Ensemble Kalman Filter Strategies for Efficient Data Assimilation in Geosciences

Simone Spada, Anna Teruzzi, Gianpiero Cossarini

sspada@ogs.it

The EnKF Workshop 2024
17 – 19 June, 2024

Did you know that…

…your DA performances are widely affected by:

- Sampling
- Tuning

Did you know that…

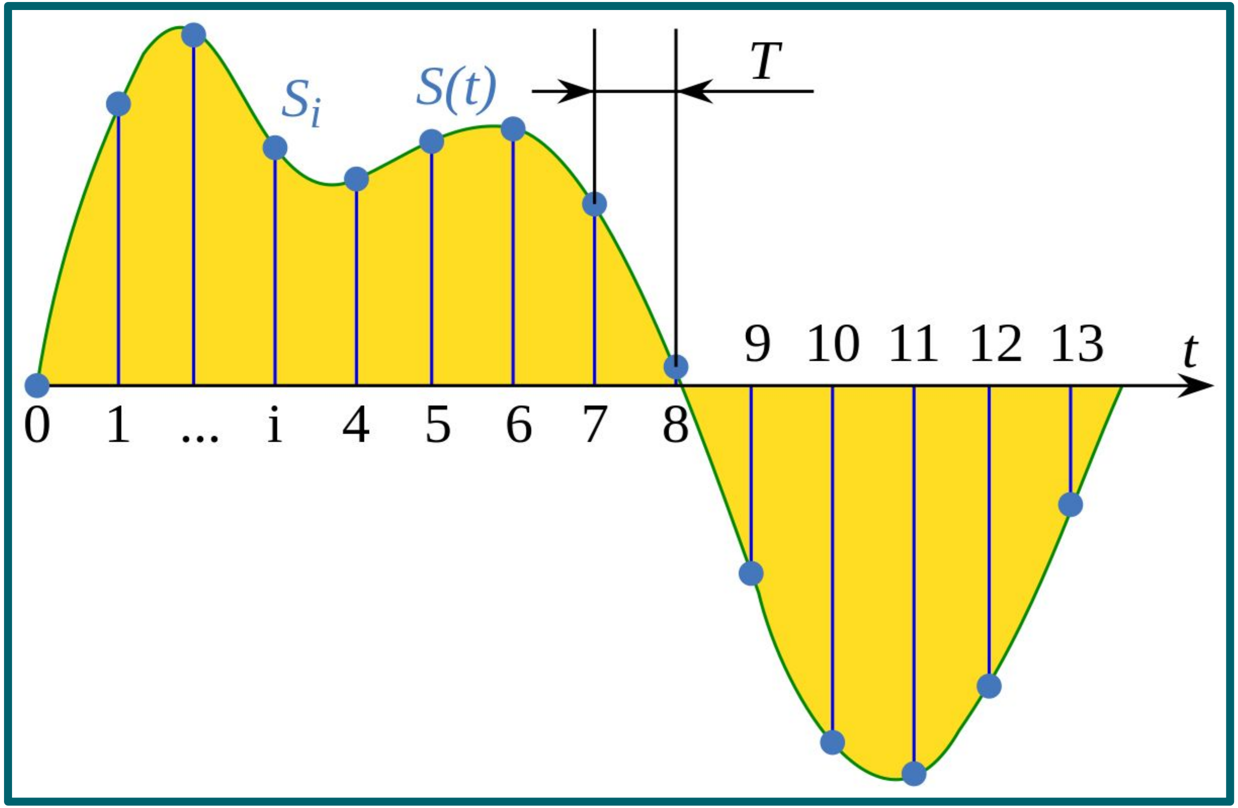
…your DA performances are widely affected by:

- Sampling
- Tuning

It's general,
it's for everyone,
it's for **you!**

OGS

Sampling



Sampling (a look to the past)

The **second-order-exact sampling**
Pham 1996, Pham 2001

*used in **SEIK**, **ETKF** and other square root filters*

The covariance **P** is approximated by a base **L** and a small symetric matrix **A**:

$$\mathbf{P} \approx \mathbf{L}\,\mathbf{A}\,\mathbf{L}^T$$

The sampling matrix **X** (i.e., the ensemble anomalies) is:

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

where $\mathbf{S}^2 = \mathbf{A}$, $\boldsymbol{\Omega}\,\boldsymbol{\Omega}^T = \mathbf{I}$, $\boldsymbol{\Omega}\,\mathbf{1} = \mathbf{0}$.

The sampling matches statistical moments up to **order 2**:

$$\mathbf{X}\,\mathbf{1} = \mathbf{0}, \quad (1/\mathrm{EnsSize})\,\mathbf{X}\,\mathbf{X}^T = \mathbf{L}\,\mathbf{A}\,\mathbf{L}^T$$

OGS

Sampling (a look to the past)

The **second-order-exact sampling**
Pham 1996, Pham 2001

*used in **SEIK**, **ETKF** and other square root filters*

The covariance **P** is approximated by a base **L** and a small symetric matrix **A**:

$$\mathbf{P} \approx \mathbf{L}\,\mathbf{A}\,\mathbf{L}^T$$

The sampling matrix **X** (i.e., the ensemble anomalies) is:

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

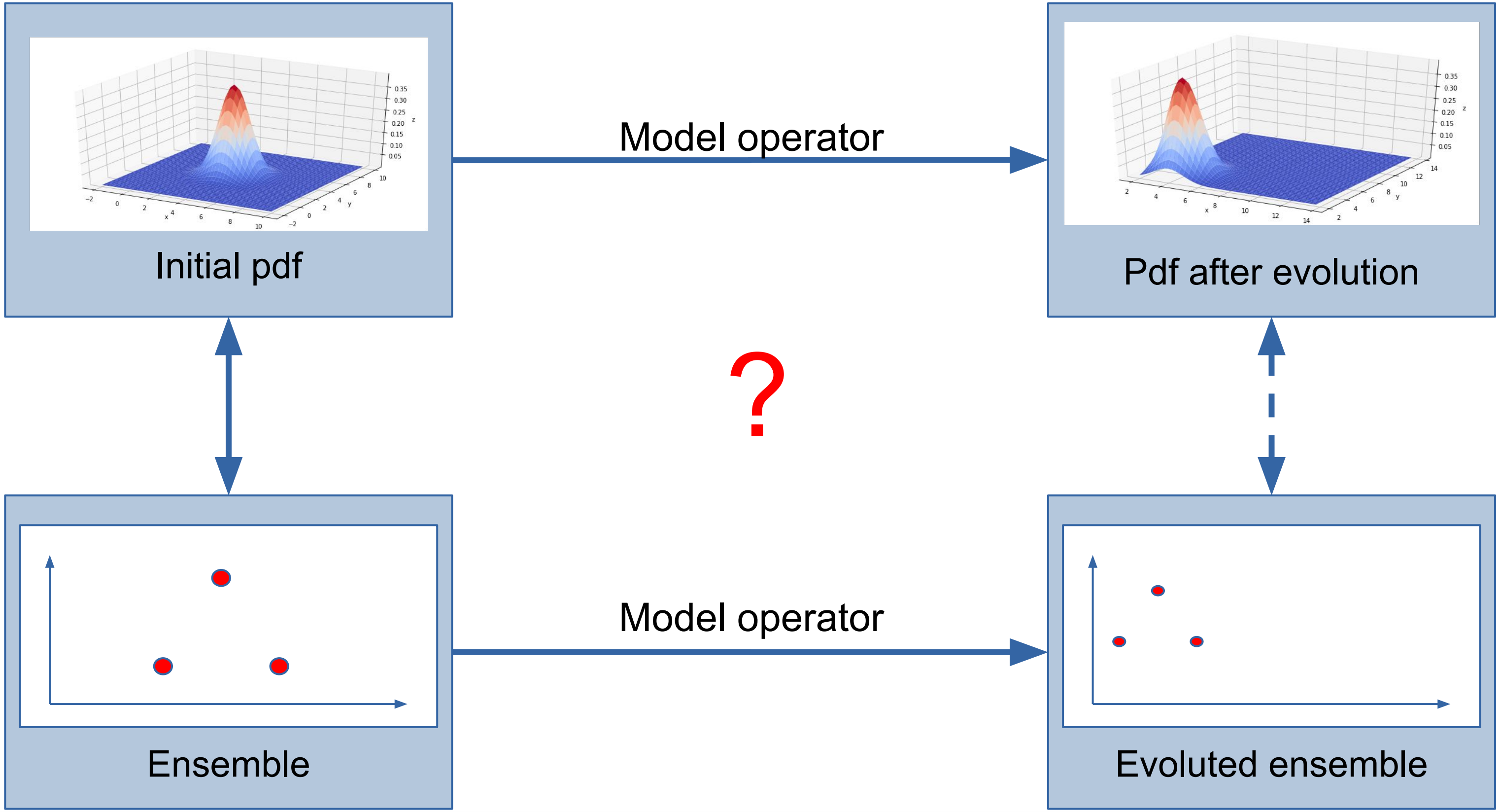
where $\quad \mathbf{S}^2 = \mathbf{A}, \quad \boldsymbol{\Omega}\,\boldsymbol{\Omega}^T = \mathbf{I}, \quad \boldsymbol{\Omega}\,\mathbf{1} = \mathbf{0}.$

The sampling matches statistical moments up to **order 2**:

$$\mathbf{X}\,\mathbf{1} = \mathbf{0}, \quad (1/\mathrm{EnsSize})\ \mathbf{X}\,\mathbf{X}^T = \mathbf{L}\,\mathbf{A}\,\mathbf{L}^T$$

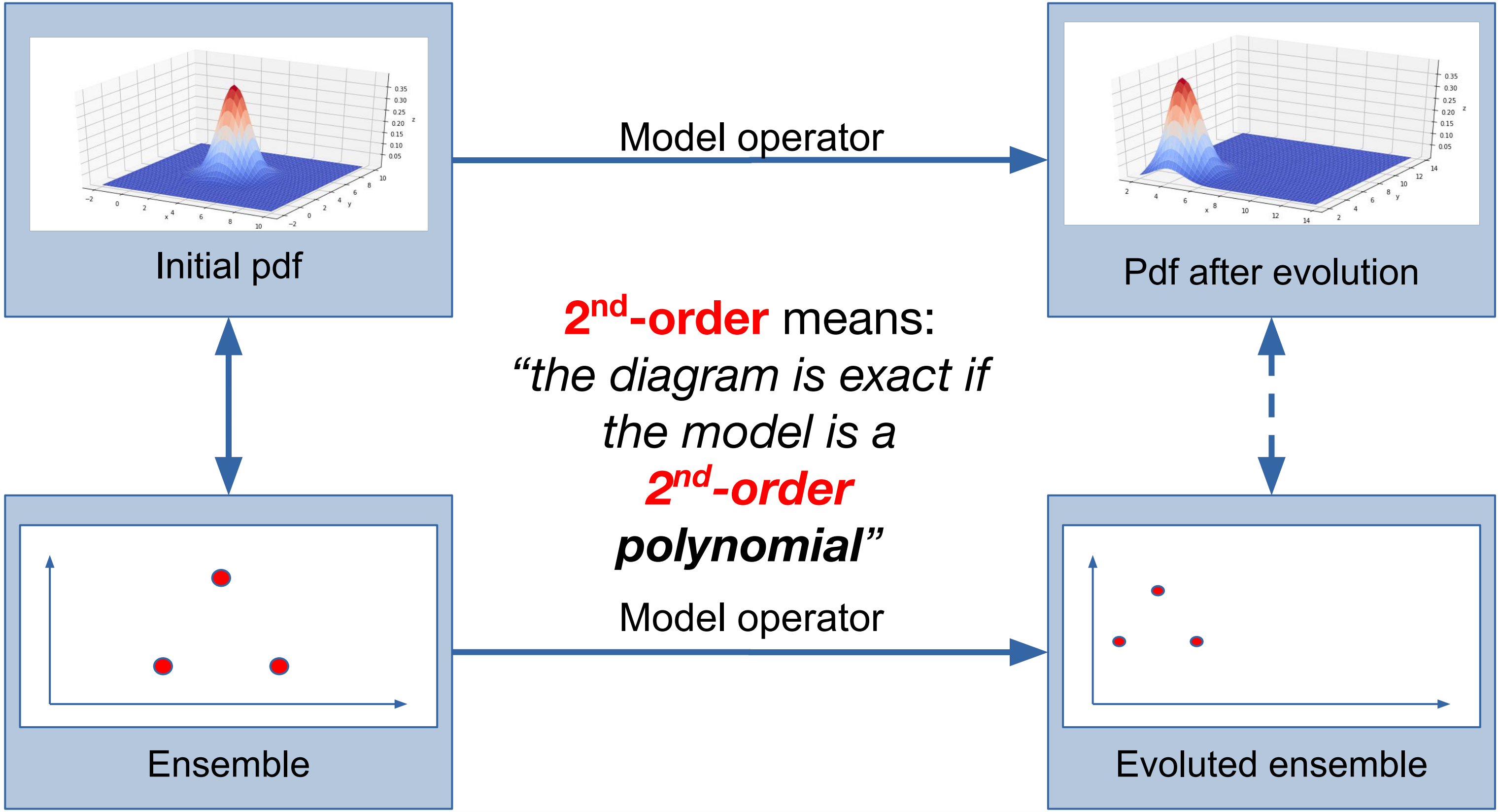
OGS

Sampling order



OGS

Sampling order



2nd-order means:
"the diagram is exact if the model is a ***2nd-order polynomial"***

Sampling (a look to the past)

The **second-order-exact sampling**
Pham 1996, Pham 2001

*used in **SEIK**, **ETKF** and other square root filters*

The covariance **P** is approximated by a base **L** and a small symetric matrix **A**:

$$\mathbf{P} \approx \mathbf{L}\,\mathbf{A}\,\mathbf{L}^T$$

The sampling matrix **X** (i.e., the ensemble anomalies) is:

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

where $\mathbf{S}^2 = \mathbf{A}, \quad \boldsymbol{\Omega}\,\boldsymbol{\Omega}^T = \mathbf{I}, \quad \boldsymbol{\Omega}\,\mathbf{1} = \mathbf{0}.$

The sampling matches statistical moments up to **order 2**:

$$\mathbf{X}\,\mathbf{1} = \mathbf{0}, \quad (1/\mathrm{EnsSize})\,\mathbf{X}\,\mathbf{X}^T = \mathbf{L}\,\mathbf{A}\,\mathbf{L}^T$$

Sampling (a look to the past)

The **second-order-exact sampling**
Pham 1996, Pham 2001

*used in **SEIK**, **ETKF** and other square root filters*

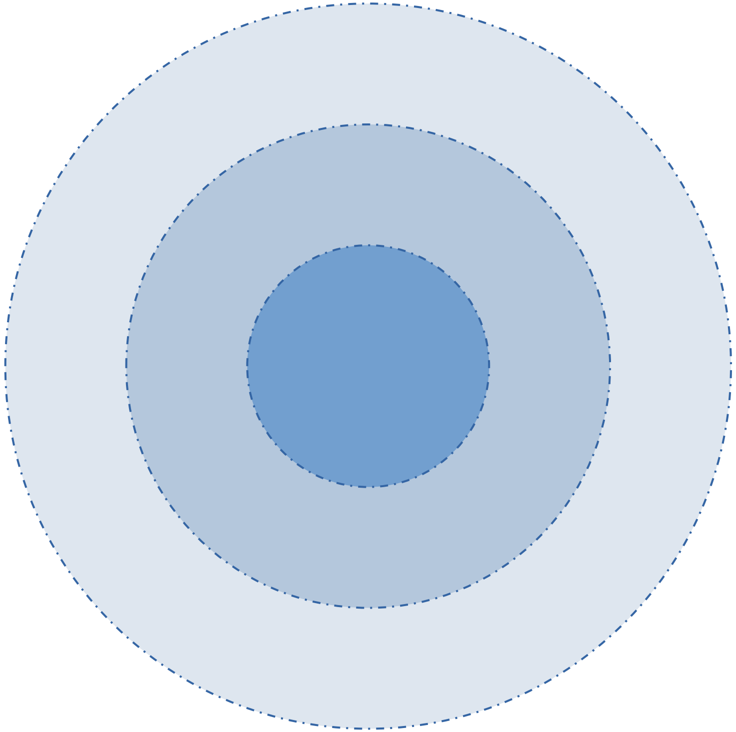
The covariance **P** is approximated by a base **L** and a small symetric matrix **A**:

The … :

This sampling method is exact if
the model is a second-order polynomial

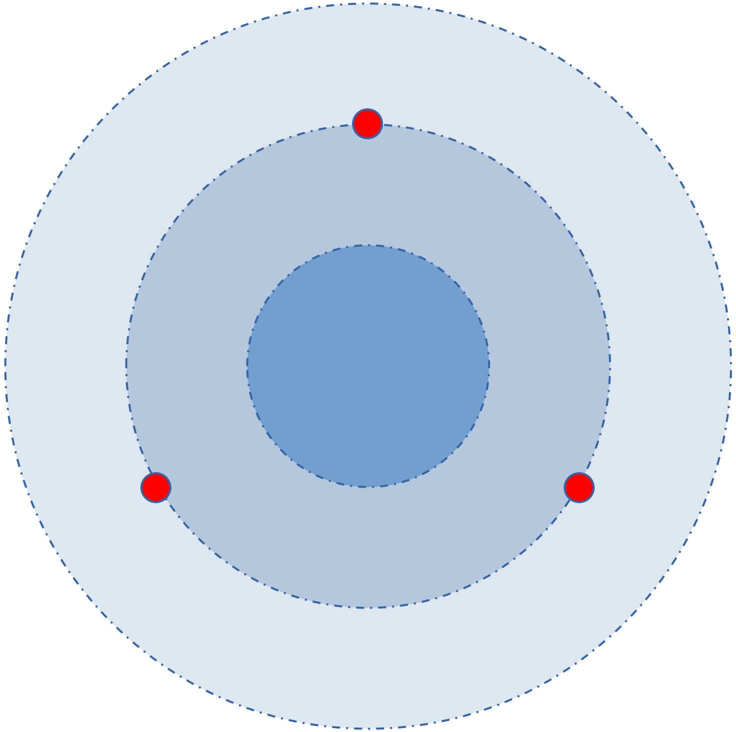
The … **2**:

Higher order requires more ensemble members



- Shady areas represent a Gaussian distribution.

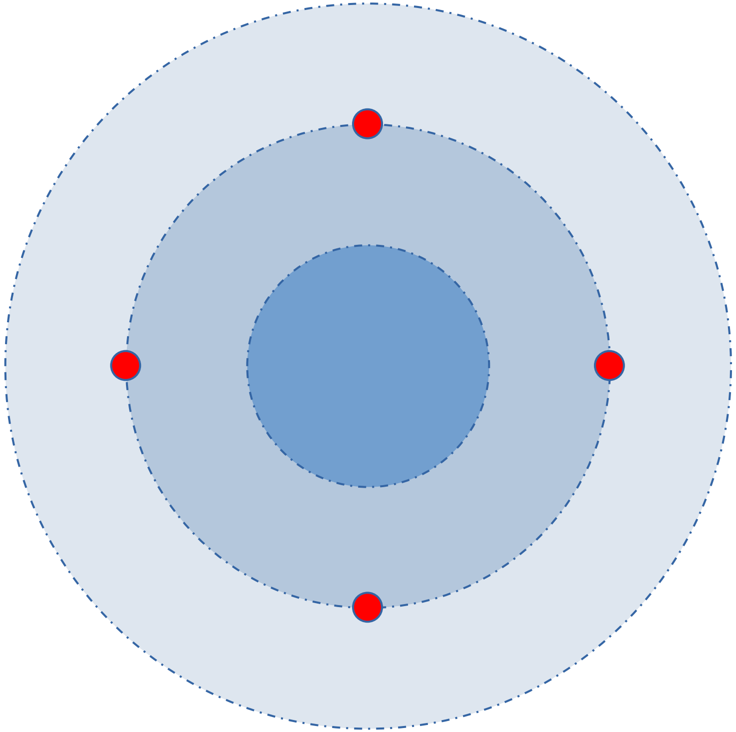
Higher order requires more ensemble members



- Shady areas represent a Gaussian distribution.

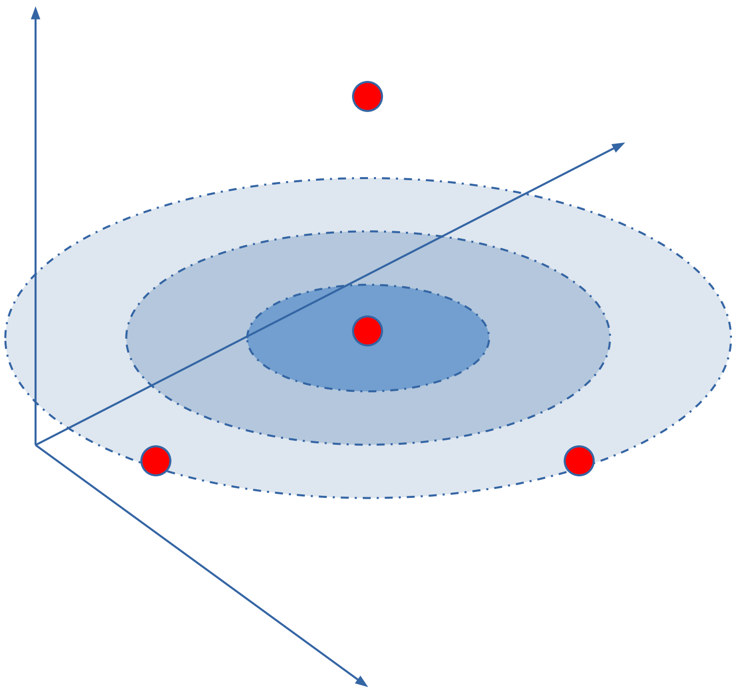
- 3 ensemble members:
  2nd-order sampling

Higher order requires more ensemble members



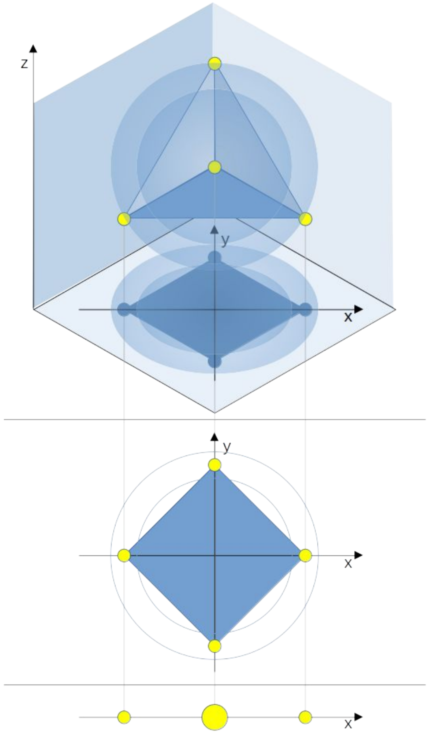
- Shady areas represent a Gaussian distribution.
- 3 ensemble members:
  2^{nd}-order sampling
- 4 ensemble members:
  3^{rd}-order sampling

Higher order requires more ensemble members



- Shady areas represent a Gaussian distribution.
- 3 ensemble members:
  2nd-order sampling
- 4 ensemble members:
  3rd-order sampling
- 4 ensemble members in 3D space:
  usual 2nd-order sampling

The high-order sampling idea



4 members in 3D
(2^{nd}-order approximation)

that project in

4 members in 2D
(3^{rd}-order approximation)

that project in

3 weighted members in 1D
(5^{th}-order approximation)

Improved precision

by

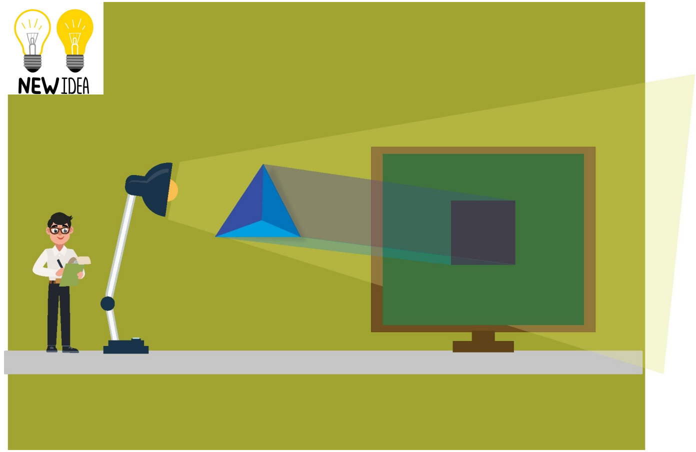
rising order
in the most relevant
PCA components

NO more members

NO higher
computational cost

OGS

Enhance your sampling method



The **high-order sampling**
Spada et al. 2024
(https://doi.org/10.5194/gmd-2023-170)

*used in **GHOSH***

$$\mathbf{P} \approx \mathbf{L}\,\mathbf{A}\,\mathbf{L}^T, \qquad \mathbf{S}^2 = \mathbf{A}$$

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

where **W** is the diagonal matrix of the ensemble weights,
$\mathbf{S}\,\mathbf{L}^T\,\mathbf{L}\,\mathbf{S} = \mathbf{E}\,\mathbf{D}\,\mathbf{E}^T$ is an eigendecomposition with decreasing eigenvalues,
$\mathbf{\Omega}_h$ is an orthogonal matrix encoding statistical moments.

The sampling matches statistical moments up to an **arbitrary high order** (limited by ensemble size) in the principal error components.

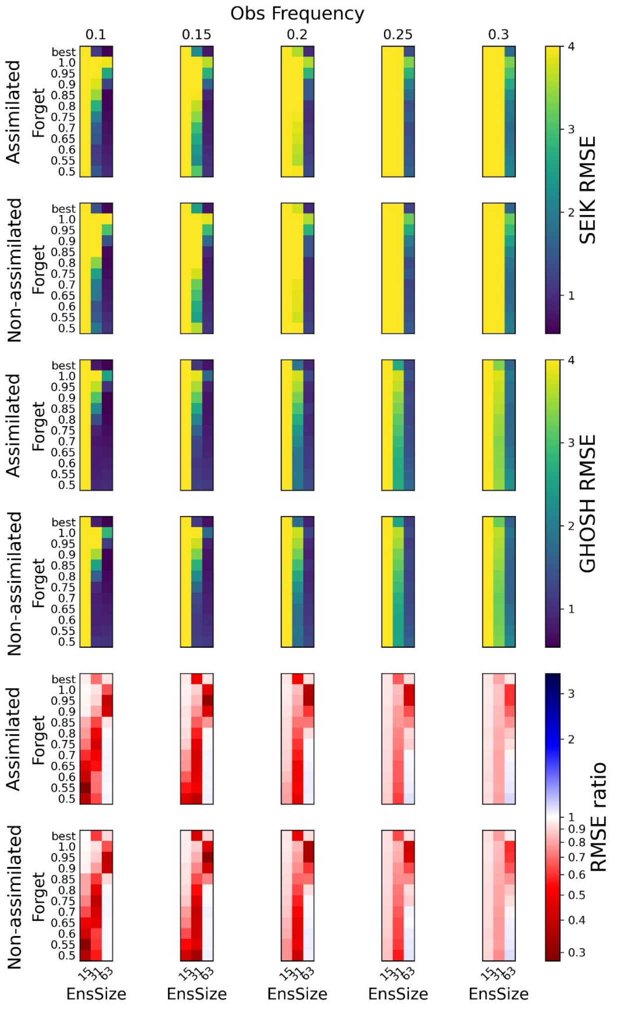
Twin experiment: SEIK vs GHOSH

Toy model: Lorenz96 (62 variables)
Observations: odd variables only
Observation error: Gaussian noise (standard deviation is 1)
Time between observations: 0.1 to 0.3 time units
Ensemble Size: 15, 31 and 63
Inflation (forgetting factor): 0.5 to 1.0
Experiments: 400 experiments for each configuration, randomly changing truth, observations and initial conditions, for a total of 66000 tests.

Twin experiment: SEIK vs GHOSH

- GHOSH **always improves RMSE** (up to 70% reduction),
- GHOSH converges for larger intervals between observations (0.25 and 0.3)
- GHOSH is more stable and needs less inflation
- GHOSH has **no higher computational cost** than SEIK

Very similar results also with the two-scale Lorenz05 model



OGS

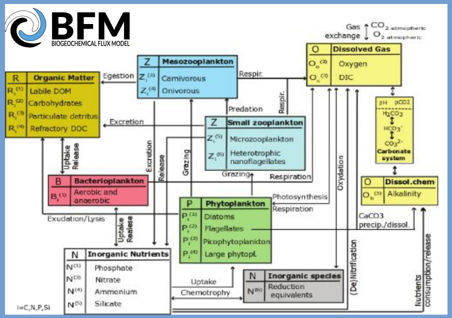
Realistic 3D test

Setup:

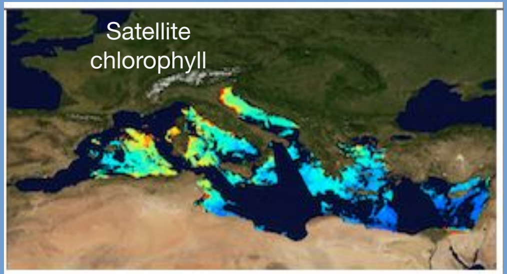
- Mediterranean Sea
- 1-year simulations
- 1/4° horizontal resolution
- 16 ensemble members
- RMSD to independent data
- 18 tests with different parameters (e.g., inflation and sampling order)

Results:

- Up to **45% RMSD reduction** in a non-assimialted variable (nitrate)



Model (BGC + transport): BFM + OGSTM



Observations: Satellite chlorophyll

OGS

Tuning

OGS

Tuning what?

Model parameters
and
initial conditions

Filter parameters, e.g.,
inflation
and
observation error

Tuning what?

Model parameters and initial conditions	Filter parameters, e.g., inflation and observation error
❌	✅

If you have a prior, leave it to filters and sampling methods

Tuning how?

We need an index to optimize.
It must be general, data-driven and it should make sense.

Likelihood
The classic

Cross-entropy
The trendy

Kullback–Leibler divergence
The nerd

Tuning how?

We need an index to optimize.
It must be general, data-driven and it should make sense.

✅ Likelihood
The classic

Cross-entropy
The trendy ✅

Kullback–Leibler divergence
The nerd ✅

Yes, they are all equivalent!

The auto-tuning minimization

Recall: $\mathbf{P} \approx \mathbf{L}\,\mathbf{A}\,\mathbf{L}^T$.

$\mathbf{P}_{like} = \mathbf{P}_\mathbf{H} + \mathbf{R} = \mathbf{L}_\mathbf{H}\,\mathbf{A}\,\mathbf{L}_\mathbf{H}^T + \mathbf{R}$,
where $\mathbf{L}_\mathbf{H}$ is the projection of $\mathbf{L}$ in observation space.

Given that $\mathbf{y}$ is the observation,
$\mathbf{y}_f$ is the forecasted observation
and $\mathbf{d} = \mathbf{y}\text{-}\mathbf{y}_f$

$$Loss = |\mathbf{P}_{like}| + \mathbf{d}^T\,\mathbf{P}_{like}^{-1}\,\mathbf{d}$$

OGS

The auto-tuning minimization

Recall: $\mathbf{P} \approx \mathbf{L}\,\mathbf{A}\,\mathbf{L}^T$.

$\mathbf{P}_{like} = \mathbf{P}_\mathbf{H} + \mathbf{R} = \mathbf{L}_\mathbf{H}\,\mathbf{A}\,\mathbf{L}_\mathbf{H}^T + \mathbf{R}$,
where $\mathbf{L}_\mathbf{H}$ is the projection of $\mathbf{L}$ in observation space.

Given that $\mathbf{y}$ is the observation,
$\mathbf{y}_f$ is the forecasted observation
and $\mathbf{d} = \mathbf{y}\text{-}\mathbf{y}_f$

$$Loss = |\mathbf{P}_{like}| + \mathbf{d}^T\,\mathbf{P}_{like}^{-1}\,\mathbf{d}$$

It can be computed lightning fast by projecting in ensemble space

OGS

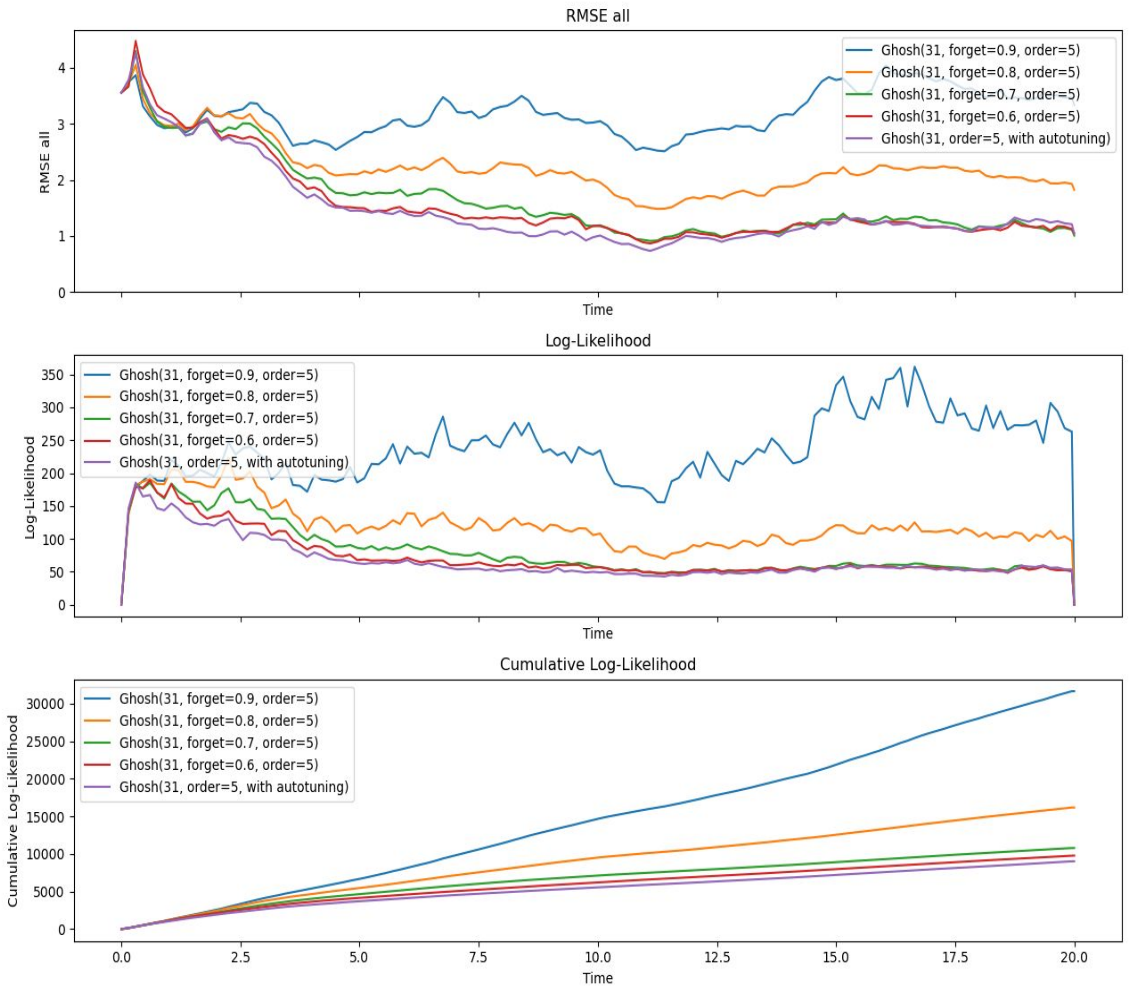
Twin experiment #1 (100-tests average)

Auto-tuning:

- forgetting factor

Results:

The filter with auto-tuning (purple) **converge faster than the best** tuned filter.



OGS

Twin experiment #2 (100-tests average)

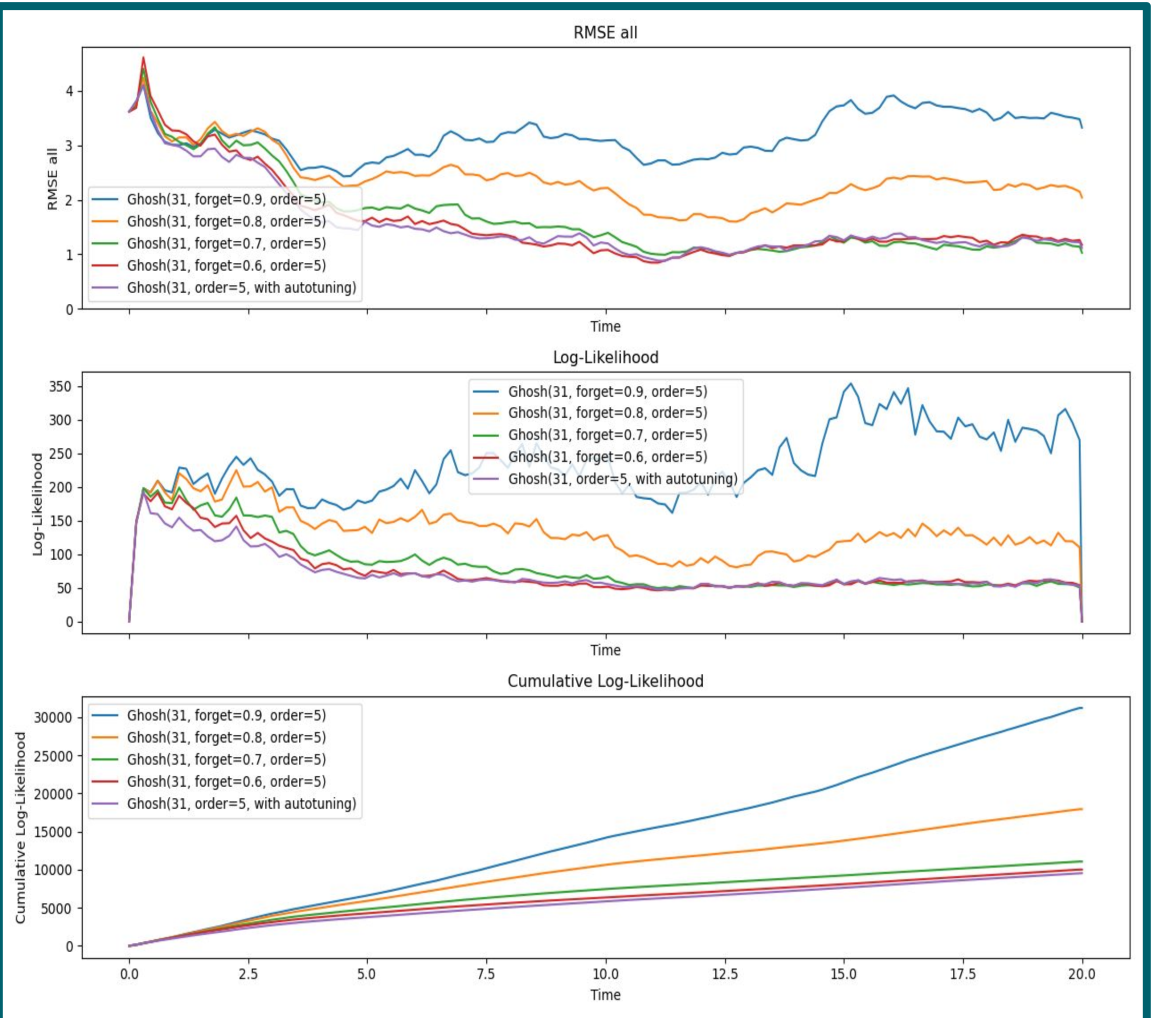
Auto-tuning:

- forgetting factor and
- observation error

Only the purple filter must guess the observation error.

Results:

The filter with auto-tuning (purple) is **as good as the best** tuned filter.



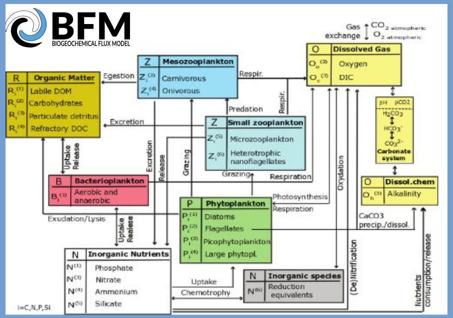
OGS

Auto-tuning 3D implementation

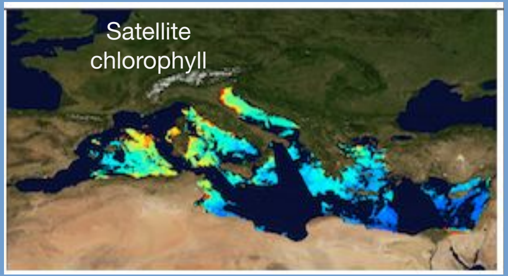
Setup:

- Mediterranean Sea
- 1-year simulations
- 1/24° horizontal resolution
- 24 ensemble members
- 3k cores x 150h =
  450k core hours per run!

Milions of core hours saved!



Model (BGC + transport):
BFM + OGSTM



Observations:
Satellite chlorophyll

Take home messages

Sampling:

- the **high-order sampling** and the **GHOSH** filter **significantly improve performance**,
- with near the **same computational cost**.

Tuning:

- the likelihood-based **auto-tuning** **saves time** (and money),
- while granting the **best performances**.

OGS

THANK YOU!

Take a look at:

"GHOSH v1.0.0: a novel Gauss-Hermite High-Order Sampling Hybrid filter for computationally efficient data assimilation in geosciences"

https://doi.org/10.5194/gmd-2023-170

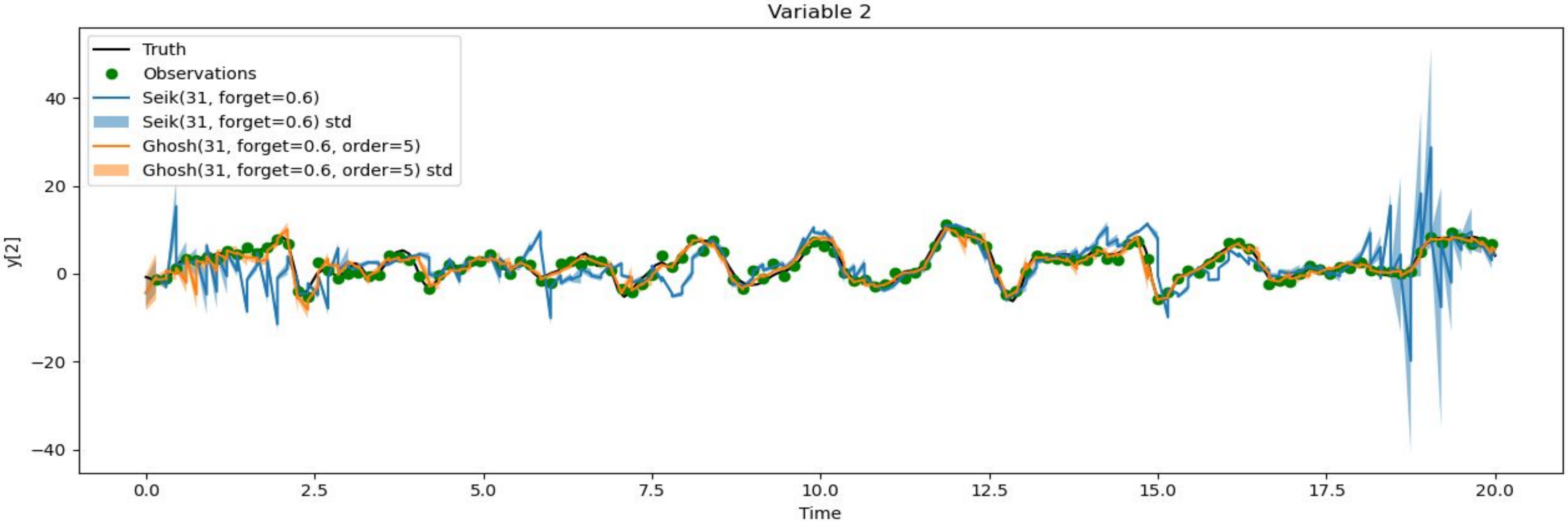
Simone Spada, Anna Teruzzi, Gianpiero Cossarini

sspada@ogs.it

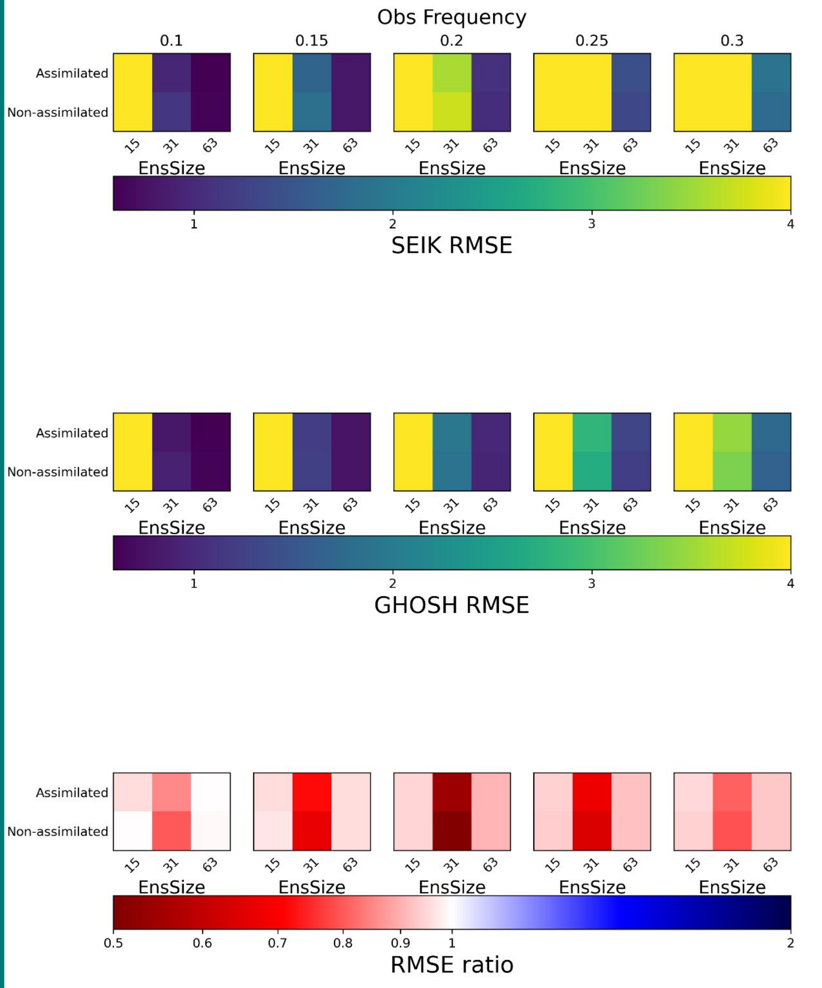
SEAMLESS

The EnKF Workshop 2024
17 – 19 June, 2024

Instabilities

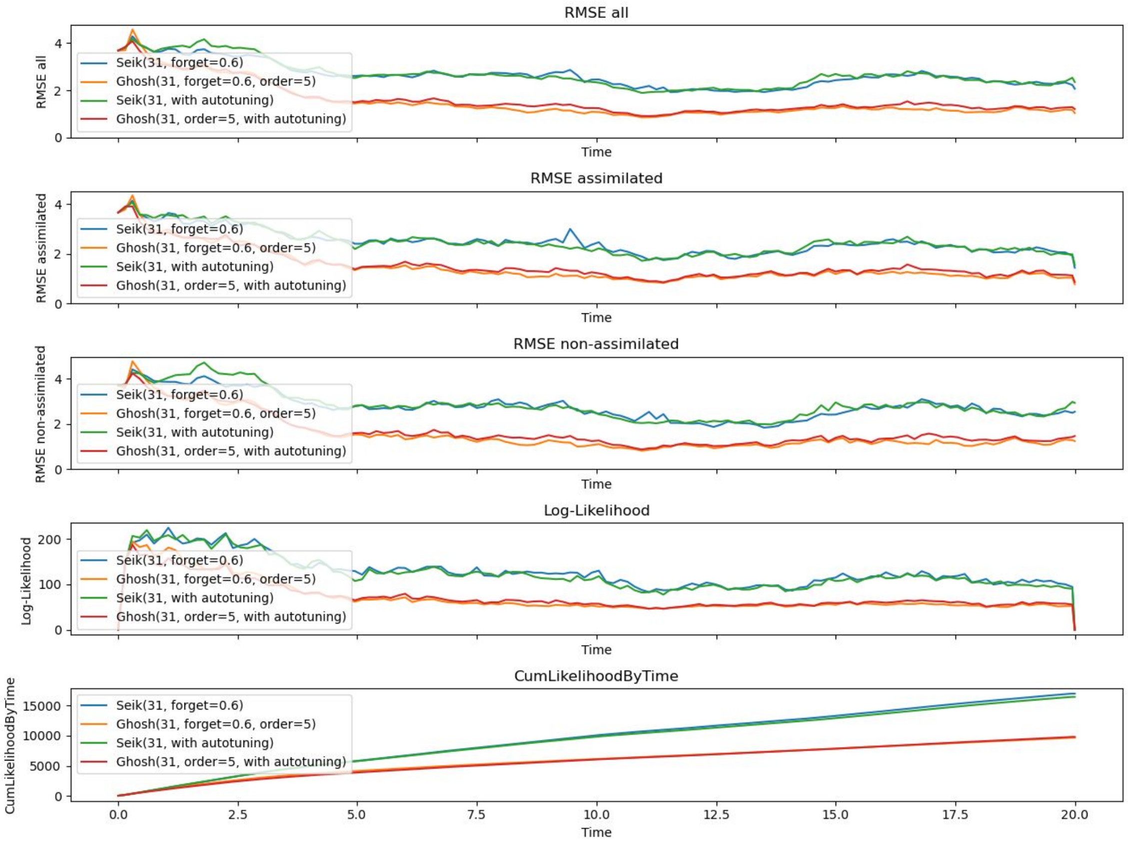


Long runs



Obs Frequency
0.1
0.15
0.2
0.25
0.3
Assimilated
Non-assimilated
15
31
63
EnsSize
1
2
3
4
SEIK RMSE
GHOSH RMSE
0.5
0.6
0.7
0.8
0.9
1
2
RMSE ratio

Auto-tuning SEIK and GHOSH



OGS