Response to Referee 1

for "Representation learning with unconditional denoising diffusion models for dynamical systems"

Finn, T. S., Disson, L., Farchi, A., Bocquet, M., and Durand, C.

24th May 2024

RC: Reviewer Comment, AR: Author Response

RC: This research paper presents a study on using denoising diffusion models for data-driven representation learning of dynamical systems. The research demonstrates the utility of such networks with the Lorenz 63 system, showing that the trained network can produce samples almost indistinguishable from those on the attractor, indicating the network has learned an internal representation of the system. This representation is then used for surrogate modeling and generating ensembles out of a deterministic run. Overall I found this paper very well written and the contribution of introducing diffusion model into dynamical systems in geoscience novel and of clear contribution. Here lists my comments before I can recommend acceptance of this manuscript:

AR: We thank Dr. Cheng for the constructive feedback on our manuscript, especially with the remarks related to small dimensionality of the here-tested Lorenz 1963 system. In the following, we will discuss the raised comments and indicate what we will change in the revised manuscript.

RC: If I understand correctly, the objective of this study is to explore the possibility of using diffusion model for high-dimension systems in geoscience. The numerical experiments are carried out using a three dimensional Lorenz model. To enhance the discussion, It would be beneficial if the authors could explain how generalizable their approach is to a high-dimensional spatial temporal system (e.g. by adding CNN or transformer layers for feature extractions (encoding) and decoding etc).

AR: The goal of this study is foremost to give a proof-of-concept on representation learning for dynamical systems with denoising diffusion models. We tackle the question,

what would happen if we would have much more data and much more parameters in denoising diffusion models than in the system. Consequently, we selected the Lorenz 1963 as system of interest as the system has only three dimensions and we can easily generate millions of data points and train networks with millions of parameters. Our results indicate that in such settings, denoising diffusion models can generalize to the system and be used for downstream tasks. While there is still the question if the generalization also holds for high-dimensional and large-scale systems, the results gives us hope that it can be the case. To take this comment into account, we will strengthen in the abstract and introduction the proof-of-concept character of the study. In Sect. 5 (Summary and Discussion), we will additionally elucidate more on an outlook how these might hold for higher-dimensional systems, e.g., hypothesizing what happens if we would add transformer layers.

RC: As a consequence of the small dimension, the "latent space" in your diffusion model (256) is much larger the one of the physics space (3). Therefore, you have little risk in losing any information when using the denoising network for surrogate modelling. The authors may consider adding a baseline of transfer learning from an untrained (randomly initialized denoising NN) in Fig 7. The authors have shown the results of untrained NN in Tab 3 but only with a linear fine-tuning. What happens if you fine-tune with a non-linear NN of an untrained denoising NN?

AR: The dimensionality of the feature space (avoiding latent space to circumvent issues with the latent/noised space from diffusion models) spanned by the denoising diffusion model is indeed much larger than the dimensionality of the system, a consequence of the study's character as proof-of-concept. Independently, the question that we answer is if this features space can be used for surrogate modelling. The trained diffusion model has the "right" features for surrogate modelling, whereas a randomly initialized model fails to have them. The correct features for surrogate modelling are hence learned and not by chance. Consequently, features that are needed to generate states on the system's attractor seem to be useful for surrogate modelling. We deliberately neglected the baseline of the untrained diffusion model in Fig. 7: the surrogate model with the untrained feature extractor rapidly converges to a nRMSE of 1 as also visible in Table 3. Including this baseline would not provide additional information to the table and could distract from the main message of the Figure that the trained features are more stable than random Fourier features. Consequently, in the revised version of manuscript, we will still omit this baseline from the Figure. For completeness, we nevertheless include the baseline in the modified Fig. 1 of this answer. As the linear probing already shows, the features extracted by the untrained diffusion model are unaligned to the dynamical system, hence, we neglected the experiment with the small NN. In the revised version of the manuscript, we will include the scores for this experiment.

RC: In figure 7, it seems that the dense neural network with two layers trained

from scratch outperforms your transfer learning from the diffusion model. Is that the case? In fact, results in tab 3 also show that the model trained from scratch (dense *3 and resnet) performs similarly to the fine-tuning from your diffusion model? The authors may want to add some comments regarding this

AR: Figure 7 shows the performance over long lead times and used here to show the stability of the surrogate models. Since the models were trained for lead times of 0.1 MTU, we cannot expect that they perform as well for very long lead times. To improve the performance therein, one could apply autoregressive training steps as often done in surrogate models for the atmosphere, e.g. in GraphCast. Furthermore, the difference between the transfer learned surrogate model and the surrogate models learned from scratch are in fact smaller than the difference caused by different random seeds and might be a result from chance. Consequently, we stay at our claim that transfer learning can perform better than NNs trained from scratch. To nevertheless take the comment into account, we will add something like *"Since the models are trained for lead times of* 0.1 MTU *without autoregressive steps, their performance for longer lead times is heavily impacted by randomness as shown by the spread between seeds in Fig. 7. The difference between the NN models is much smaller than the effect of randomness, which makes it difficult to discriminate if the differences are by chance"* to the explanation of Fig. 7 in the manuscript.

RC: Page 3, 'generative training is rarely used for pre-training and representation learning of high-dimensional systems'. There are some works tried to use diffusion model for contrastive models, e.g,

- **Yang, X. and Wang, X., 2023. Diffusion model as representation learner. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 18938-18949).**
- **Mittal, S., Abstreiter, K., Bauer, S., Schölkopf, B. and Mehrjou, A., 2023, July. Diffusion based representation learning. In International Conference on Machine Learning (pp. 24963-24982). PMLR.**

The authors may want to include some references and discuss the difference/similarity compared to the method used in this paper. This paper is probably the first one to propose diffusion-based representation learning in dynamical systems(?)

AR: The intention of this specific sentence was to show the gap. We understand that this sentence might be missleading and will change it to "Since training deep generative models remains difficult yet, generative training is less often used for pre-training and representation learning of high-dimensional systems than other methods like contrastive learning (e.g., SimCLR from Chen et al., 2020)." A smaller literature review is given in the paragraph afterwards. Caused by the timeliness of the topic, we have however missed these recent publications and we will add them to the literature review, thanks for pointing to them. Hence, we will add: "Concurrently

to our study, Mittal el al., 2023 and Yang and Wang, 2023, propose to directly use denoising difufsion models for representation learning from images. However, to our knowledge, we are the first introducing these models for representation learning from dynamical systems."

RC: Page 9, 'show that this representation is entangled' why it is important for the learned features to be entangled?

AR: Entangled features are more difficult to interpret and also more difficult to use in downstream tasks, as indicated by the need of features from several pseudo times steps and a small NN for surrogate modelling. Consequently, this can be seen as one drawback of the learned representation.

RC: Page 11, check the sentence 'As we will see later, the bigger the Because of the state-dependency, the resulting distribution is implicitly represented by the ensemble and could extend beyond a Gaussian assumption'

AR: Thank you for spotting this incomplete sentence and left over from the internal revision process. We will remove the part "As we will see later, the bigger the " since it is covered in the next paragraph.

RC: Page 13, it seems that you have used a lot of training samples (1.6*E7) for your diffusion model for the Lorenz system of dimension 3. I was wondering if a standard surrogate model will require that much. That is saying maybe a standard surrogate model can outperform the diffusion-based one with less training data. I am curious to see the authors' thought.

AR: We used this many samples to be unconstrained from the training dataset size. It is very likely that much less training samples are needed for surrogate modelling and representation learning, yet, we do not know when forecast performance drops. While many samples might be needed to learn a representation, we agree with you that standard surrogate models need much less samples as they are more specialized. The premise of representation learning is however that the learned features can be then transferred to other problems like surrogate modelling, where we would need much less training samples than for representation learning, and possibly even less than for standard surrogate models. Since the forecast error of surrogate models in the Lorenz 1963 system is very low, this hypothesis should be tested with higher-dimensional and more difficult systems.

RC: fig 5 (a) and 1(b). if I understand correctly, the x-axis is the pseudo time instead of the real time in the dynamical system. if it is the case, it would be benificial to add an x-axis label to avoid any confusion.

AR: Yes the axis is in pseudo-time and we will add this labeling to avoid confusion, thanks for spotting this.

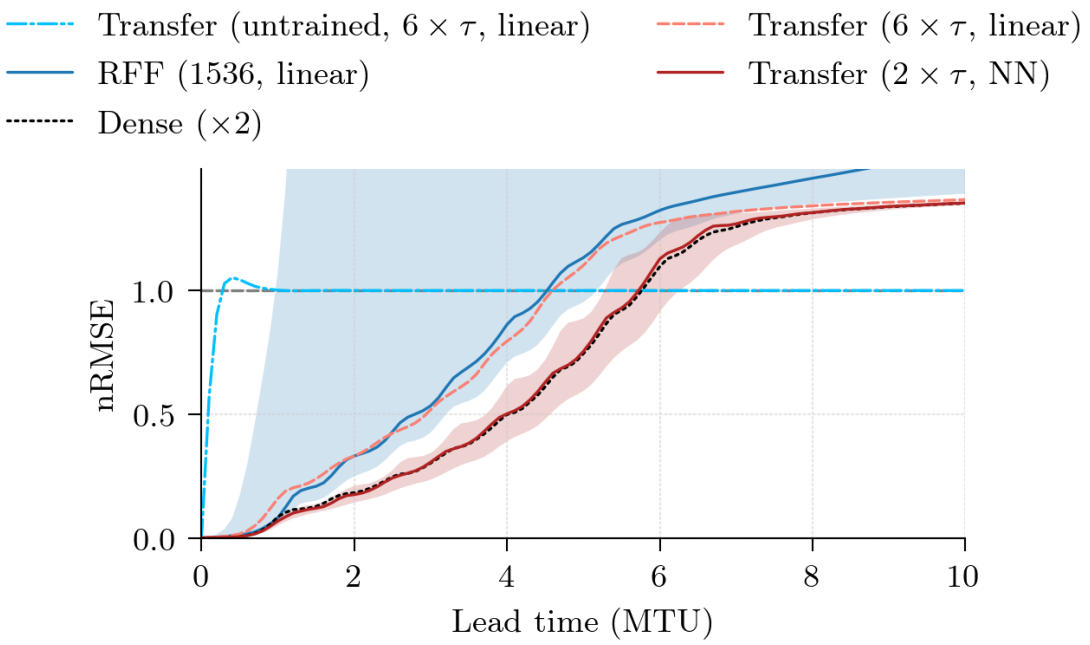

Figure 1: The normalized root-mean-squared error (nRMSE) as function of integration time steps for random Fourier features (RFF) with 1536 features and a linear regression, a *dense* neural network with two layers trained from scratch, and transfer learned models (Transfer) with features from six tipseudo-time steps with a linear regression and from two pseudo-time steps with a neural network. Shown is the median across ten different random seeds. Additionally, for the RFF (1536, linear) and the Transfer ($2 \times \tau$, NN) experiments, the 5th and 95th percentile is depicted as shading.