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.

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).

- 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?)

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