Beyond one iteration of machine learning and data assimilation steps for learning meteorological models?

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

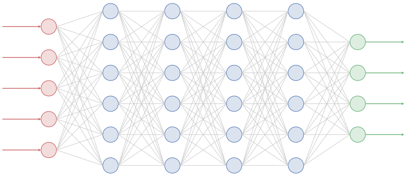
[†] CEREA, École des Ponts and EDF R&D, Île-De-France, France
[‡] ECMWF, Reading, United Kingdom.

Learning data-driven models from 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 from the loss function:

$$\mathcal{J}(\mathbf{p}) = \sum_{k=1}^{K} \left\| \mathbf{x}_k - M_k\left(\mathbf{p}, \mathbf{x}_{k-1}\right) \right\|^2$$

▶ The surrogate model to be learned $\mathcal{M}$ 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 goesciences, observations are usually sparse and noisy: we need data assimilation!

Machine learning for the geosciences with sparse and noisy observations

- A rigorous Bayesian formalism for this problem:[1]

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

- This resembles a typical *weak-constraint 4D-Var* cost function!

- Machine learning limit

If the physical system is fully and directly observed, i.e. $\mathbf{H}_k \equiv \mathbf{I}$, and if 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, $\mathcal{J}(\mathbf{p},\mathbf{x}_{0:K})$ becomes

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

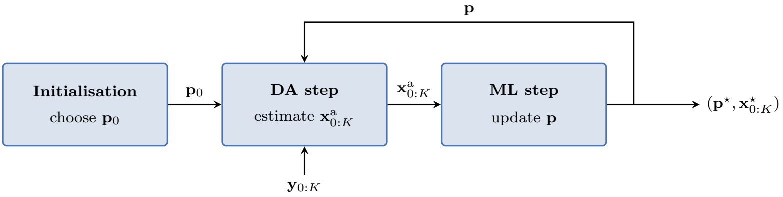
[1] [Bocquet et al. 2019; Bocquet et al. 2020; Brajard et al. 2020] in the wake of [Hsieh et al. 1998; Abarbanel et al. 2018]

Machine learning for the geosciences with sparse and noisy observations

- We need to minimise this cost function on both states and parameters:

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

- *DA* is used to estimate the state and then *ML* is used to estimate the model (*coordinate descent*):



DA: 4D-Var, WC 4D-Var, EnKS, IEnKS, etc. and *ML*: Neural network.

- This DA standpoint is remarkable as it allows for *noisy and partial observations* of the physical system.

Machine learning for the geosciences with sparse and noisy observations

▶ 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:

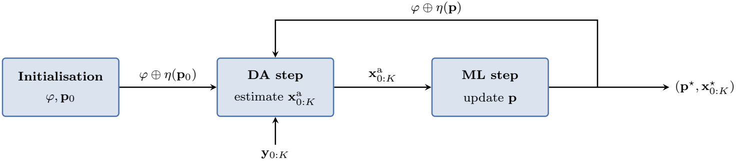
$$\begin{aligned}\mathcal{J}(\mathbf{p}, \mathbf{x}_{0:K}, \mathbf{Q}_{1:K}) &= -\ln p(\mathbf{p}, \mathbf{x}_{0:K}, \mathbf{Q}_{1:K} | \mathbf{y}_{0:K}, \mathbf{R}_{0:K}) \\ &= \frac{1}{2}\sum_{k=0}^{K}\left\{\|\mathbf{y}_k - H_k(\mathbf{x}_k)\|^2_{\mathbf{R}_k^{-1}} + \ln|\mathbf{R}_k|\right\} \\ &\quad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

▶ This problem can (almost) fully be solved from a Bayesian standpoint using the empirical Expectation-Maximisation algorithm with an ensemble smoother[2].
⟶ Very successful on low-order models[3], but it has a significant numerical cost.

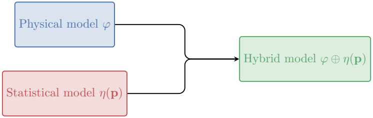
[2] [Ghahramani et al. 1999; Nguyen et al. 2019; Bocquet et al. 2020]
[3] [Bocquet et al. 2020]

Offline model correction

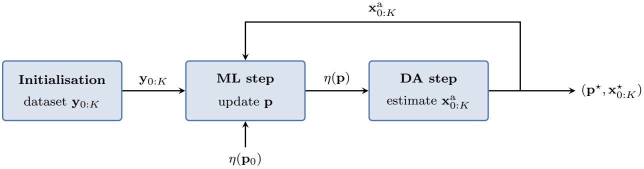
- *Partial and noisy observations* of a physical model with a *proxy of the dynamics* φ:



where the *hybrid* physical/statistical model is a combination of the proxy φ-model and a NN model:



- *Fully observed* system whose *dynamics are unknown*:



Loop order

▶ *Loop order–$1/2$*: This is traditional DA. The output is the analysed trajectory $\mathbf{x}^{\mathrm{a}}_{0:K}$ over the DA window. In an incremental operational setup, the usable output are the *analysis increments* $\mathbf{x}^{\mathrm{a}}_{k} - \mathbf{x}^{\mathrm{f}}_{k}$.

$$\longrightarrow [\mathbf{x}^{\mathrm{a}}_{0:K}]_1$$

▶ *Loop order–1*: The analysed trajectory/analysis increments are used to train a DL model η, possibly with the help of a known physical model φ.

$$\longrightarrow \varphi \oplus \eta_1$$

▶ *Loop order–$3/2$*: Within an hybrid model configuration the DL correction to the model can be used to carry out a re-analysis with the observations $\mathbf{y}_{0:K}$. We implemented this in several configurations from low-order model to realistic models (within OOPS).

$$\longrightarrow [\mathbf{x}^{\mathrm{a}}_{0:K}]_2$$

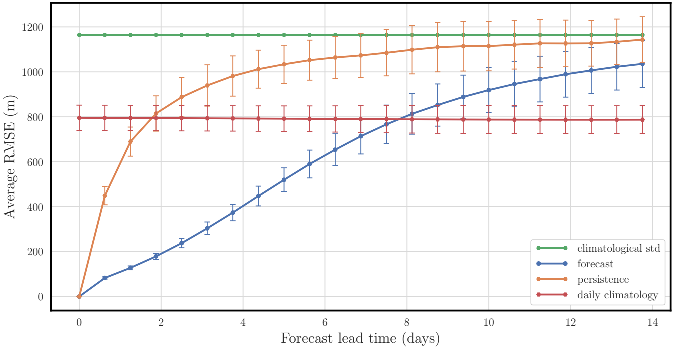
▶ *Loop order–2*: Using these new analysis increments, we can re-train the DL model to obtain an improved model correction. We showed this to be efficient on low-order model but not yet on more realistic, possibly already well-estimated, fields such as ERA-5 dataset.

$$\longrightarrow \varphi \oplus \eta_2$$

▶ *Loop order–$5/2$*, ...

Example of loop order–1 and $3/2$

- Learning a purely data-driven meteorological model from ERA-5 reanalysis
  - True model: A selection of ERA-5 fields in 1979-2018 at 0.5625°.[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]

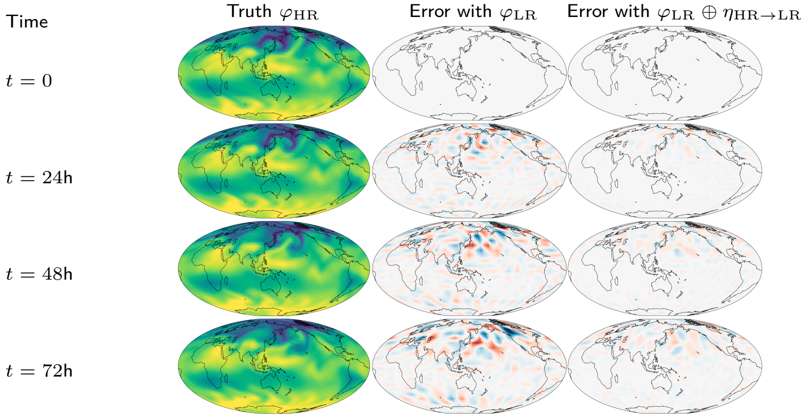


- η has also successfully been tested with DA, hence at *loop order–$3/2$*.

[4] [Rasp et al. 2020]
[5] [Bocquet et al. 2022]

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]

Online model error correction

▶ So far, the model error has been learned *offline*: a long analysis trajectory is required.

▶ We now investigate the possibility to perform *online* learning, *i.e.* improving the correction as new observations become available.

▶ To do this, we use the formalism of DA to estimate both the state and the NN parameters:[8]

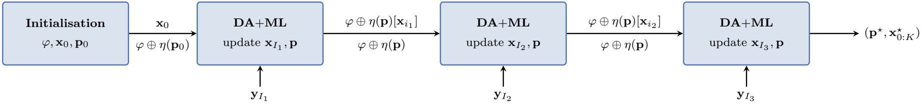
$$\mathcal{J}(\mathbf{p},\mathbf{x}) = \frac{1}{2}\left\|\mathbf{x}-\mathbf{x}^{\mathrm{b}}\right\|^2 + \frac{1}{2}\left\|\mathbf{p}-\mathbf{p}^{\mathrm{b}}\right\|^2_{\mathbf{B}_{\mathrm{p}}-1} + \frac{1}{2}\sum_{k=0}^{K}\left\|\mathbf{y}_k - \mathcal{H}_k\circ\mathcal{M}_k(\mathbf{p},\mathbf{x})\right\|^2_{\mathbf{R}_k^{-1}}.$$

▶ Potential cross-covariance between state and NN parameters are neglected in the prior.

[8] [Farchi et al. 2021; Bocquet et al. 2022]

Online model error correction

▶ Sequential DA, yet another approximate solution to the original Bayesian problem:



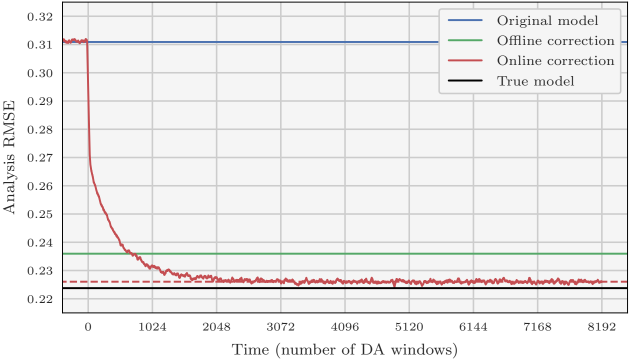
▶ Information is flowing from one window to the next using the prior for the state $\mathbf{x}^\mathrm{b}$ and for the NN parameters $\mathbf{p}^\mathrm{b}$.

▶ Already been investigated with an EnKF, with promising solutions but limitations as well.[9]

[9] [Bocquet et al. 2021; Malartic et al. 2022a]

Numerical illustration with the two-scale Lorenz model

- We use the tendency correction approach, with the same simple CNN as before, and still using 4D-Var.[10]



- 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!

[10] [Farchi et al. 2021]

Conclusions

- *Main messages:*
  - Bayesian DA view on joint state and model estimation.
    DA can address goals assigned to ML but with partial & noisy observations.
  - Offline or online optimisation strategies.
  - Loop order-N correction successful on 1D models (L96, L05III, L96i, mL96).

- *In progress: more ambitious models and datasets at higher loop order*
  - Application to the Marshall-Molteni 3-layer QG model on the sphere [order–$3/2$]
  - Application to the ERA5 and CMIP data (WeatherBench[11]-like challenge) [order–2]
  - Application to the ECMWF IFS [order–$3/2$][12]
    $\longrightarrow$ *Alban Farchi's talk and Marcin Chrust's poster*
  - Application to sea-ice surrogate modelling [order–$1/2$ & order–1]: Schmidt Futures/VESRI/SASIP project
    $\longrightarrow$ *Tobias Finn's and Simon Driscoll's posters*

[11] [Rasp et al. 2020]

[12] [Farchi et al. 2022]

References

[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* 3 (2021), pp. 305–330.

[3] 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.

[4] 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.

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

[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] 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.

[8] A. Farchi et al. "Online model error correction with neural networks in the incremental 4D-Var framework". In: *arXiv preprint arXiv:2210.13817* (2022).

[9] 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.

[10] 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.

[11] Q. Malartic, A. Farchi, and M. Bocquet. "State, global, and local parameter estimation using local ensemble Kalman filters: Applications to online machine learning of chaotic dynamics". In: *Q. J. R. Meteorol. Soc.* 148 (2022), pp. 2167–2193.

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

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

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

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