Data-driven reconstruction of chaotic dynamics using data assimilation and machine learning

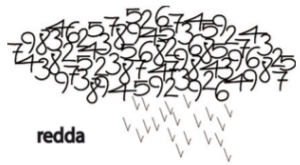
Marc Bocquet[1], Julien Brajard[2,3], Alberto Carrassi[3,4] & Laurent Bertino[3]

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

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

(3) Nansen Environmental and Remote Sensing Center

(4) University of Utrecht



Outline

From model error to the absence of a model

▶ Data assimilation and model error

Numerical predictions in geophysics based on data assimilation crucially depends on both initial condition and model error [Magnusson et al., 2013]. There are methods to mitigate model error:

- additive noise (weak parametrisation) [Trémolet, 2006; Raanes et al., 2015; Sakov et al. 2018]
- estimation of uncertain model parameters
- physically-driven stochastic perturbations [e.g., Buizza et al., 1999], stochastic subgrid parametrisations [e.g., Resseguier et al., 2017], inflation [e.g., Raanes et al., 2019]

▶ Data-driven forecast of a physical system

One step further: renounce physically-based models and use massive observation

- use data assimilation together with analogues [Lguensat et al., 2017]
- use diffusion maps for a spectral representation of datasets [e.g., Harlim, 2018]
- use neural networks (NNs), echo states networks, & deep learning [e.g., Park et al., 1994; Pathak et al, 2017; Dueben et al., 2018]

Building a surrogate model

- Learning the dynamics of a model from its output
  - more explicit (possibly with NNs) representations of the dynamics using specific regressors [e.g., Paduart et al., 2010; Brunton et al. 2016]
  - design NNs that mimic integration schemes [Wang and Lin, 1998; Fablet et al., 2018; Long et al., 2018]

- Our goal
  - Use a data assimilation framework to infer both a surrogate model and the state trajectory within a time window over which the reference model is only partially & noisily observed.

Outline

ODE representation for the surrogate model

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

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

where

- $\mathbf{A}$ is a matrix of coefficients of size $N_x \times N_p$
- $\mathbf{r}(\mathbf{x})$ is a vector of nonlinear regressors of size N_p. For instance, for one-dimensional spatial systems and up to bilinear order:

$$\mathbf{r}(\mathbf{x}) = \left[1, \{x_n\}_{0 \leqslant n < N_x}, \{x_n x_m\}_{0 \leqslant n \leqslant m < N_x}\right].$$

  A priori, $N_p = \binom{N_x+1}{2} = \frac{1}{2}(N_x+1)(N_x+2)$ such regressors.

$\longrightarrow$ Intractable in high-dimension! (typically $N_x \approx 10^6$ and beyond)

Assumptions and symmetries

▶ Locality

Physical locality of the physics: all multivariate monomials in the ODEs have variables x_n that belong to a stencil, i.e. a local arrangement of grid points around a given node.

- s_n is the stencil around node n, the pattern being the same for all nodes.
- the set of required monomials can therefore be reduced to (in 1D)

$$\mathbf{r}(\mathbf{x}) = \left[1, \{x_n\}_{0 \leqslant n < N_x}, \{x_n x_m\}_{0 \leqslant n \leqslant m < N_x, m \in s_n}\right].$$

  In 1D and with a stencil of size $2L+1$, there are $N_p = 1 + N_x(2+L)$ monomials.

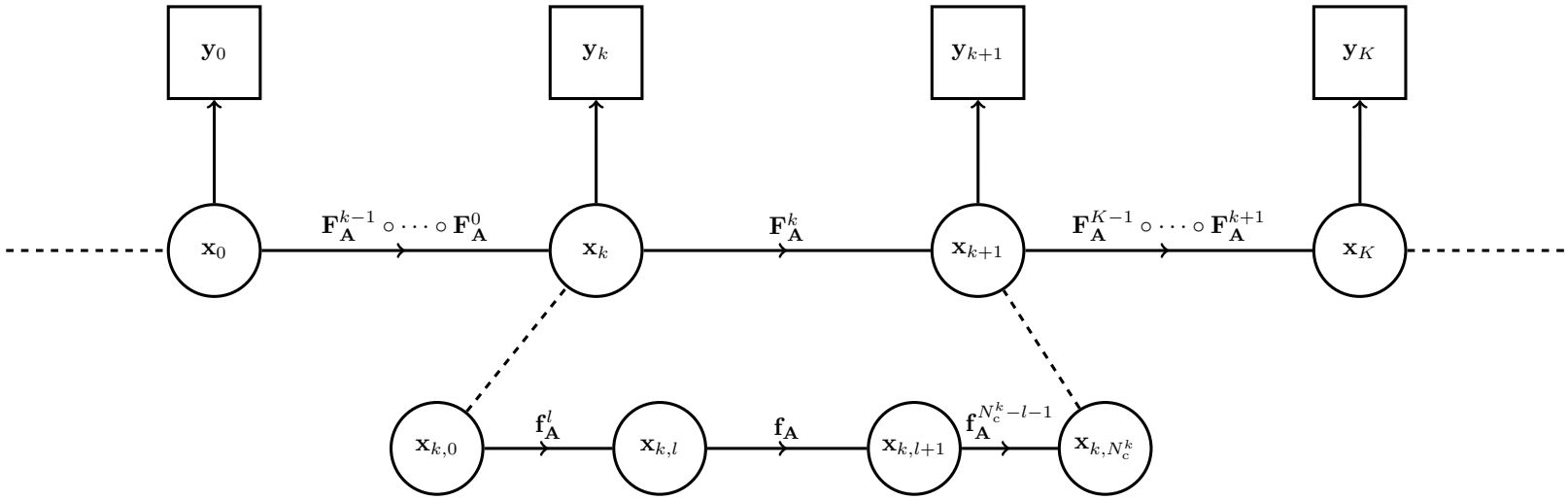
- $\mathbf{A}$ becomes sparse and can be squeezed into a dense rearrangement of $\mathbf{A}$. In 1D and with a stencil of size $2L+1$, the size of the dense $\mathbf{A}$ is

$$N_x \times N_a \quad \text{where} \quad N_a = \sum_{l=L+1}^{2L+2} l = \frac{3}{2}(L+1)(L+2).$$

▶ Homogeneity

Moreover, we can additionally assume translational invariance. In that case $\mathbf{A}$ becomes a vector of size N_a.

Integration scheme and cycling



- Compositions of integration schemes:

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

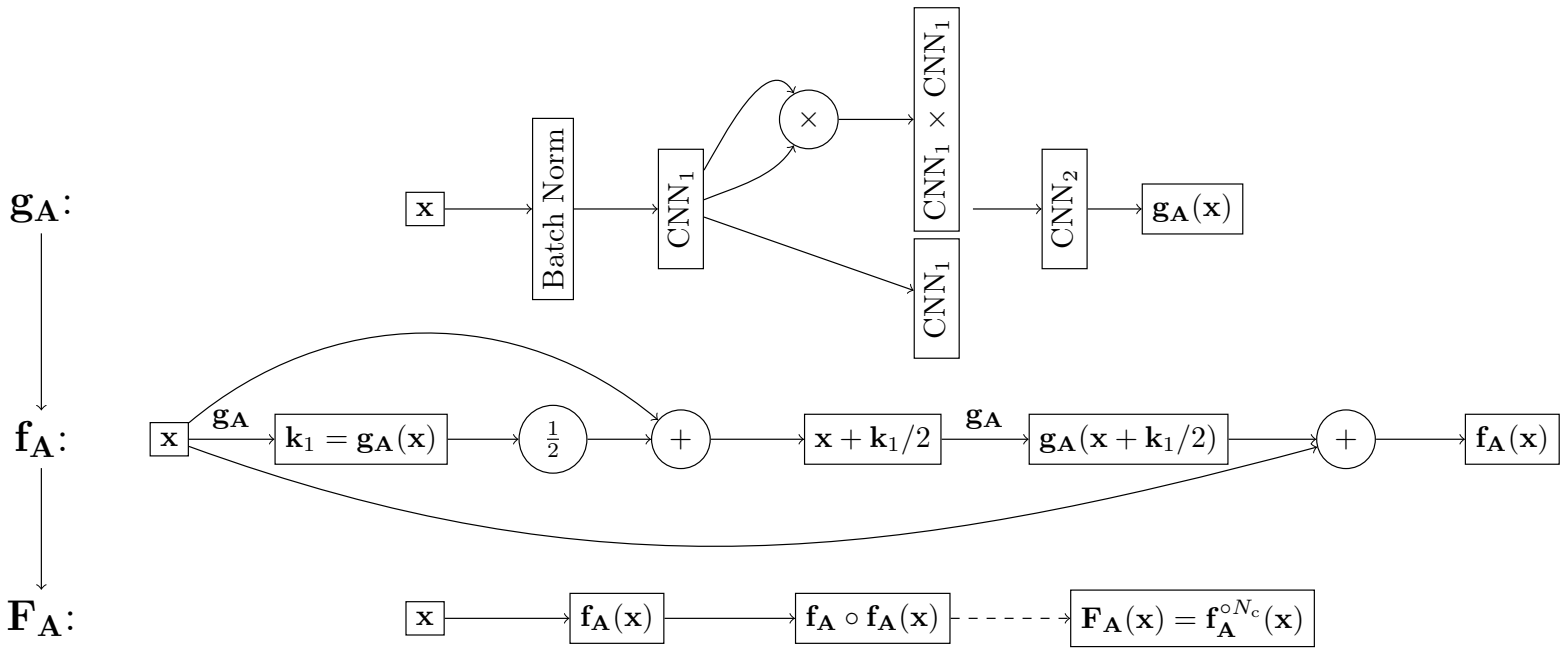
- Choosing a Runge-Kutta method as integration scheme:

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

Outline

Neural network models

- We tested many simple architectures, all following the structure of N_c Runge-Kutta schemes, with linear or nonlinear activation functions:
  - The thin algebraic representation above does not rely on ML libraries (TensorFlow, PyTorch, etc.). It was also implemented as an NN.
  - 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:

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

$\longrightarrow$ 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 and Tang 1998; Abarbanel et al. 2018].

Bayesian analysis of the joint problem

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

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

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

(2) ▶ The optimisation of $\mathcal{J}(\mathbf{A}, \mathbf{x}_{0:K} | \mathbf{Q}_{1:K})$ can be solved using a coordinated descent.

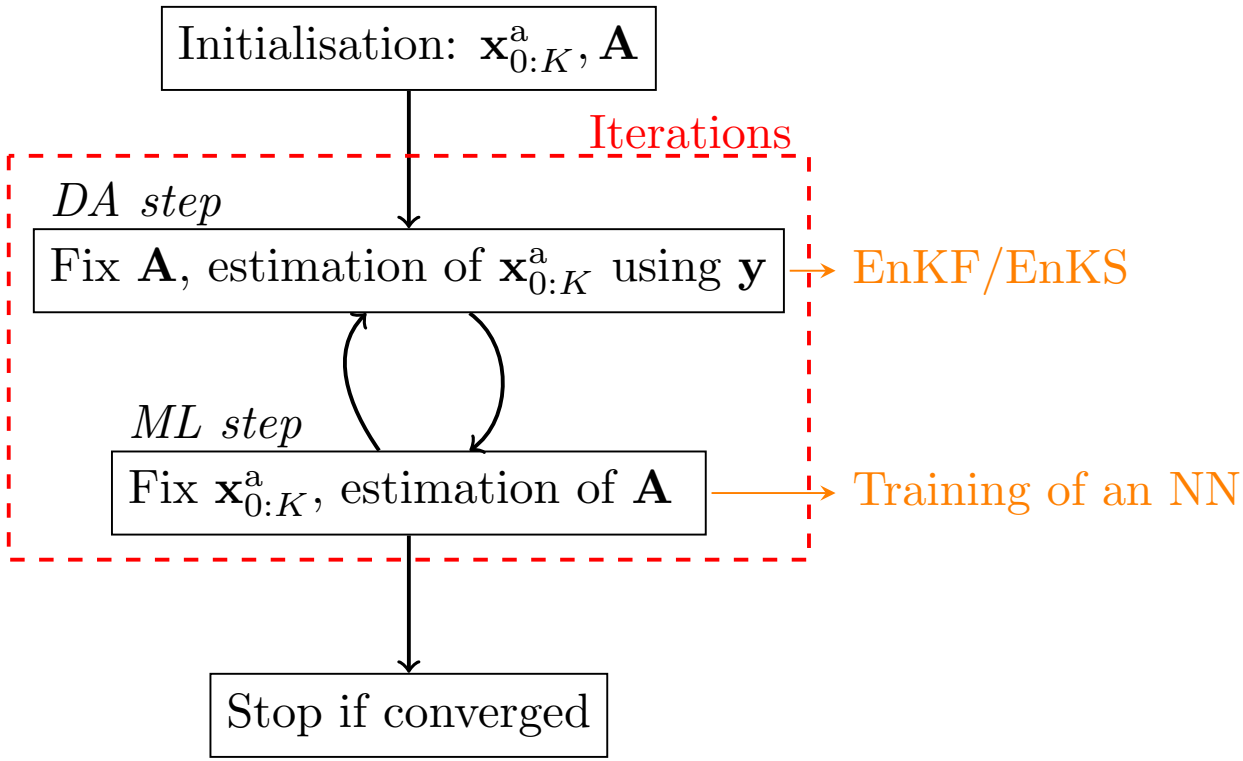
- ▶ For $\mathcal{J}(\mathbf{A}, \mathbf{x}_{0:K} | \mathbf{Q}_{1:K})$: using a weak constraint 4D-Var for $\mathbf{x}_{0:K}$ and a variational optimisation problem for $\mathbf{A}$ [Bocquet et al. 2019b].

- ▶ For $\mathcal{J}(\mathbf{A}, \mathbf{x}_{0:K} | \mathbf{Q}_{1:K})$: using an EnKF for $\mathbf{x}_{0:K}$ and a variational optimisation problem for $\mathbf{A}$ [Brajard et al. 2019].

Bayesian analysis of the joint problem

- Coordinated descent of [Brajard et al. 2019].

Hybrid data assimilation and machine learning techniques.



- The coordinated descent algorithm is interpreted as an Expectation-Maximisation (EM) algorithm by [Nguyen et al. 2019].

Bayesian analysis of the marginal problem

- Looking only for the dynamics and its model error:

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

- A solution is provided by the EM algorithm. Applying it for the reconstruction of a dynamical system has been suggested in [Ghahramani and Roweis 1999], using an extended Kalman smoother, or for the estimation of subgrid stochastic processes in [Pulido et al. 2018] using an ensemble Kalman smoother.

- Here we solve for the MAP of $p(\mathbf{A}, \mathbf{Q}_{1:K} | \mathbf{y}_{0:K}, \mathbf{R}_{0:K})$ using iterations over:

(1) - Expectation/DA step: EnKS over a long period $[t_0, t_K]$

(2) - A coordinated descent over

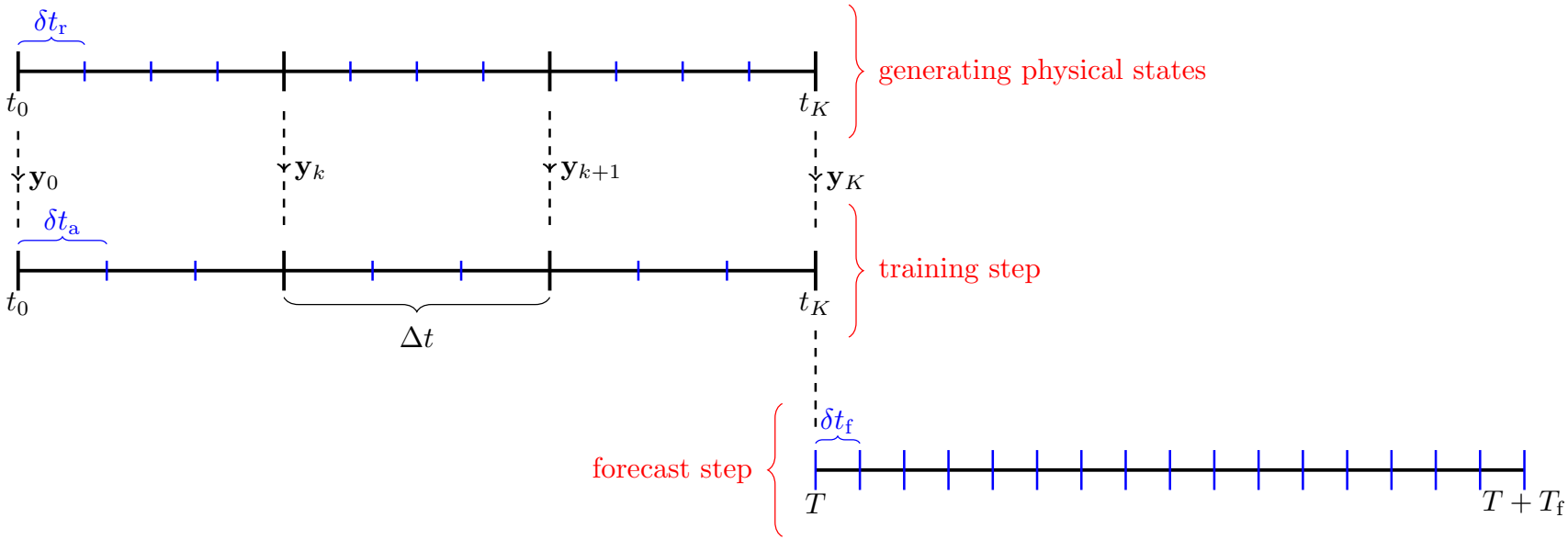
- (i) ML/deep learning step: variational solution of $\mathbf{A}$
- (ii) Maximisation step: variational solution of $\mathbf{Q}_{1:K}$

[Bocquet et al. 2019a]

Outline

Experiment plan

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



- Metrics of comparison:
  - Model: ODE coefficients norm $\|\mathbf{A}_{\mathrm{a}} - \mathbf{A}_{\mathrm{r}}\|_{\infty}$.
  - NRMSE between the reference and the surrogate forecasts as a function of the lead time (averaged over many initial conditions).
  - Lyapunov spectrum.
  - Power spectrum density.

Identifiable model and perfect observations

- Inferring the dynamics from dense & noiseless observations of identifiable models
  - The Lorenz 63 model (L63, 3 variables):

$$\frac{\mathrm{d}x_0}{\mathrm{d}t} = \sigma(x_1 - x_0),$$
$$\frac{\mathrm{d}x_1}{\mathrm{d}t} = \rho x_0 - x_1 - x_0 x_2,$$
$$\frac{\mathrm{d}x_2}{\mathrm{d}t} = \rho x_0 x_1 - \beta x_2,$$

  $\longrightarrow \|\mathbf{A}_\mathrm{a} - \mathbf{A}_\mathrm{r}\|_\infty \sim 10^{-13}$ close to perfect reconstruction at machine precision.

  - The Lorenz 96 model (L96, 40 variables)

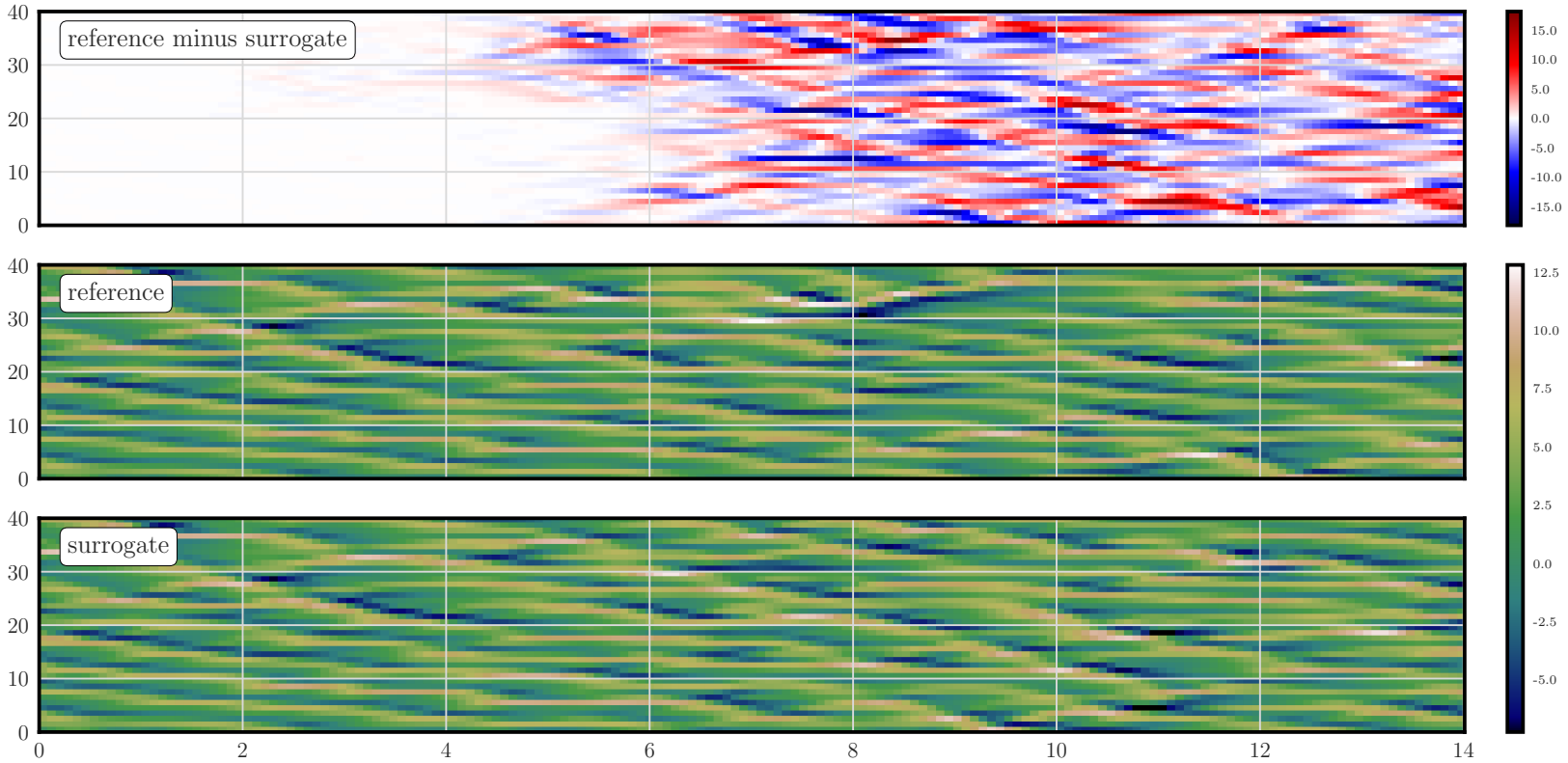
$$\frac{\mathrm{d}x_n}{\mathrm{d}t} = (x_{n+1} - x_{n-2})x_{n-1} - x_n + F,$$

  $\longrightarrow \|\mathbf{A}_\mathrm{a} - \mathbf{A}_\mathrm{r}\|_\infty \sim 10^{-13}$ close to perfect reconstruction at machine precision.

Non-identifiable model and perfect observations

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

The Lorenz 96 model (40 variables). Surrogate model based on an RK2 scheme.
Analysis of the modelling depth as a function of N_c.

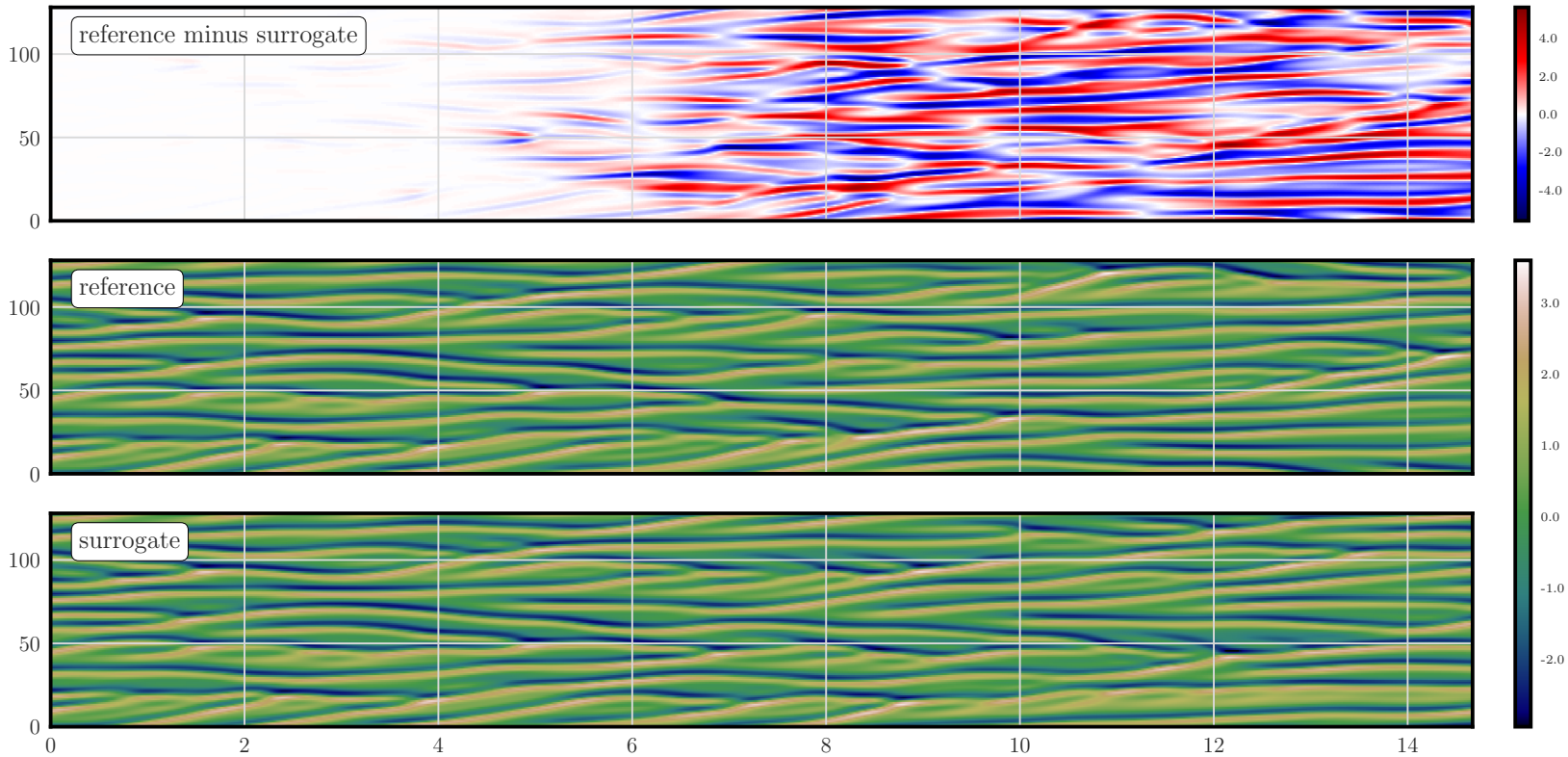


Non-identifiable model and perfect observations

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

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

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

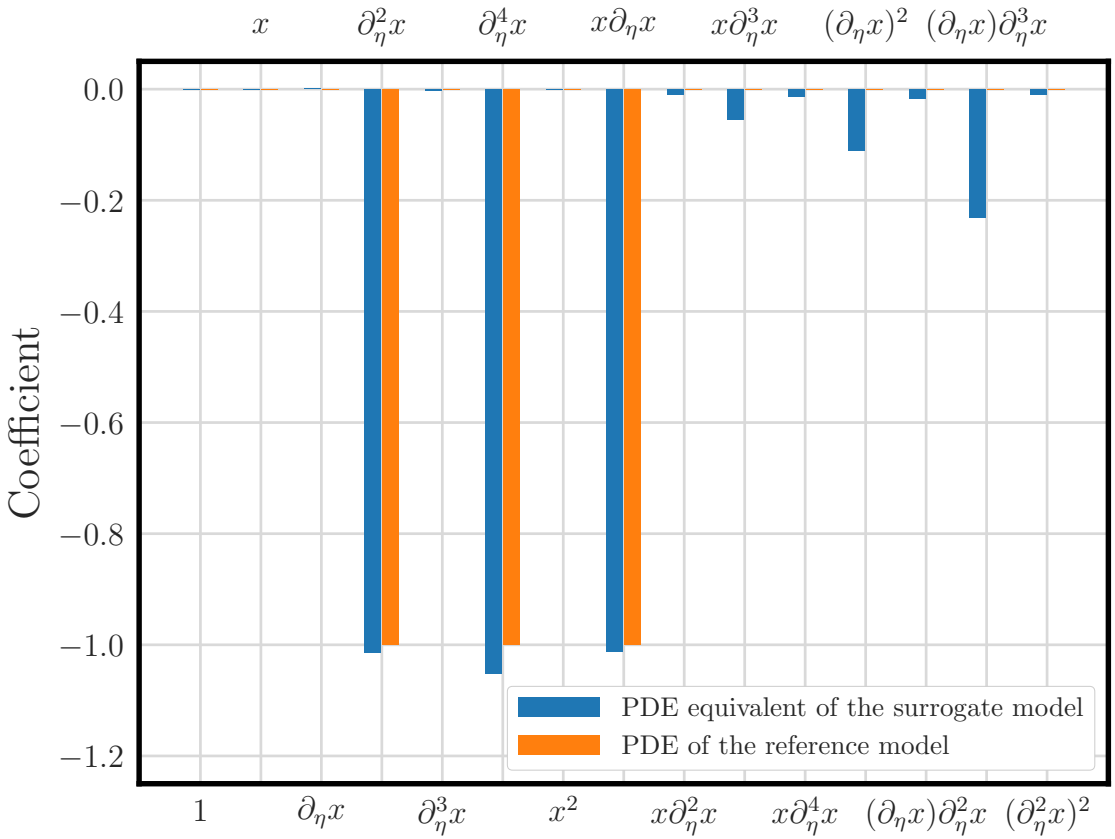


Non-identifiable model and perfect observations

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

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

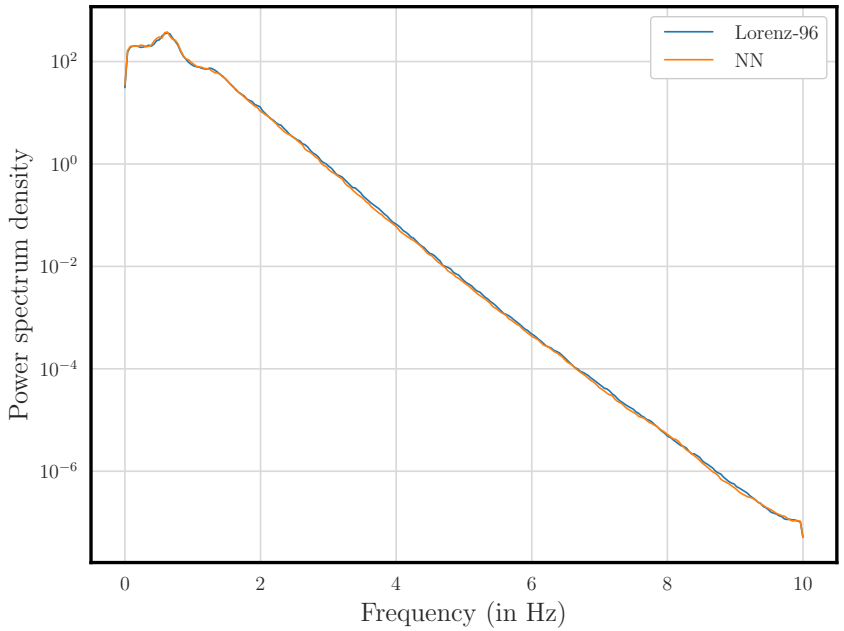
$$\frac{\partial u}{\partial t} = -u\frac{\partial u}{\partial x} - \frac{\partial^2 u}{\partial x^2} - \frac{\partial^4 u}{\partial x^4},$$

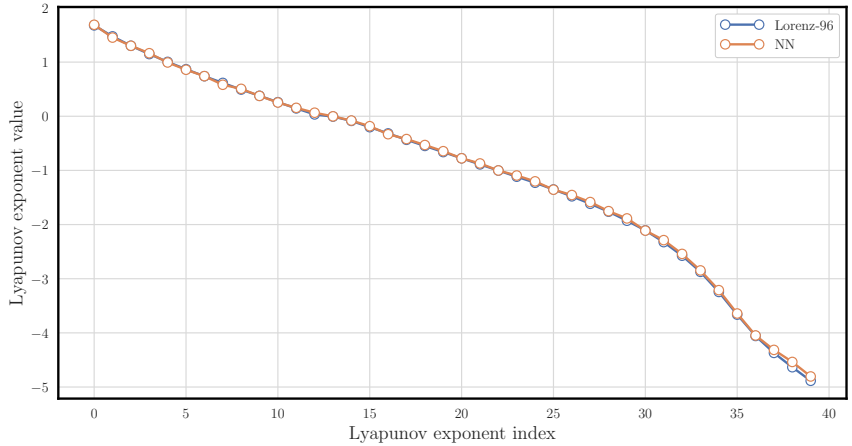


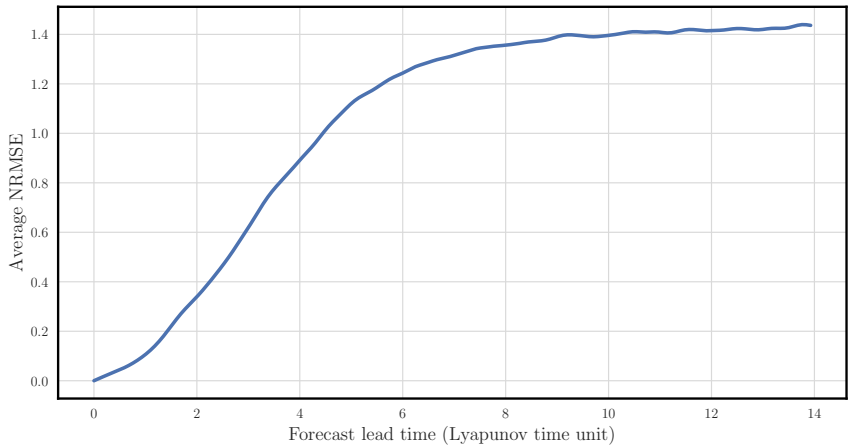
Identifiable model and imperfect observations

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

  - Fully observed L96.
  - Significantly noisy observations $\mathbf{R} = \mathbf{I}$
  - Long window $K = 10^4$, with $\Delta t = 0.05$
  - RK4 residual convolutional NN
  - 10 EM iterations
  - Takes a few mins on a GTX 1070 Ti





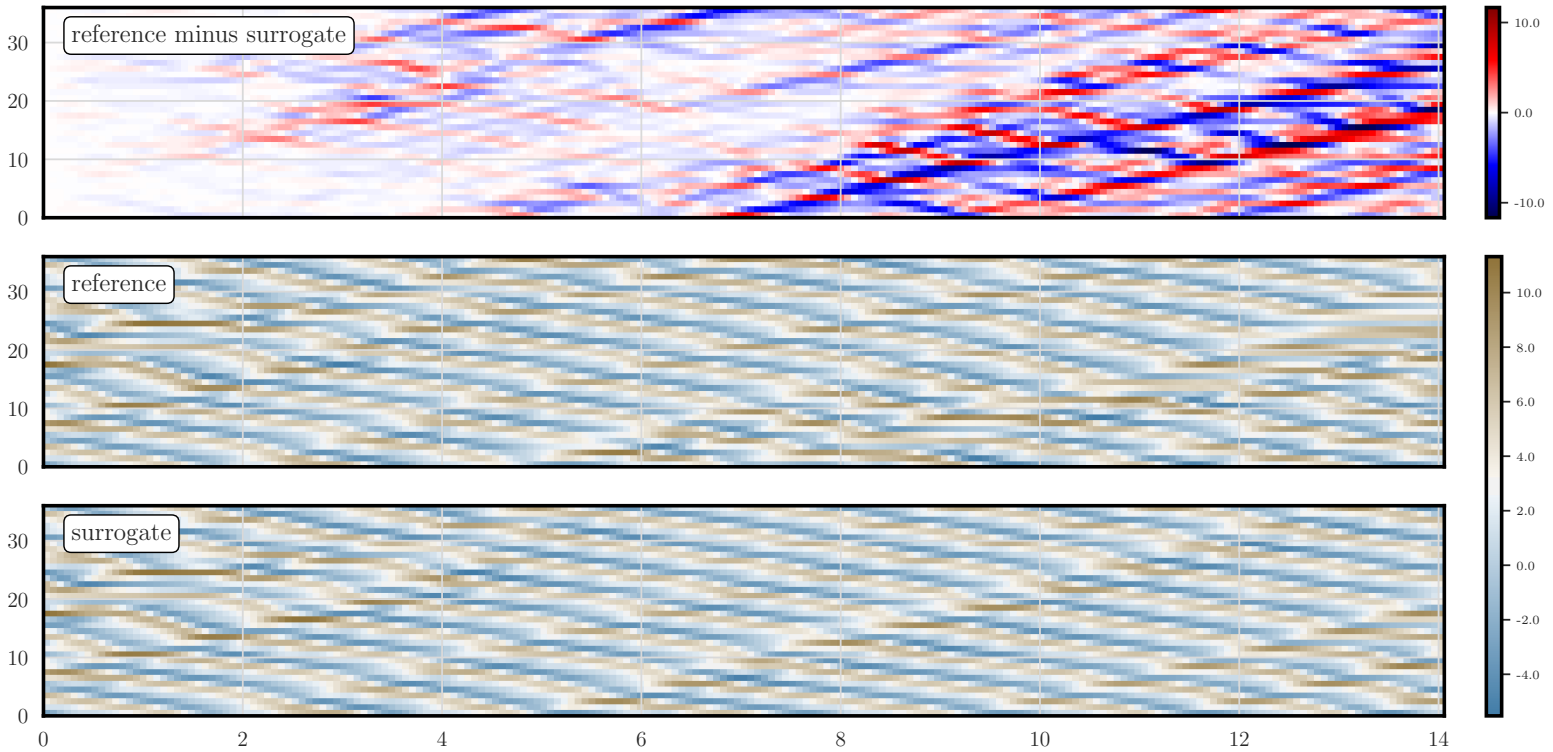


Non-identifiable model and imperfect observations

- The Lorenz 05III (two-scale) model (36 slow & 360 fast variables).

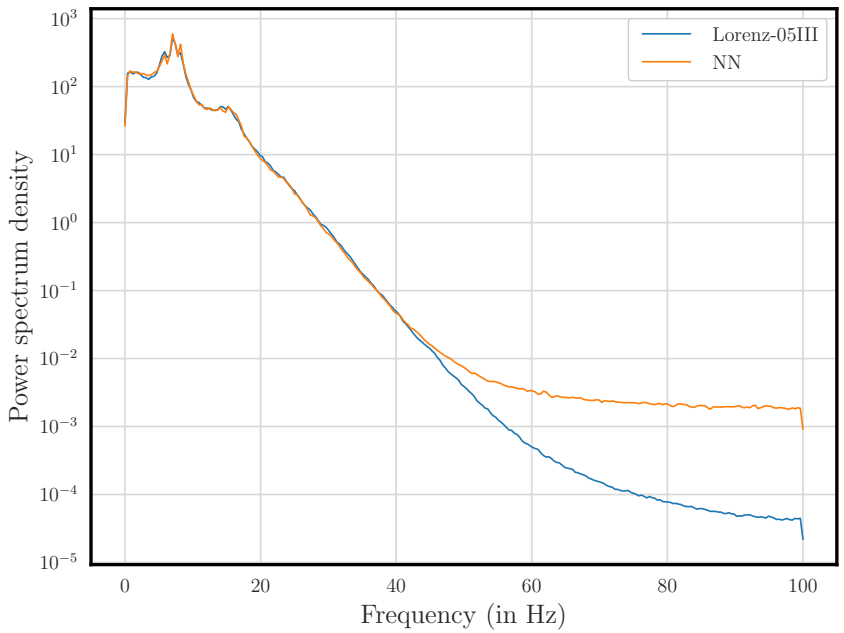
$$\frac{\mathrm{d}x_n}{\mathrm{d}t} = \psi_n^+(\mathbf{x}) + F - h\frac{c}{b}\sum_{m=0}^{9} u_{m+10n},$$

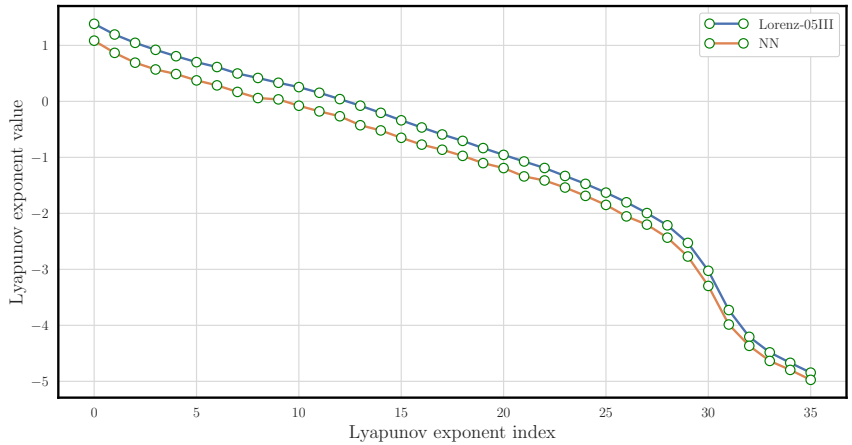
$$\frac{\mathrm{d}u_m}{\mathrm{d}t} = \frac{c}{b}\psi_m^-(b\mathbf{u}) + h\frac{c}{b}x_{m/10}, \quad \text{with} \quad \psi_n^\pm(\mathbf{x}) = x_{n\mp1}(x_{n\pm1} - x_{n\mp2}) - x_n,$$

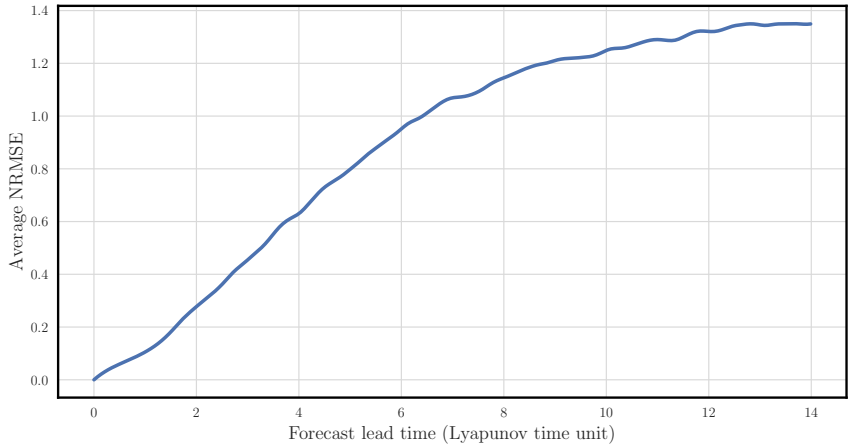


Non-identifiable model and imperfect observations

- Fully observed L05III.
- Significantly noisy observations $\mathbf{R} = \mathbf{I}$
- Long window $K = 10^4$, with $\Delta t = 0.05$
- RK4 residual convolutional NN
- 10 EM iterations
- Takes a few mins on a GTX 1070 Ti







[Bocquet et al. 2019a]

Outline

Conclusions

- Main messages:
  - Bayesian DA view on state and model estimation.
    DA can address goals assigned to ML but with partial & noisy observations.
  - Numerical costs of high-dimensional systems significantly reduced by locality and homogeneity assumptions.
  - Full EM technique (not only coordinated descent) successful.
  - The method can handle very long training windows.
  - Successful on various 1D low-order models (L63, L96, KS, L05III) in presence of partial observation with significant noise.

- Open questions and technical hardships (non-exhaustive):
  - Non-autonomous dynamics?
  - Implicit integration schemes?
  - Online learning scheme?
  - More complex models?

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 et al. "Bayesian inference of dynamics from partial and noisy observations using data assimilation and machine learning". In: (2019). in preparation.

[3] 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.

[4] 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: *Geosci. Model Dev. Discuss.* 2019 (2019), pp. 1–21.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

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

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

[12] 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.

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

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

References II

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

[16] 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.