X} \leftarrow \tilde{\boldsymbol{U}}_{\text{samp}}\ \tilde{\boldsymbol{O}}\tilde{\boldsymbol{M}}$ ▷ avec $\tilde{\boldsymbol{U}}_{\text{samp}}$ la matrice $\boldsymbol{U}_{\text{samp}}$ tronquée à la colonne $r_{\boldsymbol{\Sigma}}$

return $\boldsymbol{R_X}$ $=0$

Comparaison numérique entre le KETKF à noyau linéaire et l'ETKF classique



Les deux algorithmes l'équivalence des méthodes comme prévu par la théorie.

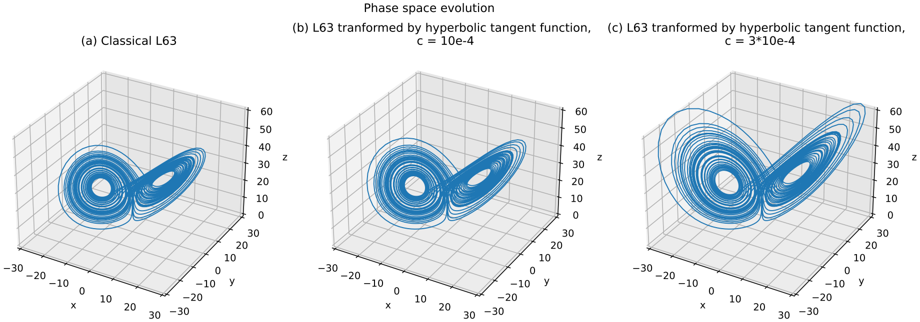
Impact du prétraitement

figures/impact_pretraitement.png

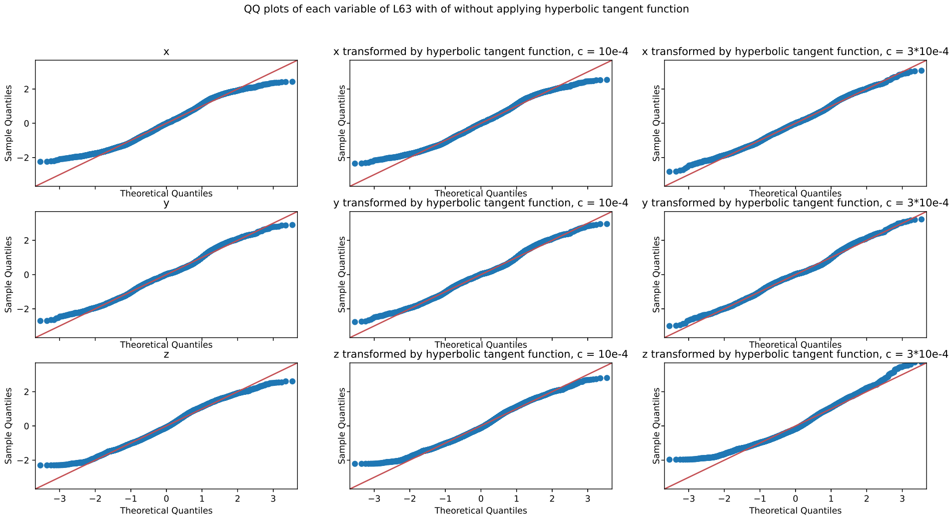
Impact du noyau tangent hyperbolique

figures/hyptanfun.png

Impact du noyau tangent hyperbolique



Impact du noyau tangent hyperbolique



Résultats LKETKF

figures/tests_LKETKF_clustersdis_plsnoyaux.png

Une piste pour traiter les discontinuités du LKETKF

figures/cluster_sup_4.png

Extension des clusters avec chevauchement

Enjeux actuels

Contributions multiples des points sur les bords dans $\boldsymbol{K}_{\text{loc}}$.
⇒ Trouver un moyen de pondérer les contributions tout en maintenant la semi-définie positivité de $\boldsymbol{K}_{\text{loc}}$.