Learning data assimilation from artificial intelligence

Are ensemble-based data assimilation methods necessary for accurate filtering?

Marc Bocquet,[*]

Alban Farchi,[*,†] Tobias Finn,[*] Charlotte Durand,[*] Sibo Cheng,[*]
Yumeng Chen,[‖] Ivo Pasmans,[‖] Alberto Carrassi[§]

[*] CEREA, ENPC, EDF R&D, Institut Polytechnique de Paris, Île-de-France, France
[‖] Department of Meteorology and National Centre for Earth Observation, University of Reading, United-Kingdom
[§] Department of Physics and Astronomy, University of Bologna, Italy
[†] ECMWF, Reading, United Kingdom

Outline

Outline

Sequential data assimilation for chaotic dynamics

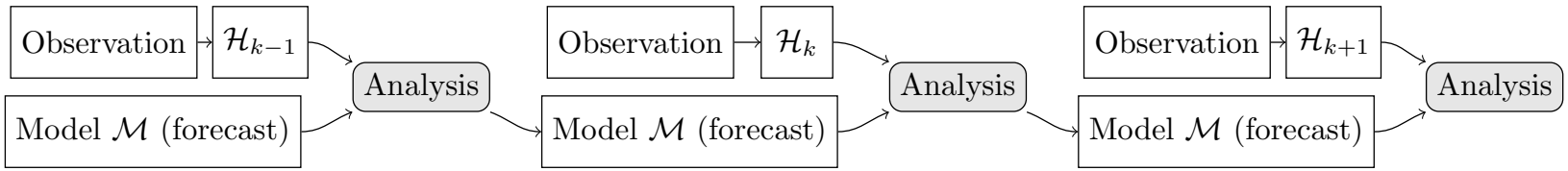
- Here, data assimilation (DA) methods are formulated from

$$\mathbf{x}_{k+1} = \mathcal{M}(\mathbf{x}_k), \tag{1a}$$

$$\mathbf{y}_k = \mathcal{H}_k(\mathbf{x}_k) + \boldsymbol{\varepsilon}_k, \qquad \boldsymbol{\varepsilon}_k \sim N(\mathbf{0}, \mathbf{R}_k), \tag{1b}$$

where $\mathcal{M}$ is the *autonomous* evolution model, $\mathbf{x}_k$ is the state vector at time τ_k, $\mathbf{y}_k$ is the observation vector, $\mathcal{H}_k$ is the observation operator, $\boldsymbol{\varepsilon}_k$ is the observation error, assumed to be additive, unbiased, white in time, and Gaussian of covariance matrix $\mathbf{R}_k$.

- DA for geofluids has to be *sequential* in time because (i) observations need to be assimilated *as they arrive* to update the state estimation, (ii) applied to *chaotic dynamics*, typical errors have an exponential growth.

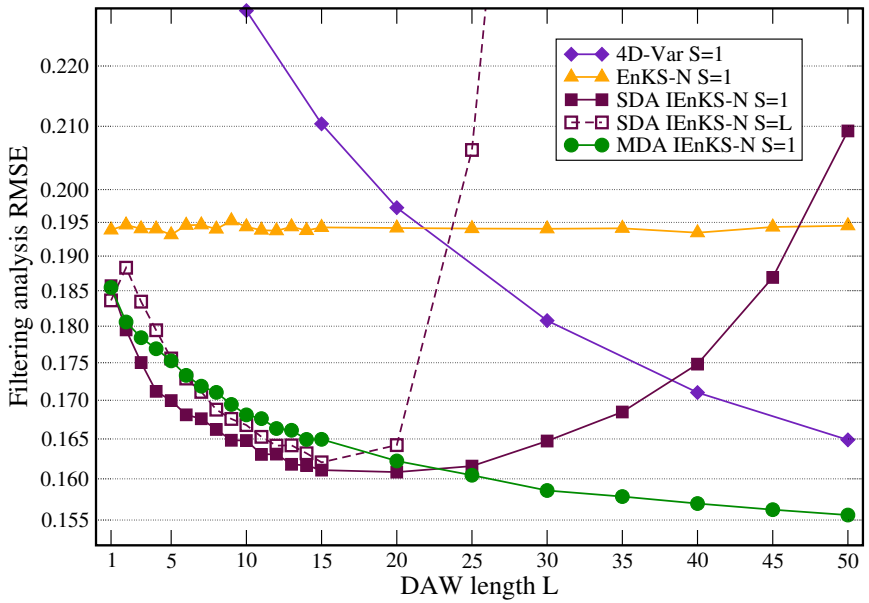


The edge of ensemble filtering methods

▶ The variational methods (3D–Var, 4D–Var): can handle nonlinearity of the operators, asynchronous observations, but *cannot handle the errors of the day.*

▶ The ensemble filtering methods (EnKFs): can only handle weak nonlinearity of the operators, cannot handle asynchronous observations, *can handle the errors of the day* through the ensemble . . . but requires regularisation of the error covariances estimate.

▶ Testing the EnKF ($N_e = 20$), 4D–Var, and IEnKS ($N_e = 20$) variants with the chaotic 40–variable Lorenz 96 model [Bocquet et al. 2013]:



▶ In mild nonlinear regime, the EnKF significantly outperforms the (basic) 4D–Var with moderately large DA windows because it captures the *errors of the day*.

Outline

Our focus: learning the analysis

- Let us assume that $\mathcal{M}$ *is known*, that the Jacobian of $\mathcal{H}_k$ is $\mathbf{H}_k$, and that we wish to learn an *incremental analysis operator* $a_{\boldsymbol{\theta}}$, typically a neural network parametrised by $\boldsymbol{\theta}$.

- If $\mathbf{E}_k^{\mathrm{a}}, \mathbf{E}_k^{\mathrm{f}} \in \mathbb{R}^{N_{\mathrm{x}} \times N_{\mathrm{e}}}$ are the analysis and forecast ensemble matrices at time τ_k, $a_{\boldsymbol{\theta}}$ is defined via the (ensemble) update:

$$\mathbf{E}_k^{\mathrm{a}} = \mathbf{E}_k^{\mathrm{f}} + a_{\boldsymbol{\theta}} \left(\mathbf{E}_k^{\mathrm{f}}, \mathbf{H}_k^{\intercal} \mathbf{R}_k^{-1} \boldsymbol{\delta}_k \right), \tag{2a}$$

where $\boldsymbol{\delta}_k$, the innovation at time τ_k, is defined by

$$\boldsymbol{\delta}_k \stackrel{\Delta}{=} \mathbf{y}_k - \mathcal{H}_k \left(\bar{\mathbf{x}}_k^{\mathrm{f}} \right), \quad \bar{\mathbf{x}}_k^{\mathrm{f}} \stackrel{\Delta}{=} \frac{1}{N_{\mathrm{e}}} \sum_{i=1}^{N_{\mathrm{e}}} \mathbf{x}_k^{\mathrm{f},i}. \tag{2b}$$

$\longrightarrow$ Notice our trick: $a_{\boldsymbol{\theta}} \left(\mathbf{E}_k^{\mathrm{f}}, \boldsymbol{\delta}_k \right) \longrightarrow a_{\boldsymbol{\theta}} \left(\mathbf{E}_k^{\mathrm{f}}, \mathbf{H}_k^{\intercal} \mathbf{R}_k^{-1} \boldsymbol{\delta}_k \right)$, i.e., *uplift of observational information in state space.*

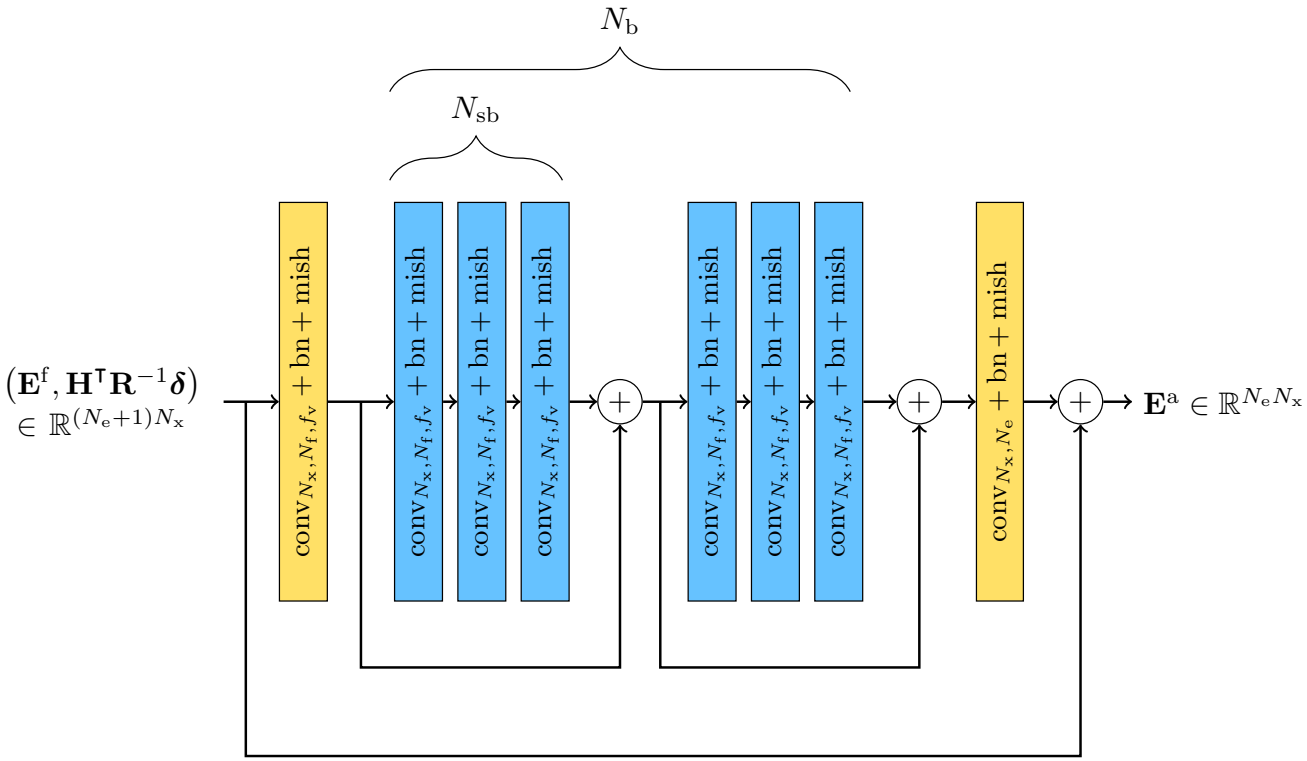
- The DA forecast step propagates the analysis ensemble, member-wise:

$$\mathbf{E}_{k+1}^{\mathrm{f}} = \mathcal{M} \left(\mathbf{E}_k^{\mathrm{a}} \right). \tag{3}$$

- The $a_{\boldsymbol{\theta}}$–based sequential DA will be called *DAN* in the following.

Neural network architecture

- We choose $a_{\boldsymbol{\theta}}$ to have a simple residual convolutional neural network (CNN) architecture.



Architecture of the residual convolutional network, where $N_{\mathrm{b}} = 2$, $N_{\mathrm{sb}} = 3$. $\mathrm{conv}_{N_1,N_2,f}$ is a generic one-dimensional convolutional layer of dimension N_1, with N_2 filters of kernel size f.

Training scheme – 1/2

- Literature (focused on *sequential data assimilation*):
  - Learning the analysis of sequential DA is not new [Härter et al. 2012; Cintra et al. 2018], though barely explored.
  - Learning key components of the analysis in the (En)KF [H. Hoang et al. 1994; S. Hoang et al. 1998] possibly leveraging auto-differentiable structure [Haarnoja et al. 2016; Chen et al. 2022; Luk et al. 2024] was also investigated.
  - Only two key papers so far focused on a *non-parametrised* analysis using backpropagation *through the DA cycles*: [McCabe et al. 2021; Boudier et al. 2023].

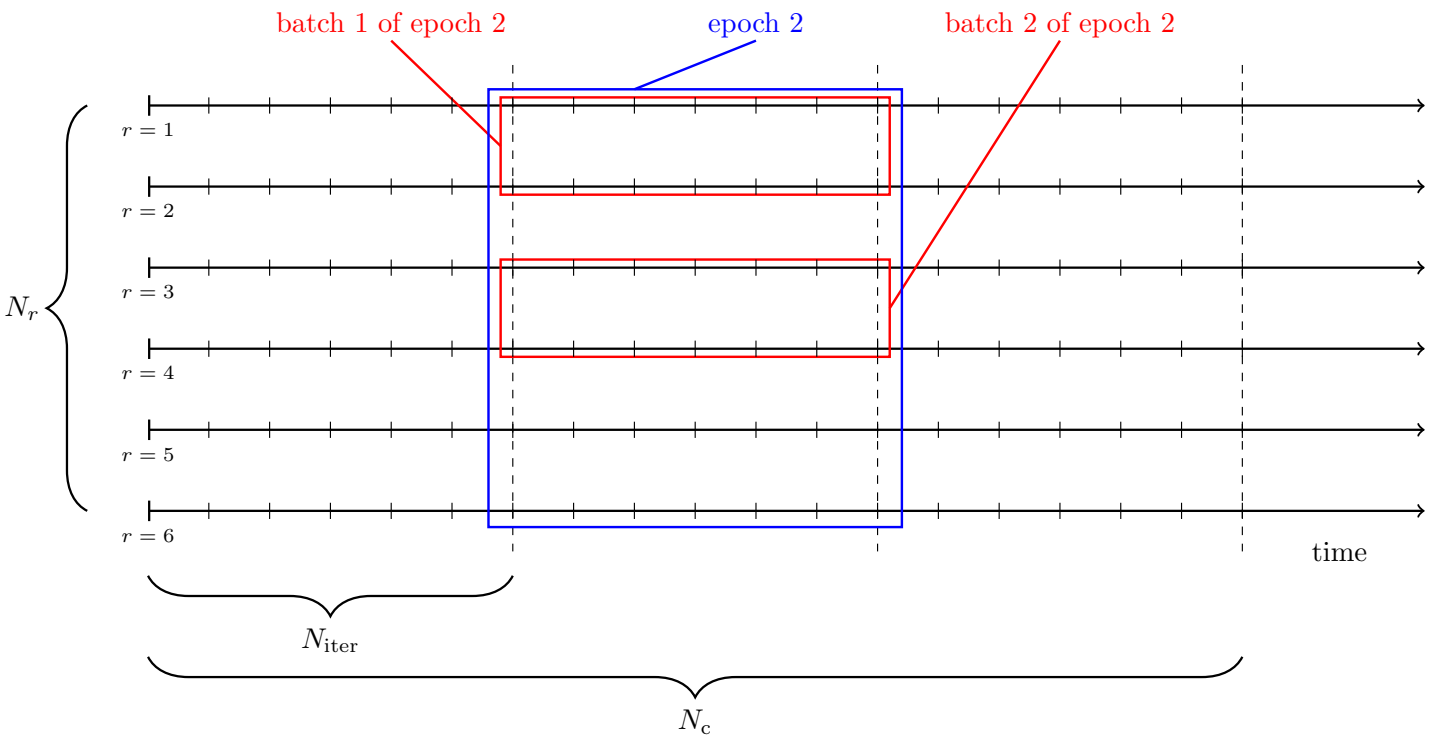
- Our training loss (supervised learning):[1]

We consider N_{r}, N_{c} cycle-long ensemble DA runs, based on N_{r} independent concurrent trajectories of the dynamics $\mathbf{x}_k^{\mathrm{t},r}$ and as many sequences of observation vectors $\mathbf{y}_k^r$.
The analysis ensemble is $\mathbf{x}_k^{\mathrm{a},i,r} \in \mathbb{R}^{N_{\mathrm{x}} \times N_{\mathrm{e}}}$. The *loss function* is defined by

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{r=1}^{N_{\mathrm{r}}} \sum_{k=1}^{N_{\mathrm{c}}} \left\| \mathbf{x}_k^{\mathrm{t},r} - \bar{\mathbf{x}}_k^{\mathrm{a},r}(\boldsymbol{\theta}) \right\|^2, \quad \bar{\mathbf{x}}_k^{\mathrm{a},r} \triangleq \frac{1}{N_{\mathrm{e}}} \sum_{i=1}^{N_{\mathrm{e}}} \mathbf{x}_k^{\mathrm{a},i,r}. \tag{4}$$

[1] [Bocquet et al. 2024]

Training scheme – 2/2



Structure of the dataset organised as a function of time, trajectory sample, batches and epochs.

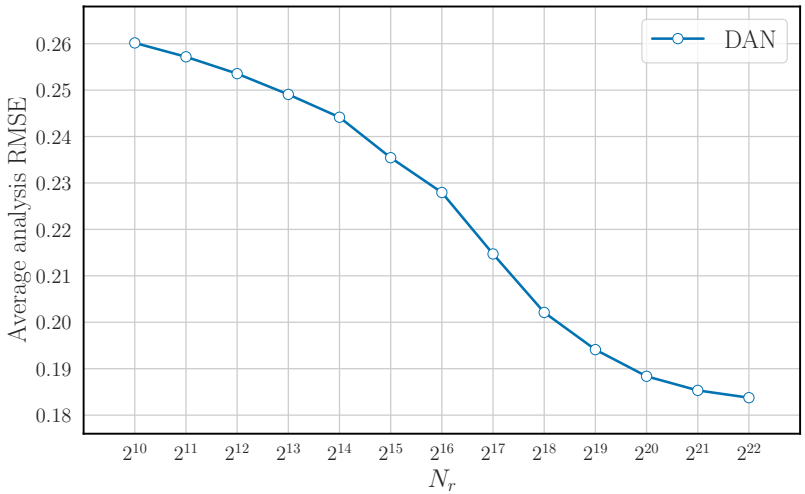
- Like [McCabe et al. 2021; Boudier et al. 2023], we use *truncated backpropagation through time TBPTT* [Tang et al. 2018; Aicher et al. 2020].

- For numerical efficiency, we choose to generate the samples *online*, as the training progresses, i.e. an *infinite training dataset!*

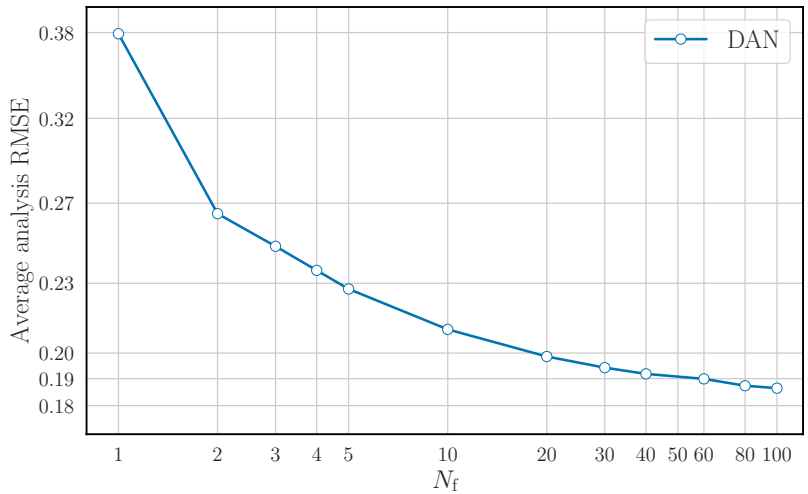
Outline

Hyperparameter sensitivity analysis

- Sensitivity analysis on key hyperparameters such as the number of trajectories N_{r} in the dataset, and the architecture parameters (N_{f}, N_{b}, N_{sb}) using the standard Lorenz 96 DA configuration ($\mathcal{H} = \mathbf{I}_{\mathrm{x}}$, $\mathbf{R} = \mathbf{I}_{\mathrm{x}}$).



Filtering performance
vs the number of trajectories N_{r}



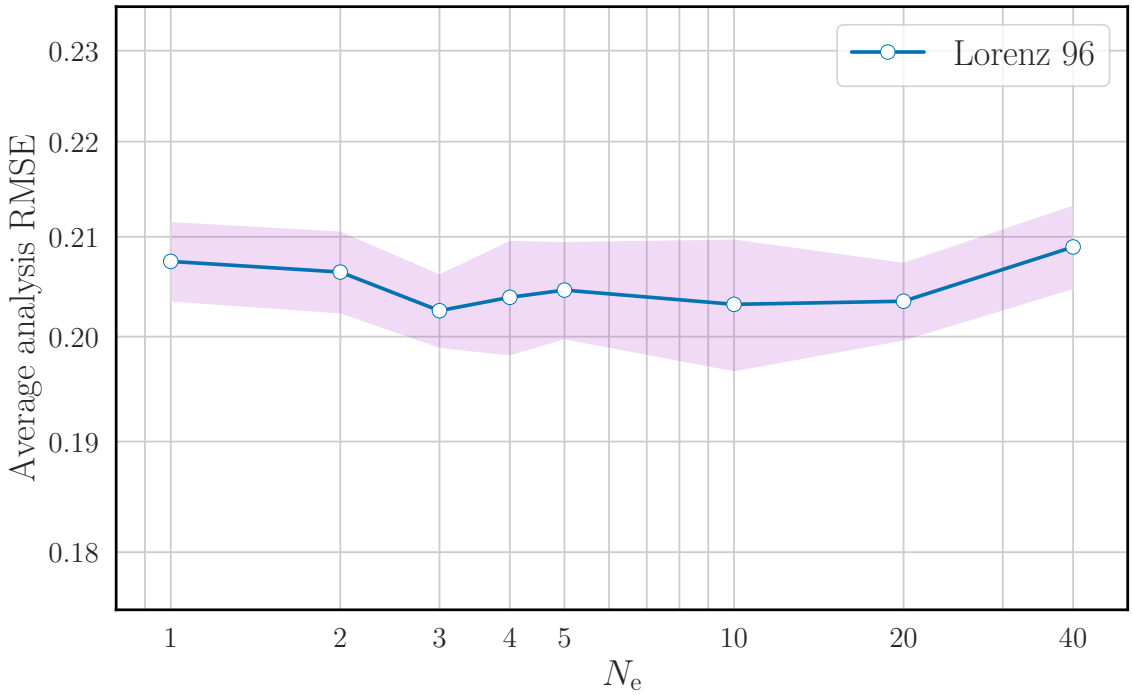
Filtering performance
vs the number of filters N_{f}

- The learned DA scheme *yields EnKF-like accuracy!*

- Compromise between $a_{\boldsymbol{\theta}}$'s size and its accuracy: $N_{\mathrm{r}} = 2^{18}$, $N_{\mathrm{f}} = 40$, $N_{\mathrm{b}} = 5$, $N_{\mathrm{sb}} = 5$.

Sensitivity to the ensemble size

▶ First key observation: The performance of $a_{\boldsymbol{\theta}}$ barely depends on the ensemble size N_{e}. Hence localisation is irrelevant and unnecessary.

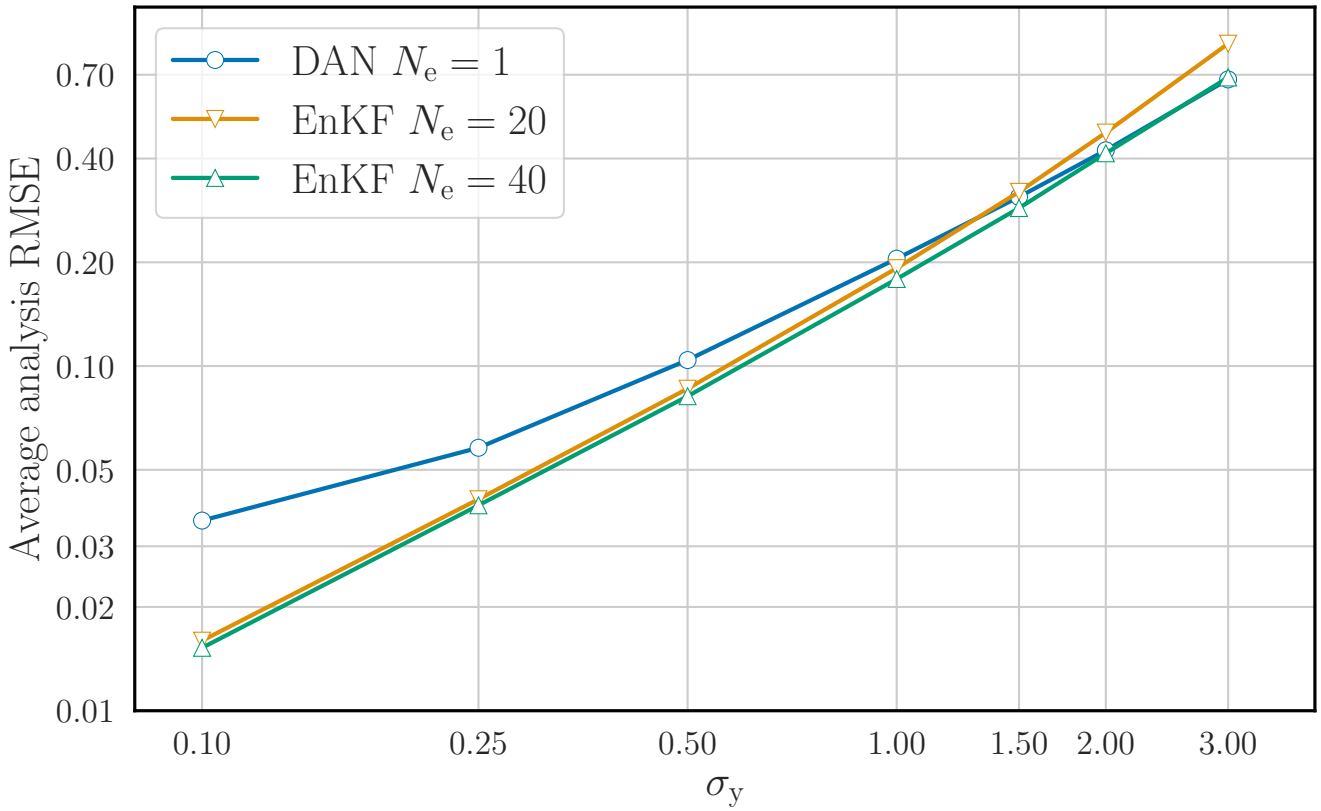


▶ Second key observation: $a_{\boldsymbol{\theta}}$ does not require inflation and is incredibly robust to noise (as we shall see it applies its own inflation).

▶ Explanation from the optimisation standpoint: *feature collapse* of $a_{\boldsymbol{\theta}}$ with respect to N_{e} in the training. Potential better solution when $N_{\mathrm{e}} > 1$, but $a_{\boldsymbol{\theta}}$ with $N_{\mathrm{e}} = 1$ is as accurate as the EnKF!

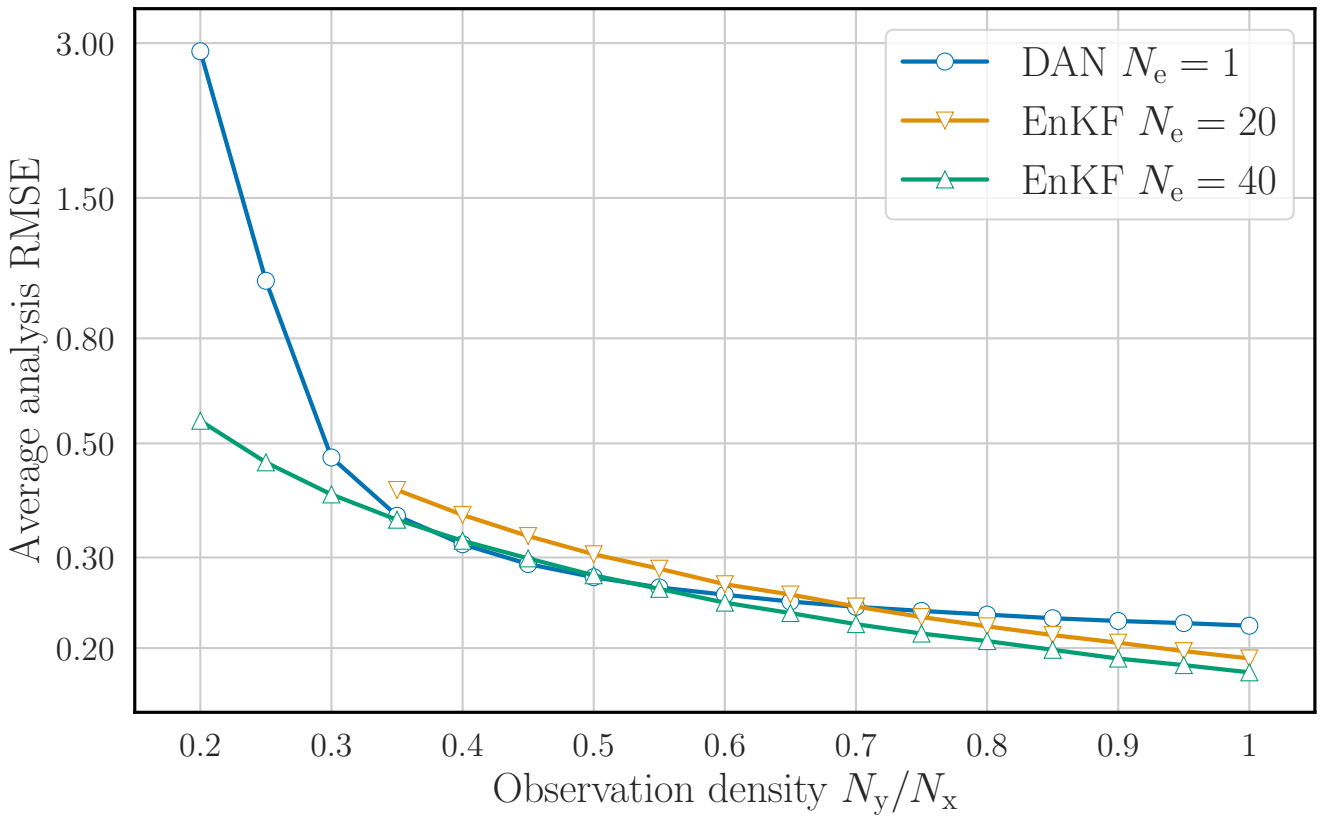
Sensitivity to observation error magnitude

- Next, we carry out a series of experiments that are not central to our message here but further ground the *viability of such learned* $a_{\boldsymbol{\theta}}$ (assuming here $\mathbf{R}_k \overset{\Delta}{=} \mathbf{I}_{\mathrm{x}}$).
- Impact of the *observation noise magnitude* on the data assimilation tests:



Sensitivity to observation sparsity

- Impact of the *sparsity of the observation dataset* on the data assimilation tests:

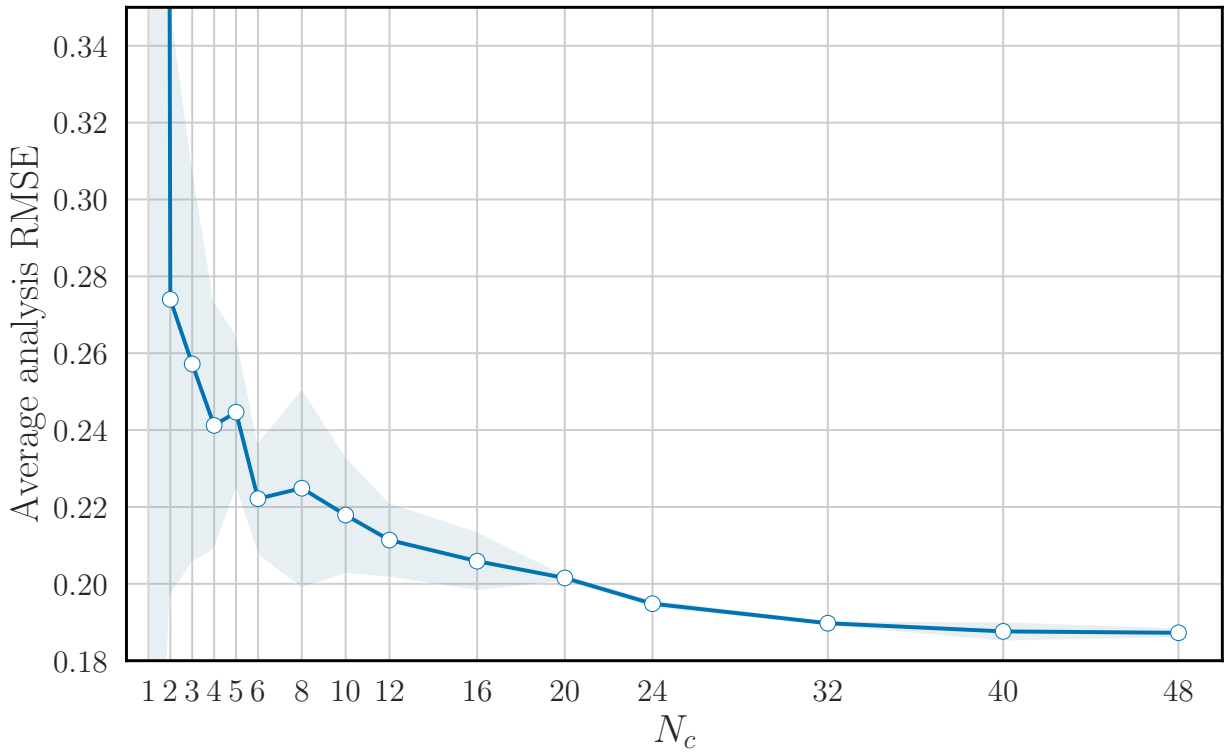


- $a_{\boldsymbol{\theta}}$ trained with time-dependent, random, observation numbers and positions.

Outline

Sensitivity to the training depth N_c

- Training through $N_c = 1$ cycles cannot learn about the direct impact of the dynamics on DA.
- Training through N_c chained cycles is expected to be crucial to the accuracy and robustness of the learned $a_{\boldsymbol{\theta}}$.



- Training depth does matter!
  As expected, $N_c \geq 2$ cycles are required to see significant benefits.

Semi-supervised learning

- What if we do not have access to the truth $\mathbf{x}_k^{\mathrm{t}}$ but to the observations only $\mathbf{y}_k$?
- Assume (i) $\mathcal{H}_k$ is linear, (ii) $\mathbf{y}_k \perp \mathbf{y}_{k+1}$, and the estimator $\mathbf{z}_{k+1}^{\boldsymbol{\theta}}$ only depends on $(\mathbf{x}_k, \mathbf{y}_k)$.
- We define the semi-supervised loss function as

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{k=1}^{N_{\mathrm{c}}} \left\| \mathbf{y}_k - \mathbf{H}_k \mathbf{z}_k^{\boldsymbol{\theta}} \right\|^2 = \sum_{k=1}^{N_{\mathrm{c}}} \mathcal{L}_k(\boldsymbol{\theta}). \tag{5}$$

But we have from the above assumptions:

$$\mathbb{E}_{\mathbf{y}} \left[\mathcal{L}_k(\boldsymbol{\theta}) \right] = \mathbb{E}_{\mathbf{y}} \left[\left\| \mathbf{y}_k - \mathbf{H}_k \mathbf{x}_k^{\mathrm{t}} \right\|^2 \right] + \mathbb{E}_{\mathbf{y}} \left[\left\| \mathbf{H}_k \left(\mathbf{x}_k^{\mathrm{t}} - \mathbf{z}_k^{\boldsymbol{\theta}} \right) \right\|^2 \right] \tag{6a}$$

$$= \mathrm{Cst} + \mathbb{E}_{\mathbf{y}} \left[\left\| \mathbf{H}_k \left(\mathbf{x}_k^{\mathrm{t}} - \mathbf{z}_k^{\boldsymbol{\theta}} \right) \right\|^2 \right] \tag{6b}$$

- Hence, generalising [McCabe et al. 2021] to non-trivial $\mathbf{H}_k$, we can learn $\mathbf{z}_k^{\boldsymbol{\theta}}$ from the observation only, with further assumptions on $\{\mathbf{H}_k\}_{k=1,\ldots,K}$. For instance, we can choose $\mathbf{z}_k^{\boldsymbol{\theta}}$ such that:

$$\mathcal{L}_k(\boldsymbol{\theta}) = \left\| \mathbf{y}_{k+1} - \mathbf{H}_{k+1} \mathcal{M} \left\{ \mathbf{x}_k^{\mathrm{f}} + a_{\boldsymbol{\theta}} \left(\mathbf{x}_k^{\mathrm{f}}, \mathbf{H}_k^{\intercal} \mathbf{R}_k^{-1} \left(\mathbf{y}_k - \mathbf{H}_k \mathbf{x}_k^{\mathrm{f}} \right) \right) \right\} \right\|^2. \tag{7}$$

Outline

Consequences and further checks

- Hence, from now on, we will focus on the mode: $\boxed{N_{\mathrm{e}} = 1}$.

- Recall

$$\mathbf{E}_k^{\mathrm{a}} = \mathbf{E}_k^{\mathrm{f}} + a_{\boldsymbol{\theta}} \left(\mathbf{E}_k^{\mathrm{f}}, \mathbf{H}_k^{\intercal} \mathbf{R}_k^{-1} \boldsymbol{\delta}_k \right). \tag{8}$$

- Performance of $a_{\boldsymbol{\theta}}$ compared to baselines such as optimally tuned 3D-Var, the learned optimal linear filter, optimally tuned EnKF:

DA method	well-tuned classical	DL-based	aRMSE
EnKF-N, $N_{\mathrm{e}} = 20$	yes		0.191
EnKF-N, $N_{\mathrm{e}} = 40$	yes		0.179
3D-Var	yes		0.40
$a_{\boldsymbol{\theta}}$, $N_{\mathrm{e}} = 1$, $N_{\mathrm{f}} = 40$		yes	0.191
$a_{\boldsymbol{\theta}}$, $N_{\mathrm{e}} = 1$, $N_{\mathrm{f}} = 100$		yes	0.185
linear $a_{\boldsymbol{\theta}}$, $N_{\mathrm{e}} = 1$, $N_{\mathrm{f}} = 40$		yes	0.384
simplified $\hat{a}_{\boldsymbol{\theta}}$, $N_{\mathrm{e}} = 1$, $N_{\mathrm{f}} = 40$		yes	0.382

where the simplified Ansatz $\hat{a}_{\boldsymbol{\theta}}$ is defined through

$$\mathbf{E}_k^{\mathrm{a}} = \mathbf{E}_k^{\mathrm{f}} + \hat{a}_{\boldsymbol{\theta}} \left(\mathbf{H}_k^{\intercal} \mathbf{R}_k^{-1} \boldsymbol{\delta}_k \right). \tag{9}$$

Operator expansion of the analysis

- We look for *a classical Kalman update that would be a good match to* $a_{\boldsymbol{\theta}}$ seen as a mathematical map, at least for small analysis increments.

- To that end, we define the time-dependent normalised scalar anomalies

$$b_k = \frac{1}{\sqrt{N_{\mathrm{x}}}} \left\| a_{\boldsymbol{\theta}}(\mathbf{x}_k, \mathbf{0}) \right\|, \tag{10}$$

along with the associated mean bias b and the standard deviation s of b_k in time.

- Next, expanding with respect to the innovation, the following functional form for $a_{\boldsymbol{\theta}}$ is assumed:

$$a_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{H}^{\mathsf{T}}\mathbf{R}^{-1}\boldsymbol{\delta}) \approx \mathbf{K}(\mathbf{x}) \cdot \boldsymbol{\delta}, \tag{11}$$

owing to the fact that no state update is needed when the innovation vanishes, and only keeping the leading order term in $\boldsymbol{\delta}$.

Identifying the operators in the expansion

- Innovations $\{\boldsymbol{\delta}_j\}_{j=1,\dots,N_\mathrm{p}}$ are sampled from $\boldsymbol{\delta}_j \sim N(\mathbf{0}, \mathbf{R})$.

This yields a set of corresponding incremental updates $\left\{\mathbf{a}_j = a_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{H}^\intercal \mathbf{R}^{-1} \boldsymbol{\delta}_j)\right\}_{j=1,\dots,N_\mathrm{p}}$.
$\mathbf{K}(\mathbf{x})$ is then estimated with the least squares problem

$$\mathcal{L}_{\mathbf{x}}(\mathbf{K}) = \sum_{j=1}^{N_\mathrm{p}} \left\| \mathbf{a}_j - \bar{\mathbf{a}} - \mathbf{K}(\mathbf{x}) \cdot \left(\boldsymbol{\delta}_j - \bar{\boldsymbol{\delta}}\right) \right\|^2, \tag{12}$$

where $\bar{\mathbf{a}} = N_\mathrm{p}^{-1} \sum_{j=1}^{N_\mathrm{p}} \mathbf{a}_j$ and $\bar{\boldsymbol{\delta}} = N_\mathrm{p}^{-1} \sum_{j=1}^{N_\mathrm{p}} \boldsymbol{\delta}_j$.

- Within the *best linear unbiased estimator* framework, $\mathbf{K}$ is related to $\mathbf{P}^\mathrm{a}$ through $\mathbf{K} = \mathbf{P}^\mathrm{a} \mathbf{H}^\intercal \mathbf{R}^{-1}$ so that from Eq. (11),

$$a_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{H}^\intercal \mathbf{R}^{-1} \boldsymbol{\delta}) \approx \mathbf{P}^\mathrm{a} \mathbf{H}^\intercal \mathbf{R}^{-1} \boldsymbol{\delta}, \tag{13}$$

which suggests that an expansion in the second variable $\boldsymbol{\zeta} \in \mathbb{R}^{N_\mathrm{x}}$ of $a_{\boldsymbol{\theta}}$ yields

$$a_{\boldsymbol{\theta}}(\mathbf{x}, \boldsymbol{\zeta}) \approx \mathbf{P}^\mathrm{a}(\mathbf{x}) \cdot \boldsymbol{\zeta}. \tag{14}$$

Hence, we can obtain a numerical estimation of an equivalent $\mathbf{P}^\mathrm{a}(\mathbf{x})$.

What is learned? Supporting numerical results

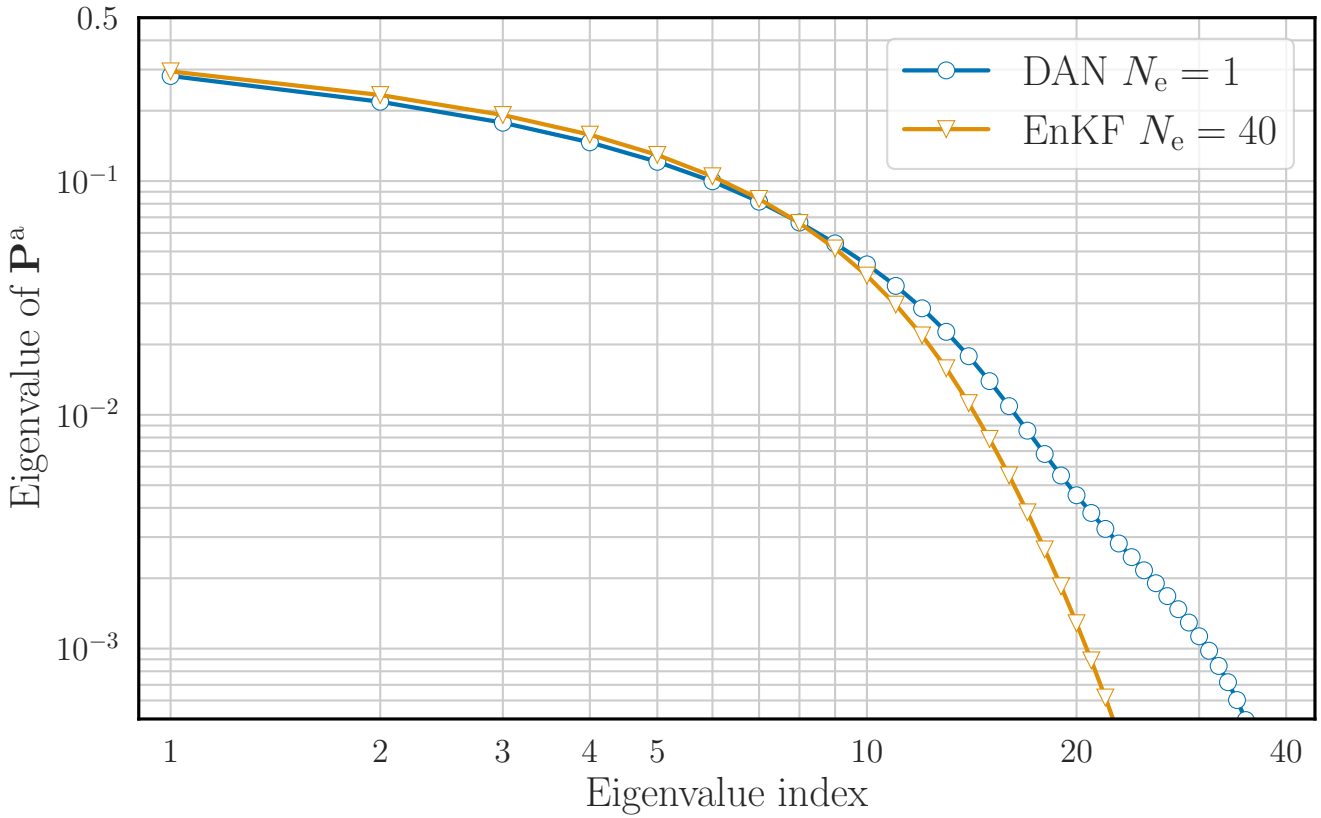
▶ We obtain $b \simeq 5 \times 10^{-3}$ and $s \simeq 10^{-3}$, which are indeed very small compared to the typical aRMSE of an either DAN or EnKF run, i.e., 0.20.

▶ The surrogate $\mathbf{P}^{\mathrm{a}}$, denoted $\mathbf{P}^{\mathrm{a}}_{\mathrm{DAN}}$ and estimated from Eq. (14), is compared to that of a concurrent well-tuned EnKF with $N_{\mathrm{e}} = 40$, whose analysis error covariance matrix is $\mathbf{P}^{\mathrm{a}}_{\mathrm{EnKF}}$.

$\longrightarrow$ The time-averaged Bures–Wasserstein distance distance between $\mathbf{P}^{\mathrm{a}}_{\mathrm{DAN}}$ and $\mathbf{P}^{\mathrm{a}}_{\mathrm{EnKF}}$ is 0.013 whereas it is 0.048 between $\mathbf{P}^{\mathrm{a}}_{\mathrm{DAN}}$ and $(0.40)^2\mathbf{I}_{\mathrm{x}}$, which approximates $\mathbf{P}^{\mathrm{a}}$ of a well-tuned 3D-Var.

What is learned? Supporting numerical results

- The time-averaged eigenspectra of $\mathbf{P}^{\mathrm{a}}_{\mathrm{DAN}}$ and $\mathbf{P}^{\mathrm{a}}_{\mathrm{EnKF}}$:



- They are remarkably close to each other for the first 10 modes. Beyond these modes the $a_{\boldsymbol{\theta}}$ operator is likely to selectively apply *some multiplicative inflation*, as one would expect from such stable DA runs.

Main interpretation

▶ Conclusion 1: $a_{\boldsymbol{\theta}}$ depends on the innovation but also directly on $\mathbf{x}_k^{\mathrm{f}}$ when $N_{\mathrm{e}} = 1$, as opposed to the incremental update of the EnKF: *$a_{\boldsymbol{\theta}}$ extracts important information from $\mathbf{x}_k^{\mathrm{f}}$.*

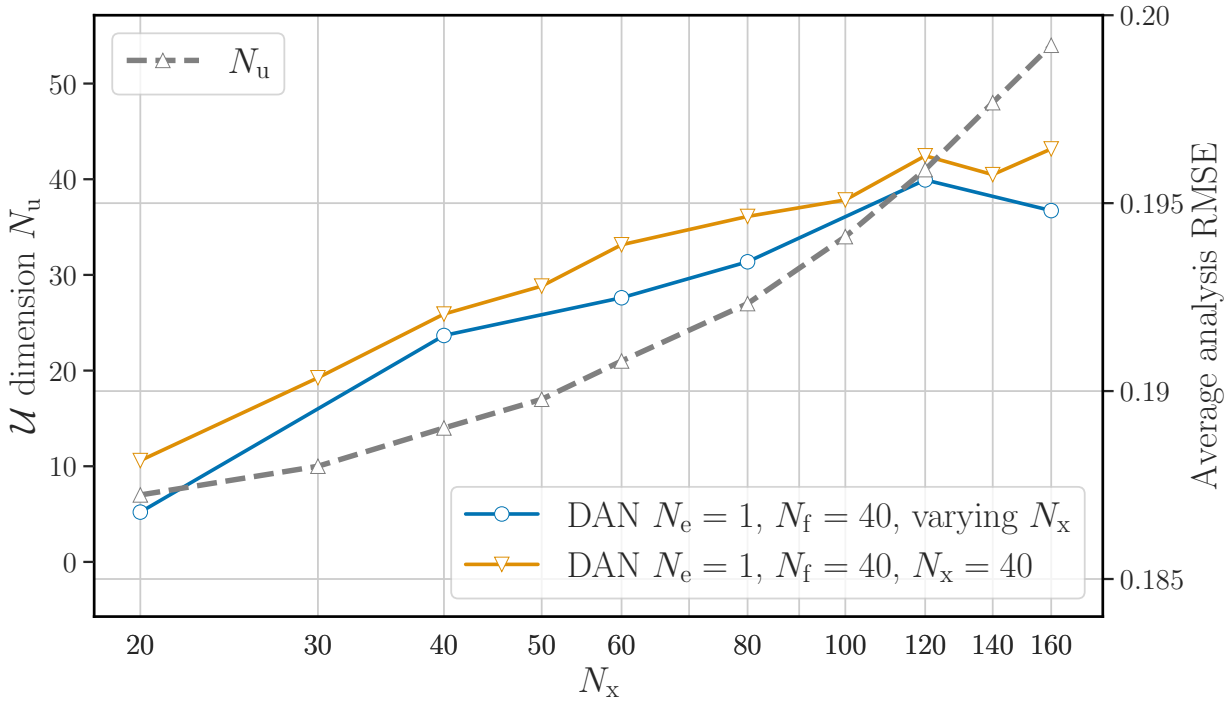
▶ Conclusion 2: $a_{\boldsymbol{\theta}}$ manages to assess $\mathbf{P}_{\mathrm{DAN}}^{\mathrm{a}}$ with $N_{\mathrm{e}} = 1$ which is very close to $\mathbf{P}_{\mathrm{EnKF}}^{\mathrm{a}}$ with $N_{\mathrm{e}} = 40$, for the dominant axes, and applies multiplicative inflation on the less unstable modes.[2] We conclude that *$a_{\boldsymbol{\theta}}$ directly learns about the dynamics features.* Hence, for $a_{\boldsymbol{\theta}}$, critical pieces of information on $\mathbf{P}_k^{\mathrm{a}}$ are encoded, and thus exploitable, in $\mathbf{x}_k^{\mathrm{f}}$ alone.

▶ Explanation, conclusion 3: Furthermore, if the DA run (the forecast and analysis cycle) is considered as an ergodic dynamical system of its own,[3] the *multiplicative ergodic theorem* guarantees the existence of a mapping between $\mathbf{x}_k^{\mathrm{f}}$ and $\mathbf{P}_k^{\mathrm{a}}$ that $a_{\boldsymbol{\theta}}$ is able to guess.
We believe that a generalised variant of *the multiplicative ergodic theorem for non-autonomous random dynamics should be applicable*.[4]

[2] [Bocquet et al. 2015]
[3] [Carrassi et al. 2008]
[4] [Arnold 1998; Flandoli et al. 2021]

Locality and scalability

▶ $a_{\boldsymbol{\theta}}$ is now trained without changing the architecture and the hyperparameters ($N_{\mathrm{f}} = 40$), but with a changing state space dimension $N_{\mathrm{x}} \in [20, 160]$. Almost as good as well tuned EnKFs with changing dimension N_{x} and $N_{\mathrm{e}} = N_{\mathrm{x}}$!



⟶ We conjecture that $a_{\boldsymbol{\theta}}$ extracts *local* pieces of information from $\mathbf{x}_k^{\mathrm{f}}$.

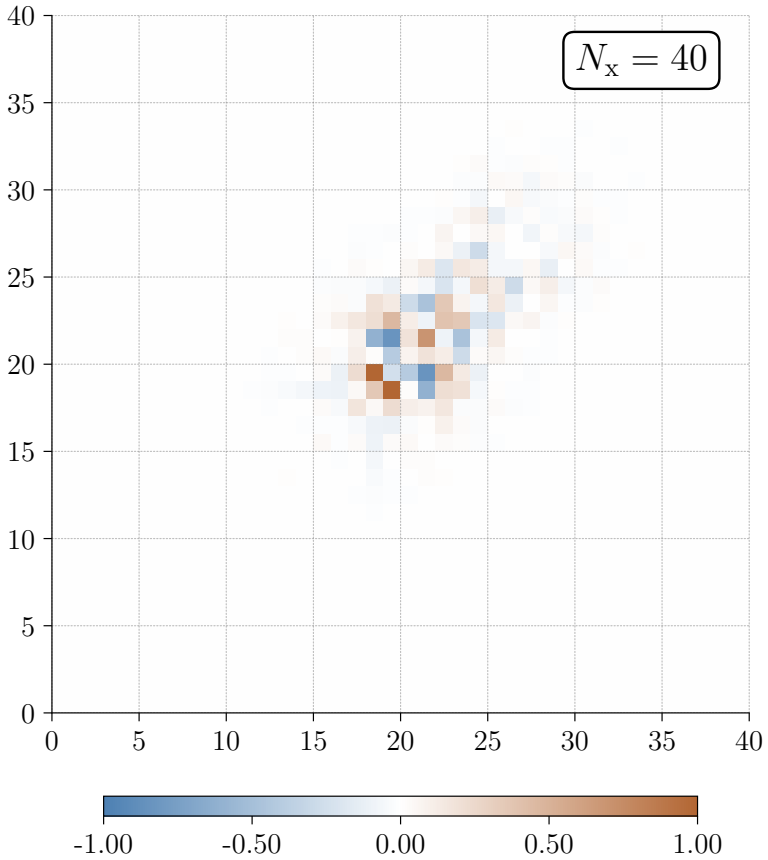
▶ $a_{\boldsymbol{\theta}}$, learned from Lorenz 96 with $N_{\mathrm{x}} = 40$ is now tested on Lorenz 96 models with N_{x} ranging from 20 to 160 (same weights and biases!). The performance is still on par with retraining!
We called this a *transdimensional transfer*.

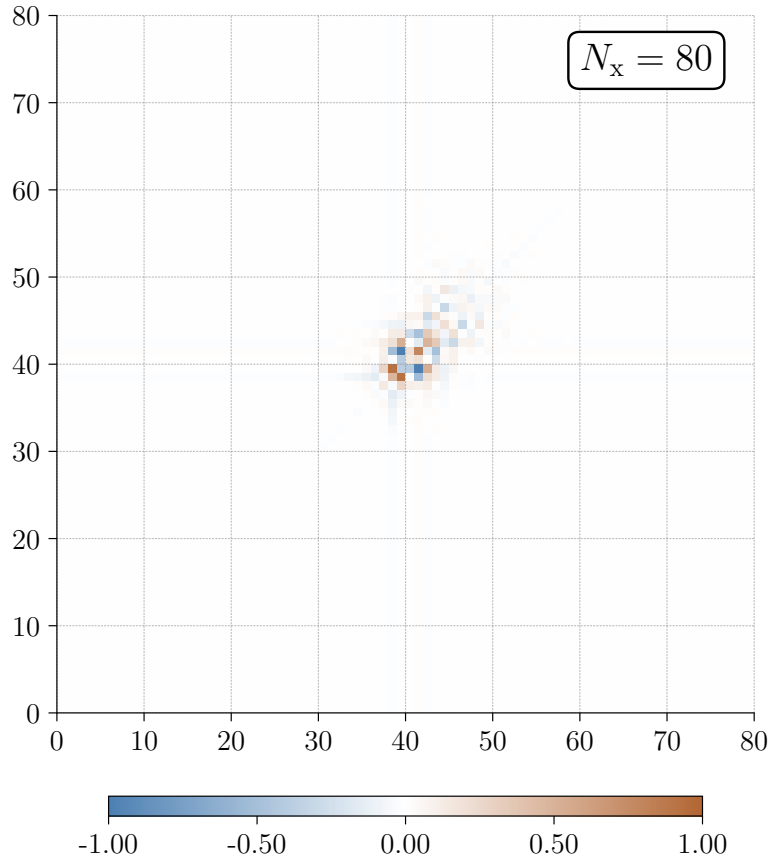
Locality and scalability

▶ These local patterns (for $a_{\boldsymbol{\theta}}$, not $\mathcal{M}$) can be pictured from the *mean marginal analysis error covariance matrix*:

$$\mathbf{S} = \left\langle \mathbf{C} : \left[\nabla_{\mathbf{x}} \nabla_{\zeta} a_{\boldsymbol{\theta}}(\mathbf{x}, \boldsymbol{\zeta})_{|\boldsymbol{\zeta}=\mathbf{0}} \right] \right\rangle_{\mathbf{x} \in \mathcal{T}} = \left\langle \mathbf{C} : \left[\nabla_{\mathbf{x}} \mathbf{P}^{\mathrm{a}}(\mathbf{x}) \right] \right\rangle_{\mathbf{x} \in \mathcal{T}}, \tag{15}$$

where $\mathcal{T}$ is a long L96 trajectory, and $\mathbf{C}$ is a tensor that leverages translational invariance of the L96 model: $[\mathbf{C}]_{ij}^{nmk} = \frac{1}{N_{\mathrm{x}}} \delta_{n,i+k} \delta_{m,j+k}$.

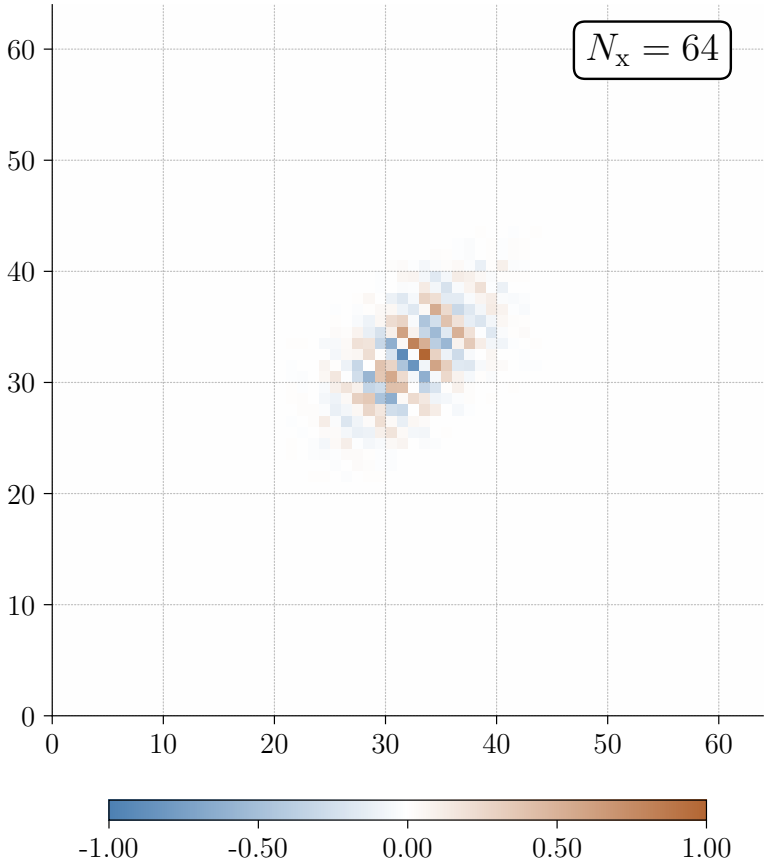


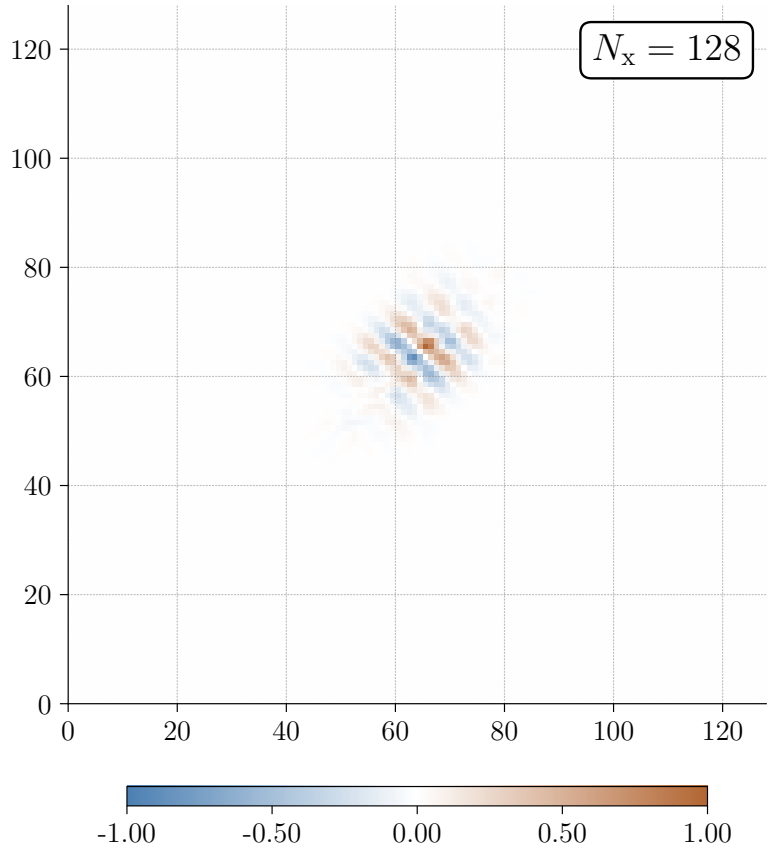


Application to the Kuramoto-Sivashinski model

- The results are very similar to those of the Lorenz 96 model.

Mean marginal analysis error covariance matrix:





Outline

Conclusion: addendum and perspectives

- We have carried similar numerical experiments with a *single-layer QG model on the sphere*, with similar conclusions.

- Will such *multiplicative ergodic theorem* still be valid in more anisotropic, non-autonomous, forced, multivariate, heterogeneously observed systems?

- In any case, this promotes a rethinking of the popular sequential DA schemes for chaotic dynamics.

$\longrightarrow$ Talk (mainly) based on Bocquet et al., Chaos, 2024.

References I

[1] C. Aicher, N. J. Foti, and E. B. Fox. "Adaptively Truncating Backpropagation Through Time to Control Gradient Bias". In: *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*. Ed. by Ryan P. Adams and Vibhav Gogate. Vol. 115. Proceedings of Machine Learning Research. PMLR, 22–25 Jul 2020, pp. 799–808.

[2] L. Arnold. *Random Dynamical Systems*. Springer Berlin, Heidelberg, 1998, p. 586.

[3] M. Bocquet, P. N. Raanes, and A. Hannart. "Expanding the validity of the ensemble Kalman filter without the intrinsic need for inflation". In: *Nonlin. Processes Geophys.* 22 (2015), pp. 645–662.

[4] M. Bocquet and P. Sakov. "Joint state and parameter estimation with an iterative ensemble Kalman smoother". In: *Nonlin. Processes Geophys.* 20 (2013), pp. 803–818.

[5] M. Bocquet et al. "Accurate deep learning-based filtering for chaotic dynamics by identifying instabilities without an ensemble". In: *Chaos* 29 (2024), p. 091104.

[6] P. Boudier et al. "Data Assimilation Networks". In: *J. Adv. Model. Earth Syst.* 15 (2023), e2022MS003353.

[7] A. Carrassi et al. "Data assimilation as a nonlinear dynamical systems problem: Stability and convergence of the prediction-assimilation system". In: *Chaos* 18 (2008), p. 023112.

[8] Y. Chen, D. Sanz-Alonso, and R. Willett. "Autodifferentiable Ensemble Kalman Filters". In: *SIAM J. Math. Data Sci.* 4 (2022), pp. 801–833.

[9] R. S. Cintra and H. F. de Campos Velho. "Data assimilation by artificial neural networks for an atmospheric general circulation model". In: *Advanced applications for artificial neural networks*. Ed. by A. ElShahat. IntechOpen, 2018. Chap. 17, pp. 265–286.

[10] F. Flandoli and E. Tonello. *An introduction to random dynamical systems for climate*. 2021.

[11] T. Haarnoja et al. "Backprop KF: Learning Discriminative Deterministic State Estimators". In: *Advances in Neural Information Processing Systems*. Ed. by D. Lee et al. Vol. 29. Curran Associates, Inc., 2016.

[12] T. P. Härter and H. F. de Campos Velho. "Data Assimilation Procedure by Recurrent Neural Network". In: *Eng. Appl. Comput. Fluid Mech.* 6 (2012), pp. 224–233.

[13] H.S. Hoang, P. De Mey, and O. Talagrand. "A simple adaptive algorithm of stochastic approximation type for system parameter and state estimation". In: *Proceedings of 1994 33rd IEEE Conference on Decision and Control*. Vol. 1. 1994, 747–752 vol.1.

[14] S. Hoang et al. "Adaptive filtering: application to satellite data assimilation in oceanography". In: *Dynam. Atmos. Ocean* 27 (1998), pp. 257–281.

[15] E. Luk et al. *Learning Optimal Filters Using Variational Inference*. 2024. arXiv: 2406.18066 [cs.LG].

References II

[16] M. McCabe and J. Brown. "Learning to Assimilate in Chaotic Dynamical Systems". In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato et al. Vol. 34. Curran Associates, Inc., 2021, pp. 12237–12250.

[17] H. Tang and J. Glass. "On Training Recurrent Networks with Truncated Backpropagation Through time in Speech Recognition". In: *2018 IEEE Spoken Language Technology Workshop (SLT)*. 2018, pp. 48–55.