Response to reviewer 1

1 Summary

Overall, the authors have addressed my comments. After reading the manuscript I still have a major comment on the manuscript regarding how the authors have reported their results. Overall, I think that the manuscript deserves publication but I would like to get a clarification from the author regarding why there are inconsistencies in the results between the different versions of the manuscript.

We appreciate you very much for the suggestions that improved our work. We performed sensitivity tests and re-ran prognostic simulations for better consistency based on the reviewers' suggestions. As you mentioned, there are inconsistencies in the results between the different versions of the manuscript. We try to make a clarification as follows.

2 Major Comments

I might be misinterpreting what the authors have been writing, but I think that the authors might be using some inconsistent reporting when they present their data. Mostly I am worried since during the evolution of the paper between iterations, there are some inconsistencies in the text and figures, and at least to my understanding the authors did not clarify these differences. I suggest that the authors would clarify these inconsistencies. My main concern is rooted in how the authors report and write about their methodology of trial and error and in the large differences between the results presented in the first version of the manuscript and the recent version of the manuscript.

- *Results in Figures 12 and 14. The results presented in figure 12 shows overall that NNCAM is performing better than CAM in terms of latitudinal distribution of precipitation. However, in the first version of the manuscript a similar figure was shown (figure 11 in the first version) and in the first version NNCAM was performing substantially worse than CAM. How come this result was changed? Did the authors change something that caused this difference, and if yes where was it reported? I know that now the analysis is performed over more years, but I doubt that this can explain the large differences between these results. (A similar comment on figure 14 that is now different compared to figure 13 in the first version of the manuscript).*

Reply: Thanks for this comment. In all versions of manuscripts, the prognostic validation uses the same NNCAM, and only the model start date is different.

In the 1st version of the manuscript, NNCAM started with the SPCAM checkpoint on July 1, 1998, and SPCAM and CAM started on January 1, 1998. We chose the next four years, from January 1, 1999 to December 31, 2002 for evaluation. And this SPCAM checkpoint was made by running SPCAM from 1998-01-01 to 1998-07-01.

As requested by the Reviewer 1, in the 2nd and 3rd versions of the manuscripts, the start time of NNCAM, CAM, and SPCAM was all January 1, 1998. All models were run for 6 years, with the first year for spin up and the next 5 years (January 1, 1999, to December 31, 2003) for evaluation and comparison.

By applying a different way of the startup, the simulations results show some differences in the land-surface precipitation (the right column in Figure 12) and the MJO spectra signals (Figure 14), which is an interesting point and should be studied in the future.

Results from online tests. In the first version of the manuscript the authors wrote that: "prepared 50 groups of ResMLPs with similar R2 as candidate models using different train samples and epochs. Secondly, we conducted comprehensive prognostic tests on these candidate neural networks and obtained the feasible NN-Parameterization schemes that can support NNCAMs stable simulation for multiple years." From that statement it is clear that the authors have run online simulations with at least 50 NNs and some of them were stable. Since this is an important part of the authors' work (where they also argue that a fully connected DNN are not stable and therefore they use a different architecture) in the first revision I requested that they give more details about how many networks were stable and how many were not (for each type of NN that they tried). In the response to my first review the authors wrote that they add:

"Figure 4 shows the offline validation versus the number of prognostic steps that our NNCAM can run. First, the DNN parameterizations are less accurate than the ResDNN ones in terms of offline validation accuracy. As a results, all the DNN parameterizations cannot run stably longer than half a year in prognostic tests. For the ResDNNs (blue dots and black inverted triangles), the less well-trained ones with high MSE crash within half a year simulation. However, when the offline MSE of ResDNN decreases to a certain level the ResDNN parameterization may run stably for long periods. In Figure 4, we observed 4 ResDNNs can run stably."

In the figure that the authors added they show 3 DNN and 21 ResDNNs. In the initial version of their manuscript the author wrote that they have prepared and tested 50 groups of networks so I am not sure why they report only on part of their results.

Furthermore, since I thought that the comparison between 21 runs (out of the 21 runs of ResDNN only 4 of the runs were stable) and 3 runs of DNN cannot support the claim that DNNs aren't stable is not established since only very few networks DNN were examined. In a response for my comment the authors included in their latest response:

"We tested all 37 NN sets (27 ResDNN sets and 10 DNN sets) in the sensitivity tests. As shown in Figure 4, there are 10 ResDNN sets that can sustain simulations of longer than 10 years. Figure 4 is intended to show the relationship between the MSEh and the stability, not to prove that the ResDNN is better than the DNN in terms of stability."

What I found strange is that they have added 3 additional simulations with the ResDNN case, and found all additional simulations to be stable (I can count only 24 ResDNN and not 27; BTW, the addition of DNNs makes the argument much better). I find this *result of adding 3 additional ResDNN simulations and finding them to be all stable is not very likely given that they previously reported to find that only 4 out of 21 simulations were stable but maybe I am missing something here, so it would be great if the authors could clarify.*

I would encourage the authors to include all the results they have obtained and reported in the first version of the manuscript (e.g., could the authors provide the results of the 50 simulations they reported in the first version of the manuscript? Is there a reason that these results are not included?)

Reply: Thanks for the question. It should be noted that all trainsets used in the 1st, 2nd, and 3rd versions of the manuscripts are split and sampled by time on the same original data which are stored open-soured on Zenodo as netCDF files (https://doi.org/10.5281/zenodo.5625616). Furthermore, we use the same set of ResDNNs for prognostic validation in all versions of the manuscripts. This set of ResDNNs has been open-sourced at Zenodo (https://doi.org/10.5281/zenodo.5596273).

To make our sensitivity tests more convincing in later versions of the manuscripts, we reorganized and described the trainset according to the suggestions by Reviewer 1. Especially, the testset was changed to the SPCAM's simulation in 2000 which is well separated from the trainset. To ensure the rigor of the sensitivity test, we did not use the first 50 NN sets because the conditions (datasets for training and evaluation) to build these NN sets are not guaranteed to be the same as the later 37 NN sets in sensitivity tests. In the 2nd revision of the manuscripts, we just arbitrarily selected 21 NN sets (18 ResDNN sets and 3 DNN sets) from the 37 NN sets and performed prognostic runs according to the settings in section 3.1. As requested by Reviewer 1, in the 3rd version of the manuscript, we included all the 37 NN sets by performing prognostic runs over the rest NN sets (9 ResDNN sets and 7 DNN sets) for sensitivity tests.

Among the 27 ResDNN sets, 10 ResDNN sets are stable, shown in the Figure 4 with close MSE_h in x-axis and large prognostic steps marked by infinite line, leading to several (the results of four sets of ResDNNs) points overlapping. We replotted Figure 4 and showed three overlapping points above the infinity line.

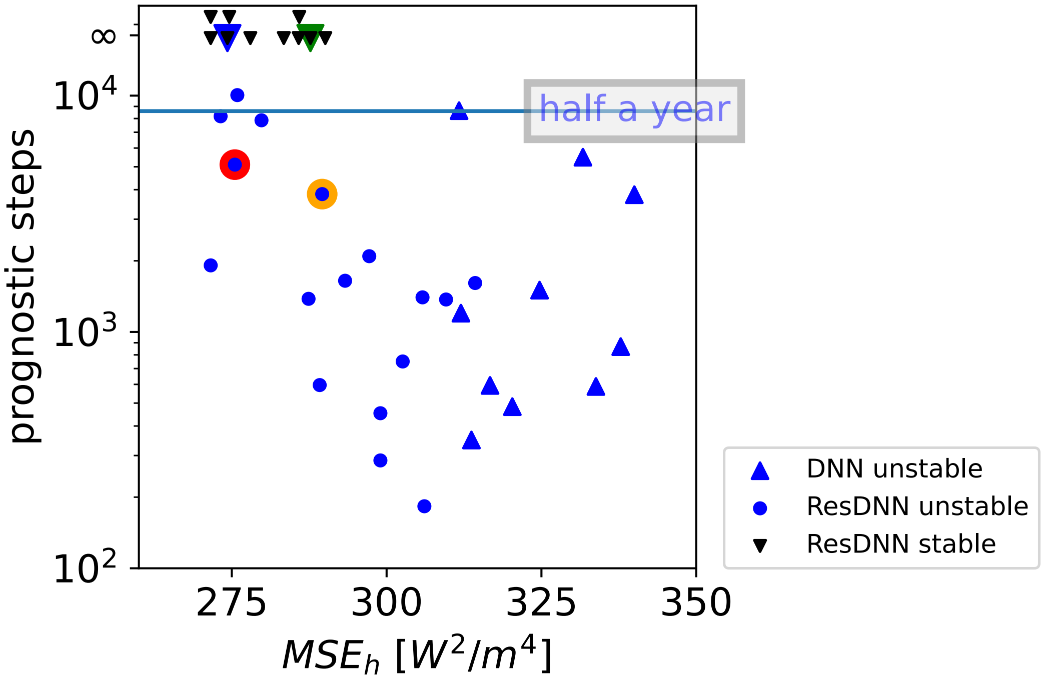

Figure 4. The mean square error of the offline moist static energy vs. the prognostic steps. The black inverted triangles (the three black inverted triangles above the infinity line to avoid overlapping) denote stable NN coupled prognostic simulations that last for more than 10 years. The blue dots denote unstable simulations, and the blue triangles denote unstable DNNs. The dots with colored outlines are shown in Figure 5 for the time evolution of the globally averaged energy.

3 minor comments

- *lines 100-102: The authors write that the Yuval and O'gorman 2020 used random forest to predict fluxes to ensure physical constraints, but this is not correct. As far as I understand the usage of random forest allows to ensure linear physical constraints because RF just averages over samples.*

Reply: Thanks for pointing this out. We have changed them in the latest revised manuscript in line 89-91 as: *"Yuval and O'Gorman (2020) used the random forest algorithm to develop an ML parameterization based on training data from a high-resolution idealized 3-D model with a setup on the equatorial beta plane. They used two independent random forests to separately emulate different processes. Later, Yuval et al. (2021) ensured the physical constrains by using an NN parameterization with a special structure to predict the subgrid fluxes instead of tendencies"*

- *line 168: the authors use the terms qrs and qrl but they never define these.*

Reply: Thanks for pointing this out. We have removed the terms *qrs* and *qrl* in line 153.

Response to reviewer 2

1 Summary

I thank the authors for the revised manuscript and replies. I think the article is much clearer with the advantages and disadvantages of the neural network models. I have a few minor points below but otherwise I would be happy for the manuscript to be accepted.

In my opinion, the tool that the authors develop to do coupling in parallel simulations is worthy of more detail and publicity. I would suggest that the authors package their tool up with clear documentation. If this is well done, and the tool has the advantages they suggest over existing coupling methods, then researchers in the field will be keen to adopt the tool.

Thank you very much for your constructive comments. We have uploaded the coupler NN-GCM in https://doi.org/10.5281/zenodo.5596273 together with the NNCAM model. Proper instructions were added to the root directory of open-sourced codes and described in the manuscript. We are making further development on it and would like to share the tool with the community. We reply to your comments as follows.

2 minor comments

R1 is an interesting figure. I think it should be included in the supplementary material and a reference that it is there be added to Figure 7. I understand that the authors wanted figure 7 to compare with previous articles but I think R1 is a valuable addition too.

Reply: Thanks for the comment. We have added the Figure R1 in the supplementary as the Figure S6 combined with a reference information in the figure captions.

L295: "10 DNN sets and dozens" - please give the exact number here, imprecision is not helpful.

Reply: Thanks. We have made clear that there are 27 ResDNN sets in the latest manuscript in line 295.

L304: g & Cp no longer features in these equations.

Reply: Thanks for the comment. *"Cp"* is no longer used in the equation and we have removed it in line 304.