Geoscientific Model Development Discussions

Geosci. Model Dev. Discuss., referee comment RC2
https://doi.org/10.5194/gmd-2021-299-RC2, 2021

Comment on gmd-2021-299

Anonymous Referee #2

Referee comment on "Stable climate simulations using a realistic general circulation model with neural network parameterizations for atmospheric moist physics and radiation processes" by Xin Wang et al., Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2021-299-RC2, 2021

In this manuscript the authors use neural networks to emulate the grid-box mean output of a superparametrization scheme, which predicts the sub-grid tendencies for moist physics and radiative heating. After a period of offline training the authors develop a new coupling approach for online testing. In online testing they find some evidence of improvements over the existing CAM5, i.e. a closer fit to the SPCAM approach.

There are several interesting ideas in this manuscript and some impressive technical developments in the coupling framework. However, I do not currently feel the manuscript is near acceptance for publication. My main issue is that the online testing analysis is not consistent, and does not persuade the reader that NNCAM is an improvement over the existing CAM5 model. Given that NNCAM is slower than the normal CAM5 parametrizations I think it is important to show that NNCAM provides an improvement. If this is not possible, then instead the authors could focus more effort on establishing whether any offline metrics provide a better indicator of online stability. Below I will detail my comments further. I hope the authors will take these on board, as I think this manuscript could make for an interesting and useful paper.

- The section on online performance analysis is a weak point. I think it is important to standardise the measurement periods used by the CAM5, SPCAM and NNCAM. Showing CAM5 and NNCAM as deviations away from the "truth" of SPCAM would ease the process of comparison. In many of the figures it is unclear that NNCAM is an improvement on CAM5, which begs the question of the purpose of the networks. I also think that showing plots of global metrics against time would help identify drift in the models. Given that this paper has a climate motivation, examining this behaviour seems crucial. I also think there is insufficient analysis of the effects of emulating radiative heating. It would be interesting to see some global maps of average 2m temperature to see the effects of the surface fluxes and near-surface heating rates.

- The authors highlight that online stability is not a given for coupling of parametrization schemes. This is a really interesting and important point for this field of study. However the authors' proposed solution is trial and error, suggesting that short term stability is a good predictor of long term stability. I would like to see more detailed analysis of whether there are good offline measures that can guide online stability. The authors suggest that improving R2 scores are not fully correlated with stability. Can one find a different metric that is better correlated with stability, analysing the results the authors have already conducted? I would be interested to see if mean-squared-error, mean bias or some measure of worst error were better predictors. If I understand the work correctly, you train four networks in your SPCAM. When you test stability are you swapping these four networks individually, or swapping all four together? This might shed light on which components were more important for stability. I think studying these points could provide great insight into the problem.

- If I understand the training/validation/testing split correctly, these are random subsets in space and time from the 1998/9 dataset. If so, I do not think this is a safe method for ensuring no overfitting, as this does not take into account spatial/temporal correlations. I think the total dataset should be split by time only, with temporal gaps between training, validation and testing to ensure independence. This might explain why NN with better R2 values provide less stable answers, if there is overfitting on the dataset.

There is very little discussion about the benefits and downsides to superparametrization. It is my understanding that there is very limited (if any) evidence that superparametrization actually improves model climate versus typical parametrization schemes. I think it is worth stating this, or if the authors disagree, provide citations.

Are the only parametrizations within the CAM model those in the superparametrization? e.g. there is no parametrization for sub-grid orographic gravity waves.

I suggest re-ordering manuscript to explain coupling before explaining results. The results section makes reference to coupled testing without explaining how this is achieved.

L135: Where does the variability originate in the CRMs? Are they initialised with different perturbations of the larger-scale conditions? If there is stochasticity in the system? It would be good to state this if true.

L166: "as output the NN-Parameterization". I think this should be "as outputs from the NN-Parameterization".

L167 "is critical to improve the performance of the NNCAM". I could find no further discussion of this. It sounds like a very interesting point. Please expand.

L190: Are you training to maximise R2? If not, what is your function to minimise/maximise?

L195: Have you tested this theory of mutual interference? I would have thought that training two different models to predict the TOA and surface fluxes would introduce physical inconsistencies. These are not separate pieces of physics.

L214: "a well-fit is necessary". This was unclear and could be better written.

L229: Is the "best performance" network based upon the best performance in offline or online testing?

L242: The online coupler sounds like an interesting solution of value to the wider scientific community. Are the authors planning to share this as a stand-alone piece of software?

L278: I do not understand "reaches half the speed of CAM5". Are the authors comparing to the speed of CAM5 with the normal parametrization schemes? By half the speed to they mean it will take twice the time to simulate the same period?

L279: Have the authors profiled how much time is spent communicating data versus doing ML inference? This would be very interesting to see.

L280: If I have understood correctly the authors carry out the online testing on the same time period that the NN was trained on. Has any effort been made to ensure independence between the training and testing data?

L305: "tunned" -> "tuned"

L320: The authors run for 10 years but only analyse 4 years of data. So their only expectation of the final 5 years is for the model to not crash. I do not think this is an appropriately strict assessment for their NN models. I think examining model drift is exactly the important test of a NN. If not, what is the purpose of the model that the authors are building?

L325: It seems a very strange choice to not use the same periods for each of the models being tested. I understand that there are computational costs to be accounted for, why not assess each model for the 1998-2001 period?

L600 Table 2. "Number of samples trained per iteration". Are the authors referring to batch size here? "Number of rounds to traverse the data set". Sorry, this is unclear to me. Is this stating that the training dataset contains 50 batches of 1024?

L620: "Note: Spatial averaging of MSE is performed before calculating *R2*." This is unclear. Could the authors please explain further.

Figure 7: It would be very interesting to also plot the R2 values for the CAM5 parametrization as a model for the SP.

Figure 8: There appear to be negative R2 values in portions of the globe. This is a worryingly low skill for the model.

Figure 9: I think this figure could strongly benefit from a companion figure where the differences from the SPCAM run are shown for both CAM5 and NNCAM. Otherwise is it challenging to decipher if NNCAM lies closer to the SPCAM mean state than CAM5.

I also think it would be very interesting to compare all of these runs to the ERA5 state of the atmosphere for those years. This would go towards answering the question of whether SP is an improvement over CAM5.

Figure 10: As with figure 9, I think showing the differences would add significant information.

Figure 11: My interpretation of this plot is that NNCAM is a worse model of SPCAM than CAM5. Do the authors agree, and if so, why do they think this is true?

Figure 14: As with figure 11. It is not clear that NNCAM has succeeded at this task.