Using machine learning and data assimilation to learn both dynamics and state

Marc Bocquet[1], Julien Brajard[2,3], Alberto Carrassi[4,5] & Laurent Bertino[2]
Alban Farchi[1], Quentin Malartic[1]
and many fruitful discussions with Massimo Bonavita[6] and Patrick Laloyaux[6]

(1) CEREA, joint laboratory École des Ponts ParisTech and EDF R&D, Université Paris-Est

(2) Nansen Environmental and Remote Sensing Center

(3) Sorbonne University, CNRS-IRD-MNHN, LOCEAN

(4) Department of meteorology, University of Reading, United Kingdom

(5) Mathematical institute, University of Utrecht, The Netherlands

(6) ECMWF, Reading, United Kingdom

Outline

From model error to the absence of a model

- At crossroads between:

  Data Assimilation (DA), Machine Learning (ML) and Dynamical Systems (DS)

- Goal: Estimate autonomous chaotic dynamics from partial and noisy observations

  $\longrightarrow$ Surrogate model

- Subgoal 1: Develop a Bayesian framework for this estimation problem.
- Subgoal 2: Estimate and minimize the errors attached to the estimation.
- Subgoal 3: What about online (sequential) learning?

- References connected to data-driven reconstruction of the dynamics in DA and ML:
  [Park et al. 1994; Wang et al. 1998; Paduart et al. 2010; Lguensat et al. 2017; Pathak et al. 2017; Harlim 2018; Dueben et al. 2018; Long et al. 2018; Fablet et al. 2018; Vlachas et al. 2020; Brunton et al. 2016]
  and many more since the beginning of 2020.

Outline

ODE representation for the surrogate model

▶ Ordinary differential equations (ODEs) representation of the surrogate dynamics

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = \boldsymbol{\phi}_{\mathbf{A}}(\mathbf{x}),$$

where $\mathbf{A}$ is a set of N_p coefficients.

▶ We need:

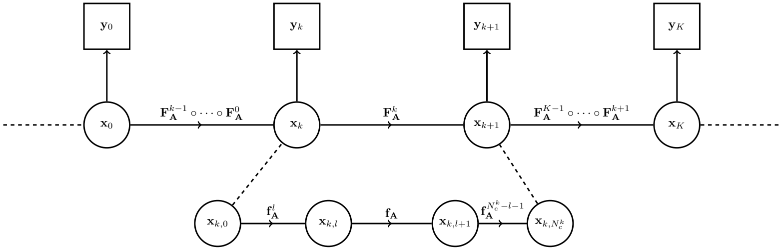
- to choose a numerical scheme to integrate in time the tendency $\boldsymbol{\phi}_{\mathbf{A}}$ and be able to build resolvent of the surrogate dynamics $\mathbf{x}_{k+1} = \mathbf{F}_{\mathbf{A}}(\mathbf{x}_k)$,
- to specify the tendency $\mathbf{x} \mapsto \boldsymbol{\phi}_{\mathbf{A}}(\mathbf{x})$.

▶ Going beyond in this work, we wish to account for (surrogate) model error, so that the surrogate model representation is actually an SDE:

$$\mathrm{d}\mathbf{x} = \boldsymbol{\phi}_{\mathbf{A}}(\mathbf{x})\mathrm{d}t + \sqrt{\mathbf{Q}}\mathrm{d}\mathbf{W}(t),$$

with $\mathbf{W}(t)$ an N_x-dimensional Wiener process.

Integration scheme and cycling



- Choosing a Runge-Kutta method as integration scheme:

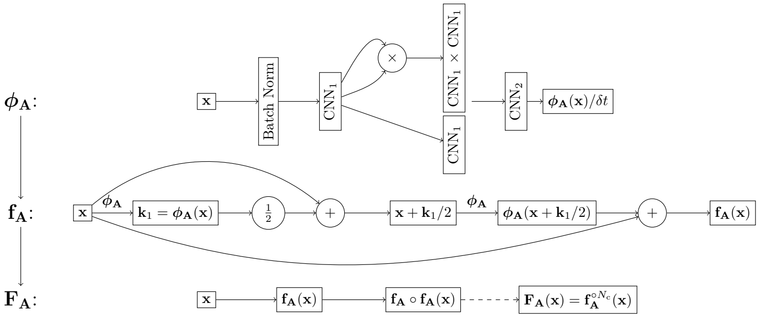
$$\mathbf{f}_\mathbf{A}(\mathbf{x}) = \mathbf{x} + h \sum_{i=0}^{N_{\mathrm{RK}}-1} \beta_i \mathbf{k}_i, \qquad \mathbf{k}_i = \boldsymbol{\phi}_\mathbf{A}\left(\mathbf{x} + h \sum_{j=0}^{i-1} \alpha_{i,j} \mathbf{k}_j\right).$$

- Compositions of integration schemes:

$$\mathbf{x}_{k+1} = \mathbf{F}_\mathbf{A}^k(\mathbf{x}_k) \quad \text{where} \quad \mathbf{F}_\mathbf{A}^k \equiv \mathbf{f}_\mathbf{A}^{N_c^k} \equiv \underbrace{\mathbf{f}_\mathbf{A} \circ \ldots \circ \mathbf{f}_\mathbf{A}}_{N_c^k \text{ times}},$$

Neural network models

- We tested many simple architectures, all following the structure of N_c explicit Runge-Kutta schemes, with linear or nonlinear activation functions:
  - $\phi_{\mathbf{A}}$: minimal representation (as few parameters as possible), or based on a NN (with potentially many parameters) implemented in TensorFlow 2.x
  - Convolutional layers were used for local, homogeneous systems.
  - Locally connected convolutional layers were used for local, heterogeneous systems.



Outline

Bayesian analysis of the joint problem

- **Bayesian view** on state and model estimation:

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

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

$$\mathcal{J}(\mathbf{A}, \mathbf{x}_{0:K}, \mathbf{Q}_{1:K}) = \frac{1}{2} \sum_{k=0}^{K} \left\{ \| \mathbf{y}_k - \mathbf{H}_k(\mathbf{x}_k) \|^2_{\mathbf{R}_k^{-1}} + \ln |\mathbf{R}_k| \right\}$$
$$+ \frac{1}{2} \sum_{k=1}^{K} \left\{ \left\| \mathbf{x}_k - \mathbf{F}_{\mathbf{A}}^{k-1}(\mathbf{x}_{k-1}) \right\|^2_{\mathbf{Q}_k^{-1}} + \ln |\mathbf{Q}_k| \right\} - \ln p(\mathbf{x}_0, \mathbf{A}, \mathbf{Q}_{1:K}).$$

⟶ This is a (4D) variational problem.
⟶ Allows to rigorously handle partial and noisy observations.

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

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

Similar outcome or improved upon [Hsieh et al. 1998; Abarbanel et al. 2018].

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

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

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

which is not as general as $\mathcal{J}(\mathbf{A}, \mathbf{x}_{0:K}, \mathbf{Q}_{1:K})$.

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

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

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

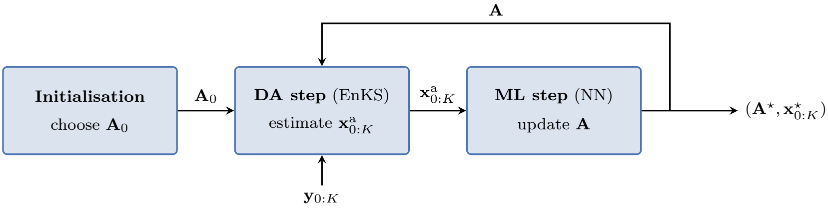
(2) ▶ $\mathcal{J}(\mathbf{A}, \mathbf{x}_{0:K}|\mathbf{Q}_{1:K})$ is minimized using a coordinate descent:

- ▶ using a weak constraint 4D-Var for $\mathbf{x}_{0:K}$ and a variational subproblem for $\mathbf{A}$ [Bocquet et al. 2019].

- ▶ using a (higher-dimensional) strong constraint 4D-Var for $\mathbf{x}_{0:K}$ and a variational subproblem for $\mathbf{A}$ [Bocquet et al. 2019].

- ▶ using an EnKF/EnKS for $\mathbf{x}_{0:K}$ and a variational subproblem for $\mathbf{A}$ [Brajard et al. 2020a; Bocquet et al. 2020b].

⟶ 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 $p(\mathbf{A}, \mathbf{Q}_{1:K}|\mathbf{y}_{0:K}, \mathbf{R}_{0:K})$:

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

yields the loss function

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

- 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] but their approximation is in practice a coordinate descent where model error is not controlled.

Algorithm for an approximate solution of the marginal problem (1/2)

▶ The expectation step: EnKS over a long period $[t_0, t_K]$ which accounts for model error (SQRT-CORE scheme). The outputs are $\overline{\mathbf{x}}_{0:K}^{(j)}$ and $\mathbf{Q}^{(j+1)}$ computed online by accumulating over the time window.

$$\mathbf{Q}^{(j+1)} = \frac{1}{K N_\mathrm{e}} \sum_{i=1}^{N_\mathrm{e}} \sum_{k=1}^{K} \left(\mathbf{x}_{k,i}^{(j)} - \mathbf{F}_{\mathbf{A}^{(j)}}^{k-1}(\mathbf{x}_{k-1,i}^{(j)})\right) \left(\mathbf{x}_{k,i}^{(j)} - \mathbf{F}_{\mathbf{A}^{(j)}}^{k-1}(\mathbf{x}_{k-1,i}^{(j)})\right)^\top .$$

Hence this expectation step anticipates the maximization step when estimating $\mathbf{Q}^{(j+1)}$.

Algorithm for an approximate solution of the marginal problem (2/2)

▶ The maximization step: Minimize ($\mathbf{Q}^{(j+1)}$ is fixed):

$$\begin{aligned}\mathcal{L}^{(j)}(\mathbf{A},\mathbf{Q}^{(j+1)}) =& -\ln p(\overline{\mathbf{x}}^{(j)}_{0:K},\mathbf{y}_{0:K},\mathbf{A},\mathbf{Q}^{(j+1)},\mathbf{R}_{0:K}) \\ =& \frac{1}{2}\sum_{k=1}^{K}\left\{\left\|\overline{\mathbf{x}}^{(j)}_{k}-\mathbf{F}^{k-1}_{\mathbf{A}}(\overline{\mathbf{x}}^{(j)}_{k-1})\right\|^2_{\mathbf{Q}^{(j+1)}{}^{-1}}+\ln\left|\mathbf{Q}^{(j+1)}\right|\right\} \\ & -\ln p(\overline{\mathbf{x}}^{(j)}_{0},\mathbf{A},\mathbf{Q}^{(j+1)})+\cdots.\end{aligned}$$

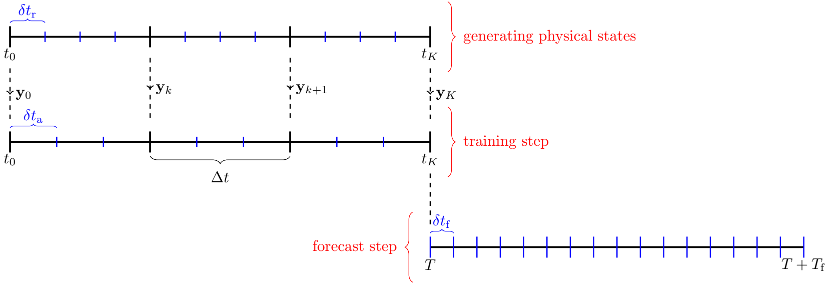
Note the use of the ensemble mean instead of the ensemble (main approximation).

▶ No iteration in the maximization step over $\mathbf{A}$ and $\mathbf{Q}$ (should be fine if $\mathbf{Q}=q\mathbf{I}_{\mathrm{x}}$).

▶ In practice, almost numerically equivalent to the full EM scheme (see [Bocquet et al. 2020b] for details).

Experiment plan

- The reference model, the surrogate model and the forecasting system

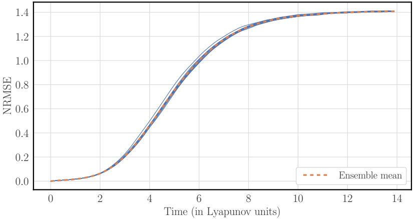


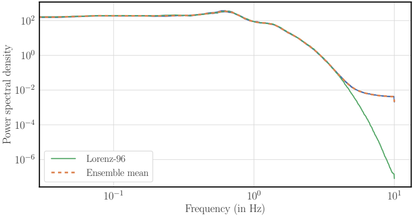
- Metrics of comparison:
  - Forecast skill [FS]: Normalized RMSE (NRMSE) between the reference and surrogate forecasts as a function of lead time (averaged over initial conditions).
  - Lyapunov spectrum [LS].
  - Power spectrum density [PSD].

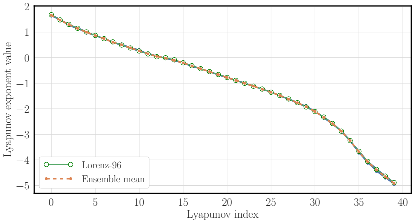
Application to the one-scale Lorenz-96 model

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

  - CNN+RK4
  - 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



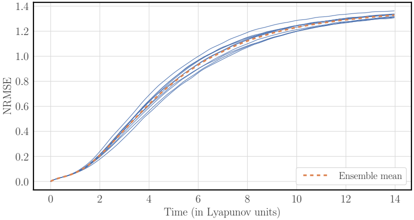


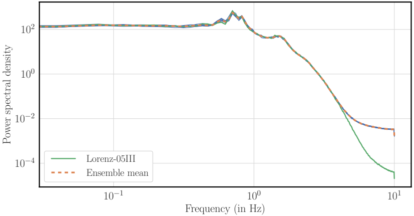


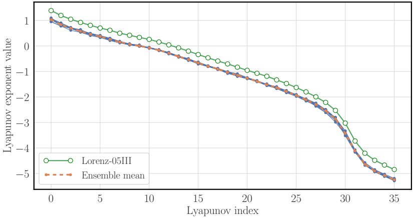
Application to the two-scale Lorenz-05III model

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

  - CNN+RK4
  - 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







Outline

Online learning scheme: principle

▶ So far, the learning was offline, i.e. based on variational technique using all available data. Can one design a sequential (online) 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.
A few hypotheses could be relaxed.
All parameters of the model are hereafter noted $\mathbf{A} \longrightarrow \mathbf{p} \in \mathbb{R}^{N_\mathrm{p}}$.

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

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

Just a (more ambitious) parameter estimation problem!?

Online learning scheme: EnKF update

- We use the augmented state formalism with ensemble Kalman filters (EnKFs):
  1. global deterministic EnKFs (EnSRF, ETKF),
  2. local EnKFs (LEnSRF, LETKF),
  3. global iterative EnKF (IEnKF).

- The ensemble update (analysis) can be decomposed into a two-step scheme:
  1. Update the state part of the ensemble $\mathbf{E}_\mathrm{x}^\mathrm{f} \longrightarrow \mathbf{E}_\mathrm{x}^\mathrm{a}$ using an EnKF.
  2. Update the parameter part of the ensemble:

$$\mathbf{E}_\mathrm{p}^\mathrm{a} = \mathbf{E}_\mathrm{p}^\mathrm{f} + \mathbf{B}_\mathrm{px}\mathbf{B}_\mathrm{xx}^{-1}\left(\mathbf{E}_\mathrm{x}^\mathrm{a} - \mathbf{E}_\mathrm{x}^\mathrm{f}\right),$$

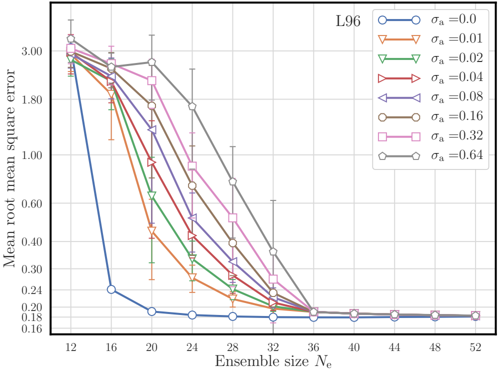
  which can be computed:
  (i) solving the linear system $\mathbf{B}_\mathrm{xx}\boldsymbol{\Delta} = \mathbf{E}_\mathrm{x}^\mathrm{a} - \mathbf{E}_\mathrm{x}^\mathrm{f}$, and
  (ii) updating $\mathbf{E}_\mathrm{p}^\mathrm{a} = \mathbf{E}_\mathrm{p}^\mathrm{f} + \mathbf{B}_\mathrm{px}\boldsymbol{\Delta}$.

The augmented dynamics

▶ Augmented dynamics (model persistence or Brownian motion):

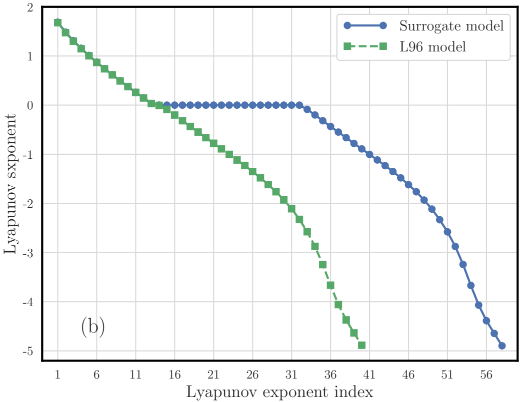
$$\begin{bmatrix} \mathbf{x}_k \\ \mathbf{p}_k \end{bmatrix} \mapsto \begin{bmatrix} \mathbf{F}^k(\mathbf{x}_k, \mathbf{p}_k) \\ \mathbf{p}_k \end{bmatrix}$$

▶ 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: $N_\mathrm{e} \gtrapprox N_0 + N_\mathrm{p} + 1.$



The augmented dynamics: Asymptotic properties

- Lyapunov spectra for the true and augmented L96.



Local EnKF-ML

- Covariance localisation/LEnSRF:

$$\mathbf{E}_\mathrm{x}^\mathrm{f} \mapsto \mathbf{E}_\mathrm{x}^\mathrm{a} \quad \text{(LEnSRF)}, \qquad \mathbf{E}_\mathrm{p}^\mathrm{a} = \mathbf{E}_\mathrm{p}^\mathrm{f} + \mathbf{B}_\mathrm{px}\mathbf{B}_\mathrm{xx}^{-1}\left(\mathbf{E}_\mathrm{x}^\mathrm{a} - \mathbf{E}_\mathrm{x}^\mathrm{f}\right) \quad \text{(global)},$$

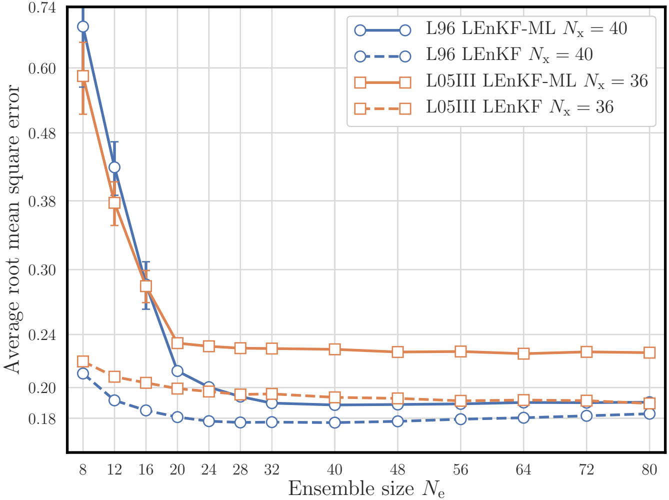
which can be computed:
(i) solving the linear system $\mathbf{B}_\mathrm{xx}\boldsymbol{\Delta} = \mathbf{E}_\mathrm{x}^\mathrm{a} - \mathbf{E}_\mathrm{x}^\mathrm{f}$, and
(ii) updating $\mathbf{E}_\mathrm{p}^\mathrm{a} = \mathbf{E}_\mathrm{p}^\mathrm{f} + \mathbf{B}_\mathrm{px}\boldsymbol{\Delta}$,
with $\mathbf{B}_\mathrm{xx} = \mathbf{C}_\mathrm{xx} \circ \left(\mathbf{X}_\mathrm{x}^\mathrm{f}\left(\mathbf{X}_\mathrm{x}^\mathrm{f}\right)^\top\right)$, $\mathbf{B}_\mathrm{px} = \zeta\mathbf{X}_\mathrm{p}^\mathrm{f}\left(\mathbf{X}_\mathrm{x}^\mathrm{f}\right)^\top$,
and ζ is the *tapering state/parameter coefficient* [Ruckstuhl et al. 2018].

- Domain localisation/LETKF:

$$\mathbf{E}_\mathrm{x}^\mathrm{f} \mapsto \mathbf{E}_\mathrm{x}^\mathrm{a} \quad \text{(LETKF)}, \qquad \mathbf{E}_\mathrm{p}^\mathrm{a} = \mathbf{E}_\mathrm{p}^\mathrm{f} + \zeta\mathbf{X}_\mathrm{p}^\mathrm{f}\left(\mathbf{X}_\mathrm{x}^\mathrm{f}\right)^\dagger\left(\mathbf{E}_\mathrm{x}^\mathrm{a} - \mathbf{E}_\mathrm{x}^\mathrm{f}\right) \quad \text{(local)},$$

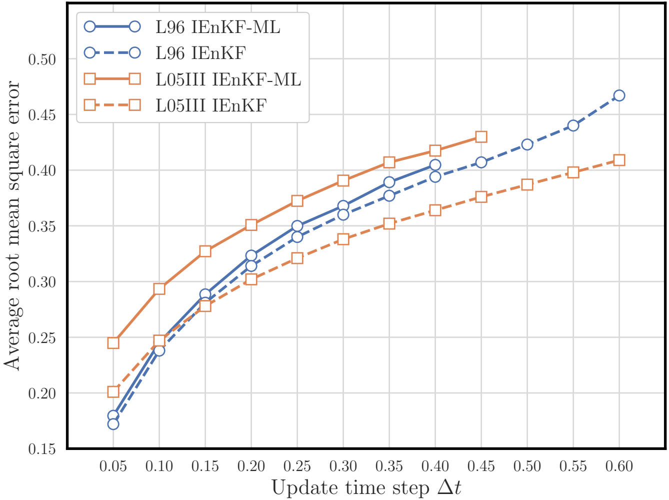
Numerics:

- LEnSRF and LEnSRF-ML applied to the L96 and L05III models.



Numerics:

▶ To deal with stronger model nonlinearities, we developed the IEnKF-ML variant of the IEnKF [Sakov et al. 2012; Bocquet et al. 2012; Sakov et al. 2018].

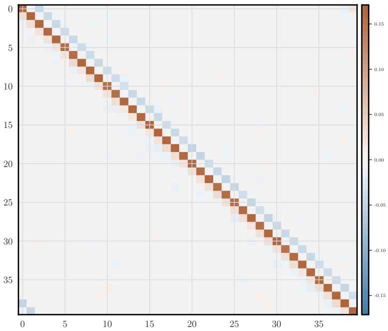


Outline

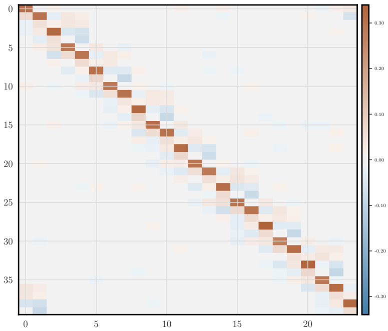
Learning 3D-Var

▶ The Bayesian formalism presented earlier can be extended to learning some key elements of the DA scheme (observed but unknown state, model known or unknown).

▶ Application to learning the Kalman gain equivalent for a 3D-Var scheme applied to the (known here) L96 model, using significantly noisy and partial observations of the state trajectory (24/40).



Fully observed



Partially observed

Outline

Conclusions

Almost all results presented here are from [Bocquet et al. 2019; Brajard et al. 2020a; Bocquet et al. 2020b; Bocquet et al. 2020a].

- Main messages:
  - Bayesian DA view on state and model estimation.
    DA can address goals assigned to ML but with partial & noisy observations.
  - The EM technique, full or approximate, is successful. Only coordinate minimization was shown to be successful so far in such context.
  - The method can handle very long training windows within a variational framework.
  - Online EnKFs-ML can also be used to sequentially estimate both state and model.
  - Successful on various 1D low-order models (L63, L96, KS, L05III).
  - Part of the DA scheme itself could additionally be learned.

- Open questions and technical hardships (non-exhaustive):
  - Non-autonomous dynamics?
  - Implicit integration schemes?
  - More complex models?
  - Surrogate model=knowledge-based+NN? $\longrightarrow$ A. Carrassi talk + A. Farchi poster.

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] M. Bocquet, A. Farchi, and Q. Malartic. "Online learning of both state and dynamics using ensemble Kalman filters". In: *Foundations of Data Science* 0 (2020). Accepted for publication, pp. 00–00.

[3] M. Bocquet and P. Sakov. "Combining inflation-free and iterative ensemble Kalman filters for strongly nonlinear systems". In: *Nonlin. Processes Geophys.* 19 (2012), pp. 383–399.

[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] 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.

[7] J. Brajard et al. "Combining data assimilation and machine learning to infer unresolved scale parametrisation". In: *Philosophical Transactions A* 0 (2020). Accepted for publication, p. 0. eprint: `arXiv:2009.04318`.

[8] S. L. Brunton, J. L. Proctor, and J. N. Kutz. "Discovering governing equations from data by sparse identification of nonlinear dynamical systems". In: *PNAS* (2016), p. 201517384.

[9] A. P. Dempster, N. M. Laird, and D. B. Rubin. "Maximum likelihood from incomplete data via the EM algorithm". In: *Journal of the Royal Statistical Society. Series B* (1977), pp. 1–38.

[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] R. Fablet, S. Ouala, and C. Herzet. "Bilinear residual neural network for the identification and forecasting of dynamical systems". In: *EUSIPCO 2018, European Signal Processing Conference*. Rome, Italy, 2018, pp. 1–5.

[12] A. Farchi et al. "Using machine learning to correct model error in data assimilation and forecast applications". In: *Q. J. R. Meteorol. Soc.* 0 (2020). Submitted, p. 0.

[13] 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.

References II

[14] J. Harlim. *Data-driven computational methods: parameter and operator estimations*. Cambridge University Press, Cambridge, 2018, p. 158.

[15] 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.

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

[17] Y. A. LeCun et al. "Efficient backprop". In: *Neural networks: Tricks of the trade*. Springer, 2012, pp. 9–48.

[18] R. Lguensat et al. "The Analog Data Assimilation". In: *Mon. Wea. Rev.* 145 (2017), pp. 4093–4107.

[19] Z. Long et al. "PDE-Net: Learning PDEs from Data". In: *Proceedings of the 35th International Conference on Machine Learning*. 2018.

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

[21] J. Paduart et al. "Identification of nonlinear systems using polynomial nonlinear state space models". In: *Automatica* 46 (2010), pp. 647–656.

[22] D. C. Park and Y. Zhu. "Bilinear recurrent neural network". In: *Neural Networks, 1994. IEEE World Congress on Computational Intelligence., 1994 IEEE International Conference on*. Vol. 3. 1994, pp. 1459–1464.

[23] J. Pathak et al. "Model-Free Prediction of Large Spatiotemporally Chaotic Systems from Data: A Reservoir Computing Approach". In: *Phys. Rev. Lett.* 120 (2018), p. 024102.

[24] J. Pathak et al. "Using machine learning to replicate chaotic attractors and calculate Lyapunov exponents from data". In: *Chaos* 27 (2017), p. 121102.

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

[26] P. N. Raanes, A. Carrassi, and L. Bertino. "Extending the square root method to account for additive forecast noise in ensemble methods". In: *Mon. Wea. Rev.* 143 (2015), pp. 3857–38730.

[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] P. Sakov, J.-M. Haussaire, and M. Bocquet. "An iterative ensemble Kalman filter in presence of additive model error". In: *Q. J. R. Meteorol. Soc.* 144 (2018), pp. 1297–1309.

References III

[30] P. Sakov, D. S. Oliver, and L. Bertino. “An iterative EnKF for strongly nonlinear systems”. In: *Mon. Wea. Rev.* 140 (2012), pp. 1988–2004.

[31] P. R. Vlachas et al. “Backpropagation algorithms and Reservoir Computing in Recurrent Neural Networks for the forecasting of complex spatiotemporal dynamics”. In: *Neural Networks* 126 (2020), pp. 191–217.

[32] Y.-J. Wang and C.-T. Lin. “Runge-Kutta neural network for identification of dynamical systems in high accuracy”. In: *IEEE Transactions on Neural Networks* 9 (1998), pp. 294–307.

[33] G. C. G. Wei and M. A. Tanner. “A Monte Carlo implementation of the EM algorithm and the poor man’s data augmentation algorithms”. In: *Journal of the American Statistical Association* 85 (1990), pp. 699–704.