Response to Reviewer 2

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.

This work presents a data-driven method to infer a linear stochastic model from a partially observed system. This work is well-written and contains interesting parts, especially the not-so-common effort to explain the dynamics of the latent (embedding) space. Nevertheless simplifications made in the work do reduce a lot the impact of this paper. Also, there is very little novelty in the approach. The principle of alternating between DA and a data-driven model has been already applied, in more challenging settings (noisy/sparse observations, model with more dimensions). The fact to have a variable that is never-observed has also already been tested. The originality of the approach to have a stochastic model and to explain the latent space is not very developed.

Thank you for your general comment. Indeed, the combination of data assimilation and machine learning is not new. However, the introduction of latent variables in this context is new (to the best of our knowledge). This is the key methodological contribution of our manuscript.

As the Reviewer mentioned, the explanation of the latent space and the stochasticity of the model are the main points of our work. In the new version of the manuscript, these points are discussed in more detail, especially by the addition of 3 new figures: Fig. 3 (right panel), Fig. 4, and Fig. 5.

Other general comments:
- the justification of the setting and the approach is not convincing to me (see my comments about the abstract and the introduction), and I fail to foresee the real application of the approach. Maybe rephrasing the last part of the conclusion and putting it in the introduction instead could help regarding that matter.

Thank you very much for your suggestion. We followed the Reviewer's advice, by rephrasing the end of the conclusion and putting it in the introduction, l. 13: “In geophysics, even if one has the perfect knowledge of the studied dynamical system, it remains difficult to predict because of the existence of nonlinear processes (Lorenz, 1963). Beyond this important difficulty, achieving this perfect knowledge of the system is often impossible. Consequently, the governing differential equations are often not known in full because of their complexity, in particular regarding scale-interactions (e.g., turbulent closures are often assumed rather than "known" per se). On top of these two major difficulties, the state of the system is not and cannot be exhaustively observed. Potentially crucial components are and might remain partly or fully out of reach of proper monitoring (e.g., deep ocean or small scale features).

Predicting a partially observed and partially known system is therefore a key issue in current geophysics and in particular for ocean, climate and atmospheric sciences."

- The data-driven model used is linear. It is acknowledged by the authors in the conclusion, but it is one limit of the approach. Maybe the linear approach works because the setting is simple enough (low dimension, weakly non-linear). But also, I wonder if the interpretability of the latent space is precisely related to the choice of the linear model (maybe with a non-linear model, there is no need for a latent space to emulate observed variables...)

This is an important remark and a discussion is now given in l. 52: "The proposed methodology is based on an important assumption: the surrogate model is linear. Although it can be considered as a disadvantage compared to nonlinear models, this linear assumption also has interesting properties. Indeed, nonlinear model combined with state-augmentation is a very broad family of model and may lead to identifiability issues. Using a linear dynamics already leads to a very flexible family of model since the latent variable may describe nonlinearities and include for example any transformation of the observed or non-observed components of a dynamical model. Furthermore, it allows a rigorous estimation of the parameters using well established statistical algorithms which can be run at a low computational cost."

We also added a perspective of work, l. 244, related to this comment: "In future works, we plan to compare the global and local linear approaches (i.e., fix or adaptive linear surrogate model). We also plan to compare them to nonlinear surrogate models, based on neural network architectures with latent information encoded in an augmented space or in hidden layers (e.g., LSTM)."

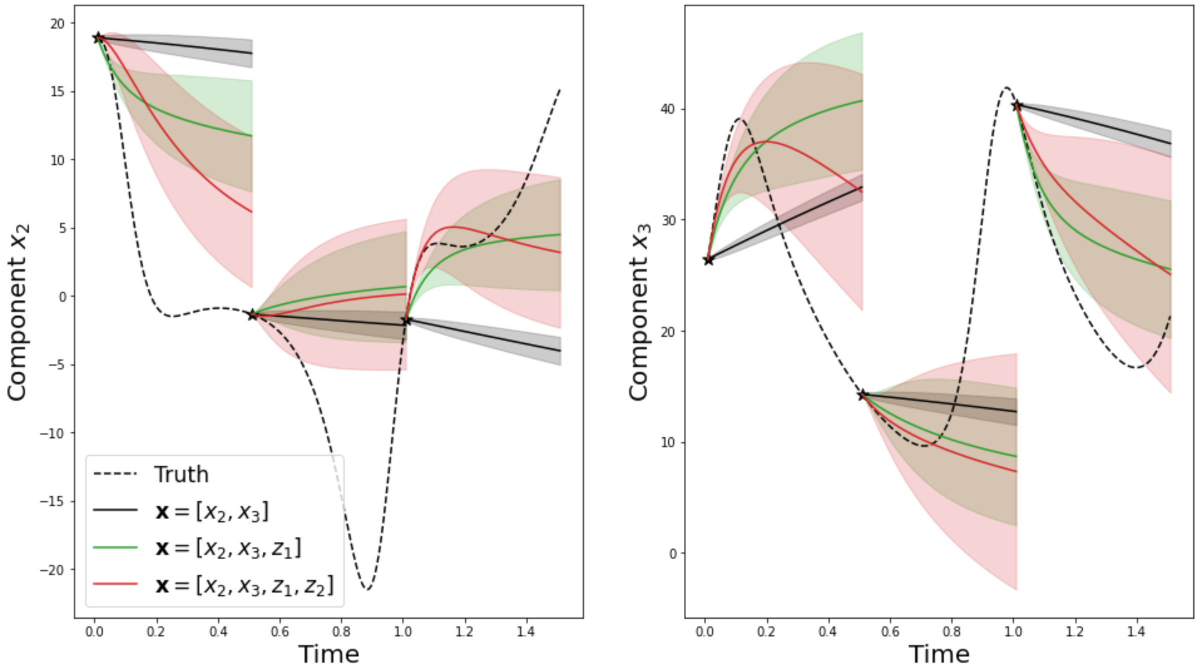
- The experiment is done on the Lorenz 63 model, which is very low-dimensional (3) and weakly non-linear. See for example: https://raspstephan.github.io/blog/lorenz-96-is-too-easy/# There are toy models (L96, QG) that could display more interesting behaviors for this methodology.

Thank you very much for this proposition. We added this clarification in l. 58: "The proposed methodology is evaluated on a low-dimensional and weakly nonlinear chaotic model. As this paper is a proof of concept, a linear surrogate model is certainly well suited for this situation."

We added some perspective of work, l. 247: "In this paper, we have demonstrated the feasibility of the method on an idealized and comprehensive problem, using the Lorenz-63 system. In the future, we plan to apply the methodology to more challenging problems, like the Lorenz-96 system or a quasi-geostrophic model. For the application on real data, we plan to use a database of observed climate indices and try to find latent variables that help to make data-driven predictions."

- The forecast is evaluated only in the next time step, which is again a very easy case. How would behave the forecast over several time steps?

Thanks for your remark. We have added a new Fig. 4 showing statistical forecasts over several time steps.

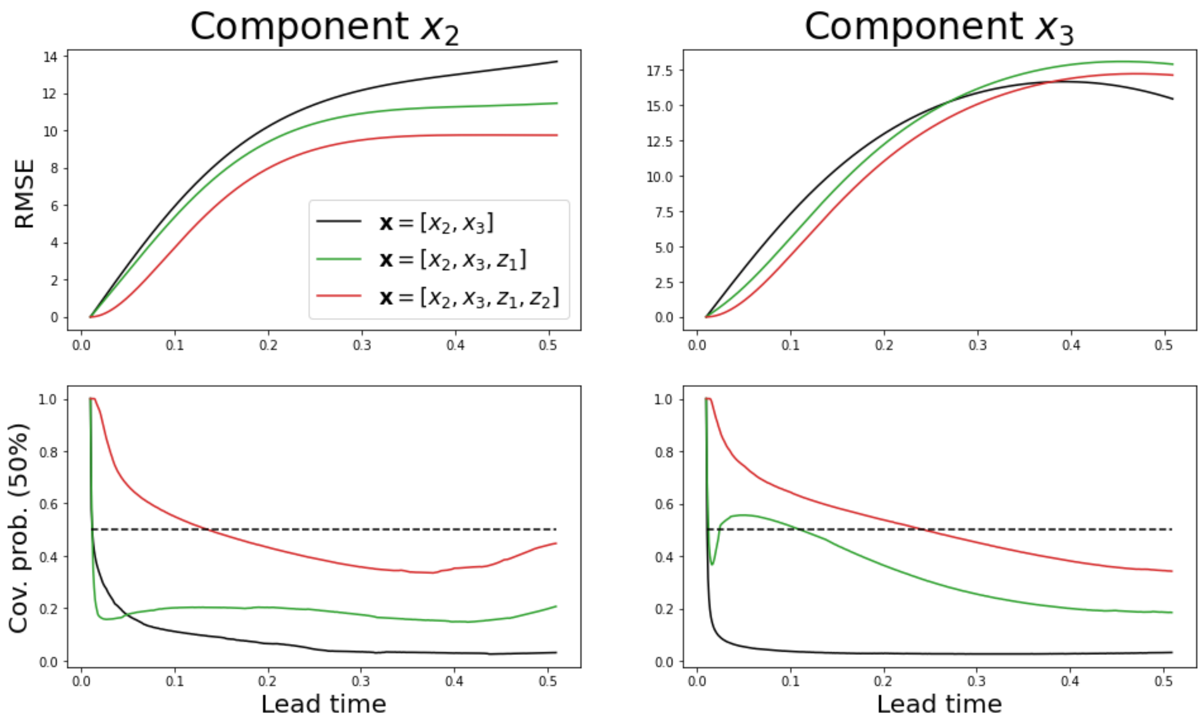


The caption of new Fig. 4 reads: "Example of three statistical forecasts of x2 (left) and x3 (left) with their 50% prediction interval using 3 different linear operators with: no hidden component (dashed black), one hidden component (green), and two hidden components (red). These predictions are obtained using sequential statistical forecasts, as explained in Eqs. (8), on an independent test dataset."

Then, a discussion related to this Fig. 4 has been added, l. 200: "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."

- The method is interestingly stochastic, but no ensemble metrics are used to evaluate the work which would have been interesting.

We have also introduced a new Fig. 5, showing the evolution of the RMSE and the coverage probability, a simple metric to evaluate the estimated prediction intervals.



The caption of new Fig. 5 reads: "Root Mean Square Error (top) and 50% coverage probability (bottom) as a function of the lead time (x-axis) for the reconstruction of the components x2 (left) and x3 (right). These metrics are evaluated on an independent test dataset."

Then, a discussion related to this Fig. 5 has been added, l. 204: "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."

Specific comments:

Abstract: The 2 first sentences of the abstract is a justification of the approach. Due to the limited size of the abstract, this justification cannot be extended making it too simplistic: 1) It is true that defining a set of equations is difficult, but I would say that a bigger issue is the resolution of the existing set of equations given that coefficients are unknowns and that a discretization is needed for the numerical resolution, which introduces some errors. 2) If we follow the narrative, it is well justified that we should cope with "imperfect equations". But here, the choice is to assume that no equation is known. The fact that those two points are overlooked makes the narrative a bit too simple to be convincing for me. I would suggest starting right away with what you want to achieve in the abstract and having an extended justification in the introduction.

Thank you for your comment, but we want to keep these first sentences at the beginning of the summary. Although they are overly simplistic, we believe they provide important context for this study. In addition, the summary with these sentences is 198 words, which does not exceed the limit of 200 words. However, we agree with your comments and have taken them into account in the new version of the introduction, between l. 13 and l. 20.

L13: "governing differential equations are not necessarily known": I would like to see examples of that. I think that, even if some equations are known, a fully data-driven system can be justified, but here, this core question is eluded: What is the range of applications of a purely data-driven model from partial observations?

To develop this point, we added the following paragraph, l. 21: "A typical example of such a framework is the use of climate indices (e.g., Global Mean Temperature, Niño 3.4 index, North Atlantic Oscillation index) and the study of their links and their dynamics. In this context, the direct relationship between those indices is unknown, even if their more indirect and complex relation exist, through the full knowledge of the climate dynamics. Also, it is highly possible that climate indices are dependent on components of the climate that are not currently considered as key indices, and so are not fully monitored. However, these key indices could be sufficient to describe the most important aspect of climate, leading to accurate and reliable predictions, and enabling cost-effective adaptation and mitigation."

L21: "All the approaches cited above are assuming that the full state of the system is observed, which is a strong assumption." This is misleading. The papers above (at least Fablet, Bocquet and Brajard) assume that observations are noisy and sparse, but indeed each variable has a non-null probability to be observed. Is it what you mean by "the full state is observed?" There are also many works done in the case a variable is never observed, e.g.: https://arxiv.org/pdf/2102.07819.pdf

This mistake has been also detected by Reviewer 1. Based on his/her suggestion, we replaced "all" by "many" in l. 35. Moreover, we added a reference to the paper Wikner et al. 2021, 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 seems relevant (see e.g., Wikner et al. 2021)." Thanks for the suggestion.

Figure 1. My understanding is that the paper aims at going one step forward into learning a data-driven model from a realistic setting (by assuming that the state is not fully observed), but it assumes later on that a part of the state is always observed with a very small error. To me, this is a very strong assumption, even stronger than assumptions made by the existing cited papers. So again, I don't see what application is targeted by this work.

We clarified this point, 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." The objective of this paper is to retrieve dynamical information that is never observed. But the observed data is assumed to be of good quality (i.e., not noisy).

L110: the "sequential methodology": Is there a theoretical reason to add sequentially the hidden components or is this mainly practical? How do you see that applied with high-dimensional systems in which, e.g., 10^5 variables are non-observed?

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 (results not shown), with similar performance as the sequential procedure described above. In this all at once case, the interpretation of the retrieved components is not as informative, thus we focus on the sequential case." We decided to retain the iterative strategy, especially to introduce and explain clearly Eq. (6a) and (6b).

Regarding the application to high-dimensional systems, we refer to the perspective of work, between l. 247 and l. 250.

L140: This part is, in my opinion, the most interesting part. But I miss some details to fully understand what is done (see below)

Your next comment is related to this one. See my response below.

Eq 7: How do you derive those equations? Is it by trials/error or is does it correspond to theoretical reasons?

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 as regressors 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.

L150 "correspond to a3 ≈ 0 and to a2 ≈ 0, respectively": sorry I don't get the "respectively" here, in which case a3 is 0 and in which case a2 is 0?

The explanation is easier to follow with the new Fig. 3 (right panel). It now reads, l. 181: "As shown in Fig. 3 (right panel), the minimum likelihood corresponds to a3 = 0 and the maximum likelihood corresponds to a2 = 0."

L150-151: "This suggests that xdot3 is more important than xdot2": Why is that? you still have b2 coefficient associated with xdot2…

To better understand this sentence, we added l. 182: "Thus, the likelihood when z1 = a3 ẋ3 is higher than when z1 = a2 ẋ2." This is why it is said that ẋ3 is more important than ẋ2 in the reconstruction of the latent variables.

L153-155: sorry I have read this part several times, and I still don't understand. What does it mean that "the algorithm focuses on the estimation of a_2" I don't see where is the estimation of a_2 in the algorithm and I don't understand what is meant by "focus".

Sorry for that. We agree that it was not ideally explained. We rephrased it, l. 186: “It means that if a3 = 0 when considering only z1, then b3 ≠ 0 when introducing z2. In terms of forecast performance, this is similar to a2 = 0 and b2 ≠ 0, because the likelihoods converge to the same value (red lines after 30 iterations).”

L158: The term "model-driven" is misleading. The data-driven model is also a model.

“Model-driven” has been replaced by “physics-driven” everywhere.

L175: "the dynamical evolution of the system is retrieved with our methodology. "This is a strong assertion since by construction the evolution of x2 and x3 are observed and you test the forecast skill over only one time-step.

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

End of the introduction: I think it would be nice to have part of these comments in the introduction, justify the approach.

Indeed, it is better to give the context in the introduction. The beginning of the introduction has been entirely rewritten, between l. 13 and l. 26.