Combining machine learning and data assimilation to learn dynamics from sparse and noisy observations

Marc Bocquet[†], Quentin Malartic[†], Alban Farchi[†],
Tobias Finn[†], Charlotte Durand[†],
Massimo Bonavita[‡], Patrick Laloyaux[‡], Marcin Chrust[‡],
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.

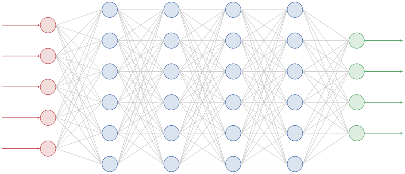
Outline

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 from the loss function:

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

- 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]

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

- This resembles a typical *weak-constraint 4D-Var* cost function!
- This DA standpoint is remarkable as it allows for noisy an partial observations on the physical system.
- *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

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

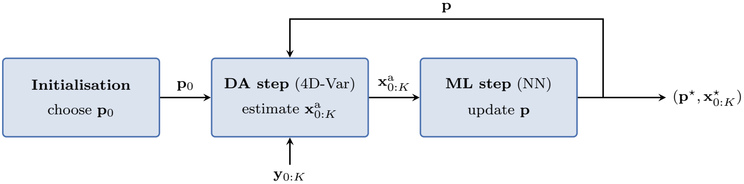
[1] [Bocquet et al. 2019; Bocquet et al. 2020; Brajard et al. 2020] in the wake of [Hsieh et al. 1998; Abarbanel et al. 2018]

Machine learning for the geosciences with sparse and noisy observations

- We need to minimise this cost function on both states and parameters:[2]

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

- *DA* is used to estimate the state and then *ML* is used to estimate the model:



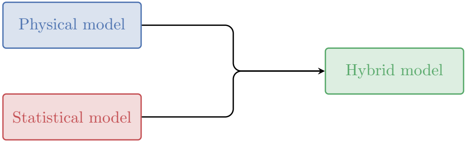
- This DA standpoint is remarkable as it allows for noisy an partial observations on the physical system.
- The problem can (almost) fully be solved from a Bayesian standpoint using the empirical Expectation-Maximization algorithm with an ensemble smoother[3]. But it has a significant numerical cost.

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

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

Hybrid models

- Even though NWP models are not perfect, they are already quite good!
- 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:[4]



- 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.*

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

Model integration and surrogate model architecture

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

$$\frac{\partial \mathbf{x}}{\partial t} = \phi(\mathbf{x}). \tag{1}$$

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

$$\mathbf{x}(t + \delta t) = \mathcal{F}\Big(\mathbf{x}(t)\Big). \tag{2}$$

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

$$\mathcal{M} : \mathbf{x}_k \mapsto \mathbf{x}_{k+1} = \mathcal{F} \circ \cdots \circ \mathcal{F}(\mathbf{x}_k). \tag{3}$$

Resolvent correction

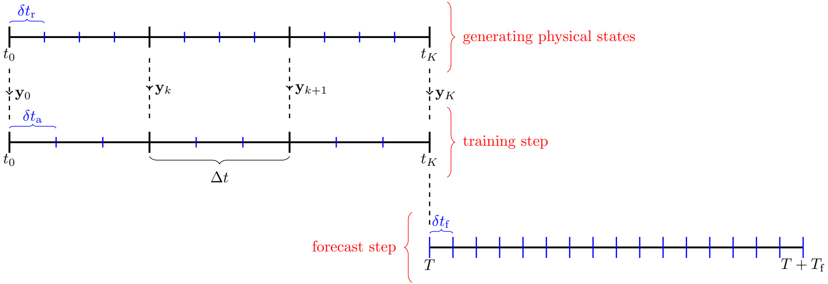
- Physical model and of NN are *independent*.
- NN must predict the analysis increments.
- Resulting hybrid model not suited for short-term predictions.
- For DA, need to assume *linear growth of errors in time* to rescale correction.

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.

Experiment plan

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



- Metrics of comparison:
  - Model: ODE coefficients norm $\|\mathbf{p}_{\mathrm{a}} - \mathbf{p}_{\mathrm{r}}\|_{\infty}$, when the reference parameters $\mathbf{p}_{\mathrm{r}}$ are known.
  - Forecast skill [FS]: Normalized RMSE (NRMSE) between the reference and the surrogate forecasts as a function of the lead time (averaged over many initial conditions).
  - Lyapunov spectrum [LS].
  - Power spectrum density [PSD].

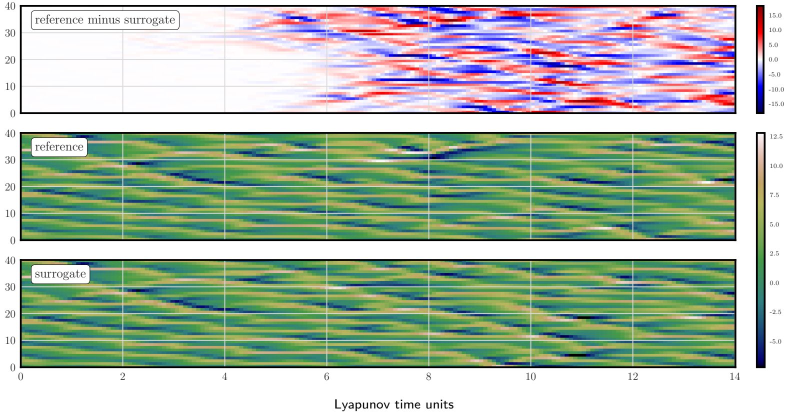
Almost identifiable model and perfect observations

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

The Lorenz 96 model (40 variables)

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

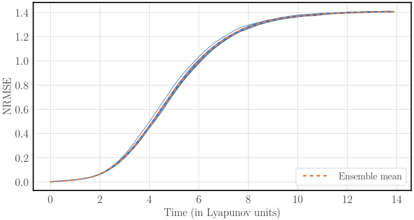
Surrogate model based on an RK2 scheme.

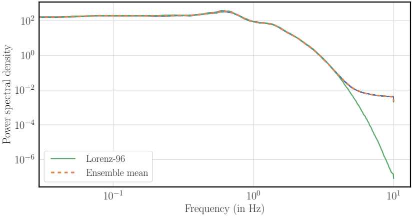


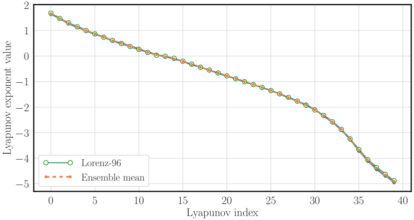
Almost identifiable model and imperfect observations

- 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





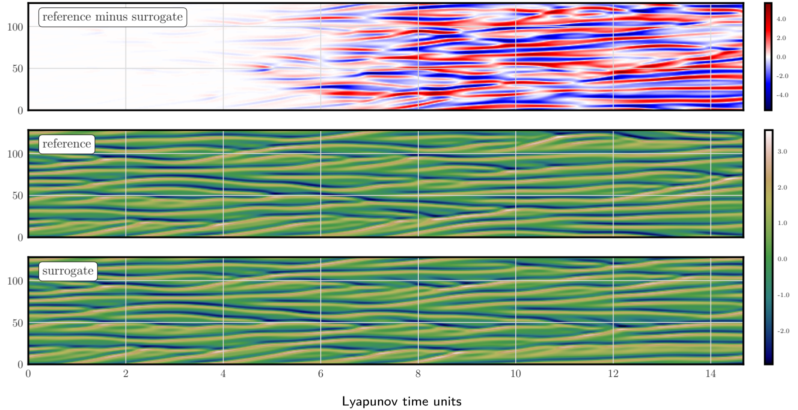


Not so identifiable model and perfect observations

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

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

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

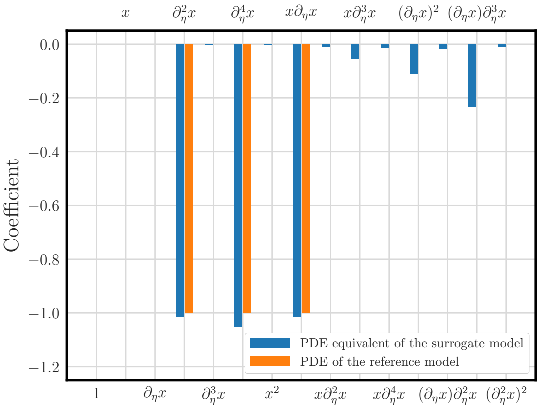


Not so identifiable model and perfect observations

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

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

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

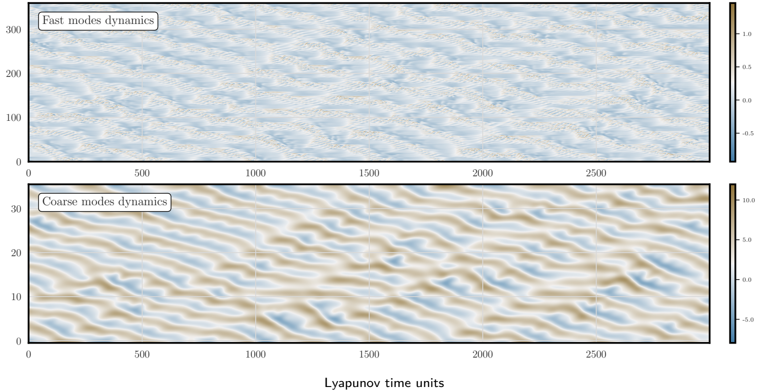


Two-scale Lorenz model (L05III)

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

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

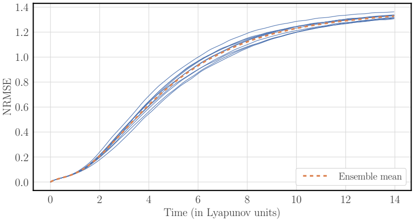
$$\frac{\mathrm{d}u_m}{\mathrm{d}t} = \frac{c}{b}\psi_m^-(b\mathbf{u}) + h\frac{c}{b}x_{m/10}, \quad \text{with} \quad \psi_n^\pm(\mathbf{x}) = x_{n\mp1}(x_{n\pm1} - x_{n\mp2}) - x_n,$$

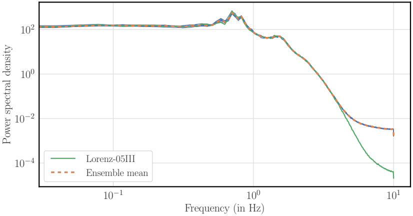


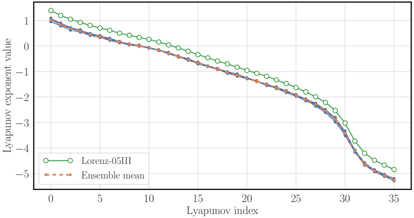
Non-identifiable model and imperfect observations

- 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

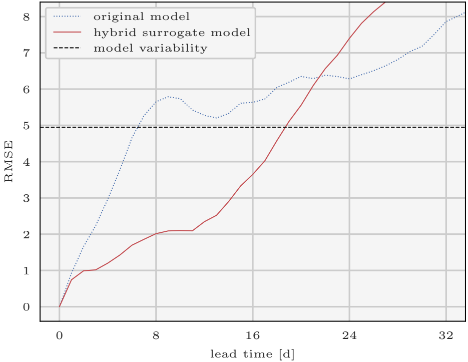






Data assimilation with the surrogate model of L0III (order 1.5 of the loop)

- The non-corrected model is the one-scale Lorenz system.
- 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

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.
- To do this, we use the formalism of DA to estimate both the state and the NN parameters:[5]

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

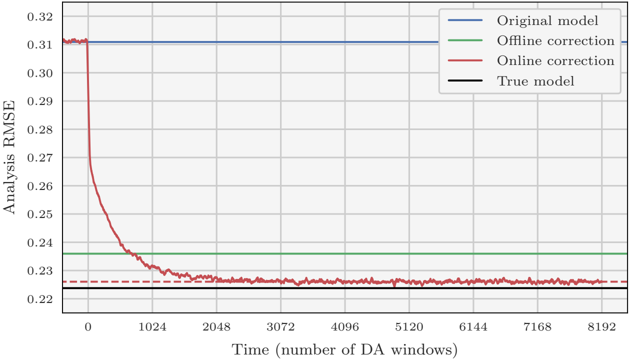
- For simplicity, we have neglected potential cross-covariance between state and NN parameters in the prior.
- Information is flowing from one window to the next using the prior for the state $\mathbf{x}^{\mathrm{b}}$ and for the NN parameters $\mathbf{p}^{\mathrm{b}}$.

- Already been investigated with an EnKF, with solutions.[6]

[5] [Farchi et al. 2021a]
[6] [Bocquet et al. 2021; Malartic et al. 2022]

Numerical illustration with the same two-scale Lorenz system

- We use the tendency correction approach, with the same simple CNN as before, and still using 4D-Var.[7]

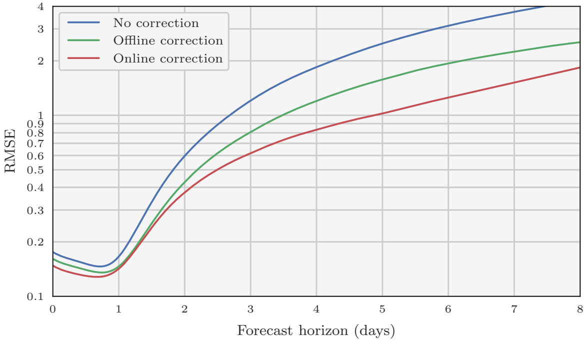


- The online correction steadily improves the model.
- At some point, the online correction *gets more accurate* than the offline correction.
- Eventually, the improvement saturates. The analysis error is similar to that obtained with the true model!

[7] [Farchi et al. 2021a]

Online learning: towards an operational implementation with OOPS

- *Development of a fortran NN library* to interact with the fortran implementation of the forecast model.
- *Interfacing the NN library with OOPS* to estimate the NN parameters with DA.
- Simplifications of the NN correction:
  - the correction is additive, and added after each integration step (close to tendency correction);
  - the correction is computed independently for each atmospheric column[8].
  - the correction is computed at the start of the DA window and not updated during the window;
  - in practice, it requires only *small adjustments* to the current WC 4D-Var already implemented.
- Demonstration with OOPS-QG with promising results, implementation with OOPS-IFS in progress.



[8] [Bonavita et al. 2020]

Online learning with a LEnKF: Augmented state vector

- 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]:

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

- Beware that nonlocal observations require covariance localisation!

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

- Summary of the EnKF-ML family of algorithms we built:[9]

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 algorithm	LEnSRF-ML [Bocquet et al. 2021] new algorithm	L^2EnSRF-ML not discussed
State + global & local param.	LETKF-HML new algorithm	LEnSRF-HML new algorithm	L^2EnSRF-HML new algorithm

[9] new algorithms: [Bocquet et al. 2021; Malartic et al. 2022], see also [Ruckstuhl et al. 2018]

Conclusions

▶ *Main messages:*

- Bayesian DA view on joint state and model estimation.
  DA can address goals assigned to ML but with partial & noisy observations.
- Successful on 1D and 2D low-order models (L96, L05III, L96i, mL96, OOPS QG).

▶ *In progress: more ambitious models and datasets*

- Application to the Marshall-Molteni 3-layer QG model on the sphere
- Application to the ERA5 and CMIP data (WeatherBench[10]-like)
- Application to the ECMWF IFS
- Application to sea-ice surrogate modelling: Schmidt Futures/VESRI/SASIP project

[10] [Rasp et al. 2020]

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* 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. Bonavita and P. Laloyaux. "Machine Learning for Model Error Inference and Correction". In: *J. Adv. Model. Earth Syst.* 12 (2020), e2020MS002232.

[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: *Phil. Trans. R. Soc. A* 379 (2021), p. 20200086.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

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

References II

[15] 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.

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

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

[18] 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.

[19] 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.

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