Using Probabilistic Machine Learning to Better Model Temporal Patterns in Parameterizations: a case study with the Lorenz 96 model.

October 2022

The manuscript under consideration implements Recurrent Neural Networks for stochastic parameterization of the Lorenz 96 Model. The authors compare the performance of their model in emulation of the Lorenz 96 to a GAN and a baseline polynomial. Their goal is to capture temporal patterns of this idealized atmosphere better and generalize to out-of-sample atmospheric states. They see this work as building on the more naive addition of red noise to parameterizations for a probabilistic framework. As evidence of this, they present approximated Kull-Back Liebler Divergences and likelihood scores from the various models examined.

There are a number of aspects to this manuscript that make it appealing. It uses modern machine learning techniques (LSTM) to try and address a problem (stochasticity) that previous implementations of machine learning in this space have struggled with. However, this manuscript has significant shortcomings. The most significant of these are the evaluation metrics and baselines the authors use to attempt to prove the significance of their findings. In their present form, it is not clear if we can gain any insight from them. Furthermore, crucial details about model training are missing. There are general issues with readability including incomplete figures, undefined terms, and poorly defined concepts that would make this article difficult for a general audience. For these reasons, I recommend Major Revisions prior to consideration for publication in the Journal of European Geosciences Union (EGU). Below, I give recommendations to address the manuscript's shortcomings

Contents

1 Major Issues

1.1 Missing context and clarity in manuscript setup

The authors do a nice job of framing the issue that small-scale and unresolved processes are the greatest drivers of uncertainty in the cloud-climate feedback. However, given its importance to the premise of the manuscript, the authors need to give an explanation of what exactly "red noise" is rather than just saying it is important to parameterization. Likewise for "white noise" mentioned but not explained (and possibly blue noise as well). This will make the work more approachable to a more general audience. There are other examples of this throughout the manuscript but one more to point out is that, especially given it is the model of choice for the authors, the difference between traditional RNNs and LSTM should be explained more, and how specifically it is helpful in this framework as well as terms like "gated".

1.2 clarity in Training procedure

The authors do include some important information regarding the development of their model but more information is needed for reproducibility. The authors should clarify the process of hyper-parameter tuning that lead to their final hyper-parameter choices. What was tried and what wasn't? The authors should also clarify if they did a training/validation/test split or if they just do a training/validation split – and if the latter they should provide some justification for this decision. Additionally, did the authors experiment with different training/validation splits before settling on the final one, and if not a justification should be provided for the absence of this test.

I appreciate the authors details in Section 4.4 about computational costs between different models. But in terms of reproducibility, it would be helpful to include the amount of gpu/cpu/tpu hours required to train and tune these models.

1.3 Climate valuation Metric lacks justification

I think the idea of approximating the true continuous distributions through a discrete one is likely appropriate here. But it needs proper justification and explanation (and citations). The authors are deriving the approach through "vector quantization" [2, 4], something they should introduce and explain. But through this vector quantization, the authors are not getting a true KL divergence, but rather estimating the lower bound of the true KL Divergence and their equation in 4.2 (not numbered) should be adjusted to reflect this for accuracy [1].

Because this is an estimation it requires additional justification. The principle (as though be included in this explanation) is that the more "bins" the more accurate the approximation of the lower bound of the KL Divergence. The authors need to justify the number of bins. More specifically they need to include either a graph or table (would recommend including this in the appendix rather than the main text) showing how the approximated KL Divergence changes as bin count is increased and at what bin count the approximated KL Divergence converges. This is the bin count the authors should use. As it stands now we cannot interpret any significance from the results because we don't know if the approximation of the KL Divergence they use is accurate or not.

Additionally, they need to confirm that Figures 3, 5, and 7 are built off of common bins. It would also be appropriate for the authors to acknowledge they are not the first to use this approach, it has been used in both supervised and unsupervised frameworks [1, 3]. Lastly, given that the KL Divergence is non0symmetric, there is an argument that the authors should symmetrize their approximated KL Divergences into Jensen-Shannon Divergences.

1.4 Missing information

There are a number of places in the manuscript missing citations, context, or other information or where figures are not complete. I encourage the authors to do a thorough proofread but I will include some of them below.

2 Technical Edits

(Line 25-27) [Add citation to paper with an idealized model.]

(Lines 31) [Add a reference to a section/figure that supports this claim.]

(Line 65) [Please be careful about calling any neural networks "deterministic" For most, the stochastic gradient descent and random seedings in initialization can/do add some variance that makes the label "deterministic" inaccurate.]

(Lines 73-73) [Ciation needed for the claim about minmax loss.]

(Lines 104-105) [Is there a physical justification for this in terms of phenomena in the atmosphere? If so, would be helpful to mention that here.]

(Line 105) [Word "good" is opinionated and vague.]

(Line 159) [Language is unscientific.]

(161 and elsewhere) [Would suggest changing the term "truth data" to the more widely used "target data" to avoid confusion.]

(Line 163) [Why is the GAN only trained on F = 20? If the authors are trying to claim their model is superior, should it not be trained similarly to the RNN. Otherwise, the scope of the claim needs to be limited to moving beyond the specific paper the GAN was from rather than the approach itself.]

(Figure 2) [Need y axis label and title]

(Figure 3) [Remove ticks on the y-axis of plot b. Add titles.]

(Figures 3, 5, 7) [It is difficult to observe anything quantitative in these plots (as the authors themselves state in line 193). If the authors want to keep them, I would suggest moving to the Appendix.]

(Line 209) [Explicitely define PCx as "Principle Component".]

(Figure 4) [Add a title. Remove x ticks on a,b. Remove y ticks on b, d. This will allow the plots to be a bit bigger. Make the colorbar discrete to match the plots]

(Figure 5) [Remove y ticks from b. Add a title.).]

(Figure 7) [Remove y ticks from b. Add a title.).]

(Figure 5) [Remove y ticks from b, d and x ticks from a.c. Add a title. Set same y axis max on all four plots.).]

(Lines 259-263) [Add citation.]

(Lines 270) [Reference the appropriate table.]

(Lines 277) [Reference the appropriate table.]

(Lines 310-312) [Would suggest including references to diffusion models or VAEs here.]

References

[1] J. Duchi. Lecture notes for statistics 311/electrical engineering 377. *Stanford*, 2:23, 2016.

[2] R. Gray. Vector quantization. *IEEE Assp Magazine*, 1(2):4–29, 1984.

[3] G. Mooers, M. Pritchard, T. Beucler, P. Srivastava, H. Mangipudi, L. Peng, P. Gentine, and S. Mandt. Comparing storm resolving models and climates via unsupervised machine learning, 2022.

[4] S. Rabanser, S. Günnemann, and Z. Lipton. Failing loudly: An empirical study of methods for detecting dataset shift. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.