A Gaussian high-order sampling hybrid filter for biogeochemical data assimilation: application to chlorophyll satellite data

Simone Spada, Anna Teruzzi, Stefano Salon, Stefano Maset, Gianpiero Cossarini

sspada@ogs.it

The Internationl EnKF Workshop 2022
30 May – 2 June, 2022

Overview

- **Introduction**
  (which are the main elements of novelty in this work?)

- **Motivation**
  (why do you need it?)

- **High-Order Sampling formulation**
  (how does it work?)

- **Implementation and results**
  (does it work well?)

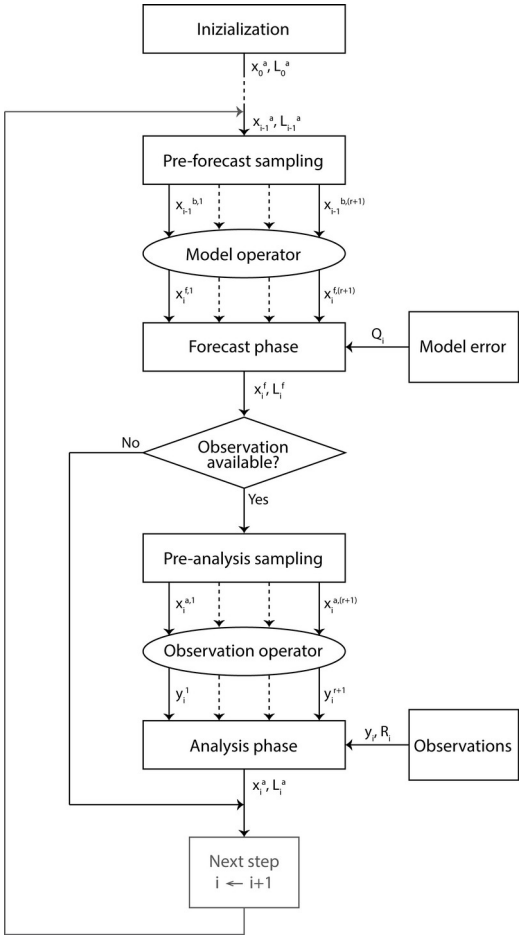
Introduction

The **Gaussian High-Order Sampling Hybrid** filter (**GHOSH**) is a novel ensemble filter developed at OGS. Main features:

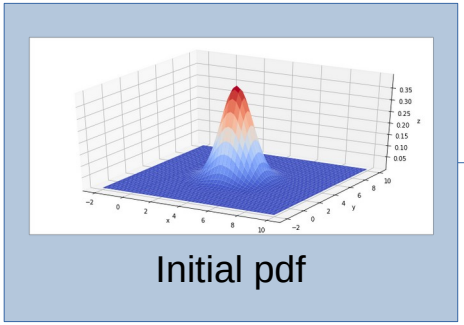
1) High-order sampling method for weighted ensembles

2) Resampling before assimilation

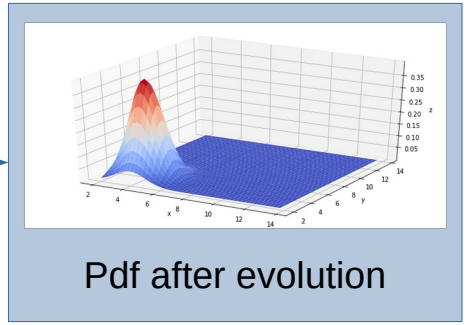
3) Non-orthogonal model error projection into the ensemble subspace (not presented here)



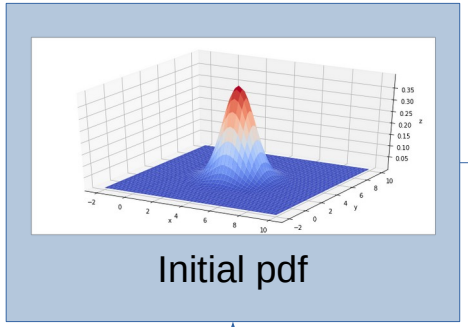
Motivation



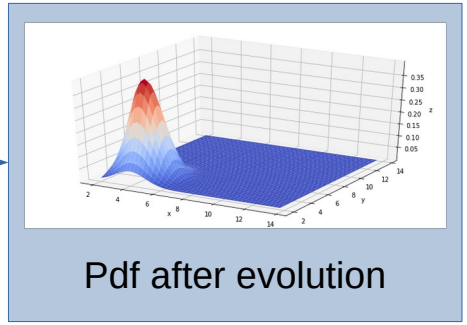
Model operator

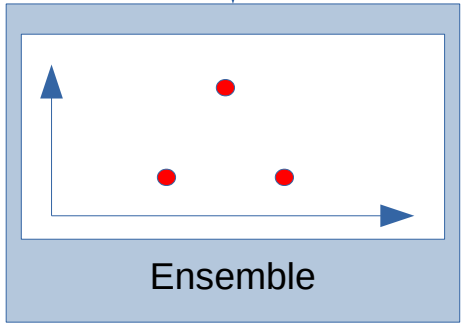


Motivation

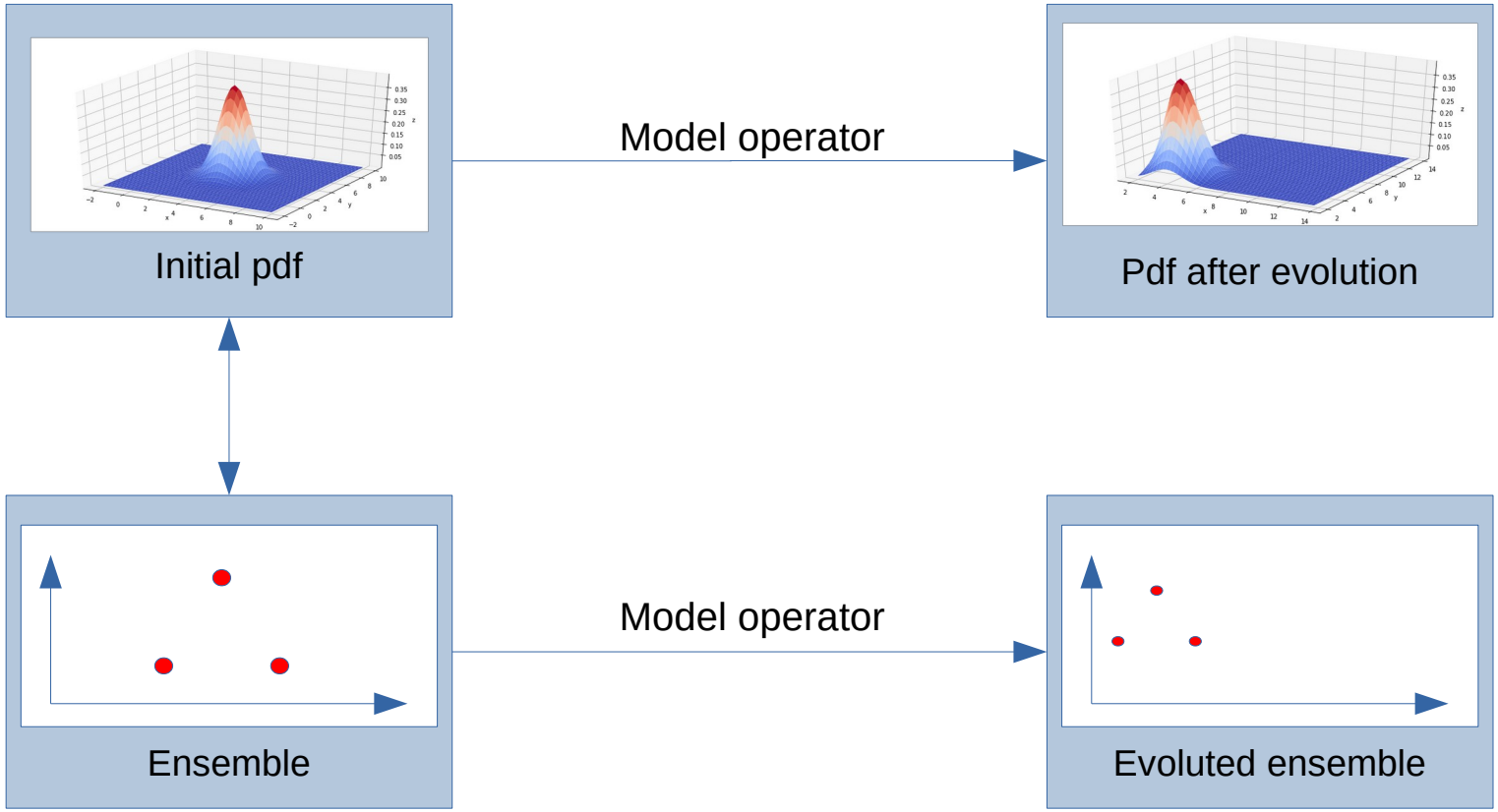


Model operator

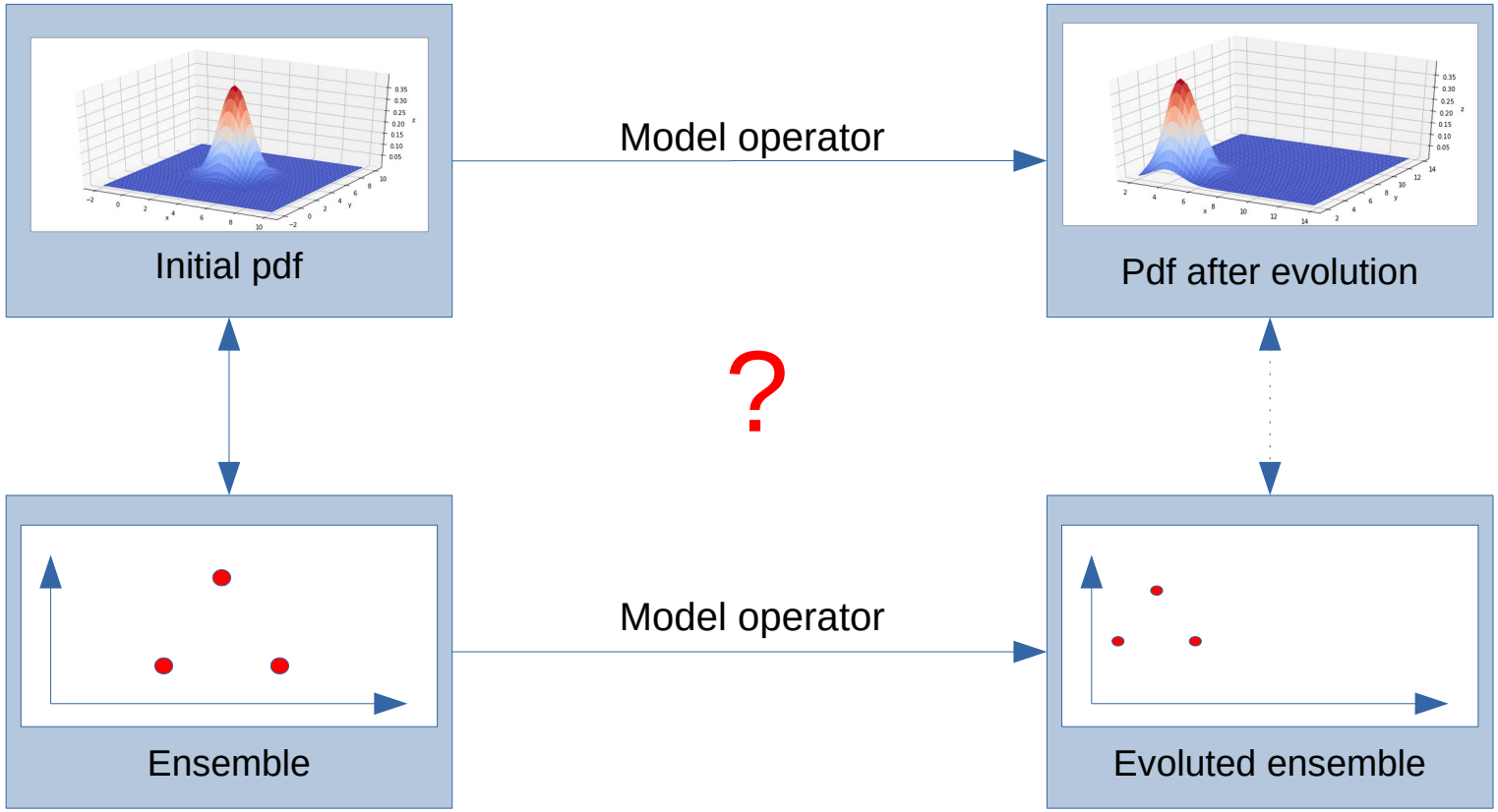




Motivation


Initial pdf
Model operator
Pdf after evolution
Ensemble
Model operator
Evoluted ensemble

Motivation


Initial pdf
Model operator
Pdf after evolution
?
Ensemble
Model operator
Evoluted ensemble

Motivation

After the evolution, is the ensemble representative of the pdf?

- If the model is **linear**, the evolved ensemble is as good as the initial ensemble (“good” = same statistical moments as the pdf).

- If the model is a **second order polynomial**, the initial pdf and ensemble must have same mean and covariance to ensure that the evolved pdf and ensemble have the same mean. This is what is done in square root filters (e.g., SEIK, ETKF, ESTKF etc.), but…

- …the model is usually a more **general function** M,

$$M(\bar{x} + x) = M(\bar{x}) + M'(\bar{x})x + M''(\bar{x})x^2 + O(x^3),$$

  thus an error term appears and it can be relevant!

The high-order sampling

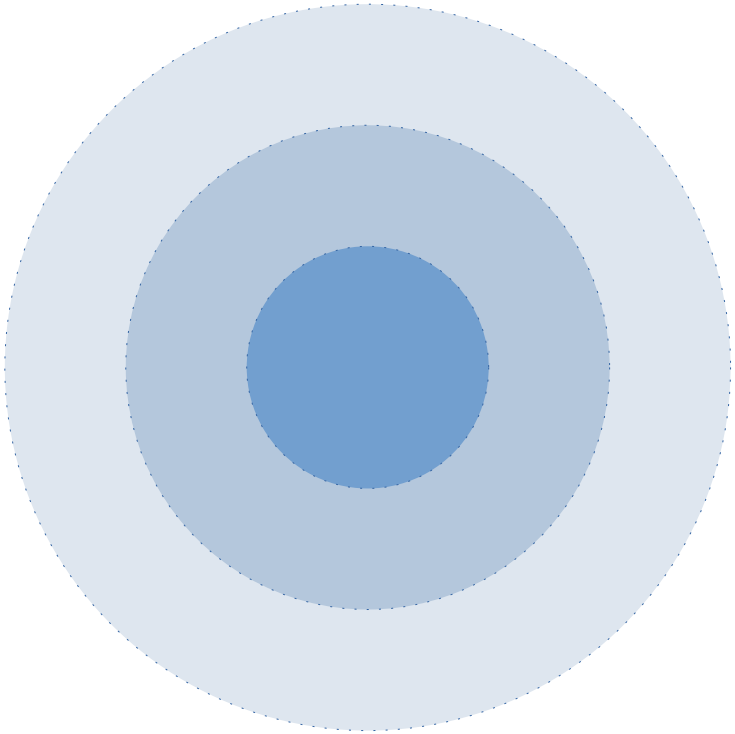
How can we build a better ensemble?

If the initial ensemble is good enough (same statistical moments up to order *n*), then the mean of the evoluted ensemble is exact up to order *n* (i.e., the first *n* terms of the Taylor expansion of M are taken into account).

There is still an error but **higher order implies smaller errors**. This is true for the estimation of the covariance too.

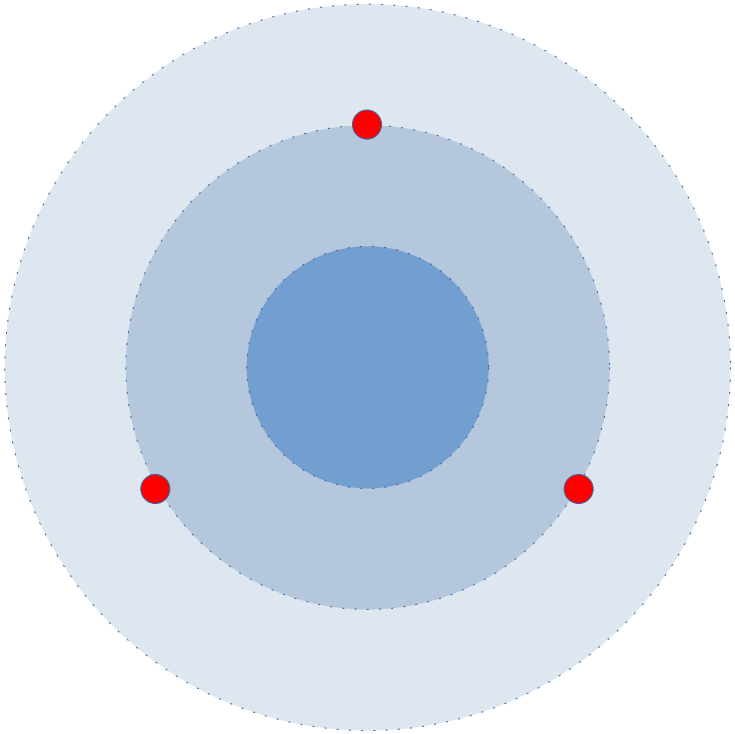
=> We need an ensemble matching higher moments, not only mean and covariance

The high-order sampling



Shady areas represent a Gaussian distribution.

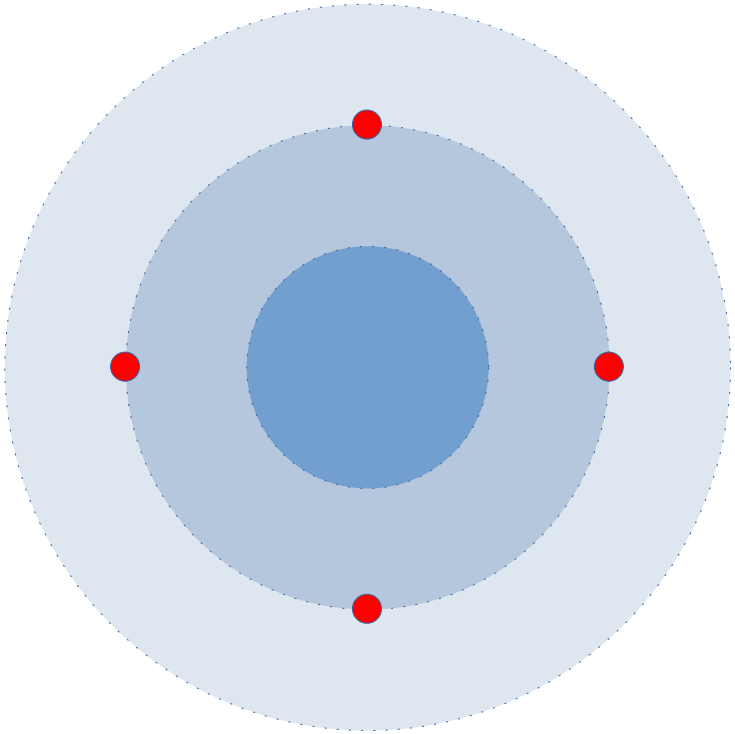
The high-order sampling



Shady areas represent a Gaussian distribution.

3 ensemble members:
2nd order sampling

The high-order sampling

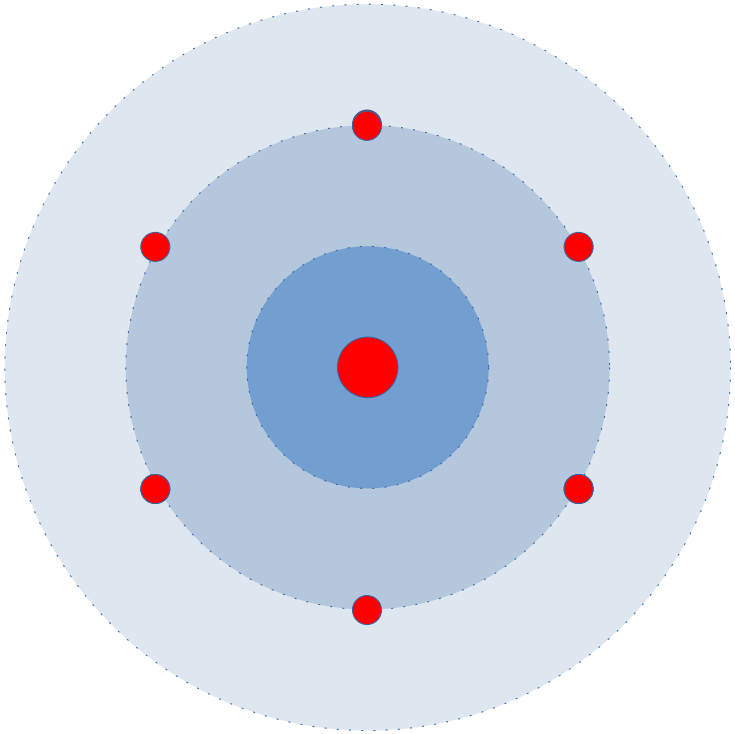


Shady areas represent a Gaussian distribution.

3 ensemble members:
2nd order sampling

4 ensemble members:
3rd order sampling

The high-order sampling



Shady areas represent a Gaussian distribution.

3 ensemble members:
2^{nd} order sampling

4 ensemble members:
3^{rd} order sampling

7 weighted ensemble members:
5^{th} order sampling

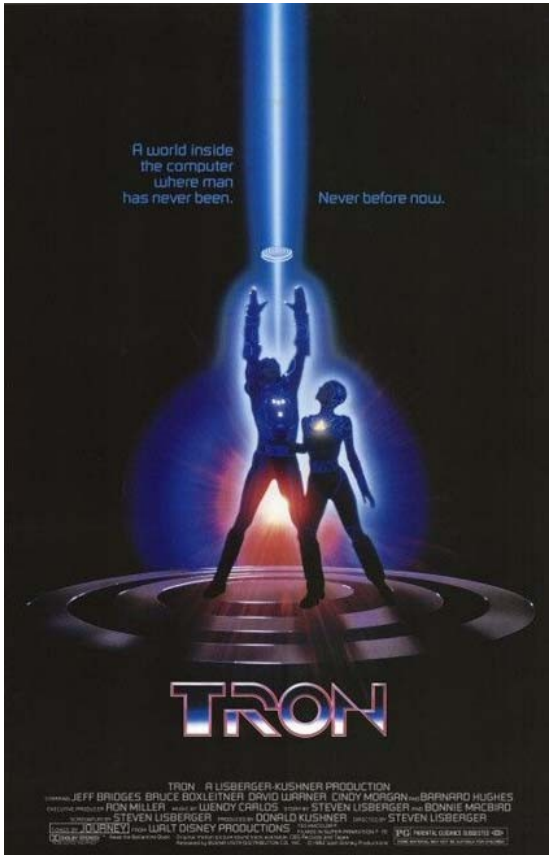
The high-order sampling

There are 2 **problems**:

1) The moment matching equations form a big (depending on the sampling order *n* and the ensemble size) **non-linear system**. Solving non-linear system is not easy.

2) Square root filters use *d* + *1* ensemble members to represent an error subspace of dimension *d*. A **larger ensemble** (size > *d* + *1*) is needed to match moments higher than order 2 in the same error subspace.

The high-order sampling



Like a man says,
“there’s no problems,
only solutions.”
[cit. TRON, 1982]

The high-order sampling

There are 2 **solutions**:

1) The moment matching equations form a big (depending on the sampling order *n* and the ensemble size) **non-linear system**. Solving non-linear system is not easy.
   Just solve it once, offline!

2) Square root filters use *d* + *1* ensemble members to represent an error subspace of dimension *d*. A **larger ensemble** (size > *d* + *1*) is needed to match moments higher than order 2 in the same error subspace.
   Use a weighted ensemble with the same number of members, loosing accuracy only where it is less relevant.

The high-order sampling

A little bit of story: the second order exact sampling
(used in SEIK, ETKF, ESTKF and other square root filters)

The covariance **P** is approximated using a base **L** and a (smaller then **P**) symetric matrix **A**. We are looking for **X** (i.e., the ensemble anomalies matrix), such that:

$$(1/\text{EnsSize})\ \mathbf{X}\,\mathbf{X}^T = \mathbf{L}\,\mathbf{A}\,\mathbf{L}^T \approx \mathbf{P}$$

A 2nd order solution is given by:

$$\mathbf{X} = \text{sqrt}(\text{EnsSize})\ \mathbf{L}\,\mathbf{S}\,\mathbf{\Omega}$$

$$\mathbf{S}^2 = \mathbf{A}$$

$$\mathbf{\Omega}\,\mathbf{\Omega}^T = \mathbf{I}, \qquad \mathbf{\Omega}\,\mathbf{1} = \mathbf{0}$$

The high-order sampling

A little bit of story: the second order exact sampling
(used in SEIK, ETKF, ESTKF and other square root filters)

The covariance **P** is approximated using a base **L** and a (smaller then **P**) symetric matrix **A**. We are looking for **X** (i.e., the ensemble anomalies matrix), such that:

$$(1/\text{EnsSize})\ \mathbf{X}\,\mathbf{X}^T = \mathbf{L}\,\mathbf{A}\,\mathbf{L}^T \approx \mathbf{P}$$

A 2nd order solution is given by:

$$\mathbf{X} = \text{sqrt}(\text{EnsSize})\ \mathbf{L}\,\mathbf{S}\,\mathbf{\Omega}$$

$$\mathbf{S}^2 = \mathbf{A}$$

$$\mathbf{\Omega}\,\mathbf{\Omega}^T = \mathbf{I}, \qquad \mathbf{\Omega}\,\mathbf{1} = \mathbf{0}$$

Ω is (up to a factor) a 2nd order sampling of a standard Gaussian

The high-order sampling

A little bit of story: the second order exact sampling
(used in SEIK, ETKF, ESTKF and other square root filters)

The covariance **P** is approximated using a base **L** and a (smaller then **P**) symetric matrix **A**. We are looking for **X** (i.e., the ensemble anomalies matrix), such that:

$$(1/\text{EnsSize})\ \mathbf{X}\,\mathbf{X}^T = \mathbf{L}\,\mathbf{A}\,\mathbf{L}^T \approx \mathbf{P}$$

A 2nd order solution is given by:

$$\mathbf{X} = \text{sqrt}(\text{EnsSize})\,\mathbf{L}\,\mathbf{S}\,\mathbf{\Omega}$$

$$\mathbf{S}^2 = \mathbf{A}$$

$$\mathbf{\Omega}\,\mathbf{\Omega}^T = \mathbf{I}, \qquad \mathbf{\Omega}\,\mathbf{1} = \mathbf{0}$$

L S is the projection in the error subspace

Ω is (up to a factor) a 2nd order sampling of a standard Gaussian

The high-order sampling

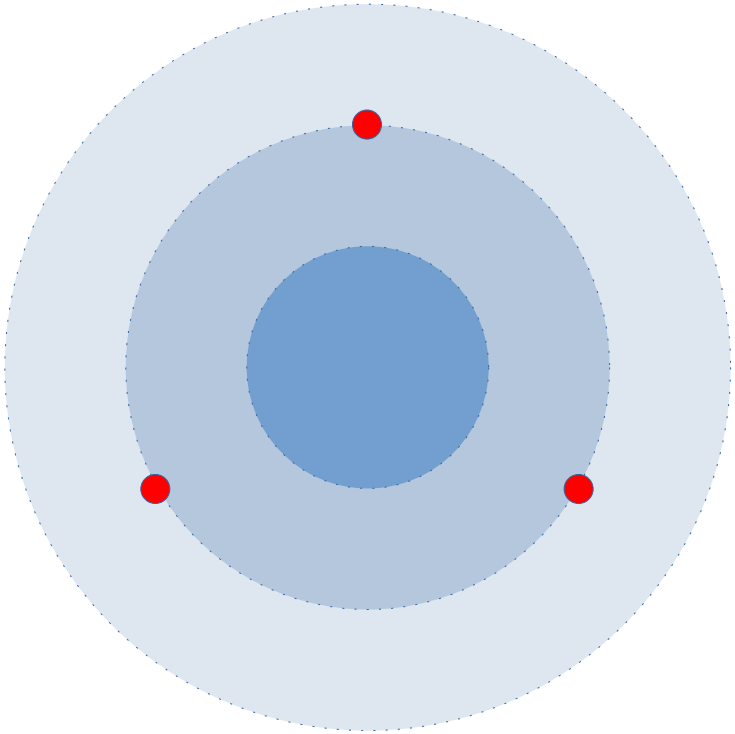
To take advantage of the high-order sampling, the $\mathbf{\Omega}$ matrix is factorized in 2 components.

$$\mathbf{X} = \mathbf{L}\,\mathbf{S}\,\mathbf{\Omega}_{rnd}\,\mathbf{\Omega}_{h}$$

$\mathbf{\Omega}_{rnd}$ is a random ortogonal matrix which capture the randomness of $\mathbf{\Omega}$, $\mathbf{\Omega}_{h}$ is the fixed matrix of the sampling obtained by solving the moment matching non-linear system in a standard case.

=> The non-linear system is solved only once!

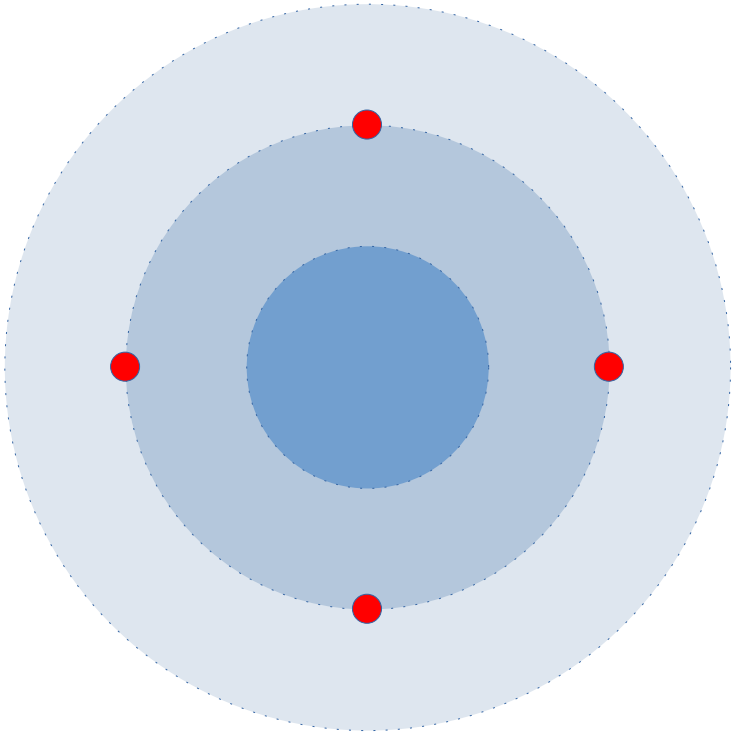
The high-order sampling



The ensemble size problem

In 2 dimensions, just 3 ensemble members are needed for a 2nd order sampling

The high-order sampling

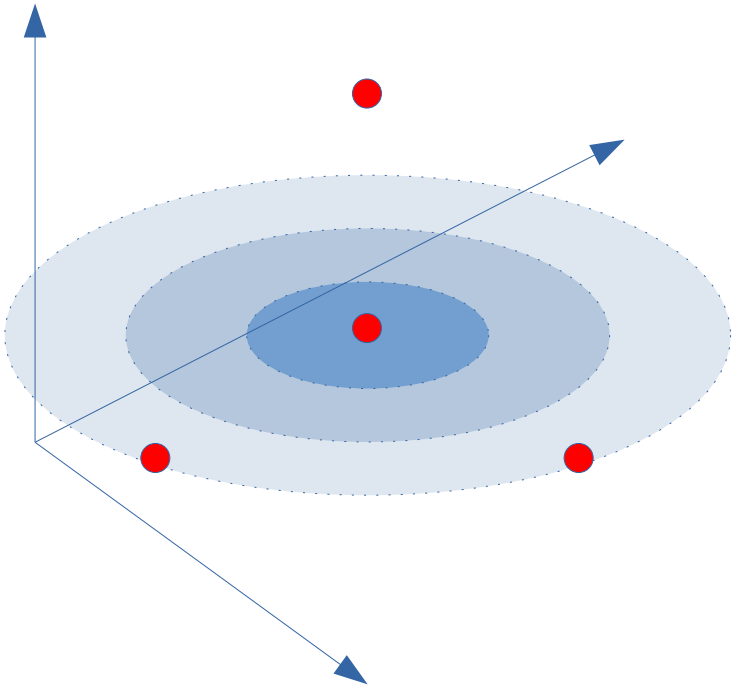


The ensemble size problem

In 2 dimensions, just 3 ensemble members are needed for a 2nd order sampling

and it is impossible to achieve a 3rd order sampling with less than 4 ensemble members

The high-order sampling



The ensemble size problem

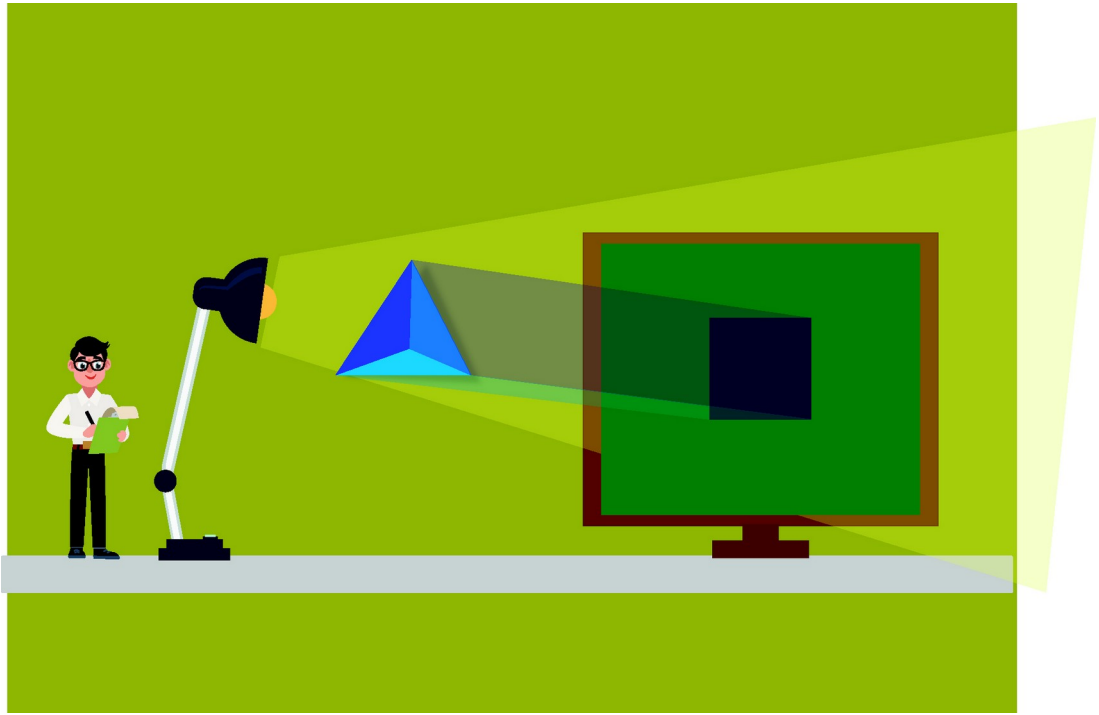
In 2 dimensions, just 3 ensemble members are needed for a 2nd order sampling

and it is impossible to achieve a 3rd order sampling with less than 4 ensemble members

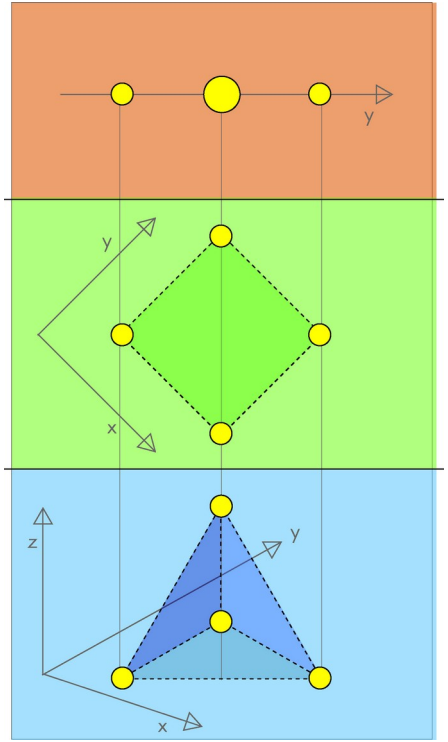
but what if those 4 ensemble members came from a 2nd order sampling in **3 dimensions** instead of 2?

The high-order sampling

It is like playing with shadows!



The high-order sampling

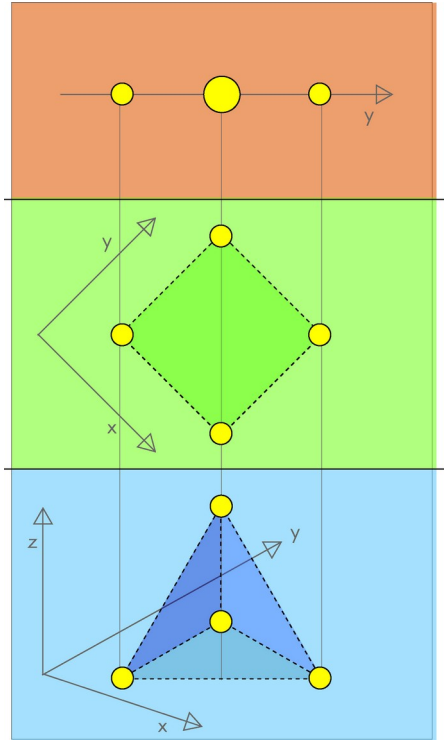


A 5^{th} order sampling in 1 dimension is achieved with 3 weighted ensemble members, which can be obtained by projecting:

a 3^{rd} order sampling in 2 dimensions with 4 ensemble members, which can be obtained by projecting:

a 2^{nd} order sampling in 3 dimensions with 4 ensemble members.

The high-order sampling



A 5^{th} order sampling in 1 dimension is achieved with 3 weighted ensemble members, which can be obtained by projecting:

a 3^{rd} order sampling in 2 dimensions with 4 ensemble members, which can be obtained by projecting:

a 2^{nd} order sampling in 3 dimensions with 4 ensemble members.

It's time for **principal component analysis**!

The high-order sampling

Using this method, in the **GHOSH** filter, the most relevant components of the covariance $\mathbf{P} \approx \mathbf{L}\,\mathbf{S}^2\,\mathbf{L}^T$, are approximated with an order higher than 2.

$$\mathbf{X} = \mathbf{L}\,\mathbf{S}\,\mathbf{E}\,\boldsymbol{\Omega}$$

where $\mathbf{S}\,\mathbf{L}^T\,\mathbf{L}\,\mathbf{S} = \mathbf{E}\,\mathbf{D}\,\mathbf{E}^T$ is an eigenvalue decomposition with eigenvalues in decreasing order

and $\boldsymbol{\Omega}$ is built taking into account both randomness and the sampling with higher order in the first dimensions.

=> Ensemble size and error subspace same as in square root filters

The high-order sampling

Using this method, in the **GHOSH** filter, the most relevant components of the covariance $\mathbf{P} \approx \mathbf{L}\,\mathbf{S}^2\,\mathbf{L}^T$, are approximated with an order higher than 2.

Old base

$$\mathbf{X} = \mathbf{L}\,\mathbf{S}\,\mathbf{E}\,\mathbf{\Omega}$$

New base

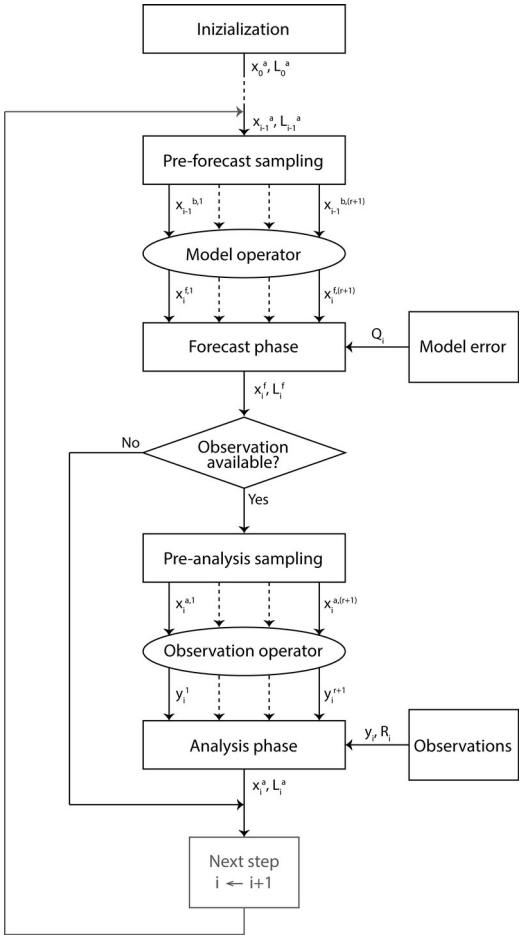
where $\mathbf{S}\,\mathbf{L}^T\,\mathbf{L}\,\mathbf{S} = \mathbf{E}\,\mathbf{D}\,\mathbf{E}^T$ is an eigenvalue decomposition with eigenvalues in decreasing order

and $\mathbf{\Omega}$ is built taking into account both randomness and the sampling with higher order in the first dimensions.

=> Ensemble size and error subspace same as in square root filters

The re-sampling before analysis

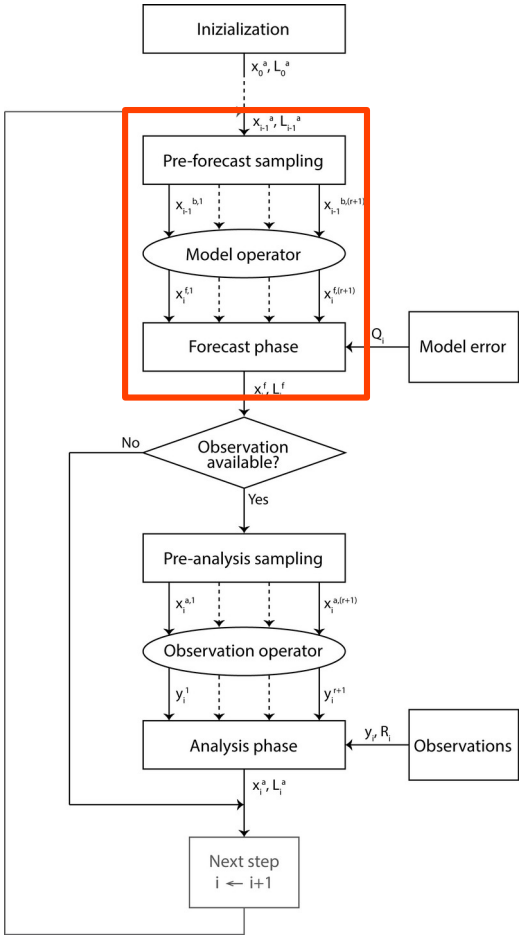
Any operator acting on a pdf can take advantage by the high-order sampling and there are 2 operators!



The re-sampling before analysis

Any operator acting on a pdf can take advantage by the high-order sampling and there are 2 operators!

In the case of model operator, the high-order sampling helps approximating the forecast pdf

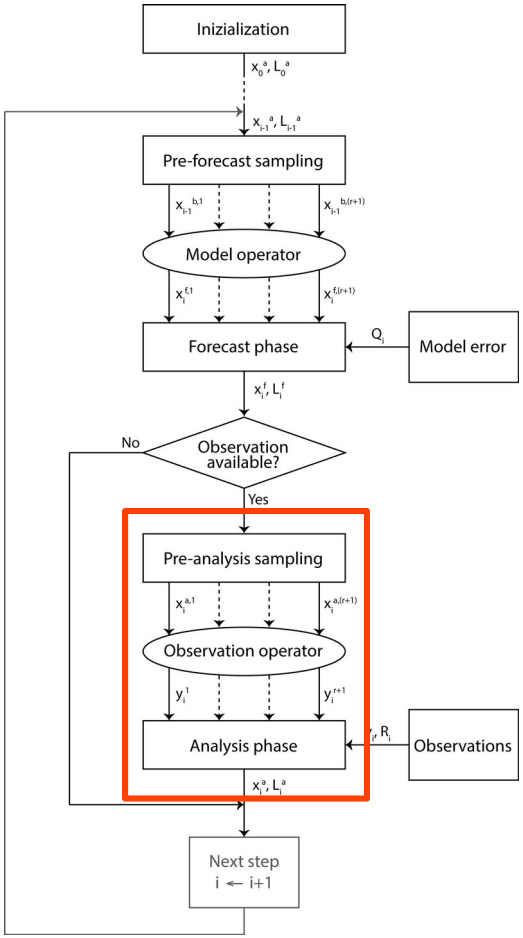


The re-sampling before analysis

Any operator acting on a pdf can take advantage by the high-order sampling and there are 2 operators!

In the case of model operator, the high-order sampling helps approximating the forecast pdf

In the case of observation operator, it helps with the likelihood pdf, used to produce the analysis



The re-sampling before analysis

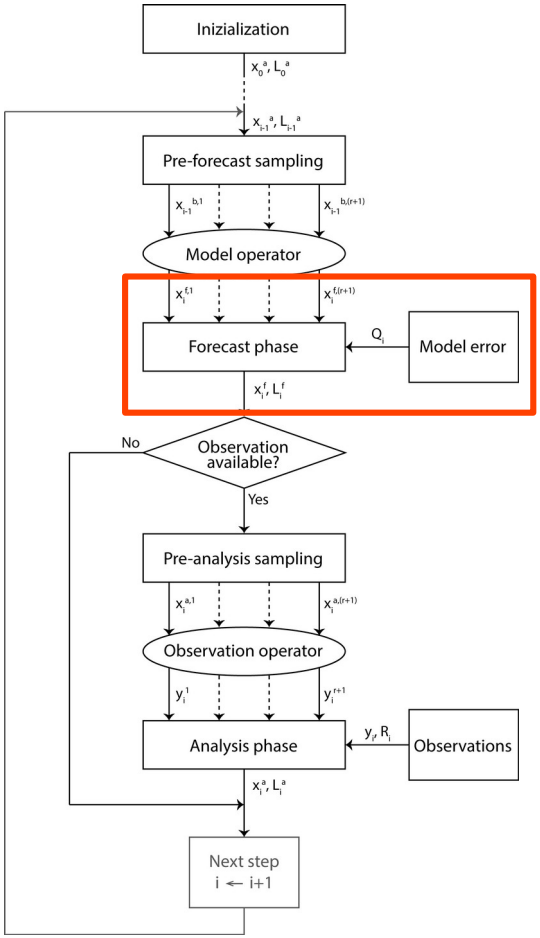
Any operator acting on a pdf can take advantage by the high-order sampling and there are 2 operators!

In the case of model operator, the high-order sampling helps approximating the forecast pdf

In the case of observation operator, it helps with the likelihood pdf, used to produce the analysis

Sampling twice is relevant because:

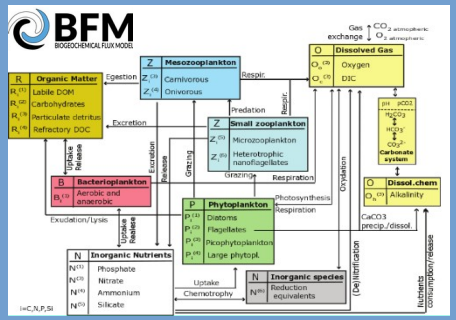
- Inflation or Model error change the forecast pdf
- H is often non linear in Biogeochemistry



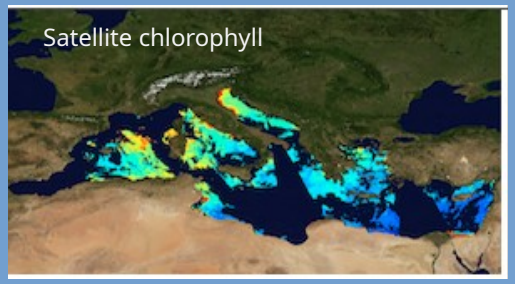
Assimilation experiment

Setup

- Mediterranean Sea
- 1-year simulations
- 1/4° horizontal resolution
- 16 ensemble members
- RMSD to independent data
- 18 tests with different parameters (e.g., inflation and sampling order)



Model (BGC + transport):
BFM + OGSTM



Observations:
Satellite chlorophyll

Assimilation experiment

Results:

- The sampling order was not strongly effective on the assimilated variable (surface chlorophyll)
- The sampling order had a significant impact in a non-assimilated variable (Nitrate), improving up to 80% the average assimilation skill.

	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6
Order 2	0.93	1.50	1.96	0.94	0.94	1.35
Order 3	0.85	1.14	2.15	0.89	0.93	1.02
Order 5	0.83	0.83	1.76	0.91	0.91	0.91

***Table**: Nitrate RMSD (mmol/m3)*

Assimilation experiment

Remarks:

- The GHOSH filter has been implemented in the parallel framework developed by OGS inside the SEAMLESS project
- The computational cost of an assimilation experiment with the GHOSH filter is equivalent to other square root filters

Future developments:

- Comparison between GHOSH and SEIK at full resolution (1/24^{o})
- A manuscript is under submission (feel free to contact me if interested in the final version: sspada@ogs.it)

THANK YOU

Simone Spada, Anna Teruzzi, Stefano Salon, Stefano Maset, Gianpiero Cossarini

sspada@ogs.it

SEAMLESS

The Internationl EnKF Workshop 2022
30 May – 2 June, 2022