Review of the paper
« Combining data assimilation and machine learning to emulate a dynamical model from sparse and noisy observations: a case study with the Lorenz 96 model»

## General comments

This paper introduces an algorithm to learn a surrogate model based on neural networks, from noisy and partial observations. The proposed algorithm is tested on the Lorenz 96 model using simulations.

This is an interesting topic but the work seems incomplete in both the methodology and its evaluation.

## Specific comments

Many choices in the algorithm seem arbitrary and should be further discussed and validated through simulations. In particular, the authors state on Page 6 "The procedure can be viewed as an expectation-maximization algorithm...". I believe that it is a good idea to use the EM machinery but I see fundamental differences between the proposed algorithm and the EM algorithm. I think that these differences may deteriorate significantly the performances of the proposed algorithm. These differences should be highlighted, discussed and their impact precisely validated through simulations. Some of these differences are highlighted below.

1. Page 4, l 25-30. "Our choice of the EnKF-N is motivated by its efficiency, its high accuracy for low-dimensional systems, and its implicit estimation of the inflation that would otherwise have had to be tuned." In the EM algorithm, the surrogate model is fitted by maximizing a likelihood function based on the smoothing distribution. Here the authors propose to use the filtering (instead of smoothing) distribution and only the conditional expectations of the filtering distribution (instead of the full conditional distribution of x(k+1) given x(k) and the observation y(1:K)). The authors motivate this choice by the fact that a smoother "is less common in the operational DA community". I agree that using the proposed approach leads to substantial simplifications, but I also feel that it may deteriorate substantially the estimate of the surrogate model. This should be assessed using numerical simulations to compare the results obtained with the proposed approach with the ones obtained with a smoother.
2. Page 4, eq. (5). Why using the norm associated to Pk here? In the M-step of the EM algorithm, the function to maximize is defined as the expectation of the full log-likelihood function with respect to the smoothing distribution. What is the link with the cost function proposed here?
3. Page 16. "A perspective of this work, which is outside the scope of this paper, would be to propose some methodology to estimate the model error statistics...". I don't agree that it is outside the scope since it may lead to substantial improvements of the proposed approach. The choice of the model error is really ad-hoc, with a diagonal matrix (although I would expect some correlations between the errors on the different components) considered as an hyperparameter with arbitrary value. I believe that the estimation of the model error should be included in the methodology before publication. The estimation of the error covariances matrices with the EM algorithm is discussed for example in Dreano et al.

I also have concerns about the numerical experiments, see below.

1. "The model is integrated over 40,000 time steps (K = 40,000)". It seems to be a huge learning sequence, unrealistic for application in DA! Please vary this value and discuss the sensitivity of the results with respect to K.
2. Figure 8, panel a. In my opinion, this is the more interesting plot. It permits a comparison between the NN model fitted on a “perfect” sequence of the true state without observational error (best surrogate of the true model based on NN) and the NN model fitted on noisy observation, which comes out the proposed algorithm. I have the feeling that a good algorithm should permit to retrieve similar NN model from perfect or noisy observations (this is what you expect to get with the EM algorithm) and this plot suggests that there are important differences between the two NN models, despite the huge learning sequence, and thus that the algorithm should be improved.
3. Section 3.2. Various scores are proposed to measure the "distance" between the true and the surrogate model. If I understand correctly, the two last criteria (Lyapunov spectrum and Power spectrum) are used to measure some long-term statistical properties of sequences simulated with the surrogate model. I think that these criteria are more indirect validation criteria of the fitted model and it should come later in the discussion. I also wonder if other criteria linked to the distribution of the stationary distribution (e.g. mean, covariance) may provide additional information and be easier to interpret.

Page 18. "One drawback of the method is the computation cost". I think that this is a really important point which should be more discussed since one the argument of using ML tools is to reduce computational costs compared to running a more physical model. The authors should explain why it is so costly and give a more precise idea of the computational costs for the numerical experiments done in the paper.

## Technical corrections

1. Writing: the authors often use "we...". Please avoid.
2. Page 4, eq. (5) and Section 4.5.1. If the idea is to mimic the EM algorithm, the choice Nf=1 is natural since it arises when writing the full likelihood function of a state space model using usual Markovian assumptions. And simulation results in Section 4.5.1 suggest indeed that it is the best choice. Why making things more complicated by introduction this "hyperparameter"? The authors may consider only the case Nf=1.
3. Figure 7, panel a. The long-term forecasts are improved when the observation noise in increased: any idea to explain this?
4. Page 5, l 8-10. "Convolutive layers apply a convolution acting locally around each grid point of the field. It is equivalent to a locality hypothesis, assuming that there are no long-range correlations between the state variables. Note that it does not discard further distance correlation arising from the time integration". Many terms have not been introduced before like "grid point", "field", "locality". The authors may introduce the context before. The last sentence is completely mysterious for me.
5. End of Page 7, "Note however that localization...". Is it really necessary to talk about localization here? If yes, please explain what it means.
6. Title of Section 4.1 "Convergence of the algorithm". I do not see any convergence here, only some criteria which decrease.
7. Page 11, l3. “The former is the RMSE of a field obtained via quadratic interpolation

without any use of a dynamical model (which is instead essential in DA)". You could also compare with a space-time interpolation without the model to be fair.

8. Some references are incorrect (e.g. the paper by Weinan).
9. Lots of papers cited in the references were written by the authors of the present paper, and, sometimes, I have the feeling that it is a bit artificial.