1 Summary

The authors have addressed some of my comments and added some sections to the manuscript. The authors did add some analysis that substantially improve the manuscript. After reading the manuscript I still have several major comments on the manuscript, and I think the manuscript is not ready for publication yet. In case the authors intend to resubmit, please clearly highlight in the response all the important changes that were made in the manuscript since I felt such changes were not highlighted enough in their previous response.

Overall, I think that the manuscript deserves publication but still needs major revisions before it can be published.

2 Major Comments

- I still think the authors overstate their achievement. The authors highlight many times that their NNCAM is more accurate than CAM. I think that in a few aspects it is true, but in many aspects it is not. The authors should highlight also the parts that NNCAM fails. e.g., the climatology (which is a crucial thing in simulations of climate) is not accurate (e.g., RMSE in temperature is 3 degrees with certain regions getting to 10 degrees in the zonal mean! which is much larger than CAM RMSE). Please clearly mention the large biasses in climatology (e.g., in the abstract). For example, I do not understand why "most importantly, NN parameterization successfully reproduces the climate variability in a superparameterizated GCM" - is it more important than achieving the correct climatology? Furthermore, the large errors in the climatology (e.g., of temperature) should be shown in the main text and not in the SI since it is a key problem in the simulations and should be highlighted. In addition, in the previous version the authors showed (Figure 11 in previous version) results for precipitation such that it was easy to see that SPCAM is less accurate than CAM in many precipitation features, please keep the figure in the manuscript.

- The authors write that a DNN cannot run stably. However, I do not think that this claim is backed up by scientific evidence. The authors ran 3 different DNN and found them unstable (though with lower accuracy they could run longer than the RESNETs). The authors tested 10s (over 50 if I understand correctly) of ResNets and found only a few of them stable. If the authors want to argue that DNN cannot be run stably, they should follow the same methodology as with the ResNets and train 10s of DNNs to show that they do not run stably. I am also not convinced yet that a that the key for stable runs is accurate NNs since some NNs are more accurate than your stable networks and they are still unstable. If the authors want to argue this they should provide better evidence. Otherwise they should clearly write that this is a hypothesis that they have not yet provided clear evidence for.

- The authors write that they do not couple the NN radiation in the sensitivity tests (lines 272-275). I am surprise by this statement as it did not exist in the previous version of the manuscript and was not mentioned in their response to my review. I am not sure I understand correctly what it means. Does the radiation scheme run in the prognostic simulations? If not, this is very strange to me because the text discusses several times (even in the implementation section) how they implemented NN radiation. Furthermore, the reason that the authors give why they do not use the NN radiation in the prognostic simulations is that it is very accurate. I do not think that this is a convincing reason (especially since if the NN radiation is very good, I think it is important to show that it works also in a prognostic test). Also, the argument that the NN radiation is very accurate is based on a zonally averaged result so please include the R^2 result before using a zonal average. Overall, the authors should show online results when their NNs change SP (including radiation). If they decide not to use their NN radiation in prognostic tests please provide a convincing explanation why (does simulations have climate drift like in the previous version of the manuscript when radiation is used? do online results similar to what is presented). I note that It is possible that I misunderstood what the authors did with radiation.

- The authors ran several stable NNs, and from what I understand they choose the "best" (line 303-305: "we still have to use the trial-and-error to filter out unstable ones and then select the best ResDNN pair for moistening and heating rate that can reduplicate the total energy time evolution of SPCAM with the least deviation"). I feel that this is a bit of cherry picking and I am not sure what to think about this process. Maybe this should be highlighted in the abstract that for no clear reason there are large differences in the results of stable NNs and that you have trained multiple networks and present only the "best" results.

3 Comments

- The authors wrote that in SPCAM it is not possible to separate different processes. Why? As far as I understand, the authors could keep track on the different processes that run in the SP (which is SAM model) and model each process separately.

- The authors write that "In our study, the NN parameterizations are tendency-based trained with realistic configuration SPCAM simulation without any physical-constrain, where stability is indeed a problem to face". I disagree with this statement. There are many constraints in the parameterization. Their SP is SAM based model. Each process (e.g., subgrid convection, microphysics etc.) has certain constraints and physical relationships within each process.

- The authors should discuss negative precipitation in detail in the manuscript. They write in thier response that 27 percent of the time NN give negative precip but they do not mention in it in the manuscript.

- In the answer to my review, there are several times the authors respond in a manner that is not related to what I have asked. (e.g., I wrote "Is there some citation that can backup the statement that MLP can generalize better than other types of networks? If not please remove statement.")

- I insist that the authors will give the results for R^2 before zonal averaging. It will help to understand how accurate the networks are.

- The authors write in their response: "Our NN parameterization is trained with the loss function of mean squared error, which is not sensitive to incorrect predictions of small values. In Figure R1b, the local variance/std is close to zero for those low skill regions. The MSE in those regions is also low but is still high compared with its variance. Therefore, when calculating R2 as 1-mse/var, many of those low std regions will have R2 close to zero." However, the authors have low skill in the tropics where STD is large.

- There are still unclear citations for me. For example the author cite Moores et al. 2020 (and no such reference exist in their bibliography).

- The authors compare the offline results of their NN to SP and of a conventional parameterization to SP. They should highlight that this is not exactly a fair comparison. Their NN was tuned to emulate the SP, and CAM parameterizations where tuned to get a better online results (so the SP is not a ground truth for CAM).

- there are still many sentences in the manuscript that do not have context - e.g., "Brenowitz et al. (2020) proposed methods to interpret and stabilize ML parameterization of convection. In their work, a wave spectra analysis tool was introduced to explain why ML coupled GCMs blew up."

- - It is unclear to me how variables were normalized - did each output was normalized separately at each level? If yes - didn't it lead to problems in regions with very little subgrid values (e.g., there should be hardly any moisture and moisture tendencies in the stratosphere - how did you deal with that)

- The paragraph from line 185-196 has some statements I do agree with, but more importantly than that -I do not see why it is necessary to be included.

- The authors write "After numerous experiments" - can they please provide in the SI the hyperparameters that they used in their search?

- The author write that their NN predicts the temperature tendency (line 213). However, temperature is not a prognostic variable in SAM or in CAM so I do not see how it makes sense to modify the temperature with the NN and not the prognostic variables. Could the authors write what are the prognostic variables they are actually changing in the simulations.

- In line 217 there is a reference to figure 1 but I think the text should not refer to this figure

- In their response - the authors write "random forest is less likely to perform as accurately as neural networks and cannot be implemented in GPUs (Yuval et al., 2021)" however, I think that this statement is not correct and RFs were already used with GPUs

- The moist static energy should include a term gz (g is the acceleration due to gravity and z is the height), and currently the moist static energy is not written correctly.

- line 313: dump->damp

- line 379: The authors refer to figure S2 (line 380) and say it shows RMSE - but I do not think it shows this. Please add RMSE to the figure. F

I thank the authors for their detailed reply and significant changes to the manuscript. I now have a better understanding of where ResDNN exceeds the performance of CAM and where it does not yet do so. I believe that the article is close to acceptance but would like to see a few more improvements.

From author replies:

"As for swapping neural networks, we do not change the neural network for the 8 radiation fluxes because they are highly accurate and well trained with a collaborate R2 above 0.98."

So for each of the dots in figure 4, the same network is used for the 8 fluxes? This information should please be included in the manuscript (or highlighted if it is already included).

"the tendencies of temperature and moisture are rather difficult to train and, if not trained well or with the right NN architecture, can seriously affect the prognostic performance and stability. So, we swap the neural networks for dqv and dT together but not individually"

Sorry. I do not understand this important point. To my understanding all the configurations producing dots in figure 4 result from the same loss function and architecture. But each time the temperature NN is trained separately from the moisture NN. In this case why could they not be interchanged? Would it not be fairly simple to take the most accurate stable configuration, the most accurate unstable configuration, and exchange networks and see the impact. I think this is a reasonable request in order to elucidate where the destabilisation is originating.

L19: It is worth stating that these biases are larger than those found in CAM5.

L176: I do not agree with the authors' assessment here. To me the calculation of moisture and temperature tendencies resulting from moist processes does not count at multi-task learning but as multi-output regression (see figure 1 of Zhang). The authors may state that they chose to split the learning of these two outputs, but I do not believe the work of Crawshaw or Zhang and Yang gives any evidence that the task will be easier by doing so. If the authors still hold that Crawshaw and Zhang & Yang provide clear evidence that this task will be harder because of the multiple outputs could they please reference where in these papers this is discussed. Both papers are surveys, without any strong overall argument that MTL is flawed or more difficult. I would request that the authors change this section, as it could negative influence future research. In my mind it is future research to establish if splitting the task between two models (a) results in lower offline scores (b) produces more stable coupled results.

Section 2.2.2. I found it hard to establish how many neural networks are built and what they are each learning. Please could this be made more clear in the text. In the original version of the manuscript there was evidence that 4 NN were built (e.g. the now removed figure 2). But I find no explanation of the change. Did the authors change their approach, if so, why?

L456: successes -> succeeds

Figure 3: Several of the numbers mentioned in the caption do not appear in the figure.

Figure 5. Is there a reason you do not include CAM in this figure?

Figure 9. This caption could now be tidied as all panels represent the same period.

Figure 10, as with 9.

Figure 11: What bias amount does the white colour represent in panel (d)? Are the colours correct in panel (c)? There seems to be a global bias of between 2 and 4mm, based on the colours, yet the number reported in the top is 0.148. Personally, I would recommend using a neutral colour, e.g. white, for errors less than some threshold, e.g. 0.5mm. Otherwise the eye is drawn to places where the bias changes from positive to negative even when such a change is miniscule. But the authors are free to ignore this suggestion if they disagree.