Dear Editor,

The authors sincerely thank you for your time and effort in handling this manuscript. We have revised the paper carefully in accordance with the reviewer's comments, and our detailed point-by-point response is attached.

--

Best regards,

The authors

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 Q1 and Q2, 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.

Reply: The authors sincerely thank this reviewer for detailed review and comments on this manuscript. Our response to each comment is given below in blue font.

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.

Reply: We thank the reviewer for this comment, with which we agree. In the revised manuscript, we have more clearly presented this limitation in both the Introduction and the Summary and Outlook sections. (***Line: 89-91, Line: 444-445***)

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.

Reply: In the revised manuscript, we have added a concise description of three architectures in the main text and a tabular summary of their structures and trainable parameter counts, in the beginning of Section 2.3. (***Line: 172-180***)

"We examined three neural networks (Table 3). All of them share the same input variables, output variables, preprocessing procedure, loss function, optimizer, and online coupling interface, so the comparison isolates the effect of network architecture as much as possible. The input has 5 channels and 30 vertical levels, the output has 2 channels and 30 vertical levels, and the hidden width is 128. The ResNet consists of one initial 1-D convolutional layer, five residual units, and one output convolutional layer, with each residual unit comprising two 1-D convolutional layers plus an element-wise shortcut addition. The plain CNN retains the same number of convolutional weight layers but omits the residual shortcuts. The MLP consists of 11 fully connected hidden layers applied to the flattened vertical column, followed by a single output layer. The trainable parameter counts are 495,618 for ResNet, 495,618 for CNN, and 192,188 for MLP. ResNet and CNN have identical parameter counts because the residual shortcut is an Add operation and introduces no trainable parameters."

Table 3 The three networks studied and their specific setup.

Architecture	Layer structure	Hidden width	Trainable parameters	Role in comparison
ResNet	1 Conv1D + 5 residual units + 1 Conv1D output layer; each residual unit contains two Conv1D layers and an additive shortcut.	128 kernels	495,618	Default MPS architecture and best-performing offline and online configuration.
CNN	1 Conv1D + 10 plain Conv1D hidden layers + 1 Conv1D output layer, without residual shortcuts.	128 kernels	495,618	Tests the effect of removing residual connections while keeping a comparable convolutional capacity.
MLP	Flattened column input followed by 11 fully connected	128 neurons	192,188	Tests a non-convolutional

	hidden layers and one output layer reshaped to vertical profiles.	fully connected architecture.

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 ⊂ CNNs ⊂ 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).

Reply: Thanks for this suggestion. We agree that "DNN" is overly general in this context. ***We have revised the term to "MLP" accordingly.***

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 Q1 and Q2 prediction (an offline evaluation would show this), or if this poor performance arises once it is combined with the GCM.

Reply: We agree that the neural-network comparison should include offline evidence before discussing online simulations. We have therefore updated Figure 2 and the associated discussion to include offline R^2 evaluations for different architectures on an independent GSRM validation experiment. The revised Figure 2 and related discussion show the latitude-model-level R^2 distributions of the selected ResNet for Q_1 and Q_2, and compares the vertical R^2 profiles among ResNet, CNN, and MLP. (***Line: 200-203***)

CNN and MLP show smaller offline R^2 values than ResNet. They still satisfied the minimum offline screening criteria based on the jointly trained Q_1, Q_2 tendencies, and were thus retained for online testing. Their lower R^2 on the independent validation set suggests an architecture-dependent limitation under the same data, preprocessing, loss function, and optimization procedure, rather than simply insufficient training.

Therefore, the online experiments are used to test how these acceptable while less accurate AI models behave online when coupled with the GCM. We have clarified this logic in the revised manuscript by presenting offline evaluation as a screening and diagnostic step (Figure 2), and online testing as a separate assessment of stability and climate fidelity.

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.

Reply: Thank you for pointing out this ambiguity. We have clarified that the "eight NNs" all refer to ResNet models with the same architecture and training configuration, but initialized with different random seeds. All eight ResNet models satisfied our predefined offline screening criteria based on validation-set MSE and R^2, and were therefore incoporated to online integration tests.

Among these eight seed-dependent ResNet realizations, two remained stable for more than six years. We selected the better-performing model as the final MPS configuration based on its online climate-performance metrics, especially precipitation RMSE. We have revised the relevant sentence to make this selection sequence explicit: training several ResNet realizations with different random seeds, applying offline screening, conducting online stability tests, and selecting the final model based on online climate performance. (***Line: 209-212***)

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.

Reply: Thanks for this suggestion. ***We have redrawn most of the figures to improve their clarity (Axis labels, unit, etc.). In particular, Figures 2, 3, 4, 5, 6, 8, 9, and 10 have been further polished.***

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.

Reply: Thanks for this insightful suggestion. ***We have added some sentences in the first paragraph of Section 2.2*** to address this issue. Correspondingly, ***we also adjusted some sentences in the second paragraph of Section 2.2***.

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 cannot 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.

Reply: We agree with this point. The statement that "convolutional layers explicitly resolve vertical couplings in profiles" reflects a design motivation rather than a proven source of success.

We have revised the sentence slightly to avoid an overly strong claim. (Line: 181-184)

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.

Reply: Thank you. We have checked and verified the consistency. ***Some wrong figure captions (e.g., Figure 6, Figure 9) and some other inconsistent statements have been corrected***.

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

Reply: ***Added.***

5. L210: Typo, "ecomonical".

Reply: Thanks. ***Fixed.***

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.

Reply: ***We have slightly revised this statement here***. (**Line: 382-385**)

"These systematic discrepancies suggest that precipitation simulations are strongly influenced by architecture-dependent behavior, including differences in optimization, information propagation, and the subsequent amplification of tendency errors through online coupling."

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.

Reply: We have revised this explanation, in the last paragraph of Section 3.2. In the revised manuscript, we avoid claiming that the fully connected MLP is inherently more regularized than the CNN. Instead, we describe the result more directly (**Line: 397-401**):

"The MLP shows better agreement in light-rain frequency despite its overall weaker precipitation intensity. This apparent paradox originates from the smoother and weaker Q_1/Q_2 responses produced by the trained MLP in the online simulations, which tend to suppress heavy precipitation while keeping the dominant light-rain category (1–10 mm/day; accounting for >78% of tropical rainfall events) closer to observations. ResNet demonstrates intermediate performance."

We now treat this as an empirical behavior of the trained MLP configuration rather than as a general regularization property of fully connected networks.

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.

Reply: We agree that interpolation does not strictly triple the effective sample size. We would like to clarify, however, that the primary motivation for applying linear interpolation is to generate data with shorter time intervals for computing Q_1 and Q_2, bringing them closer to the time step of the coarse-resolution model (1200 s); the increase in data volume is a secondary benefit.

Although the interpolated samples are correlated, this temporal-resolution alignment helps reduce the mismatch between the offline training targets and the online coupling interval. We therefore regard it as a practical way to improve training and online stability, rather than as a strict threefold increase in effective sample size. ***We have clarified this point in the revised manuscript***. (**Line: 228-231**)

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.

Reply: Thanks. ***We have now relocated and referenced this experiment at the end of Section 3.2***.

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.

Reply: Thank you. Based on your comments and our own inspection, we have revised the manuscript to improve its quality.