Response to reviewer 1

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.

Thank you for carefully reviewing our manuscript. We have tried our best to respond to all of the comments below and have made proper revisions to the manuscript. The reviewers' comments are in italics and our responses are in normal blue font. For clarification, we have made changes to the figures, and there are 15 figures in the revised manuscript, and 5 figures and 2 movies in the supplementary materials. We have revised Figures 4 and 5, added Figure S2 (a plot of mean state differences) in the last manuscript as Figure 10 now. Figure 11 shows the global distribution of the mean precipitation, and the zonally averaged precipitation is shown in Figure 12. Figures 12–14 in the previous manuscript are now Figures 13–15. We also added a plot of the cross-NN precipitation STD (Figure S1), and the original plot of the precipitation differences is now Figure S5. For the convection and combined radiation heating rate, we no longer use the phrase temperature tendency. Now, we use the tendency of the dry static energy.

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.

Reply: Thank you for your comments. For the parts where NNCAM fails, we have clearly highlighted the climate biases in the abstract as follows: *"However, there are still substantial biases in the mean states, including the temperature field in the tropopause and at high latitudes and the precipitation over tropical oceanic regions, which are larger than those in CAM5."* We have moved the plots of the temperature differences from Figure S2 to Figure 10, have added clear statements about temperature biases into Section 5.1.1 and precipitation biases into Section 5.1.2. The original Figure 11 is now Figure 12, and the TRMM precipitation has been added to supplement SPCAM and CAM5. The plot of the precipitation difference is now Figure S5. These two figures show that NNCAM produces larger precipitation simulation biases over the equatorial tropical ocean in boreal winter. However, both SPCAM and NNCAM produce better rainfall simulations than CAM5 over the land, which is related to the SPCAM's better-simulated convection variability.

As for the target model SPCAM, which uses the 2-D SAM as the SP, SPCAM does not significantly improve the climate mean states (Khairoutdinov et al., 2005; Kooperman et al., 2016). We have made such statements in lines 62-63 in the introduction and in lines 406-407 in Section 5. However, SPCAM improves the model's climate variability, as well as the tropical precipitation over the land, which is shown by Figures 12–15 and is consistent with the findings of other studies (see the SPCAM part in the introduction). The NNCAM also retains these advantages and succeeds in these aspects.

We have also added lines 407–410: *"What is remarkable about NNCAM is not its performance in simulating the mean climate, but its ability to achieve a stable multi-year prognostic simulation under a real-world global land-ocean distribution. The advantages and problems of this study will provide important references for future research on NN-based stable long-term model integrations."*

- *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.*

Reply: Thank you for the comments. 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 MSE_h and the stability, not to prove that the ResDNN is better than the DNN in terms of stability.

The stability of the NN-Parameterization still requires further study. We have also made changes to the abstract by only stating that "*We explore the relationship between the accuracy and stability by validating multiple Deep Neural Network (DNN) and ResDNN sets in prognostic runs.*" We have added a statement in lines 313–314: *" We speculate that the more accurate ResDNN sets have a higher probability of becoming stable NN-Parameterizations since all of the stable NN-Parameterizations are ResDNNs."*

- *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 R2 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.*

Reply: The reviewer may have misunderstood what we did with the radiation. This is probably due to the confusing text in the previous version of the manuscript. We used three neural networks for (1) the moistening rate, (2) the combined heating rate of the convection and radiation, and (3) the radiation fluxes at the surface and the TOA (for more details please see lines 192–198). Each of them has the same input, the same network structure and the same hyperparameters, but different outputs. Therefore, the radiative heating is part of the NN-Parameterization; and it is only not treated separately from the heating due to the moist physics. The radiation fluxes at the surface and the TOA are also predicted using the NN and are updated to the model. In addition, regarding our R^2 calculations, we directly applied the equation $R^2 = 1 - \frac{mse}{var}$ for all

of the considered fields without any preprocessing, except in Figure 7. Specifically, Figure 2 uses a *"raw"* R^2 for both time and space. Figure 7 uses the R^2 for the zonally averaged fields, and Figure 8 uses the R^2 for the time sequence at each location. The proper information has been added to the captions of the corresponding figures.

- *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.*

Reply: Thank you for pointing this out. As for the concern of cherry picking and the large differences in the results of the stable NNs, we have highlighted this in the abstract as follows: *"We explore the relationship between the accuracy and stability by validating multiple Deep Neural Network (DNN) and ResDNN sets in prognostic runs. In addition, there are significant differences in the prognostic results of the stable ResDNN sets. Therefore, trial-and-error is used to acquire the optimal ResDNN set for both high-skill and long-term stability, which we name the NN-Parameterization."* We have also added similar statements in lines 324–327: *"Apart from global averages, the prognostic results of the 10 stable ResDNN sets vary from each other in terms of the global distribution. Figure S1 shows the precipitation spread across all of the stable NN sets for the prognostic simulation from 1999 to 2003. The obvious standard deviation centers coincide with the heavy tropical precipitation areas."*

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.*

Reply: Thank you for pointing this out. We agree with the reviewer that one can separate the different processes such as advection, diffusion, microphysics, and ice sedimentation in the SP. However, in SPCAM under the multiscale model framework, the SP is a 2-D SAM and the host model is CAM5, and the accumulated tendencies across all of the processes are used to exchange information between the SAM and the host model. Therefore, to emulate the SAM in the simplest way, we directly used the accumulated tendencies without tracking the different processes. However, this is an interesting idea. We plan to study it in our future work.

• *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.

Reply: Thank you for pointing this out. We think it is a misunderstanding because of an unclear sentence. We agree with the reviewer on this issue. What we intended to say is that our NN-Parameterization does not have any imposed physical constrains.

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

Reply: Thank you for the comment. We have added the descriptions in lines 165–170. *"In this study, we used the vertical integration of the NN predicted moisture tendency as an approximation of the surface precipitation, which has also been used in previous studies (e.g., O'Gorman et al., 2018; and Han et al., 2020). In the offline validation test, we observed negative precipitation events (27% occurrence in 1-year of results). Nonetheless, 93% of the negative precipitation events had a magnitude of less than 1 mm/day. In the online prognostic runs, reasonable rainfall results (more details will be provided in Section 5) were achieved using this approximation scheme."*

• *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.")*

Reply: Thank you for the comment. We had already removed the claim that the *"MLP can generalize better"* in the last revision of the manuscript. We also briefly introduced why the DNN and the ResDNN were used in this study in lines 203–210.

• *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.*

Reply: Thank you for the comment about the R^2 calculations. Except for in Figure 7, we did not apply zonal averaging in any of the R^2 calculation (Figures 2 and 8). For Figure 7, we calculated R^2 of zonally averaged values because we compared this figure to those in Mooers et al. (2021), and thus, the same method needed to be used. We believe this is sufficient to show the accuracy of the moistening and heating. Since the reviewer expected to calculate R^2 before zonal averaging, we calculated the R^2 for every single position in the 3-D space over its own time series and then zonally averaged them into a latitude-pressure cross section in Figure R1.

• *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 R^2 as 1-mse/var, many of those low std regions will have R^2 close to zero." However, the authors have low skill in the tropics where STD is large.

Reply: Thank you for pointing this out. We agree with the reviewer that the NN-Parameterization also produces low-skill regions where the STD is large. The explanation in the last response is only for some locations near the subtropical southeast ocean areas and therefore does not apply to the regions near the equator. Therefore, we have removed this argument from the manuscript. Instead, we have added the following text in lines 390–393: *"Still, our NN-Parameterization produced low accuracy predictions along the equator over the oceans where the convection is complex and vigorous and in subtropical ocean areas where the convection is weak and concentrated at low levels. This indicates that the NN-Parameterization is still inadequate in rems of its emulation skill when simulating various types of deep and shallow convection in the tropics."*

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

Reply: Thank you for the comment. We are sorry for the typo in this citation. We actually cited Moores et al. (2021), and this has been corrected in the latest version of the manuscript.

• *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).*

Reply: We thank the reviewer for their insightful comment. The second reviewer asked for this comparison. We have added the sentence *"It should be noted that the NN-Parameterization was tuned to emulate the SP, and CAM's Parameterization was tuned to obtain close results compared to the observations."* in lines 354–355 before comparing NN-Parameterization with the conventional CAM5 parameterizations through offline tests.

• *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."*

Reply: Thank you for this comment. We have checked and corrected these sentences in the latest version of the manuscript by changing it to *"To determine why some methods can achieve stable prognostic simulations and others cannot, Brenowitz et al.*

(2020) proposed methods for interpreting and stabilizing ML parameterization for convection. In their study, a wave spectra analysis tool was introduced to explain why the ML coupled GCMs blew up." (see lines 93–95)

• *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).*

Reply: Thank you for this comment. All of the variables with vertical levels are normalized as a whole using the maximum value following the method of Han et al. (2020), rather than being normalized separately at each level. We have added this statement in line 175.

• *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.*

Reply: Thank you for the comment. We have simplified this paragraph, and now briefly and directly explain why a DNN and ResDNN were used in this study (see lines 203–210).

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

Reply: Thank you for the comment. We mainly tried to use the weight decay, learning rate, width and depth of the neural network, batch size, dropout, activation function, and other hyperparameter combinations. We have sorted this out and have added it to Table S1 in the SI.

• *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.*

Reply: Thank you for the comment. In fact, it should be the tendency of the dry static energy not the tendency of the temperature. We have changed all of the temperature tendency to the tendency of the dry static energy to represent the combined convective and radiative heating in the latest version of the manuscript. According to the appendix in Khairoutdinov and Randall (2003), the SAM predicts the changes in the liquid ice/water energy [$h_L = C_p T + gz - L_c(q_c + q_r) - L_s(q_i + q_s + q_g)$] and the total nonprecipitating water q_T (water vapor q_v + cloud water q_c + cloud ice q_i). The host GCM (CAM5) uses the tendencies of the dry static energy ds and water vapor dq_v, and it updates them to the states of the temperature, water vapor, and layer heights via the *physics_update* subroutine. There is a process in the 2-D

SAM codes that derives ds from the changes in h_L and dq_v and from the changes in q_T.

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

Reply: Thank you for the comment. We have removed the reference in Figure 1. Please see lines 230-231.

• *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.*

Reply: Thank you for the comment. Although Yuval et al. (2021) stated that *"random forest is less likely to perform as accurately as neural networks and cannot be implemented in GPUs"*, we agree that the random forest has been implemented and used on 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.*

Reply: Thank you for the insightful comment. Based on the answer *"NN predicts the temperature tendency,"* we have changed *dT* to *ds*. Therefore, we have corrected the moist static energy as follows: $h = C_pT + gz + L_vq_v = s + L_vq_v$.

• *Line 313: dump->damp*

Reply: Thank you for pointing this out. We have corrected this in the latest version of the manuscript.

• *Line 379: The authors refer to figure S2 (line 380) and say it shows RMSE - but I donot think it shows this. Please add RMSE to the figure.*

Reply: Thank you for pointing this out. Actually, the parentheses in Figure S2 are merely for the phrase *"larger differences than CAM5"*. This is indeed an unclear sentence. So we have moved Figure S2 showing the state differences to Figure 10, and we revised the sentence as follows: *"However, the multi-year mean temperature and moisture fields produced by NNCAM are more biased than those produced by CAM5, which is reflected by the larger root mean square errors (RMSEs) (Figure 9) and larger differences compared to those of CAM5 (Figure 10)."*

Reference:

Brenowitz, N. D., Beucler, T., Pritchard, M., and Bretherton, C. S.: Interpreting and Stabilizing Machine-Learning Parametrizations of Convection, Journal of the Atmospheric Sciences, 77, 4357-4375, 10.1175/jas-d-20-0082.1, 2020.

Han, Y., Zhang, G. J., Huang, X., and Wang, Y.: A Moist Physics Parameterization Based on Deep Learning, Journal of Advances in Modeling Earth Systems, 12, e2020MS002076, 10.1029/2020ms002076, 2020.

Khairoutdinov, M. F. and Randall, D. A.: Cloud Resolving Modeling of the ARM Summer 1997 IOP: Model Formulation, Results, Uncertainties, and Sensitivities, Journal of the Atmospheric Sciences, 60, 607-625, 2003.

Khairoutdinov, M., Randall, D., and DeMott, C.: Simulations of the Atmospheric General Circulation Using a Cloud-Resolving Model as a Superparameterization of Physical Processes, Journal of the Atmospheric Sciences, 62, 2136-2154, 2005.

Kooperman, G. J., Pritchard, M. S., Burt, M. A., Branson, M. D., and Randall, D. A.: Robust effects of cloud superparameterization on simulated daily rainfall intensity statistics across multiple versions of the Community Earth System Model, Journal of Advances in Modeling Earth Systems, 8, 140-165, 10.1002/2015ms000574, 2016.

Mooers, G., Pritchard, M., Beucler, T., Ott, J., Yacalis, G., Baldi, P., and Gentine, P.: Assessing the Potential of Deep Learning for Emulating Cloud Superparameterization in Climate Models With Real-Geography Boundary Conditions, Journal of Advances in Modeling Earth Systems, 13, e2020MS002385, https://doi.org/10.1029/2020MS002385, 2021.

Rasp, S., Pritchard, M. S., and Gentine, P.: Deep learning to represent subgrid processes in climate models, Proceedings of the National Academy of Sciences, 115, 9684-9689, 10.1073/pnas.1810286115, 2018.

Yuval, J., O'Gorman, P. A., and Hill, C. N.: Use of Neural Networks for Stable, Accurate and Physically Consistent Parameterization of Subgrid Atmospheric Processes With Good Performance at Reduced Precision, Geophysical Research Letters, 48, e2020GL091363, https://doi.org/10.1029/2020GL091363, 2021.

Response to reviewer 2

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.

Reply: Thank you for carefully reviewing our manuscript. We have tried our best to respond to all of the comments below and have made proper revisions to the manuscript. The reviewers' comments are in italics and our responses are in normal blue font. For clarification, we have made changes to the figures, and there are 15 figures in the revised manuscript, and 5 figures and 2 movies in the supplementary materials. We have revised Figures 4 and 5, added Figure S2 (a plot of mean state differences) in the last manuscript as Figure 10 now. Figure 11 shows the global distribution of the mean precipitation, and the zonally averaged precipitation is shown in Figure 12. Figures 12–14 in the previous manuscript are now Figures 13–15. We also added a plot of the cross-NN precipitation STD (Figure S1), and the original plot of the precipitation differences is now Figure S5. For the convection and combined radiation heating rate, we no longer use the phrase temperature tendency. Now, we use the tendency of the dry static energy.

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).

Reply: Thanks for the comment. We use the same ResDNN for predicting 8 radiation fluxes in all sensitivity tests. The sensitivity test results in Figure 4 are for the different neural networks predicting the moisture tendency dq_v and the dry static energy tendency ds. We have highlighted it in lines 298-299: *"First, we selected the best ResDNN for the radiation fluxes at the surface and the TOA that was shared in every NN set since their offline validation was exceptionally accurate with $R^2 > 0.98$ over 50 training epochs (Figure 2b)."*

"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.

Reply: Thanks for the comment. In this research, the ResDNN/DNN predicting dq_v and ds have the same hyperparameter configuration. In theory, it is possible to combine the neural networks that predict dq_v and ds under different training configurations. It is possible to influence the NN-Parameterization stability by combining the moisture NN with the dry-static energy NN for different epochs. However, the permutations and combinations will bring too many test cases (50x50). We do not have enough computing resources to perform stability tests and climate state evaluations for all these NN combinations. Even with the current method, stability can be guaranteed. As a result, we are not going to address this in this work.

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

Reply: Thank you for the question. We have made such statements in the abstract: *"However, there are still substantial biases with the hybrid ML-physical GCM in the mean states, including the temperature field in the tropopause and at high latitudes and the precipitation over tropical oceanic regions, which are larger than those in CAM5."* and also in Section 5.1.

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.

Reply: Thanks for the comment. We had already revised *"multi-target"* to *"A ResDNN set"* in the last revised manuscript. We have also written a more precise paragraph in this revision in lines 184-202. Especially, we have changed the two original references to Yu et al. (2020) for the mutual interference in gradient descending in the revised manuscript. We believe that the separated training can avoid the mutual interference in

the gradient descending. Moreover, we agree with the reviewer that more future research should be done here.

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?

Reply: Thanks for your comment. We did not change our approach in the revised manuscript. Figure 2 in the original manuscript is outdated and misleading. So we removed it and described our method verbally in Section 2.2.2. The ResDNN set contains three neural networks, which were trained to predict dq_v, ds and eight radiation fluxes at the surface and the TOA, respectively. They all have the same hyperparameter configuration.

L456: successes -> succeeds?

Reply: Thanks for pointing this out. We have corrected it in the latest revised manuscript. For details, please see line 502.

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

Reply: Thanks for pointing this out. There are some compatibility issues when compiling the pdf file. Several icons are missing in Figure 3. We have corrected them in the revised manuscript.

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

Reply: Thanks for pointing this out. We have included CAM in the revised Figure 5.

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

Reply: Thanks for pointing this out. We have revised the caption of Figure 9 and presented it in the revised manuscript.

Figure 10, as with 9.

Reply: Thanks for pointing this out. We have revised the caption. Please see the new Figure 11 in the revised manuscript.

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.

Reply: Thanks for pointing this out. It seems that we ran into some bugs in NCL with the automatic contour levels. Here we use the explicit levels with an interval of 1mm/day to plot the precipitation differences and present them in Figure S5 in the SI. According to Figure S5, NNCAM produces larger precipitation biases than CAM5 in the boreal winter and does not improve that much in the boreal summer even though NNCAM's global mean difference and root mean square error are smaller. We explicitly make such statements in lines 438-441: *"However, on a difference plot (Figure S5), NNCAM moderately underestimates the precipitation along the equator, in the Indian monsoon region, and over the Maritime Continent in the summer (Figure S5a). In the boreal winter, NNCAM simulates a weak SPCZ that is excessively separated from the ITCZ, with both precipitation centers shifted away from each other."*

Reference:

Yu, Tianhe, et al. "Gradient surgery for multi-task learning." Advances in Neural Information Processing Systems, 33 (2020): 5824-5836.

Figure R1

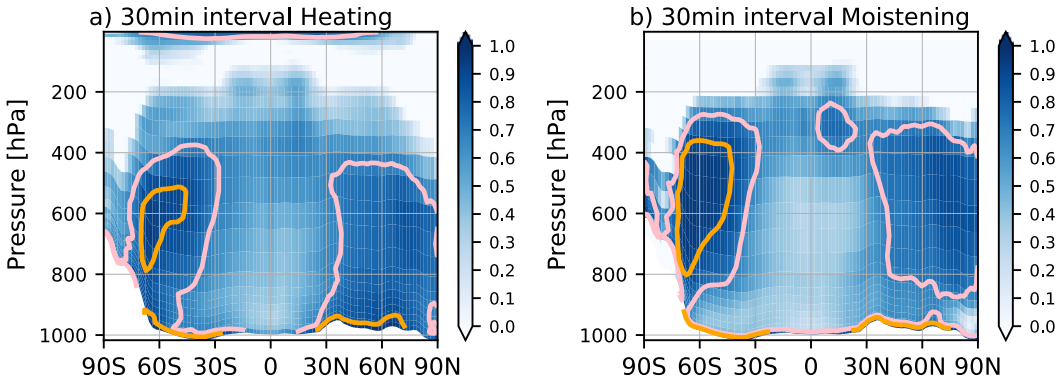

Figure R1. Latitude-pressure cross sections of the zonally averaged coefficient of determination (R^2) for (a) heating and (b) moistening predicted using the NN-Parameterization. Both are evaluated at a 30 min time step interval. Note: the areas where R^2 is greater than 0.7 are contoured in pink, and the areas where R^2 is greater than 0.9 are contoured in orange.

Table S1

Table S1. The hyperparameter search space. Note that the "Number of Hidden Layers" are used in the fully connected DNNs, "Residual blocks" are used in the ResDNNs, and each Residual block has two fully connected layers.

Hyperparameter Type	Hyperparameter Space
Number of Hidden Layers/Residual blocks	3, 4, 5, 6, 7, 8, 9
Width of Hidden Layers	64, 128, 256, 512, 1024
Learning rate	0.001, 0.002, 0.0001, 0.005
Dropout rate	0, 0.2, 0.3, 0.5
Batch size	128, 256, 512, 1024
Weight decay	0, 1e-5, 5e-4
Activation function	ReLU, Leaky-ReLU, Sigmoid
Loss function	Mean Squared Error
optimizers	Adam, SGD

Figure S1–S5

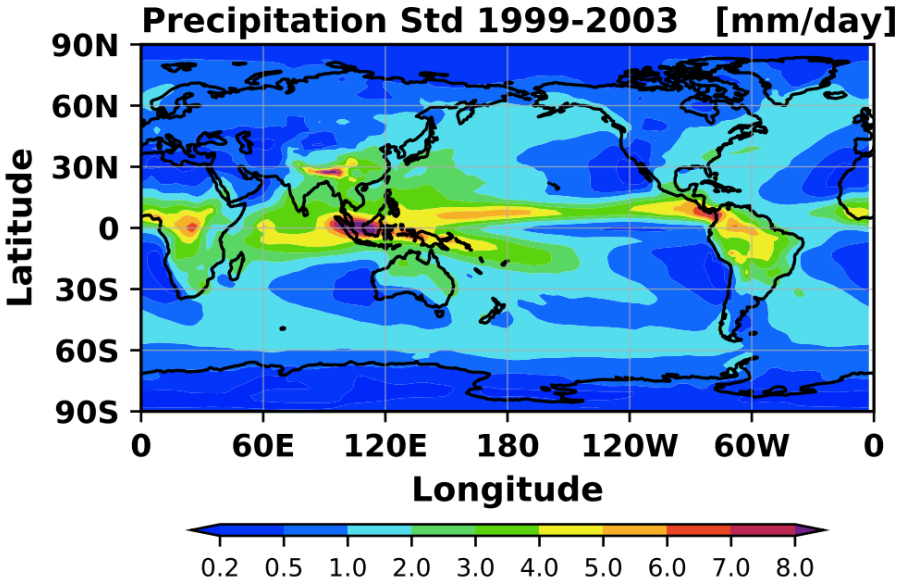

Figure S1. Spatial distribution of the precipitation STD across all 10 stable NN-Parameterizations for the prognostic simulation from 1999 to 2003.

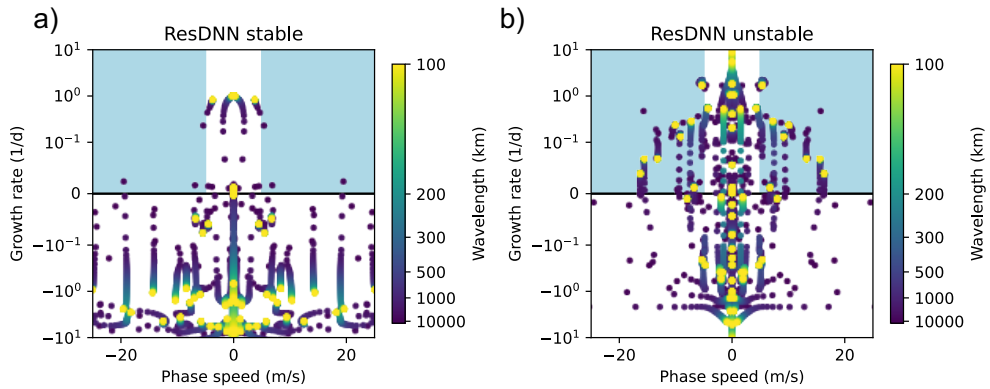

Figure S2. Wave spectra of (a) a stable NN-Parameterization and (b) an unstable parameterization. The light blue background indicates where the phase speed is greater than 5 m/s and the growth rate is positive. The stability diagrams were obtained by coupling the linear responses of the NN-parametrizations with the simplified 2-D dynamics with a chosen base state, which is the normal convection background in the long-term prognostication in (a) and the initial state for the unreal gravity wave in Move S1 in (b).

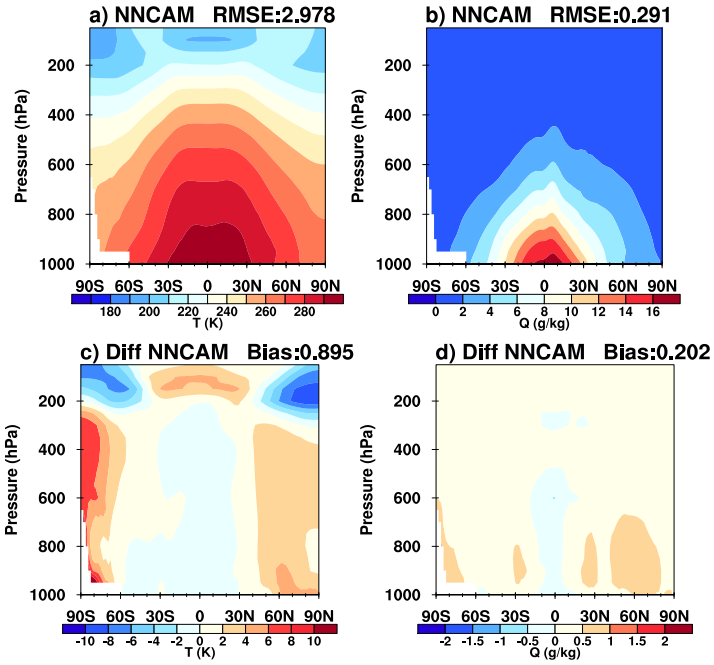

Figure S3. Latitude-pressure cross-section of the zonal and annual mean (a) temperature and (b) specific humidity for NNCAM simulated from 2004 to 2008, with their differences from the SPCAM simulation from 1999 to 2003.

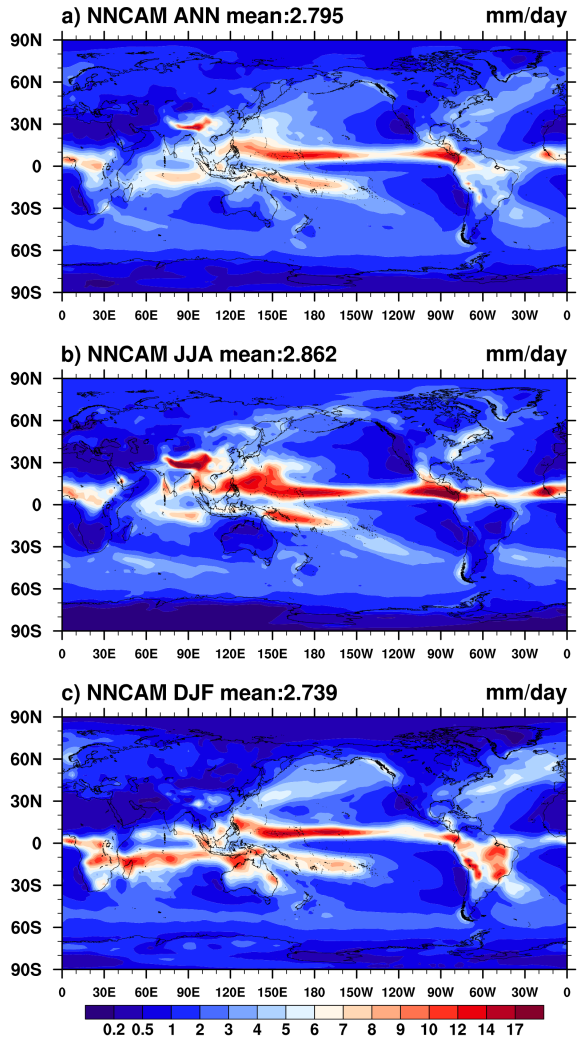

Figure S4. Global distribution of the temporal mean precipitation predicted using NNCAM from January 1, 2004, to December 31, 2008, for the (a) annual, (b) boreal summer (JJA), and (c) boreal winter (DJF).

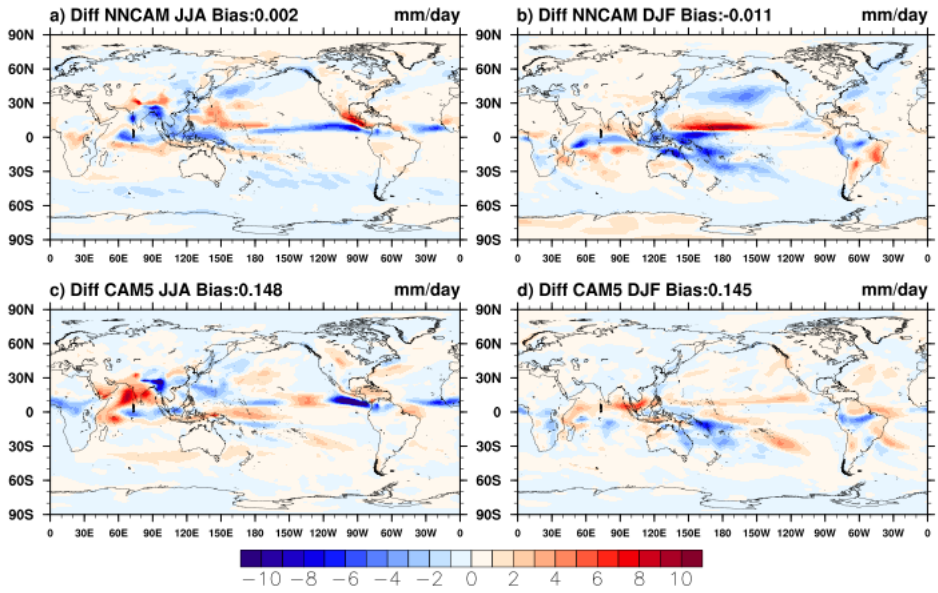

Figure S5. Global distributions of the precipitation differences between (a & b) NNCAM and SPCAM and (c & d) CAM5 and SPCAM averaged over the boreal summer (left panels) and winter (right panels).