Response to Reviewer 1

Data-Driven Reconstruction of Partially Observed Dynamical Systems, by Tandeo et al.

Note: your comments and questions are reported in this document and we use bold text for our responses.

In this manuscript, the authors derive a method to reconstruct the dynamics of a system from partial observations, in which data assimilation and machine learning steps are alternate. The data assimilation steps are used to estimate the state from observations using the surrogate model, while the machine learning steps are used to estimate the surrogate model from the data assimilation analysis. This method is the same as the one derived by Brajard et al. 2020, with the exception that, on top of this method, the authors propose a new, innovative state augmentation process. The entire method is illustrated using numerical experiments with the 3-variable Lorenz 1963 system. I am overall positive about this manuscript. The text reads very well and is easy to follow. To my knowledge, the state augmentation process is new and deserves to be published. However, I have some concerns, in particular about the methodology and about the experiments, that needs to be fixed before I can recommend publication.

Thank you for your encouraging review. Below you will find the responses to your different points.

1 General comments

1.1 How the methodology differs from that of Brajard et al. (2021)

As far as I understand, the method derived in this manuscript proposes to alternate data assimilation steps (with the ensemble Kalman smoother) and machine learning steps (with a linear regression) on a given dataset of observations until convergence. This is exactly what has been originally proposed by Brajard et al. (2020) and later formalised by Bocquet et al. (2020). Pushing further the comparison, I see only three significant differences with the original method:

- in the present method the machine learning step is restricted to linear regression, while in the original method, nonlinear regression tools (such as neural networks) are used;
- in the present method observations are assumed to be perfect (even though they are sparse), while in the original method, sparse and noisy observations are used;
- the state augmentation process added on top the data assimilation / machine learning iterations.

I do not see the first two points as a major limitations, in fact I am rather confident that the present method should also work with neural networks replacing the linear regression and with noisy observations. By contrast, the third point is in my opinion the real added value of the present work, and this should be emphasised.

Indeed, the added value is the state augmentation. This is the core of the paper, and, we feel, the innovative part. We also wanted to remind that the alternance of DA (Kalman linear) and ML (linear regression) is not new and has been proposed in the context of the Expectation Maximization (EM) algorithm. We have clarified this point in the new version of the manuscript, l. 47: "The current paper is thus an extension of (Shumway and Stoffer, 1982) to never observed components of a dynamical system, using a state augmentation strategy."

Additionally, we now mention l. 33 the following paper, which was not cited previously: Brajard et al. 2021.

Additional questions about the methodology

1. Is there a fundamental reason to use only linear regression and perfect observations? If not, I would suggest to get rid of these assumptions in the methodological section.

Indeed, there is no reason to only consider this simple case. This is now stated in l. 77: "Nonlinear and adaptive operators as well as noisy observations could be taken into account but, for the sake of simplicity, only the linear and non-noisy case is considered in this paper."

2. How does the state augmentation scale with the system dimension?

Thank you for this interesting question. We added this discussion l. 153: "This result is consistent with the effective dimension of the Lorenz-63 system, which is between two and three. Here, as the estimated dynamical model M is a linear approximation, the dimension of the augmented state and the observed components is higher than the effective one." We also point out the role of the likelihood to find the number of hidden components, l. 157: "This likelihood is useful to diagnose the optimal number of dimensions needed to emulate the dynamics of the observed components."

3. Can the additional state components be added all at once? Did you try that in the numerical experiments?

Yes, indeed it is possible to add all the latent variables at the same time. This is now clarified in l. 142: "Note that several hidden components can be added all at once, with similar performance as the sequential procedure described above (results not shown). In this all at once case, the interpretation of the retrieved components is not as informative, thus we decided to retain the sequential case." We decided to retain the iterative strategy, especially to introduce and explain clearly Eqs. (6a) and (6b).

4. In the experiments, 30 iterations seem sufficient to reach convergence. Do you have an idea how this number would scale with the system dimension?

There is no clear relationship between the number of EM iterations and the dimension of the system, and this point doesn't seem to be discussed in the literature. However, the EM algorithm has a slow (linear) asymptotic convergence speed, but is generally

efficient in the first iterations to quickly increase the likelihood, and provide a first estimate of the parameters.

5. The text is ambiguous about the data assimilation method used: 'and thus uses the classic Kalman filter and smoother equations' (L 71-72), 'by the Kalman filter' (77-78) 'a Kalman smoother is applied' (L 87) 'Kalman filter and smoother' (L 161). Kalman filter or smoother, you have to choose (I assume it is Kalman smoother).

The two Kalman recursions are necessary in this paper and we decided to keep this distinction because, as now stated in l. 98: "The Kalman filter (forward in time) is used to get the information of the likelihood, whereas the Kalman smoother (forward and backward in time) is used to get the best estimate of the state."

1.2 About the numerical experiments

The description of the experiments is incomplete, in such a way that the experiments cannot be reproduced without further assumptions. For example, what numerical method is used to integrate in time the model equations to compute the truth? Furthermore, I have a serious concern about the 'model distance' introduced by equation (6). Without further details, I assume that it is computed using the same trajectory as the training step. Using the same data for training and testing should be avoided by all means. Moreover, in this context where observation are perfect, I am not sure to see the point of this metric: observations are required to initialise the model (for the hidden components), but if we have observations, we do not need the forecasting system any more since observations are perfect... Therefore, I think that the metric used to evaluate the accuracy of the model should be reconsidered.

Thank you for those remarks about the numerical experiments. It is now stated in l. 132: "Runge-Kutta 4-5 is used to integrate the Lorenz-63 equations to generate x1, x2, and x3."

Regarding the Eq. (6) and the metric of evaluation, it has been completely modified, taking into account the remarks of the two Reviewers. It now reads, l. 189: "To compare the performance of the naive linear model M with [x2, x3] and the ones with [x2, x3, z1] or [x2, x3, z1, z2], their forecasts are evaluated. After applying the proposed algorithm, the M and Q estimated matrices are used to derive probabilistic forecast, starting from the last available observation y_t, using:

$$\mathrm{E}[\mathbf{x}_{t+1}|\mathbf{y}_1,\ldots,\mathbf{y}_t] = \widehat{\mathbf{M}}\mathrm{E}[\mathbf{x}_t|\mathbf{y}_1,\ldots,\mathbf{y}_t],$$

$$\mathrm{Cov}[\mathbf{x}_{t+1}|\mathbf{y}_1,\ldots,\mathbf{y}_t] = \widehat{\mathbf{M}}\mathrm{Cov}[\mathbf{x}_t|\mathbf{y}_1,\ldots,\mathbf{y}_t]\widehat{\mathbf{M}}^T + \widehat{\mathbf{Q}},$$

with E and Cov, the expectation and the covariance, respectively. To test the predictability of the different linear models (i.e., with or without hidden components z), a test set has been created, starting from the end of the sequence of observations (y_1, …, y_T) used in the assimilation window. This test set is also corresponding to 10^4 time steps with dt=0.001. It is used to compute two metrics, the Root Mean Square Error (RMSE) and the coverage probability at 50%. The RMSE is used to evaluate the precision of the forecasts, comparing the true x2 and x3 components to the estimated ones, whereas the coverage probability is used to evaluate the reliability of the prediction, evaluating the proportion of true trajectories falling within

the 50% prediction interval of x2 and x3. Examples of predictions are given in Fig. 4. It shows bad linear predictions of the model with only [x2, x3] (dashed black lines). As the M operator is not time-dependent, the predictions are quite similar, close to the persistence. Then, adding one (green) or two (red) hidden components in the M operators creates some nonlinearities in the forecasts.

In Fig. 5, the predictions are evaluated over the whole test dataset, for different lead times. By introducing hidden components, the RMSE decreases for both x2 and x3 components (top panels). For instance, for a lead time of 0.05, when considering two hidden components, the RMSE is halved when it is compared to the naive linear model without hidden components. The coverage probability metric is also largely improved (bottom panels). Indeed, the results with two hidden components are close to 50%, the optimal value.

To evaluate where the linear model with [x2, x3, z1, z2] performs better than the one with [x2, x3], the Euclidean distances between the forecasts (for a lead time of 0.1) and the truth are computed. Those errors are evaluated at each time step of the test dataset, in the (x2, x3) space. Based on those errors, Fig. 6 shows the relative improvement between the model without and the model with hidden components. When the two models have similar performance, values are close to 0 (white), and when the model including z1 and z2 is better, values are close to 1 (red). Figure 6 clearly shows that error reduction is not homogeneous in the attractor. The improvement is moderate in the outside of the wings of the attractor, but important in the wing-transition. It suggests that the introduction of the hidden components z1 and z2 makes it possible to provide information on the position in the attractor and thus to make better predictions, especially in bifurcation regions."

Additional questions about the experiments:

1. '10 loops of the Lorenz-63 system' (L 104-105) Do you mean 10 model time units or 10 revolutions on the model attractor? In any case, I would not say that this is a small period of time, compared to the doubling time which is 0.78 MTU.

Thanks for the suggestion, it now reads l. 133: "10 model time units of the Lorenz-63 system" and we removed "a small period of time".

2. From what I understand (L 104-106), you have access to the true x2 and x3 (no observation noise) every dt = 0.001 (which is probably the integration time step for the truth). This seems to be very strong requirements. Can you discuss this?

Yes, this is a strong requirement, but for the sake of simplicity, we decided to keep dt=0.001, showing that only two latent variables are needed. It is now explained in l. 144: "Note also that the methodology has been tested with larger dt (i.e., 0.01 and 0.1). The conclusion is that by increasing the time delay between observations, it significantly increases the number of latent variables (results not shown)."

3. What is the choice of the data assimilation window length for the ensemble Kalman smoother? Without further details, I assume that it covers the entire experiment, i.e. 10^4 observation steps. This is really huge. Can you discuss this?

L. 146, it is mentioned that: "Finally, the assimilation window length corresponds to 10^4 time steps. By reducing this length (e.g., to 10^3, 10^2, 10^1), the conclusions remain the same as for dt=0.001."

Technical comments and suggestions

L 17-18 'using Bayesian framework' → 'using a Bayesian framework' ?

Done.

L 21 'All the approaches cited above are assuming that the full state of the system is observed' This is not true: at least Tandeo et al. (2015), Lguensat et al. (2017), Bocquet et al. (2019), Brajard et al. (2020), Fablet et al. (2021) use sparse observation operators in their methods. I would replace 'All the approaches cited above' by 'Many approaches'.

Done.

L 23-24 'To deal with those strong constraints' I would replace here 'constraints' by 'assumptions' in order to avoid a potential confusion with strong-constraint methods in variational data assimilation.

Done.

L 24-26 'An option is to [...] whereas an other option is to [...]' I would suggest to also mention here the combination of data assimilation and machine learning, because (i) this is what is used in some of the previously cited papers (the ones that can handle sparse and noisy observations), and (ii) this is what is used in the present manuscript!

Thanks, this sentence now reads, l. 37: "To deal with those strong constraints, i.e., when the model is unknown and when some components of the system are never observed, combination of data assimilation and machine learning shows potential (see e.g., Wikner et al. 2021)."

L 29 'with a dynamical model (model- or data-driven)' I would replace here 'model-driven' by 'based on physical knowledge' or something like this (to avoid a model-driven model).

We replaced "model-driven" by "physics-driven", this seems to be the adequate term.

L 31 'estimation of the parameters' Which parameters?

It is now clarified, l. 46: "dynamical parameters", i.e. M and Q matrices in our case.

L 40-41 'from data assimilation, machine learning, and theory of dynamical systems'
→ 'from data assimilation, machine learning, and dynamical systems' ?

Done.

L 42 ‘from partial observations y’ In data assimilation, observations are usually noisy in addition to being partial.

We added, l. 66: “and noisy”.

L 42-46 In this paragraph, why didn’t you mention the crucial role of the background error statistics?

We added, l. 70: “But this estimation requires good estimates of model and observations error statistics (see e.g., Dreano et al., 2017; Pulido et al., 2018).”

L 48 ‘to mathematically approximate the dynamic of the system’
→ ‘to mathematically approximate the system dynamics’.

Done.

L 76 In equation (2), I would suggest to explicit the definition of L, i.e. use something like that:

$$\mathcal{L} \triangleq p(\mathbf{y}_1, \dots \mathbf{y}_T | \mathbf{x}_1^f, \dots \mathbf{y}_T^f) \propto \prod_{t=1}^{T} \cdots$$

Thanks, the mathematical definition of the likelihood has been introduced in Eq. (2), l. 104.

Furthermore, T is undefined in this equation.

L. 103, we added: “is computed using T time steps such that”.

L 78-79 ‘The innovation likelihood given in Eq. (2) is interesting because it corresponds to the squared distance between the observations and the forecast normalized by their uncertainties, represented by the covariance Σt.’ In data assimilation, this quantity is simply called ‘the likelihood’.

We prefer to keep “innovation likelihood” because different likelihoods appear in DA: the likelihood of the innovation and the total likelihood of the state-space model (see Tandeo et al. 2020, section 4, available here: https://tandeo.files.wordpress.com/2020/11/tandeo_2020_mwr.pdf).

L 89-90 ‘This random sampling is used to exploit the correlations between the components of the state vector’ I do not understand why this is necessary. Could you elaborate?

Sorry, it was not clear. It now reads, l. 117: “This random sampling is used to exploit the linear correlations between the components of the state vector, which appear in the non-diagonal terms of P^s.”

L 110 ‘After 30 iterations of the algorithm presented in section 2, the hidden component z1 is stabilized.’ Can you please explain the exact meaning of ‘stabilized’ in this context?

In l. 138, we replaced “is stabilized” by ”has converged”.

L 114 ‘this augmented state procedure is repeated’ → ‘this state augmentation process is repeated’.

We prefer, l. 141: “this state augmentation procedure is repeated”.

L 117 ‘z3 is very flat’ I would replace ‘very’ by ‘rather’ in this statement.

We prefer, l. 151: “z3 remains close to 0”.

L 125 ‘Finally, the inclusion of z3 reduces the likelihood (purple lines).’ Do you have an explanation for this phenomenon?

Thanks for this question. After investigation, we discovered that, l. 162: “Finally, due to a significant increase of the forecast covariance P^f in Eq. (2), the inclusion of z3 reduces the likelihood (purple lines). This suggests that a third variable is not needed, and is even detrimental to the skill of the reconstruction.”

L 131 In equation (6), I would explicit the dependence on time, i.e. replace dist (M) by dist(M)(t).

We hope that the new Eqs. (7) clarify this point.

L 137-138 ‘Are they correlated with the unobserved component x1 or with the observed one x2 and x3?’ → ‘Are they correlated to the unobserved component x1 or to the observed ones x2 and x3?’ ?

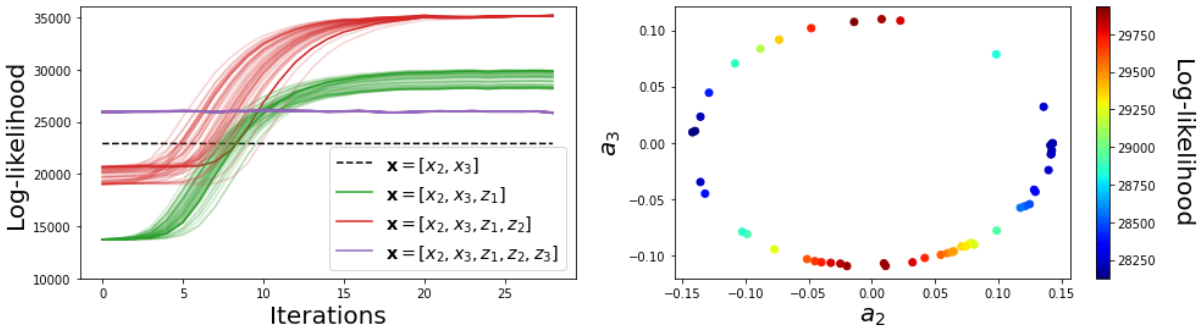
Done.

L 139 ‘It has been found that...’ How did you come up with this? As it is presented, it looks like something pulled out of a hat.

Thank you very much for this remark. It now reads, l. 169: “Using symbolic regression (i.e., using basic mathematical transformations of x2 and x3 to explain z1 and z2), it has been found that the hidden components z correspond to linear combinations of the derivatives of the observations such that: Eqs. (6)”. Sorry for this important omission.

L 47-48 ‘This is illustrated in Fig. (3), with 50 independent realizations of the proposed algorithm.’ Strictly speaking, this is not the case since a and b are not represented in this figure.

We decided to add a subfigure in the right panel of Fig. 3.



The caption of Fig. 3 now reads: "Likelihoods as a function of the iteration of the augmented Kalman procedure (left) and estimation of the a2 and a3 parameters (right). Different dynamical models are considered, from none to three hidden components in z, whereas only x2 and x3 are observed in the Lorenz-63 model. The likelihood of 50 independent realizations of the iterative and augmented Kalman procedure are shown."

L 152-153 'Then, when considering z1 and z2 (red lines), the 50 independent realizations reach the same likelihood after 30 iterations.' What about a and b? Are they similar over the 50 realisations?

Based on the new Fig. 3 (right panel), it now reads, l. 184: "Interestingly, the scatter plot between a2 and a3 shows a circular relationship. This is also the case for b2 and b3 (results not shown)."

L 153-154 'it will then focuses' → 'it will then focus'.

Done.

L 175-176 'the dynamical evolution of the system is retrieved with our methodology'. This is not clearly shown in the experiments.

We hope that the new Fig. 4 and Fig. 5 give more information about this. However, the sentence was maybe too strong and we replaced it, l. 236, by: "the dynamical evolution of the system is relatively well captured, for short lead times, with our methodology."