Lecture 2: Combining data assimilation and machine learning .. with application to surrogate modelling and model error

Marc Bocquet[†] Alban Farchi[†], Quentin Malartic[†],
Massimo Bonavita[‡], Patrick Laloyaux[‡], Marcin Chrust[‡],
Tobias Finn[†], Charlotte Durand[†]
and *last but not least* Julien Brajard, Alberto Carrassi, Laurent Bertino.

[†]CEREA, École des Ponts and EDF R&D, Île-De-France, France
[‡]ECMWF, Reading, United Kingdom.

Outline

1 Machine learning and the geosciences

2 Offline surrogate model learning
- With dense and perfect observations
- With sparse and noisy observations
- Hybrid models
- Resolvent or tendency correction
- Numerical experiments

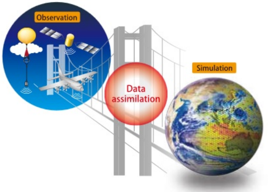
3 Online surrogate model learning
- Variational approach
- Ensemble Kalman filtering approach

4 Illustrations in the climate sciences
- Atmospheric sciences
- Sea-ice
- GHG emission retrievals

5 References

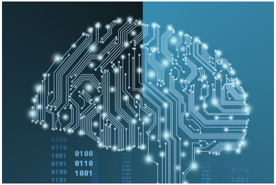
Machine learning in the geosciences

- Estimation theory–inverse problems were already key in the geosciences:
  - Sensitivity analysis,
  - Data assimilation,
  - Parameter estimation,
  - Uncertainty quantification,
  - Ensemble forecasting, etc.



Especially data assimilation (including adjoint modelling) for numerical weather prediction.

- Machine learning has started to percolate in the field of geosciences about five years ago:
  - Climate sciences,
  - Numerical weather prediction,
  - Ocean sciences,
  - Land surface and biogeochemichal processes,
  - Glaciology, sea-ice models,
  - Atmospheric chemistry, air quality, etc.



Emergence of machine learning techniques

- Why this ML tsunami?
  - New sparse representations of data that yield better and numerically affordable optimisations.
  - Relies on comprehensive deep learning libraries (Tensorflow/Keras, PyTorch/Lightening, Julia/Flux, etc.) powered by Google, Facebook, Apache, Nvidia, etc.

TensorFlow Keras flux PyTorch Lightning julia

- Why this ongoing ML hype in the geosciences and in geophysical data assimilation?
  - Huge success of deep learning (DL) in computer vision, speech recognition and AI in general. This makes it fashionable in geophysics.
  - Some of the DL models in vision, speech can be straightforwardly applied to the geosciences.
  - Forces us to reconsider difficult questions of geophysical DA (e.g., model error). Gives an alibi to reconsider those questions!
  - Above all: these libraries efficiently address one of the key issue of variational data assimilation: adjoint modelling.

What can ML bring to NWP and data assimilation?

- Advanced quality control of observations and forecasts.

- Emulate, build surrogate models for subpart of the main forecast model, for instance subgrid scale parametrisations, microphysics, convection parametrisations, etc.

- Bias correction, residual model error correction with application to forecasting and re-analysis.

- Generate tangent linear and adjoint of emulated components of the model.

- Postprocessing, downscaling: advanced and nonlinear statistical adaptation and correction, downscaling, feature detection, feature extraction.

- Improvement of existing DA schemes, especially ensemble-based methods. Substitute for the analysis, refinement and regularisation of existing DA schemes.

[Dueben et al. 2018; Reichstein et al. 2019; Bolton et al. 2019] and many others.

Outline

1. Machine learning and the geosciences

2. Offline surrogate model learning
   - With dense and perfect observations
   - With sparse and noisy observations
   - Hybrid models
   - Resolvent or tendency correction
   - Numerical experiments

3. Online surrogate model learning
   - Variational approach
   - Ensemble Kalman filtering approach

4. Illustrations in the climate sciences
   - Atmospheric sciences
   - Sea-ice
   - GHG emission retrievals

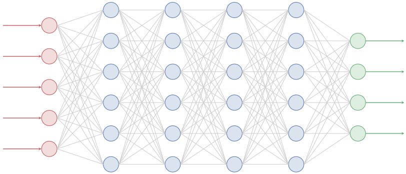
5. References

Machine learning for the geosciences with dense and perfect observations

- A typical (supervised) machine learning problem: given observations $\mathbf{y}_k$ of a system, derive a *surrogate model* of that system.

$$\mathcal{J}(\mathbf{p}) = \sum_{k=1}^{N_\mathrm{t}} \left\| \mathbf{y}_{k+1} - \mathcal{M}(\mathbf{p}, \mathbf{y}_k) \right\|^2 .$$

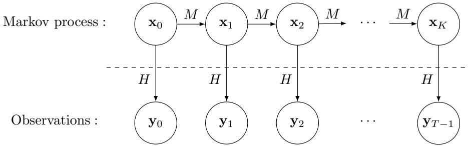
- $\mathcal{M}$ depends on a *set of coefficients* $\mathbf{p}$ (*e.g.*, the weights and biases of a neural network).



- This requires dense and perfect observations of the system.

- In the geosciences, observations are usually *sparse* and *noisy*: we need *data assimilation*!

Traditional Bayesian approach to data assimilation



▶ Bayesian justification of the weak-constraint 4D-Var

Application of Bayes' rule over a time window $[t_0, t_K]$ with batches of observations $\mathbf{y}_k$ at each time step t_k. Define $\mathbf{x}_{0:K} = \mathbf{x}_0, \ldots, \mathbf{x}_K$ and $\mathbf{y}_{0:K} = \mathbf{y}_0, \ldots, \mathbf{y}_K$.
The most general conditional pdf of interest is $p(\mathbf{x}_{0:K}|\mathbf{y}_{0:K})$ and reads:

$$p(\mathbf{x}_{0:K}|\mathbf{y}_{0:K}) \propto p(\mathbf{y}_{0:K}|\mathbf{x}_{0:K})p(\mathbf{x}_{0:K}).$$

Assuming that the observation errors are Gaussian and uncorrelated in time, with error covariance matrices $\mathbf{R}_0, \ldots, \mathbf{R}_K$, so that:

$$p(\mathbf{y}_{0:K}|\mathbf{x}_{0:K}) = \prod_{k=0}^{K} p(\mathbf{y}_k|\mathbf{x}_k) \propto \exp\left(-\frac{1}{2}\sum_{k=0}^{K} \|\mathbf{y}_k - H_k(\mathbf{x}_k)\|^2_{\mathbf{R}_k^{-1}}\right).$$

Next, we assume that the prior pdf $p(\mathbf{x}_{0:K})$ is Markovian, i.e. the state $\mathbf{x}_k$ conditional on the previous state $\mathbf{x}_{k-1}$ does not depend on all other previous past states:

$$p(\mathbf{x}_{0:K}) = p(\mathbf{x}_0)\prod_{k=1}^{K} p(\mathbf{x}_k|\mathbf{x}_{0:k-1}) = p(\mathbf{x}_0)\prod_{k=1}^{K} p(\mathbf{x}_k|\mathbf{x}_{k-1}).$$

Traditional Bayesian approach to data assimilation

▶ Bayesian justification of the weak-constraint 4D-Var

Now, we assume Gaussian statistics for the model error which are uncorrelated in time, with zero bias and error covariance matrices $\mathbf{Q}_1, \ldots, \mathbf{Q}_K$ so that:

$$p(\mathbf{x}_{0:K}) \propto p(\mathbf{x}_0) \exp\left(-\frac{1}{2}\sum_{k=1}^{K}\|\mathbf{x}_k - M_k(\mathbf{x}_{k-1})\|^2_{\mathbf{Q}_k^{-1}}\right).$$

We can assemble the likelihood and prior pieces to obtain the cost function associated to the conditional pdf $p(\mathbf{x}_{0:K}|\mathbf{y}_{0:K})$:

$$\mathcal{J}(\mathbf{x}_{0:K}) = -\ln\, p(\mathbf{x}_{0:K}|\mathbf{y}_{0:K}) \tag{1}$$

$$= -\ln p(\mathbf{x}_0) + \frac{1}{2}\sum_{k=0}^{K}\|\mathbf{y}_k - H_k(\mathbf{x}_k)\|^2_{\mathbf{R}_k^{-1}} + \frac{1}{2}\sum_{k=1}^{K}\|\mathbf{x}_k - M_k(\mathbf{x}_{k-1})\|^2_{\mathbf{Q}_k^{-1}} \tag{2}$$

Unsurprisingly, this is the cost function of the weak-constraint 4D-Var. The associated statistical assumptions explicitly assume that the model is flawed.

Bayesian inference of state trajectory and model

▶ Bayesian analysis with model parameters

We can piggyback on the previous Bayesian analysis, but now adding the model parameter vector $\mathbf{p}$:

$$p(\mathbf{x}_{0:K},\mathbf{p}|\mathbf{y}_{0:K}) \propto p(\mathbf{y}_{0:K}|\mathbf{x}_{0:K},\mathbf{p})p(\mathbf{x}_{0:K},\mathbf{p}) \propto p(\mathbf{y}_{0:K}|\mathbf{x}_{0:K},\mathbf{p})p(\mathbf{x}_{0:K}|\mathbf{p})p(\mathbf{p}),$$

which requires to introduce a prior pdf $p(\mathbf{p})$ on the parameters. In the language of Bayesian statistics, this is called a hierarchical decomposition of the conditional pdf.

As a consequence, the cost function for the state and model parameters problem is

$$\begin{aligned}\mathcal{J}(\mathbf{x}_{0:K},\mathbf{p}) &= -\ln\, p(\mathbf{x}_{0:K},\mathbf{p}|\mathbf{y}_{0:K}) \\ &= -\ln p(\mathbf{x}_0) + \frac{1}{2}\sum_{k=0}^{K}\|\mathbf{y}_k - H_k(\mathbf{x}_k)\|^2_{\mathbf{R}_k^{-1}} + \frac{1}{2}\sum_{k=1}^{K}\|\mathbf{x}_k - M_k\left(\mathbf{p},\mathbf{x}_{k-1}\right)\|^2_{\mathbf{Q}_k^{-1}} \\ &\quad - \ln p(\mathbf{p}).\end{aligned}$$

This cost function is again similar to the weak-constraint 4D-var, but (i) $\mathbf{p}$ is now part of the control variables, and (ii) there is a background term on $\mathbf{p}$ that may or may not play a role depending on the importance of the data set.

[Hsieh et al. 1998; Abarbanel et al. 2018; Bocquet et al. 2019]

Connecting data assimilation and machine learning

▶ Machine learning limit

Let us assume that the physical system is fully and directly observed, i.e. $\mathbf{H}_k \equiv \mathbf{I}$, and that the observation errors tend to zero, i.e. $\mathbf{R}_k \to \mathbf{0}$. Then the observation term in the cost function is completely frozen and imposes that $\mathbf{x}_k \simeq \mathbf{y}_k$, so that, in this limit, $\mathcal{J}(\mathbf{x}_{0:K}, \mathbf{p})$ becomes

$$\mathcal{J}(\mathbf{p}) = \frac{1}{2} \sum_{k=0}^{K} \left\| \mathbf{y}_k - M_k\left(\mathbf{p}, \mathbf{y}_{k-1}\right) \right\|^2_{\mathbf{Q}_k^{-1}} - \ln p(\mathbf{y}_0, \mathbf{p}).$$

This coincides with the tyical machine learning loss function with $\mathbf{Q}_k \equiv \mathbf{I}$.

[Bocquet et al. 2019; Bocquet et al. 2020]

Data assimilation and machine learning unification: Summary

- Bayesian view on state and model estimation:

$$p(\mathbf{p},\mathbf{Q}_{1:K},\mathbf{x}_{0:K}|\mathbf{y}_{0:K},\mathbf{R}_{0:K}) = \frac{p(\mathbf{y}_{0:K}|\mathbf{x}_{0:K},\mathbf{p},\mathbf{Q}_{1:K},\mathbf{R}_{0:K})p(\mathbf{x}_{0:K}|\mathbf{p},\mathbf{Q}_{1:K})p(\mathbf{p},\mathbf{Q}_{1:K})}{p(\mathbf{y}_{0:K},\mathbf{R}_{0:K})}.$$

- Data assimilation cost function assuming Gaussian errors and Markovian dynamics:

$$\begin{aligned}\mathcal{J}(\mathbf{p},\mathbf{x}_{0:K},\mathbf{Q}_{1:K}) =& \frac{1}{2}\sum_{k=0}^{K}\left\{\|\mathbf{y}_k - H_k(\mathbf{x}_k)\|^2_{\mathbf{R}_k^{-1}} + \ln|\mathbf{R}_k|\right\}\\ &+ \frac{1}{2}\sum_{k=1}^{K}\left\{\|\mathbf{x}_k - M_k(\mathbf{p},\mathbf{x}_{k-1})\|^2_{\mathbf{Q}_k^{-1}} + \ln|\mathbf{Q}_k|\right\}\\ &- \ln p(\mathbf{x}_0,\mathbf{p},\mathbf{Q}_{1:K}).\end{aligned}$$

$\longrightarrow$ Allows to rigorously handle partial and noisy observations.

- Typical machine learning cost function with $H_k \equiv \mathbf{I}_k$ in the limit $\mathbf{R}_k \longrightarrow \mathbf{0}$:

$$\mathcal{J}(\mathbf{p}) \approx \frac{1}{2}\sum_{k=1}^{K}\|\mathbf{y}_k - M_k(\mathbf{p},\mathbf{y}_{k-1})\|^2_{\mathbf{Q}_k^{-1}} - \ln p(\mathbf{y}_0,\mathbf{p}).$$

Bayesian analysis of the joint problem: Assuming $\mathbf{Q}_{1:K}$ is known

▶ If the $\mathbf{Q}_{1:K}$ are known, we look for the minima of

$$\mathcal{J}(\mathbf{p}, \mathbf{x}_{0:K}|\mathbf{Q}_{1:K}) = -\ln p(\mathbf{p}, \mathbf{x}_{0:K}|\mathbf{y}_{0:K}, \mathbf{R}_{0:K}, \mathbf{Q}_{1:K}).$$

▶ Numerical solution through optimization

(1) $\mathcal{J}(\mathbf{p}, \mathbf{x}_{0:K}|\mathbf{Q}_{1:K})$ can be optimized using a full variational approach:

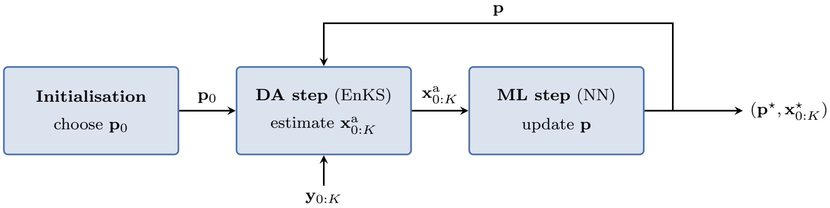
- ▶ In [Bocquet et al. 2019], $\mathcal{J}(\mathbf{p}, \mathbf{x}_{0:K}|\mathbf{Q}_{1:K})$ is minimized using a full weak-constraint 4D-Var where both $\mathbf{x}_{0:K}$ and $\mathbf{p}$ are control variables.

Bayesian analysis of the joint problem: Assuming $\mathbf{Q}_{1:K}$ is known

(2) $\mathcal{J}(\mathbf{p}, \mathbf{x}_{0:K}|\mathbf{Q}_{1:K})$ is minimized using a coordinate descent:

- using a weak constraint 4D-Var for $\mathbf{x}_{0:K}$ and a variational subproblem for $\mathbf{p}$ [Bocquet et al. 2019].
- using a (higher-dimensional) strong constraint 4D-Var for $\mathbf{x}_{0:K}$ and a variational subproblem for $\mathbf{p}$ [Bocquet et al. 2019].
- using an EnKF/EnKS for $\mathbf{x}_{0:K}$ and a variational subproblem for $\mathbf{p}$ [Brajard et al. 2020; Bocquet et al. 2020].

$\longrightarrow$ Combine data assimilation and machine learning techniques in a coordinate descent



Bayesian analysis of the marginal problem: Assuming $\mathbf{Q}_{1:K}$ is unknown

- Focusing on the marginal $p(\mathbf{p}, \mathbf{Q}_{1:K}|\mathbf{y}_{0:K}, \mathbf{R}_{0:K})$:

$$p(\mathbf{p}, \mathbf{Q}_{1:K}|\mathbf{y}_{0:K}, \mathbf{R}_{0:K}) = \int \mathrm{d}\mathbf{x}_{0:K}\, p(\mathbf{p}, \mathbf{Q}_{1:K}, \mathbf{x}_{0:K}|\mathbf{y}_{0:K}, \mathbf{R}_{0:K})$$

yields the loss function

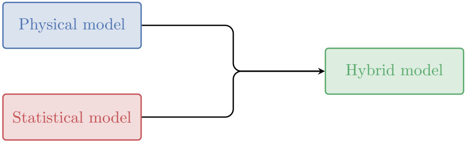
$$\mathcal{J}(\mathbf{p}, \mathbf{Q}_{1:K}) = -\ln p(\mathbf{p}, \mathbf{Q}_{1:K}|\mathbf{y}_{0:K}, \mathbf{R}_{0:K}).$$

- A MAP solution (minimum of $\mathcal{J}$) is provided by the EM algorithm. Applying it for the reconstruction of a dynamical system has been suggested in [Ghahramani et al. 1999], using an extended Kalman smoother, or for the estimation of subgrid stochastic processes in [Pulido et al. 2018] using an ensemble Kalman smoother (EnKS).

- An EM solution based on the EnKS has been suggested by [Nguyen et al. 2019] and variants of the algorithms have been successfully implemented and tested by [Bocquet et al. 2020].

Machine learning for prediction: learning model error

- Even though geophysical models are not perfect, they are sometimes already quite good (especially in NWP)!
- Instead of building a surrogate model from scratch, we use the DA-ML framework to build a *hybrid* surrogate model, with a physical part and a statistical part:[1]



- In practice, the statistical part is trained to learn the *error* of the physical model.
- In general, it is easier to train a correction model than a full model: we can use *smaller NNs* and *less training data*.
- But prone to *initialisation shocks.*

[1] [Farchi et al. 2021b; Brajard et al. 2021].

Typical architecture of a physical model

- The model is defined by a set of ODEs or PDEs which define the *tendencies*:

$$\frac{\partial \mathbf{x}}{\partial t} = \phi(\mathbf{x}). \tag{3}$$

- A numerical scheme is used to integrate the tendencies from time t to $t + \delta t$ (*e.g.*, Runge–Kutta):

$$\mathbf{x}(t + \delta t) = \mathcal{F}\Big(\mathbf{x}(t)\Big). \tag{4}$$

- Several integration steps are composed to define the *resolvent* from one analysis (or window) to the next:

$$\mathcal{M} : \mathbf{x}_k \mapsto \mathbf{x}_{k+1} = \mathcal{F} \circ \cdots \circ \mathcal{F}(\mathbf{x}_k). \tag{5}$$

Resolvent correction

- Physical model and of NN are *independent*.
- NN must predict the analysis increments.
- Resulting hybrid model not suited for short-term predictions.
- For DA, need to assume *linear growth of errors in time* to rescale correction.

Tendency correction

- Physical model and NN are *entangled*.
- Need TL of physical model to train NN!
- Resulting hybrid model suited for any prediction.
- Can be used as is for DA.

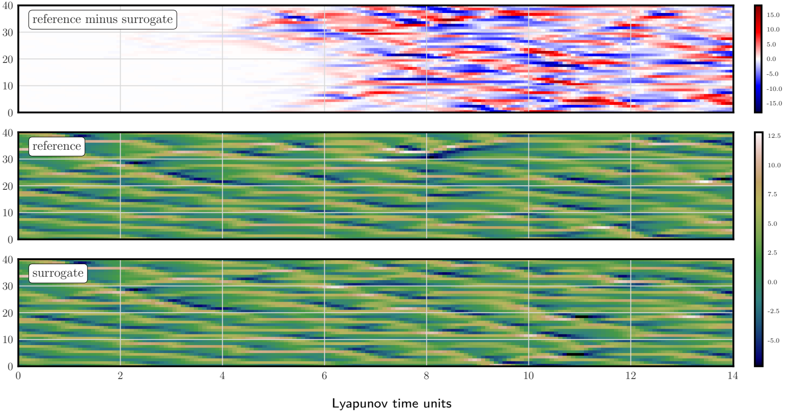
Almost identifiable model and perfect observations

▶ Inferring the dynamics from dense & noiseless observations of a non-identifiable model

The Lorenz 96 model (40 variables)

$$\frac{\mathrm{d}x_n}{\mathrm{d}t} = (x_{n+1} - x_{n-2})x_{n-1} - x_n + F,$$

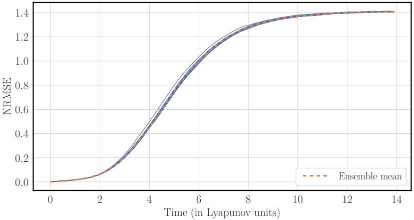
Surrogate model based on an RK2 scheme.

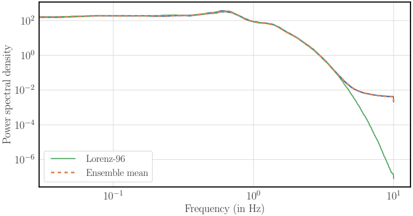


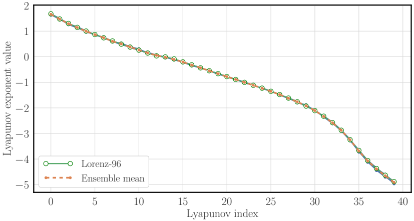
Almost identifiable model and imperfect observations

- Very good reconstruction of the long-term properties of the model (L96 model).

  - Approximate scheme
  - Fully observed
  - Significantly noisy observations $\mathbf{R} = \mathbf{I}$
  - Long window $K = 5000$, $\Delta t = 0.05$
  - EnKS with $L = 4$
  - 30 EM iterations





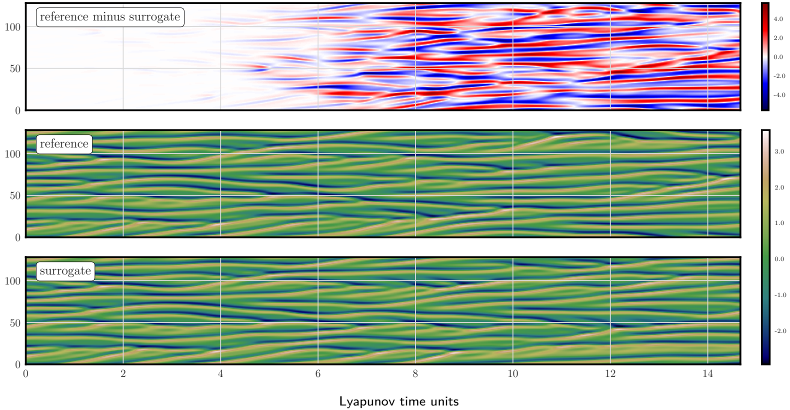


Not so identifiable model and perfect observations

▶ Inferring the dynamics from dense & noiseless observations of a non-identifiable model

The Kuramoto-Sivashinski (KS) model (continuous, 128 variables).

$$\frac{\partial u}{\partial t} = -u\frac{\partial u}{\partial x} - \frac{\partial^2 u}{\partial x^2} - \frac{\partial^4 u}{\partial x^4},$$

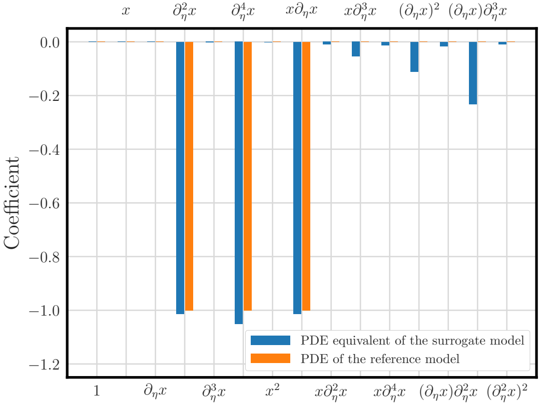


Not so identifiable model and perfect observations

▶ Inferring the dynamics from dense & noiseless observations of a non-identifiable model

The Kuramoto-Sivashinski (KS) model (continuous, 128 variables).

$$\frac{\partial u}{\partial t} = -u\frac{\partial u}{\partial x} - \frac{\partial^2 u}{\partial x^2} - \frac{\partial^4 u}{\partial x^4},$$

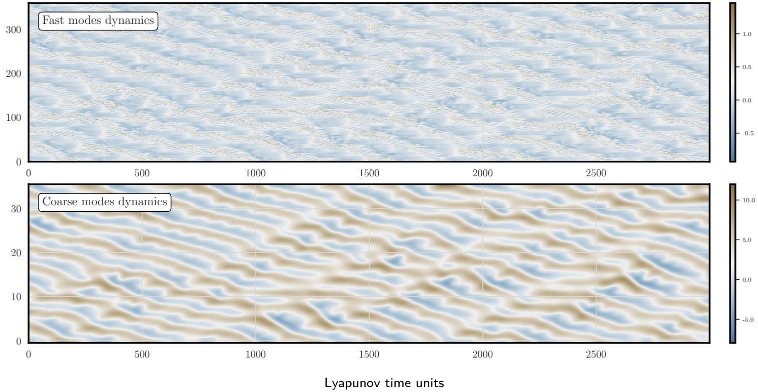


Two-scale Lorenz model (L05III)

- The two-scale Lorenz model (L05III) model: 36 slow & 360 fast variables, with equations:

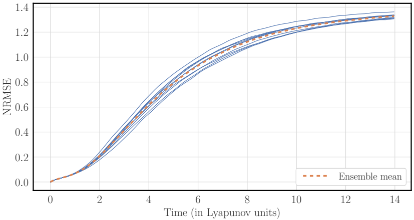
$$\frac{\mathrm{d}x_n}{\mathrm{d}t} = \psi_n^+(\mathbf{x}) + F - h\frac{c}{b}\sum_{m=0}^{9} u_{m+10n},$$

$$\frac{\mathrm{d}u_m}{\mathrm{d}t} = \frac{c}{b}\psi_m^-(b\mathbf{u}) + h\frac{c}{b}x_{m/10}, \quad \text{with} \quad \psi_n^\pm(\mathbf{x}) = x_{n\mp1}(x_{n\pm1} - x_{n\mp2}) - x_n,$$

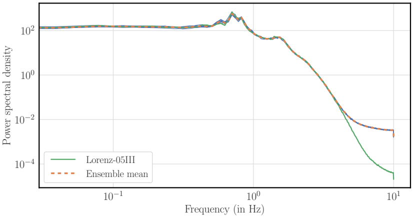


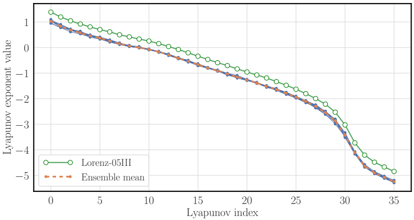
Non-identifiable model and imperfect observations

- Good reconstruction of the long-term properties of the model (L05III model).



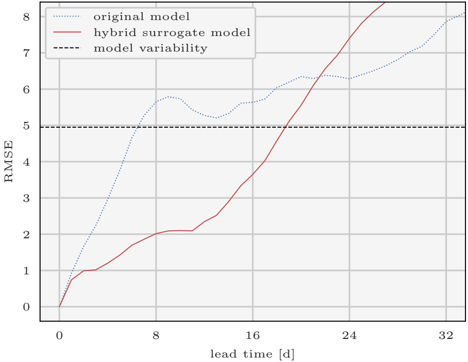
- Approximate scheme
- Observation of the coarse modes only
- Significantly noisy observations $\mathbf{R} = \mathbf{I}$
- Long window $K = 5000$, $\Delta t = 0.05$
- EnKS with $L = 4$
- 30 EM iterations





ECMWF QG model: hybrid surrogate; resolvent or tendencies?

- The non-corrected model is a perturbed ECMWF OOPS quasi-geostrophic model.
- Noisy observations are assimilated using strong-constrained *4D-Var*.
- Simple *CNNs* are trained using the 4D-Var analysis.



Data assimilation score

Model	Analysis RMSE
No correction	0.31
Resolvent correction	0.28
Tendency correction	0.24
True model	0.22

- The tendencies corr. is *more accurate* than the resolvent corr., with smaller NNs and less training data.
- The tendencies corr. benefits from the *interaction* with the physical model.
- The resolvent corr. is highly penalised (in DA) by the assumption of linear growth of errors.

Outline

1 Machine learning and the geosciences

2 Offline surrogate model learning
- With dense and perfect observations
- With sparse and noisy observations
- Hybrid models
- Resolvent or tendency correction
- Numerical experiments

3 Online surrogate model learning
- Variational approach
- Ensemble Kalman filtering approach

4 Illustrations in the climate sciences
- Atmospheric sciences
- Sea-ice
- GHG emission retrievals

5 References

Online model error correction

- So far, the model error has been learnt *offline*: the ML (or training) step first requires a long analysis trajectory.
- We now investigate the possibility to perform *online* learning, *i.e.* improving the correction as new observations become available.[2]
- To do this, we use the formalism of DA to estimate both the state and the NN parameters:

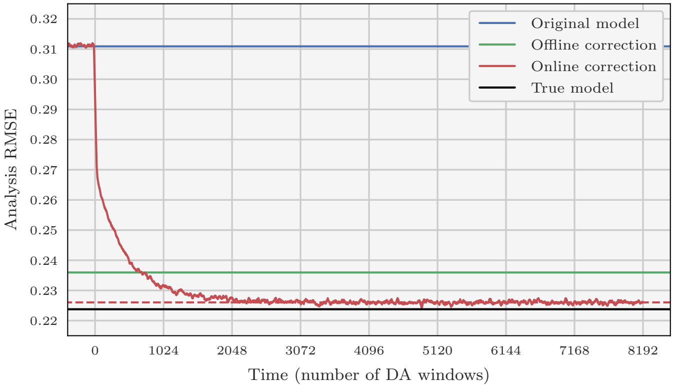
$$\mathcal{J}(\mathbf{p},\mathbf{x}) = \left\|\mathbf{x}-\mathbf{x}^{\mathrm{b}}\right\|^2_{\mathbf{B}_{\mathrm{x}}^{-1}} + \left\|\mathbf{p}-\mathbf{p}^{\mathrm{b}}\right\|^2_{\mathbf{B}_{\mathrm{p}}^{-1}} + \sum_{k=0}^{L}\left\|\mathbf{y}_k - \mathcal{H}_k\circ\mathcal{M}^k(\mathbf{p},\mathbf{x})\right\|^2_{\mathbf{R}_k^{-1}}.$$

- For simplicity, we have neglected potential cross-covariance between state and NN parameters in the prior.
- Information is flowing from one window to the next using the prior for the state $\mathbf{x}^{\mathrm{b}}$ and for the NN parameters $\mathbf{p}^{\mathrm{b}}$: sequential data assimilation

[2] [Farchi et al. 2021a]

Two-scale Lorenz system: online learning

- We use the tendency correction approach, with the same simple CNN as before, and still using 4D-Var.



- The online correction steadily improves the model.
- At some point, the online correction *gets more accurate* than the offline correction.
- Eventually, the improvement saturates. The analysis error is similar to that obtained with the true model!

Outline

1. Machine learning and the geosciences
2. Offline surrogate model learning
   - With dense and perfect observations
   - With sparse and noisy observations
   - Hybrid models
   - Resolvent or tendency correction
   - Numerical experiments
3. Online surrogate model learning
   - Variational approach
   - Ensemble Kalman filtering approach
4. Illustrations in the climate sciences
   - Atmospheric sciences
   - Sea-ice
   - GHG emission retrievals
5. References

Online learning with a LEnKF: Augmented state vector

▶ So far, learning was based on variational techniques using all available data. Can one design a sequential (online) ensemble scheme that progressively updates both the state and the model as data are collected?

▶ In the following, we make the assumptions:
(i) *autonomous* and *local* dynamics,
(ii) *homogeneous* dynamics or *heterogeneous* dynamics, or *mixed* dynamics.

▶ Parameters of the model:

$$\mathbf{p} \in \mathbb{R}^{N_\mathrm{p}} \text{ [global parameters]}, \quad \mathbf{q} \in \mathbb{R}^{N_\mathrm{q}} \text{ [local parameters]}.$$

▶ Augmented state formalism [Jazwinski 1970; Ruiz et al. 2013]:

$$\mathbf{z} = \begin{bmatrix} \mathbf{x} \\ \mathbf{p} \\ \mathbf{q} \end{bmatrix} \in \mathbb{R}^{N_\mathrm{z}}, \qquad \text{with} \quad N_\mathrm{z} = N_\mathrm{x} + N_\mathrm{p} + N_\mathrm{q}.$$

▶ Just a more ambitious parameter estimation problem!?
Yes! But we have to fill in several critical gaps of the parameter-estimation-via-EnKF literature.

Online learning with a LEnKF: difficulties

- With high-dimensional geophysical models, the use of the EnKF requires localisation.
- However, localisation in the state / local–parameter / global–parameter space is tricky!
- The assimilation of nonlocal observations radiances requires LEnSRF based on covariance localisation rather than local domains: this is even trickier!
- Ideally, one should increase the ensemble size by one for each global parameter: a challenge with deep learning!

Table: Summary of the EnKF-ML family of algorithms

Inference problem	Dom. Local. local obs. only	Cov. Local. numerically costly	Dom. + Cov. Local.
State	LETKF [Hunt et al. 2007]	LEnSRF [Whitaker et al. 2002]	L^2EnSRF [Farchi et al. 2019]
State + global param.	LETKF-ML [Bocquet et al. 2021] new implementation[3]	LEnSRF-ML [Bocquet et al. 2021] new implementation	L^2EnSRF-ML not discussed
State + global & local param.	LETKF-HML new algorithm	LEnSRF-HML new algorithm	L^2EnSRF-HML new algorithm

[3] new implementations and new algorithms: [Malartic et al. 2022a]

Outline

1. Machine learning and the geosciences

2. Offline surrogate model learning
   - With dense and perfect observations
   - With sparse and noisy observations
   - Hybrid models
   - Resolvent or tendency correction
   - Numerical experiments

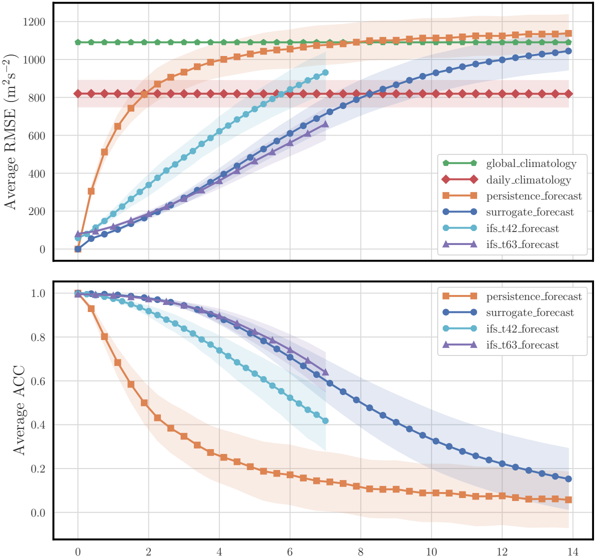
3. Online surrogate model learning
   - Variational approach
   - Ensemble Kalman filtering approach

4. Illustrations in the climate sciences
   - Atmospheric sciences
   - Sea-ice
   - GHG emission retrievals

5. References

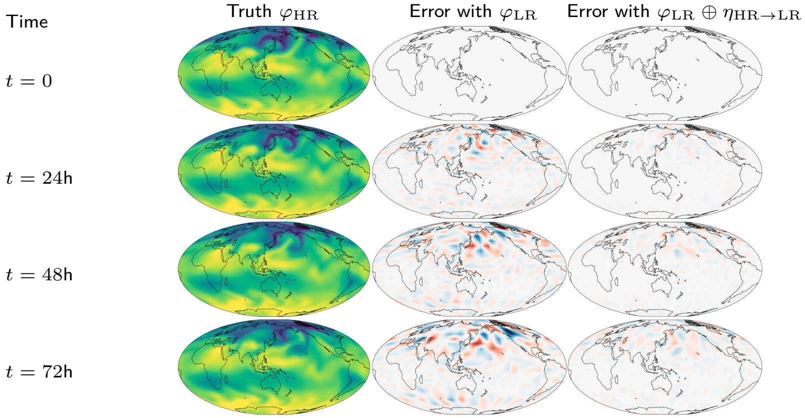
Learning a purely data-driven meterological model from ERA-5 reanalysis

▶ True model: A selection of ERA-5 fields in 1979-2018 at 5.625°.[4] ▶ DL model: Residual NN at the same resolution. ▶ Forecast skill score of the geopotential at 500hPa as a function of the forecast lead time.[5]



Example of loop order–1 and $3/2$, with an hybrid model

▶ Marshall-Molteni[6] 3-layer intermediate QG model: Learning subgrid scale parametrisation at *loop order–1* to perform more accurate forecasts at low resolution (LR) from high resolution simulations (HR).[7]

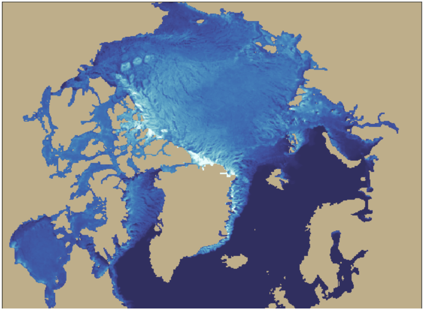


▶ $\varphi_{\mathrm{LR}} \oplus \eta_{\mathrm{HR}\to\mathrm{LR}}$ has also successfully been tested with DA, hence at *loop order–$3/2$*.

[6] [Marshall et al. 1993]
[7] [Malartic et al. 2022b]

Learning dynamics of sea-ice using neural networks

2018-01-01 03:00:00



Complex dynamics in sea-ice:

- Multifractality
- Anisotropy
- Stochasticity
- (mildly) chaotic

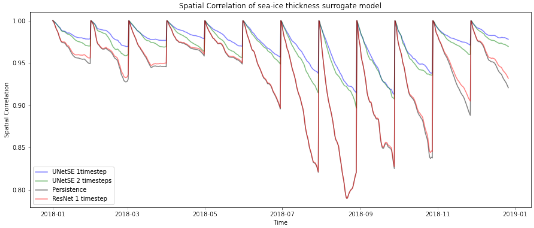
Two neural network types:

- Unet (multiscale approach)
- ResNet (residual neural network)

With partial convolutions and SE blocks.

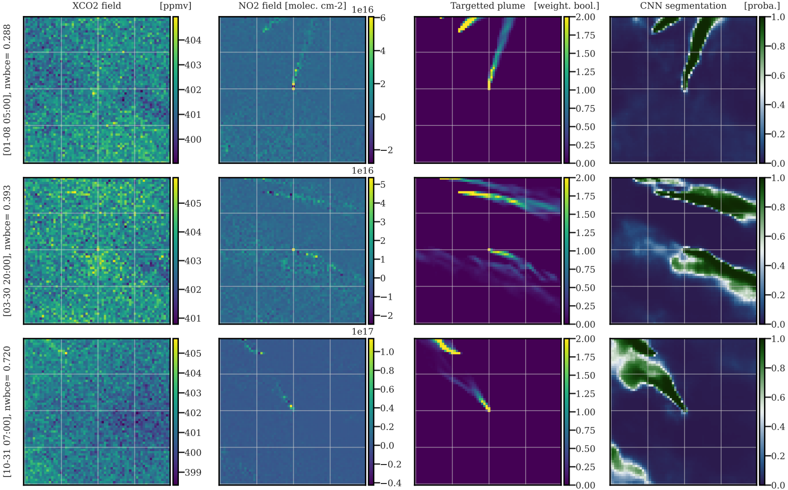
Inputs: sea-ice thickness from NeXtSIM + ERA5
Forcings: 10m air velocity, 2m air temperature and sea surface temperature
For several past timesteps
Outputs: 12h sea-ice thickness evolution

SCHMIDT FUTURES



Learning the emissions of urban plumes of greenhouse gases

- From the segmentation a GHG plume image to the inversion of the associated emission.



[Dumont Le Brazidec et al. 2022; Dumont Le Brazidec et al. 2023].

References I

[1] H. D. I. Abarbanel, P. J. Rozdeba, and S. Shirman. "Machine Learning: Deepest Learning as Statistical Data Assimilation Problems". In: *Neural Computation* 30 (2018), pp. 2025–2055.

[2] A. Aksoy, F. Zhang, and J. Nielsen-Gammon. "Ensemble-based simultaneous state and parameter estimation in a two-dimensional sea-breeze model". In: *Mon. Wea. Rev.* 134 (2006), pp. 2951–2969.

[3] M. Bocquet, A. Farchi, and Q. Malartic. "Online learning of both state and dynamics using ensemble Kalman filters". In: *Foundations of Data Science* 3 (2021), pp. 305–330.

[4] M. Bocquet et al. "Bayesian inference of chaotic dynamics by merging data assimilation, machine learning and expectation-maximization". In: *Foundations of Data Science* 2 (2020), pp. 55–80.

[5] M. Bocquet et al. "Data assimilation as a learning tool to infer ordinary differential equation representations of dynamical models". In: *Nonlin. Processes Geophys.* 26 (2019), pp. 143–162.

[6] M. Bocquet et al. "Higher-order correction to deep-learning surrogate model of meteorology from data assimilation". In: *In preparation* (2022).

[7] T. Bolton and L. Zanna. "Applications of Deep Learning to Ocean Data Inference and Subgrid Parameterization". In: *J. Adv. Model. Earth Syst.* 11 (2019), pp. 376–399.

[8] J. Brajard et al. "Combining data assimilation and machine learning to emulate a dynamical model from sparse and noisy observations: a case study with the Lorenz 96 model". In: *J. Comput. Sci.* 44 (2020), p. 101171.

[9] J. Brajard et al. "Combining data assimilation and machine learning to infer unresolved scale parametrisation". In: *Phil. Trans. R. Soc. A* 379 (2021), p. 20200086.

[10] P. D. Dueben and P. Bauer. "Challenges and design choices for global weather and climate models based on machine learning". In: *Geosci. Model Dev.* 11 (2018), pp. 3999–4009.

[11] J. Dumont Le Brazidec et al. "Deep learning applied to power plant emissions quantification using XCO_2 synthetic satellite images". In: *Geosci. Model Dev. Discuss.* 2023 (2023), pp. –.

[12] J. Dumont Le Brazidec et al. "Segmentation of XCO2 images with deep learning: application to synthetic plumes from cities and power plants". In: *Geosci. Model Dev. Discuss.* 2022 (2022), pp. 1–29.

[13] A. Farchi and M. Bocquet. "On the efficiency of covariance localisation of the ensemble Kalman filter using augmented ensembles". In: *Front. Appl. Math. Stat.* 5 (2019), p. 3.

[14] A. Farchi et al. "A comparison of combined data assimilation and machine learning methods for offline and online model error correction". In: *J. Comput. Sci.* 55 (2021), p. 101468.

References II

[15] A. Farchi et al. "Using machine learning to correct model error in data assimilation and forecast applications". In: *Q. J. R. Meteorol. Soc.* 147 (2021), pp. 3067–3084.

[16] Z. Ghahramani and S. T. Roweis. "Learning nonlinear dynamical systems using an EM algorithm". In: *Advances in neural information processing systems*. 1999, pp. 431–437.

[17] W. W. Hsieh and B. Tang. "Applying Neural Network Models to Prediction and Data Analysis in Meteorology and Oceanography". In: *Bull. Amer. Meteor. Soc.* 79 (1998), pp. 1855–1870.

[18] B. R. Hunt, E. J. Kostelich, and I. Szunyogh. "Efficient data assimilation for spatiotemporal chaos: A local ensemble transform Kalman filter". In: *Physica D* 230 (2007), pp. 112–126.

[19] A. H. Jazwinski. *Stochastic Processes and Filtering Theory*. Academic Press, New-York, 1970, p. 376.

[20] Q. Malartic, A. Farchi, and M. Bocquet. "Global and local parameter estimation using local ensemble Kalman filters: applications to online machine learning of chaotic dynamics". In: *Q. J. R. Meteorol. Soc.* 0 (2022). Accepted for publication, pp. 00–00.

[21] Q. Malartic et al. "Data assimilation and machine learning subgrid parametrisation in an advanced quasi-geostrophic model". In: (2022).

[22] J. Marshall and F. Molteni. "Toward a Dynamical Understanding of Planetary-Scale Flow Regimes". In: *J. Atmos. Sci.* 50 (1993), pp. 1792–1818.

[23] V. D. Nguyen et al. "EM-like Learning Chaotic Dynamics from Noisy and Partial Observations". In: *arXiv preprint arXiv:1903.10335* (2019).

[24] M. Pulido et al. "Stochastic parameterization identification using ensemble Kalman filtering combined with maximum likelihood methods". In: *Tellus A* 70 (2018), p. 1442099.

[25] S. Rasp et al. "WeatherBench: A Benchmark Data Set for Data-Driven Weather Forecasting". In: *J. Adv. Model. Earth Syst.* 12 (2020), e2020MS002203.

[26] M. Reichstein et al. "Deep learning and process understanding for data-driven Earth system science". In: *Nature* 566 (2019), pp. 195–204.

[27] Y. M. Ruckstuhl and T. Janjić. "Parameter and state estimation with ensemble Kalman filter based algorithms for convective-scale applications". In: *Q. J. R. Meteorol. Soc.* 144 (2018), pp. 826–841.

[28] J. J. Ruiz, M. Pulido, and T. Miyoshi. "Estimating model parameters with ensemble-based data assimilation: A Review". In: *J. Meteorol. Soc. Japan* 91 (2013), pp. 79–99.

[29] J. S. Whitaker and T. M. Hamill. "Ensemble Data Assimilation without Perturbed Observations". In: *Mon. Wea. Rev.* 130 (2002), pp. 1913–1924.

Online learning with a LEnKF: The problems

- We use the augmented state formalism with local ensemble Kalman filters (EnKFs): LEnSRF and LETKF, which are keys for scalability.

- Adequacy and inadequacy between the main LEnKF classes and the estimation of local and global parameters:

Table: Adequacy (green) and inadequacy (red) between LEnKF types and the estimation of local, global and mixed parameters. CL refers to covariance localisation and DL refers to domain localisation.

LEnKF type	Global parameters	Local parameters	Mixed set of parameters
LEnSRF (CL)	well suited localisation in parameter space?	suited numerically costly	unclear solution proposed here
LETKF (DL)	only approximate[8] solution proposed here	well suited	unclear solution proposed here

- Beware that nonlocal observations require CL!

[8] [Aksoy et al. 2006]

Table: Summary of the EnKF-ML family of algorithms

Inference problem	Dom. Local. local obs. only	Cov. Local. numerically costly	Dom. + Cov. Local.
State	LETKF [Hunt et al. 2007]	LEnSRF [Whitaker et al. 2002]	L^2EnSRF [Farchi et al. 2019]
State + global param.	LETKF-ML [Bocquet et al. 2021] new implementation[9]	LEnSRF-ML [Bocquet et al. 2021] new implementation	L^2EnSRF-ML not discussed
State + global & local param.	LETKF-HML new algorithm	LEnSRF-HML new algorithm	L^2EnSRF-HML new algorithm

Main results

New EnKF *update formula* and new LEnSRF/LETKF *algorithms* with parameter estimation:
global parameters $\longrightarrow$ LETKF-ML, LEnSRF-ML, L^2EnSRF-ML,
global and *local* parameters $\longrightarrow$ LETKF-HML, LEnSRF-HML, and L^2EnSRF-HML.

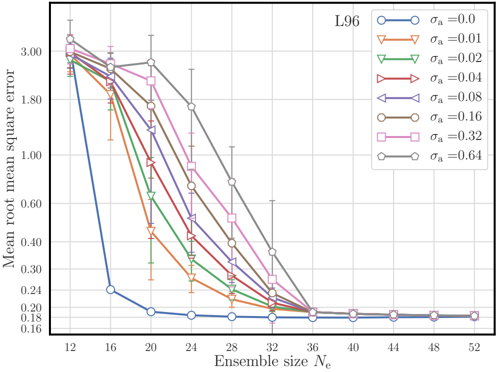
[9] new implementations and new algorithms: [Malartic et al. 2022a]

Focus on the augmented dynamics and its unstable subspace

- Augmented dynamics (model persistence or Brownian motion):

$$\begin{bmatrix} \mathbf{x}_k \\ \mathbf{p}_k \end{bmatrix} \mapsto \begin{bmatrix} \mathbf{F}^k(\mathbf{x}_k, \mathbf{p}_k) \\ \mathbf{p}_k \end{bmatrix}$$

- Assuming (i) N_0 is the dimension of the *unstable neutral subspace* of the reference dynamics, (ii) N_e is the size of the ensemble, then, in order for the augmented global EnKF (EnKF-ML) to be stable, we must have: $\boxed{N_\mathrm{e} \gtrsim N_0 + N_\mathrm{p} + 1.}$



▶ Covariance localisation in the augmented space:

$$\mathbf{B}_{\mathsf{xx}} = \boldsymbol{\rho}_{\mathsf{xx}} \circ \left[\mathbf{X}^{\mathsf{f}}_{\mathsf{x}} \left(\mathbf{X}^{\mathsf{f}}_{\mathsf{x}}\right)^{\top}\right], \qquad \mathbf{B}_{\mathsf{px}} = \boldsymbol{\rho}_{\mathsf{px}} \circ \left[\mathbf{X}^{\mathsf{f}}_{\mathsf{p}} \left(\mathbf{X}^{\mathsf{f}}_{\mathsf{x}}\right)^{\top}\right] = \mathbf{B}_{\mathsf{xp}}^{\top}, \qquad \mathbf{B}_{\mathsf{pp}} = \boldsymbol{\rho}_{\mathsf{pp}} \circ \left[\mathbf{X}^{\mathsf{f}}_{\mathsf{p}} \left(\mathbf{X}^{\mathsf{f}}_{\mathsf{p}}\right)^{\top}\right].$$

▶ The localisation matrix $\boldsymbol{\rho}_{\mathsf{xx}}$ almost certainly makes $\mathbf{B}_{\mathsf{xx}}$ positive definite.

▶ The localisation matrix $\boldsymbol{\rho}_{\mathsf{px}}$ has to be uniform with respect to space because the parameters are global. This yields[10]:

$$\boldsymbol{\rho} = \begin{bmatrix} \boldsymbol{\rho}_{\mathsf{xx}} & \mathbf{1}_{\mathsf{x}} \boldsymbol{\zeta}_{\mathsf{p}}^{\top} \\ \boldsymbol{\zeta}_{\mathsf{p}} \mathbf{1}_{\mathsf{x}}^{\top} & \boldsymbol{\rho}_{\mathsf{pp}} \end{bmatrix}, \tag{6}$$

where $\boldsymbol{\zeta}_{\mathsf{p}} \in \mathbf{R}^{N_{\mathsf{p}}}$ is a vector of tapering coefficients.

▶ The positive definitness of $\boldsymbol{\rho}$ generates constraints on $\boldsymbol{\zeta}_{\mathsf{p}}$. A sufficient condition for positive definitness of $\boldsymbol{\rho}$ is:

$$\|\boldsymbol{\zeta}_{\mathsf{p}}\| \leq \sqrt{\frac{\lambda_{\mathsf{p}}^{\min} \lambda_{\mathsf{x}}^{\min}}{N_{\mathsf{x}}}}, \tag{7}$$

where $\lambda_{\mathsf{p}}^{\min}, \lambda_{\mathsf{x}}^{\min}$ are the smallest eigenvalues of $\boldsymbol{\rho}_{\mathsf{pp}}, \boldsymbol{\rho}_{\mathsf{xx}}$, respectively.

[10] [Ruckstuhl et al. 2018; Bocquet et al. 2021; Malartic et al. 2022a]

▶ We use the LEnKF-HML on the L96i model, i.e. with unknown dynamics (global parameters) and unknown inhomogeneous forcings (40 local parameters).

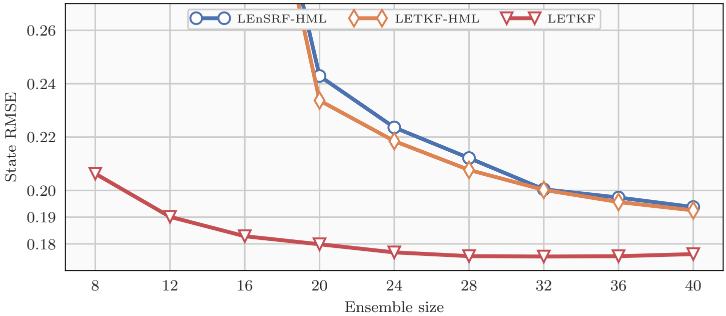


Figure: Time-averaged state analysis RMSE as a function of the ensemble size with the LEnSRF-HML (in blue) and the LETKF-HML (in yellow). For reference, the red line shows the scores obtained with the LETKF when the model is known.

▶ The mL96 model[11] is a vertical stack of $N_\mathsf{v} = 32$ coupled (atmospheric) layers, each layer being a L96 model with $N_\mathsf{h} = 40$ variables. The total state dimension is hence $N_\mathsf{x} = N_\mathsf{h} \times N_\mathsf{v} = 1280$, and the model's equations are :

$$\frac{\mathrm{d}x_{v,h}}{\mathrm{d}t} = (x_{v,h+1} - x_{v,h-2})x_{v,h-1} - x_{v,h} + F_{v,h} + \Gamma_{v+1,h} - \Gamma_{v,h}, \tag{8}$$

where $x_{v,h}$ is the h-th horizontal variable of the v-th vertical layer.

▶ The h index applies periodically in $\{1, \ldots, N_\mathsf{h}\}$. The forcing term F is inhomogeneous; it is set constant over each layer and decreases from $F_{1,h} = 8$ for the bottom layer to $F_{N_\mathsf{v},h} = 4$ for the top layer.

▶ The last two terms correspond to the vertical coupling between adjacent layers, with

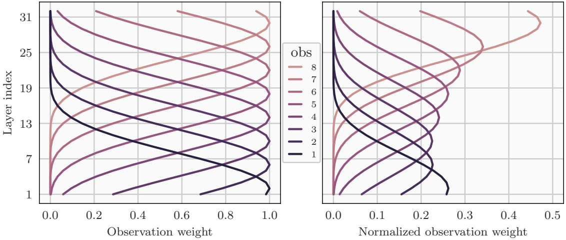
$$\Gamma_{v,h} \triangleq \begin{cases} x_{v,h} - x_{v-1,h} & \text{if } 2 \leq v \leq N_\mathsf{v}, \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

▶ We use the L^{2}EnSRF-HML on the observations of mL96, with unknown dynamics (global parameters) and unknown inhomogeneous forcings (local parameters).

[11] [Farchi et al. 2019]

Numerical illustration on the multi-layer Lorenz96 model

- Nonlocal radiance-like observations (averaging kernel for each of the 8 satellite channels without (left panel) and with (right panel) normalisation.)



- Numerical results (RMSEs):

Inference problem	N_0	Algorithm	Model	Loc.	N_e	state RMSE
1: $\mathbf{x}$	≈ 50	EnSRF	mL96		≥ 50	0.08
		L^{2}EnSRF	mL96	✓	≥ 10	0.08
2: $(\mathbf{x}, \mathbf{a}, \mathbf{f}_\mathsf{v}, \mathbf{f}_\mathsf{h})$	$\approx 50 + 88$	EnSRF-HML	sur $(\mathbf{a}, \mathbf{f}_\mathsf{v}, \mathbf{f}_\mathsf{h})$		≥ 140	0.11
		L^{2}EnSRF-HML	sur $(\mathbf{a}, \mathbf{f}_\mathsf{v}, \mathbf{f}_\mathsf{h})$	✓	50	0.12