Review: Global Climate Modeling with Improved Precipitation Characteristics by Learning Physics (GRIST-MPS v1.0) from Global Storm-Resolving Modeling

1 Summary of contributions

The paper presents an experiment developing an ML-based parametrization for global climate modeling. Training data is generated from a high-resolution storm resolving model, which is used to train a ResNet neural network. The ResNet is then used to parametrize heat and moisture tendencies in a coarser GCM. Evaluation is carried out both offline and online, with an emphasis on the ability of the final setup to more accurately represent the distribution of precipitation.

2 Strengths

1. The overall approach of using a short simulation from a storm-resolving model, training a neural network to output Q_1 and Q_2, and then plugging this parametrization into the coarser GCM is sound and interesting.

2. The overall methodlogy is well described, and the authors discuss both offline performance, stability and online performance. All of this is interesting and important to include in a work around ML parametrizations.

3. The learned parametrization clearly leads to stable simulations that reproduce realistic statistics to a similar degree as the CPS setup. Given that the parametrization is not trained or fine-tuned in any online setting (running within the GCM), this is an important achievement in itself. I find there to be a lot of nuance to the exact results for precipitation.

3 Weaknesses and comments

1. The fact that the lowest levels still have to use the CPS tendencies, rather than the ML suite, is a weakness of the work. Where this is discussed, the authors explain clearly that it is needed for stability. This limitation is however not transparent in the other parts of the paper, which clearly state that the MPS is integrated in the GCM and that it can then perform stable simulations. In reality, the MPS is used for some levels and the CPS for some, and then the system is stable. This should be more accurately presented in the introduction and in the summary and outlook.

2. The paper features a comparison between three different neural network architectures, and the authors describe this as a contribution of the work. However, there are almost no details given about the CNN and DNN models. This makes it hard to derive any takeaways from section 3.2, as I don't know what we are really comparing. The training procedures for these is described as identical to the ResNet, but the architectures are never described. Some key questions I would have are: How many and which kinds of layers do these have? How many parameters do they

have in comparison to the ResNet? This could be presented very briefly in the main text, with details given in table form in the supplementary material.

3. Some of the confusion above also stems from the names of these. In the typical machine learning terminology, these classes of models are subsets of each other as ResNets $\subset$ CNNs $\subset$ DNNs. While it is understandable that the difference between the ResNet and CNN is the presence of residual connections, I do recommend that the term DNN is changed. I can only assume that the authors here refer to a non-convolutional architecture, stacking linear layers, activation functions and (potentially) normalization layers. This is better called a Multi-Layer Perceptron (MLP) or a Fully Connected Network (FCN).

4. I miss offline evaluation results for the different neural network models. It feels quite strange to be immediately thrown into the online evaluation in section 3.2. In particular, in these results the CNN performs quite poorly, but it is unclear if this is just because it achieves larger errors on the Q_1 and Q_2 prediction (an offline evaluation would show this), or if this poor performance arises once it is combined with the GCM.

5. L187: Here eight NNs are discussed, but where does this number eight come from? Supposedly there were multiple ResNet configurations considered, and at this point eight candidates remain. The details of what these configurations were are however unclear.

6. Most of the figures lack clear axis labels, making them very hard to read independently. Sometimes the axis labels or units are written in the corner of a plot, or the x-axis labels sits under the colorbar (figure 5), but the colorbar lacks label. Quantities and units mentioned in figure captions should not compensate for the lack of axis labels. Each axis (including the colorbar) should have a label of the quantity it represents and, where relevant, its unit. For figures in a row it is sufficient if the leftmost y-axis is labeled.

Minor comments

1. L128: The paper would benefit from some additional explanation of how the residual method is applied here, to help guide the reader. This is a quite central step in the framework used in this work, so just a couple sentences to help the reader understand the different steps would make the section more pedagogical.

2. L162: It is described how convolutional layers explicitly resolve vertical couplings in profiles. It seems that rather they have the capability to do this, but we can not know for sure that a model has learned this except by inspecting the final behavior with trained parameters? Or should this be read as an observation for the final trained model? Also a linear model should be capable of doing the same, capturing interactions between different levels.

3. Some of the figure and table references seem to have the wrong number, this should be checked by the authors throughout the whole paper.

4. Bishop et al. (1995) is used in the text, but missing from the reference list.

5. L210: Typo, “ecomonical”

6. L357: “the CNN’s excessive sensitivity to localized features” is described as a reason for its poor performance. But does not the ResNet also feature convolutional layers, and should have the same sensitivity? This does not feel like a convincing explanation, as the ResNet is performing the best.

7. L370: The term “regulatization properties” is thrown around very loosely and here it is entirely unclear what this relates to. A fully-connected network is maximally expressive and has more

degrees of freedom, so should be argued is less regularized than a CNN. The current explanation is not convincing and needs to be clarified.

8. L397: Interpolating the data can not triple the *effective* sample size. It creates three times as many samples, but as these are highly correlated the *effective* sample size is less than three times as large.

9. L418: The NWP-like experiments are a valuable addition to the paper. However, it feels very surprising that they are first mentioned here, in the summary section. These should be mentioned where other results are presented.

4 Recommendation

Overall the paper presents an interesting study that should be of value to the community. However, the fact that the ML parametrization does not replace the conventional parametrization on all levels is a limitation that needs to be more accurately represented in the paper, and some clarifications are needed in order to make the neural network comparison insightful. I recommend that the paper should be accepted for publication after some minor revisions, fixing these points.