1 Summary

This paper describes the use of a neural network (NN) to emulate a super-parameterization, and the implementation of the NN parameterization in CAM with realistic boundary conditions. The authors develop a coupler between python and Fortran which allows them to easily use python-trained NNs during online CAM runs. The authors show that the architecture they use (fully connected with skip connection) combined with the idea of separate the prediction of different outputs to different networks improve the performance of the ML parameterization. The author test their trained NNs in an online CAM setting and find that some of their NN-parameterizations lead to unstable simulations, but some lead to stable simulation (although the one simulation that the results are show for has a climate drift).

I generally think that it is impressive that the authors have succeeded to run a global GCM with realistic boundary conditions with NN-parameteriztions, however I have several general comments on the manuscript that need to be addressed (and more detailed comments below):

- (a) It seems to me very disappointing that the authors do not provide any idea/hypothesis why some of their networks are stable and others are unstable. From what I understood, the strategy of the authors is to conduct an exhaustive random search for an accurate and stable parameterization. Personally, I think that from a scientific point of view this is not satisfactory. Furthermore, a previous study by Brenowitz et al. (2020) suggested a method to understand why certain parameterizations are unstable (they suggested that when the ML parameterizations are coupled to dynamics it may lead to unstable gravity waves) and certain actions that can be tested in order to remove such instability (e.g., input ablation), so I think it would be crucial to understand if there is an underlying physical reason why some parameterization go unstable (for example by testing if similar framework like Brenowitz et al. (2020) could really help in this case). Furthermore, unlike Mooers et al. (2020) that showed that stability of simulations is correlated with accuracy of the parameterization, other papers (e.g., Rasp et al. (2018). Yuval and O'Gorman (2020)) showed that even inaccurate parameterization can run stably for long periods. So I think it is still undetermined whether it is really necessary to have a very accurate parameterization scheme in order for it to run stably in a more realistic case. Therefore, I think it would be great if the authors could add their input on this since it seems that they did train also less accurate networks than the resMLP that they use. Namely, did all the non-resMLP networks were unstable? Overall, The achievement of running CAM with orography and a with a neural network parameterization is an important achievement that should be documented in the literature. However, from a scientific point of view, the contribution of this paper is a bit limited, and also it is not clear to me whether the NN parameterization performs better than traditional CAM parameterizations (in a few aspects it does, but in many other it does not).

- (b) I think that more work on the text is needed to bring it to a level that is appropriate for the journal (see comments below). There are multiple repetitions, and sometimes

sentences with little context.

- (c) There are multiple times that the authors cite works that are not relevant to what they claim. For me especially alarming since I am not familiar with many of the references in the paper, and I found such mistakes on a small subset of references that I am familiar with. Therefore, I am worried that also other references are not relevant to what they claim.

- (d) I think that some of the claims that are made in the manuscript are not supported by the results that are shown.

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

2 Comments

- Abstract - I think that the authors overstate some of their outcomes (or at least do not show explicitly in the text these results; see more comments below). Some examples: I am not sure that in their parameterization outperforms CAM in terms of spatial distribution of precipitation and MJO propagation. Of the author want to state this, I think it is necessary to show the RMSE for the spatial distribution of precipitation and to show a quantitative comparison between the MJO spectrum/propagation in CAM, SPCAM and NNCAM.

- line 15 (and also through the abstract): the authors write that they learn from a cloud resolving model. This is confusing. I think that it is important to explain that they use SPCAM, and the task at hand is an emulation of a super-parameterization rather learning from a cloud-resolving output.

- line 38: I think that one important reference regarding the biasses that are caused by different parameterization is a paper by Wilcox and Donner (2007). I might be mistaken, but this is the only case that I know where they show how two different parameterization schemes in the same model lead to a very different frequency precipitation distribution.

- line 44: Is there a citation the authors can provide for supporting the sentence “Their interaction with the atmospheric circulation affects the transport and distribution of energy and is the largest source of precipitation biases”. If not please remove.

- line 56:“The CRMs have been applied to low-resolution GCMs to replace conventional cumulus convection and cloud microphysical parameterization schemes” - the sentence is unclear

- line 57: Is there any other cases where CRMs were nested in GCM except super-parameteriztation (SP)? if not, please don’t use "such as" as it implies there are other examples

- line 61: "an order of magnitude larger compared with that for CAM." -this really depends on the CRM configuration, right? Even in the authors simulation it seems that SPCAM uses much more than factor 10 compared to CAM.

- line 66: Strange sentence, and I think it has some grammatical errors (e.g., which "the data-driven parameterization scheme", from the sentence it seems you refer to a specific scheme, and I do not think it is the case)

- line 68: "More recently," -more recently than what?

- line 72: I cannot see why the citations of Schneider et al. (2017) or of Duben and Bauer are related here. There is no convection scheme learned in these papers.

- line 74: "trained ones" - unclear

- line 76: I do not see how Rasp (2020) citation is relevant here. This citation describes a method for online learning. Nothing related to an implementation of NN in GCM.

- line 85: "They found that minor changes, either to the training dataset or in the input/output vectors, can lead to model integration instabilities." - can you give a citation. I could not find such statement in the manuscript.

- line 89: There is a very relevant paper that the authors do not refer to. Brenowitz et al. (2020) have investigated both for SPCAM and for SAM why they can lead to numerical instabilities. Please add a discussion about it - and more importantly, I think that if the community could benefit from the this work it would be if the authors could identify why some of their networks are unstable (as was done by Brenowitz et al. (2020)).

- line 94-95: maybe instability and drift prevents the application some models, but other studies (that you cite) and also (Yuval et al., 2020) did not have any problem of unstable simulations. It is not fully determined why this is the case but the two possiblities that are raised in these papers are (a) because subgrid terms were calculated more accurately in these works compared to Brenowitz and Bretherton (2019), or (b) because these works succeeded to implement physical-constraints in the ML parameterizations).

- line 102: What is "auto-learning technique".

- line 104: Unclear what is "group of NNs" - do you mean different NNs for predicting different outputs? if yes, you should mention that this was already done by Yuval et al. (2020) (although in a different model).

- line 105: The authors write: "We apply two innovative methods in neural network models: multi-target training to achieve balanced results across diverse neural network outputs and multilayer perceptron with residual blocks (ResMLP) to enhance nonlinear fitting ability." To me it is unclear if what the authors write here is really

innovative: (a) multi-target training was already done by Yuval et al. (2020), although in a different model, so I could understand that it is a slightly different context. (b) If I understand correctly, Han et al. (2020) also used a DNN with shortcuts (called ResDNN in Han et al. (2020)) and showed it performs better than a standard DNN, which if I understand correctly is exactly what the authors call ResMLP

As a side comment, I think that the term multi-target training is unclear and should be modified.

- line 111: I understand that it would be difficult to integrate complicated NNs into Fortran, but DNN (or DNN with shortcuts) should be a very simple procedure in Fortran.

- line 115: I think it is very problematic not to mention here that the simulations are stable without mentioning that there is a climate drift (although relatively very slow one).

- line 145: Why is there a distinction between 2D CRM and CRM radiation? What does it mean CRM radiation?

- line 148-151: this is a repetition of things that were already written in the intro.

- figure 1: I might be missing something but for me these figures and the text that describe this figure is confusing. Aren't you just replacing SP with NN (like what was done in several other papers?). If yes I think that a very short and concise description will be much clearer.

- line 152-154: I am not sure on what the statement is based on. I can agree that the prediction is more difficult, but why it has more numerical sensitivities (I am not even sure what it means)

- line 161: the authors use the large-scale forcing as inputs. These were not used in previous papers trying to emulate SPCAM. It would be good to mention why they introduce these inputs, and if whether the inclusion of these inputs substantially improves the offline (or online) performance.

- line 164: The authors predict also different outputs compared to previous studies that used ML to emulate a SP. e.g., they do not predict the vertical structure of radiative heating, but do predict other quantities. It would be helpful to understand what is the reason that they use different outputs (and inputs), and explain to the reader the motivation for these choices (and highlight differences from previous attempts to learn ML parameterization from SPCAM).

- line 170: How does the authors deal with a negative precipitation (both offline and online)? Can you give information what is the percentage of (both online and offline samples) with negative precip?

- line 180: Random split for the train,validation and test set is not ideal (and not the common practice) due to the time correlation between samples. In order get a reliable results for the test set what usually is done is to take the samples for each of the data sets from different time intervals. Please do that and report offline performances when the test set is taken from a time interval that was used during training. Alternatively, make a justification for this choice.

- line 185: Is there some citation that can backup the statement that MLP can generalize better than other types of networks? If not please remove statement

- General comment: I might not understand what is the NN that you used. But if your multilayer perception is just a fully connected NN please say that, and preferably change the terminology to the common one (fully connected NN)

- line 186-192: Please add graphs supporting the statements (these graphs can also go in the supplementary). Also the last sentence is unclear to me and I am not sure how it is related to the rest of the text.

- line 193: The first sentence in the paragraph is unclear to me.

- line 193-194: I disagree with this sentence about the independence - see Yuval et al. (2020) where they use the dependencies between the tendencies of moisture and thermodynamic variable to predict only one of them, and diagnose the other (because of a 1 to 1 mapping between the two for parameterizations like microphysics and sedimentation).

- line 201: I am not sure why this statement is necessary, and how it fits in the manuscript.

- figure 3: To me it seems that the fully connected NN performs almost identically to the resMLP (roughly a difference of 0.002 in R^2). Did you test several fully connected NNs in an online setup and verified that using skip connections is really what makes the difference in terms of stable simulations? If not, please confirm that fully connected NNs do not lead to similar results (since if I understand correctly, the authors argue that this is one of the important aspects of their work). Furthermore, I think that the R^2 results that presented are confusing and not the relevant ones for the reader. R^2 should be calculated over different samples without any average before since such calculation gives the idea of the real performance of the network. So please change the way R^2 is evaluated

- line 213-214: On what are you basing your statement that if the NN is underfitting NNCAM crashes quickly? Did you test 100s of NNs? 1000s NNs? 1 NN? Please show a graph supporting this evidence (similar to the one that Ott did) if you want to keep this part of the text

- line 218-219: To me, this result (more accurate NNs that crushed earlier) indicates that the accuracy is not the important part of the stability. Previous work (e.g., Rasp et al. (2018) showed that also very shallow and less accurate NNs lead to stable NNCAM simulations in an aquaplanet setting). This raises a question why the resMLP is necessary and whether fully connected network could work as well.

- table 2: Please give the full details of the training (e.g., which optimized). Furthermore, I do not think that the data has to be shown in a table.

- lines 223-229: The division between the stable and unstable group is interesting and I think here would be necessary to use the tools developed by Brenowitz et al. (2020) to check stability. If these tools cannot explain the instabilities in the simulations you are conducting, it is important for the community to know this.

- line 249: The authors claim several times during the manuscript that it is difficult to code in Fortran an NN. However, a fully connected layer is very easy to implement as it involves only matrix multiplications so it would be good to clarify that it might be difficult to use some fancy architecture, but the basic one that the authors use is very simple to code in Fortran.

- line 249: "At runtime..." This is a confusing sentence and it is unclear to me why you cite the papers (as they use a different code infrastructure so I do not understand why these citations are included here)

- line 251: "outside", you need to explain outside of what.

- line 255: "In this..." this is repeating similar statements so I do not think it is necessary

- line 265-269: Sounds like a great achievement! (I haven't tried it myself though)

- line 278: Sounds great!

- line 296 and figure 7: Please show R^2 before zonal averaging! It does not make sense to me first zonal average (and also this was not done in previous work)

- figure 8: SP is especially important in the tropics, and it seems That the skill in the tropics is very low (many regions have R^2 close to 0). Can you give some insight?

- line 301: I think that it is difficult to determine whether this is a good job or not because there is no baseline to compare to. If you could provide a baseline from CAM (for offline prediction), then it would make sense to give this statement.

- line 306: The word "fitting" should not be there as far as I understand

- line 306-308: "As suggested..." This sentence is unclear to me. How do you manually tune an NN?

- line 309-314: I think that these sentences are not well related to each other.
- line 317-318: "At the same time...." If I remember correctly this is the default choice of SPCAM so it is very confusing you are writing this as if this is something special
- line 321: The fact simulation have a climate drift should be mentioned already in the abstract.
- line 321-325: Please compare NNCAM and CAM to the relevant simulated period of SPCAM. I do not understand why you compare different periods (and different length of time intervals).
- 324: Why here (CAM) 1 year spinup and for NNCAM - half a year? I guess that this is how it is initialized?
- 324: Is it critical to use SPCAM for initialization for NNCAM? if yes please explain why? Previous studies used a coarse run (in this case CAM) to initialize the model which makes more sense to me.
- figure 9: Please show RMSE in the figure for each of the sublots - so we could compare NN to CAM. Also for other figures and claims made by the the authors - if the authors want to state that NNCAM performs better than CAM, please provide a quantitative metric for the comparison.
- figure 10: Need to quantify the accuracy - please show RMSE
- figure 11: NNCAM has less skill than CAM so it would be fair to clearly mention it. Furthermore, please use some metric for the comparison if such a comparison is made. The fact that CAM is closer to SPCAM than NNCAM in many of the fields should be also mentioned in the abstract (to me it is a bit misleading to only mention in the abstract that NNCAM performs better than CAM without stating that in many aspects CAM is better than NNCAM)
- figure 12: Would be good To avoid the noise in this plot by using bins that have similar distances in a log scale.
- figure 13: Can the authors quantify the distance between distributions? It is difficult for me to determine if NNCAM or CAM is closer to SPCAM, since NNCAM has a very different MJO structure compared to SPCAM. Furthermore, please add more contours and make sure that the colorbar isn't saturated (at the moment it seems saturated).
- line 383/figure 14: The comparison should be between SPCAM propagation and NNCAM propagation because SPCAM is the baseline. I do not see why using the value of 5 m/s is relevant. To my eyes it seems that SPCAM propagates faster than 5m/s and has similar propagation as CAM. Namely, I disagree that NNCAM propagation is closer to SPCAM propagation compared to CAM.

- figure 14: Can you describe on which data was that calculated for.
- line 395: instead of "GCM" use or "a GCM" or "GCMs"
- line 399-404: You cannot not mention the climate drift.
- line 425-427: Also I think that the relevant comparison would be how fast it would run without the coupler. As far as I understand it should work pretty fast also without the coupler (which uses also additional resources)- as the matrix multiplication operation in Fortran is optimized pretty well.
- General comment: The authors mention that they have a group of NNs that lead to stable simulations. Are all these NNs lead to similar online results? If yes, please mention this. I suggest showing the STD for precipitation and a couple of other fields among the different stable online simulations you achieved since it will show the reader that there is no "cherry picking" with the choice of the simulation you end up showing.
- references not ordered alphabetically in some cases which makes it more difficult to find them.

References

Brenowitz, N. D., T. Beucler, M. Pritchard, and C. S. Bretherton, 2020: Interpreting and stabilizing machine-learning parametrizations of convection. *arXiv preprint arXiv:2003.06549*.

Brenowitz, N. D., and C. S. Bretherton, 2019: Spatially extended tests of a neural network parametrization trained by coarse-graining. *Journal of Advances in Modeling Earth Systems*, **11**, 2727–2744.

Han, Y., G. J. Zhang, X. Huang, and Y. Wang, 2020: A moist physics parameterization based on deep learning. *Journal of Advances in Modeling Earth Systems*, e2020MS002076.

Mooers, G., M. Pritchard, T. Beucler, J. Ott, G. Yacalis, P. Baldi, and P. Gentine, 2020: Assessing the potential of deep learning for emulating cloud superparameterization in climate models with real-geography boundary conditions. *arXiv preprint arXiv:2010.12996*.

Rasp, S., 2020: Coupled online learning as a way to tackle instabilities and biases in neural network parameterizations: general algorithms and lorenz 96 case study (v1. 0). *Geoscientific Model Development*, **13 (5)**, 2185–2196.

Rasp, S., M. S. Pritchard, and P. Gentine, 2018: Deep learning to represent subgrid processes in climate models. *Proceedings of the National Academy of Sciences of the United States of America*, **115**, 9684–9689.

Schneider, T., S. Lan, A. Stuart, and J. Teixeira, 2017: Earth system modeling 2.0: A blueprint for models that learn from observations and targeted high-resolution simulations. *arXiv preprint arXiv:1709.00037*.

Wilcox, E. M., and L. J. Donner, 2007: The frequency of extreme rain events in satellite rain-rate estimates and an atmospheric general circulation model. *Journal of Climate*, **20 (1)**, 53–69.

Yuval, J., C. N. Hill, and P. A. O'Gorman, 2020: Use of neural networks for stable, accurate and physically consistent parameterization of subgrid atmospheric processes with good performance at reduced precision. *arXiv preprint arXiv:2010.09947*.

Yuval, J., and P. A. O'Gorman, 2020: Stable machine-learning parameterization of subgrid processes for climate modeling at a range of resolutions. *Nature Communications*, **11 (1)**, 1–10.