Response to reviewer 2

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.

Reply: Thanks for the careful review. We have tried our best to answer all comments below and made proper revisions to the manuscript. The reviewer's comments are in italic and our response is in normal and blue font. Please note that the Figures 1-14 are in the revised manuscript, the Figures S1-S4 are in the supplementary, the Figures R1-R5 in the last part of this response, and the Movies S1 and S2 are in the attached files. All figures mentioned in this response are listed in the last part.

- *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.*

Reply: Thanks for the comment. First, we have standardized the measurement period used by SPCAM, NNCAM and CAM5, which is from 1999-01-01 to 2003-12-31 after a year of spin up. We admit that our NNCAM still contains biases in the simulated climate mean states in climate analysis but better in climate variabilities such as extreme rainfall events and MJO. The time evolution of total energy (Figure 5) confirms no significant climate drift in NNCAM (the blue line).

Specifically for the mean climate states, NNCAM has a warmer tropopause, a colder air temperature at the polar upper levels and a warmer mid and low troposphere above polar regions than SPCAM. It is also wetter for the entire troposphere (Figure S2). On the contrary, CAM5's climatology is less deviated. NNCAM has a much warmer 2m temperature in high latitudes during boreal winter, while CAM5 only shows a slightly colder temperature in those regions (Figure R3). For global precipitation distribution, NNCAM deviates less than CAM5 in global averages in boreal summer (Figure 11a) with similar patterns and smaller RMSE (Figure 10c and 10a) but performs worse in winter accompanied by a significant underestimation along the equator (Figure 11b). Therefore, although NNCAM is still considered significant progress from the other unstable or biased NN-parameterization coupled simulations, it is admitted an experimental model. We correct the statement and don't claim it is comprehensively better than CAM5 in the revised manuscript.

Even with inevitable biases in the simulated climatology, the NN-Parameterization can still inherit the variability predicted by the resolved convection process in SP. NNCAM is significantly closer to SPCAM in simulating extreme precipitation and MJO than CAM5, with much higher heavy precipitation probability (Figure 12) and closer MJO spectrum (Figure 13) and propagating speed patterns (Figure 14). Specifically, we use the coefficient of determination (R^2) to measure the distance of spectrum and lagged coefficient between SPCAM and NNCAM or CAM5. Higher R^2 means more similarity. With an R^2 of 0.511 rather than that of 0.397 for CAM5, NNCAM indeed performs better in boreal winter MJO precipitation spectra. For the lagged coefficient with 5-year data, NNCAM performs even better in longitude evolution especially for the 200 hPa zonal wind U.

- *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.*

Reply: Thanks for the comment. We have added a series of sensitivity test in Section 3 to determine stable NN parameterizations. Although short term stability still matters the long term stabililty, we no longer use them as the only metric to select NNs. In the senstivitty test, we have adjusted the machine learning metrics for evaluating NNs, to

the mean squared error of moist static energy changing rate *(dh)* to evaluate heating and moistening rate together. This new metrics shows insightfull relationships with the max prognostic steps in Figure 4: when the offline MSE of ResDNN decreases to a certain level (e.g., $290 W^2/m^4$), the ResDNN parameterization may run stably for long periods. All less accurate NN parameterizations are unstable. After getting enough offline accuracy, we use the 2D gravity wave analysis framework (Brenowitz et al., 2020) to test stability of the NN parameterizations. The unstable ones will simulate a breaking point and with an unstable wave amplified mode (Figure S1b), while the stable ones can attenuate the perturbation energy all the time and show a stable mode (Figure S1a) in the wave spectra. This framework explains why unstable NN parameterizations crash the model but is not an a priori way to design NN models. We still have to use the trial-and-error to filter out unstable ones and then select the best NN parameterization that can reduplicate the total energy time evolution of SPCAM with the least deviation.

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 R^2 above 0.98. Different from the easily and accurately trained radiation fluxes, 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. However, we find the NN for moistening rate *dqv* is most difficult to train and possibly more important for stability.

- *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*

Reply: Thanks for the comment. We have changed the way to divide the training and offline validation dataset to ensure independence. The training dataset used by all considered NNs is 40% temporally random sampled from the 2-year SPCAM simulation from 1997-01-01 to 1998-12-31. Notably, the random sampling is only done in the time dimension but not in latitude and longitude, including all samples globally of the selecting timestep. To avoid any mix or temporal connection between the training set and offline validation set, we random sample 40% timesteps from the SPCAM simulation in the year 2000 as the offline validation set.

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.

Reply: Great question. We also think researchers should be frank with the pros and cons of SPCAM before using it as a target model. Khairoutdinov et al. (2005) shows that SPCAM produced quite "reasonable" geographical distribution of precipitation, precipitable water and cloud fraction, but has a notable precipitation bias in the Western Pacific. On the other hand, the SP substantially improves convection variability in multiple ways, including diurnal variation, probability distribution of precipitation intensity, and intraseasonal variability such as MJO.

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

Reply: Thank you for the question. The orographic gravity waves drag and vertical diffusion are computed after calling the land surface model and before the next round calculation of the dynamic core.

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.

Reply: Thank you for the comment. We actually did explain coupling before explaining results in the original manuscript. Anyway, in the revised manscript, we have added the explanation of the coupling strategy in Section 2.2.3 and 2.3, and the description of the results is in Section 4 and 5.

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.

Reply: Thank you for the question. Yes, they are kind of initializtion with different perturbations of the large-scale conditions. Specifically, at the beginning of each simulation, the SP/CRM fields in each CAM grid column are initialized by the soundings with small amplitude noise added to SP temperature fields near the surface. No noise is added at later times (Khairoutdinov et al., 2005).

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

Reply: Thanks. We applied this change in the revised manuscript as *"Also, it is important to include direct and diffuse downwelling solar radiation fluxes as output variables to force the coupled land surface model."*.

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.

Reply: Thanks for the comment. The 4 solar radiation fluxes down to surface represents the received solar energy by the land surface model. If they are not included, the land surface will not be heated up by the sun, weaking the land-sea breeze and monsoon circulations. We have expanded in the second paragraph of Section 2.2.1 in the revised manuscript.

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

Reply: Yes, we want to minimize the mean squared error for each variable not the R^2. However, R^2 is later used to measure the degree of fit between the NN predictions and the reference values generated by SPCAM.

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.

Reply: The work of Crawshaw et al. (2020) and Zhang & Yang. (2021) proved the necessity of separating forecast targets. At the beginning of the experiment, we did test using a DNN to try to predict all variables, and the DNN could hardly converge. After we separated the predicted targets *dqv, dT* and *radiation variables* according to Crawshaw et al. (2020) and Zhang & Yang . (2021), the network began to converge and obtained satisfactory results. We use one NN to train all radiation fluxes in the revised manuscript.

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

Reply: Thanks for the comment. We have added relative analysis in the new section 3.1. *"a well-fit is necessary"* is replaced with a throught analysis in sensitivity tests. First, we train NNs over a threshold of accuracy, which makes stable for long-term prognostic simulations possible. Then, we have to use trial-and-error to filter out unstable NNs and select the best for the most accuate long-term simulations.

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

Reply: Thanks for the question. *"best performance"* means best ResDNN set in 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?

Reply: Thanks for the comment. You can access it through our open resource lib (https://doi.org/10.5281/zenodo.5596273). We plan to continuously improve and optimize the online coupler in the future.

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?

Reply: Thanks for the question. *"reaches half the speed of CAM5"* means that the total simulation time of NNCAM is double of that of CAM5. When using 192 CPU Cores of our commondity cluseter computer, the SYPD of Coupler-based NNCAM (with the support of 1 GPU) is 10.0, and the SYPD of Fortran-based NNCAM using Intel Math Kernel Library but witout GPU is only 1.5.

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

Reply: Thanks for the question. To answer your question, we conducted the run of NNCAM for time breakdown. Indeed, the communication through coupler and computation of neurl networks takes almost equal time, and there is still a lot of room for performance optimization. For your concerns, we implemented a ResDNN in Fortran and tested the performance, please see the lines 295-298 in the revised manuscript.

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?

Reply: Thanks for the comment. We have changed the way to subsets the training and offline validation dataset to ensure independence. The training dataset used by all considered NNs is 80% temporally random sampled from the 2-year SPCAM simulation from 1997-01-01 to 1998-12-31. Notably, the random sampling is only done in the time dimension but not in latitude and longitude, which means once a timestep is selected, all global samples belonging to that step go “on board”. To avoid any mix or temporal connection between the training set and offline validation set, we random sample 40% timesteps from the SPCAM simulation in the year 2000 as the offline validation set.

L305: "tunned" -> "tuned"

Reply: Thanks. We have applied this change.

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?

Reply: Thanks for the question. After running NNCAM as start up, we reorganized all the results. NNCAM does simulate more biased climate states than CAM5 but has no obvious climate drift. We have shown the global distribution of temporal averaged precipitation for the last years in Figure S4. The averages are not drifted and the patterns are similar to those for the first 5 years.

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?

Reply: Thanks for pointing out this. To avoid any confusion, we choose the 5-year period from 1999-01-01 to 2003-12-31 for prognostic simulations of SPCAM, NNCAM and CAM5.

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?

Reply: Thanks for the comment. *"Number of rounds to traverse the data set"* means epochs, The description of our training process was not clear enough, and we apologize for that. We have reorganized the training process of NNs, please refer to the revised manuscript in lines 205-216.

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

Reply: Thank you for the comment. We just want to emphasize that the mean square errors from samples that globally are and weighted equally to calculate the total mean square error, and that the variance is also calculated across all samples. The we derive R^2 via $R^2 = 1 - mse/var$.

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

Reply: Thank you for the suggestion. We have added the R^2 for CAM5 parameterization as a baseline in the new Figure 7. In offline validation, it is clear that NN parameterization is closer to the SP much better than the traditional CAM5 moist parameterizations.

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

Reply: Thanks. 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. (please check Figure R1)

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.

Reply: Thank you for the suggestion. We have added the differences between SPCAM and NNCAM or CAM5 in the new ***Figure S2***. NNCAM actually simulate more biased climate states than CAM5 compared with SPCAM. We also compare the mean states of temperature and humidity from SPCAM with ERA-Interim. It turns out that SPCAM simulate a colder and wetter climate. Its improvements over CAM5 is limited in terms of climate states (Figure R4).

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

Reply: Thank you for the suggestion. We have added the differences in the new Figure 11 of the revised manuscript.

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?

Reply: Thank you for the question. We have clarified the pros and cons of NNCAM in your first major comment. NNCAM indeed carries some biases in mean states and we admitted that in the revised manuscript. The winter precipitation biases of NNCAM is most significant, we believe the spatial difference plots add more information than the zonal averaged plots. Therefore, we have replaced the old Figure 11 with the differences plot. For now, the biggest error is the underestimated rainfall along the equator in boreal winter. From the 2m temperature in Figure R4, the high latitude regions in both southern and northern (especially the northern) hemisphere are too warm in NNCAM. Therefore, anomalous northwest surface wind stress is found on the north of the equator in the western Pacific, making the ITCZ shift north in DJF (Figure R5). Also an easterly intrusion is found in the location between the ITCZ and SPCZ.

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

Reply: Thank you for the comment. In the revised new Figure 11, we plotted it with precipitation and U200 from the new 5-year simulations for SPCAM, NNCAM, and CAM5. SPCAM and NNCAM show eastern propagation signals over Indian Ocean and Maritime Continent while CAM5 shows the opposite. The R^2 for precipitation and U200 are below zero in NNCAM, but they are higher than those in CAM5 where the R^2 for U200 is as low as -0.74.

Reference:

Crawshaw, M.: Multi-task learning with deep neural networks: A survey, arXiv preprint arXiv:2009.09796, 2020.

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, 10.1175/jas3453.1, 2005.

Zhang, Y. and Yang, Q.: A Survey on Multi-Task Learning, IEEE Transactions on Knowledge and Data Engineering, 1-1, 10.1109/TKDE.2021.3070203, 2021.

Figure R1 - R5

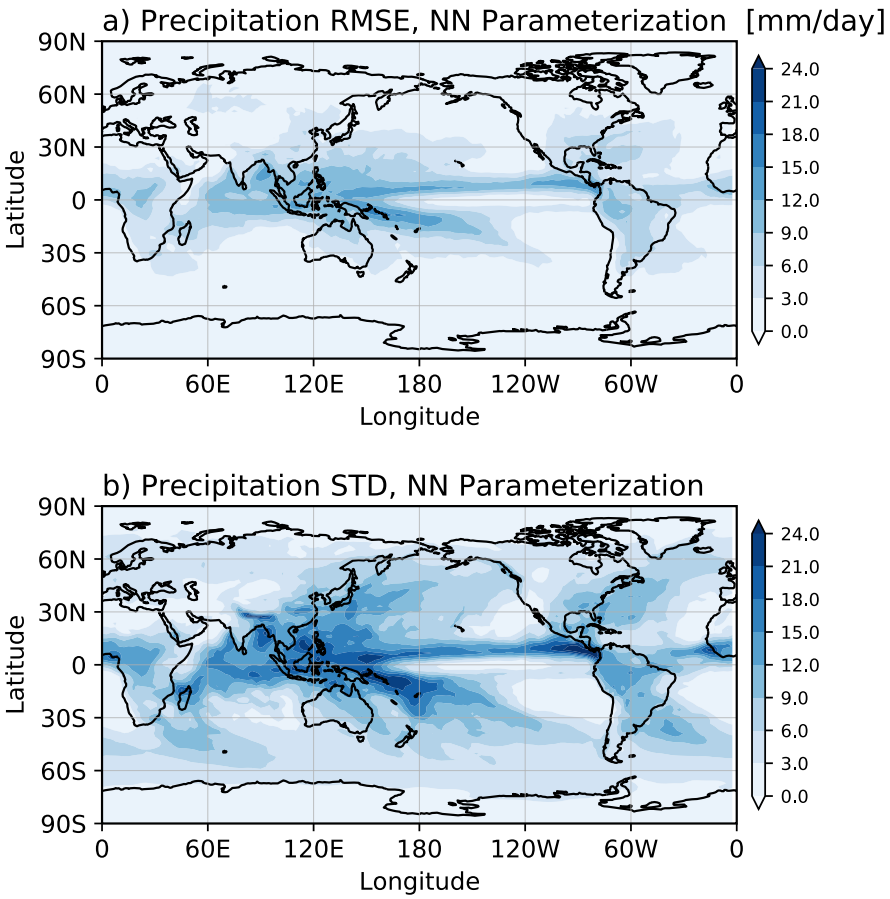

Figure R1. Spatial distribution of a) root mean square error (RMSE) and b) standard deviation (STD) of precipitation prediction.

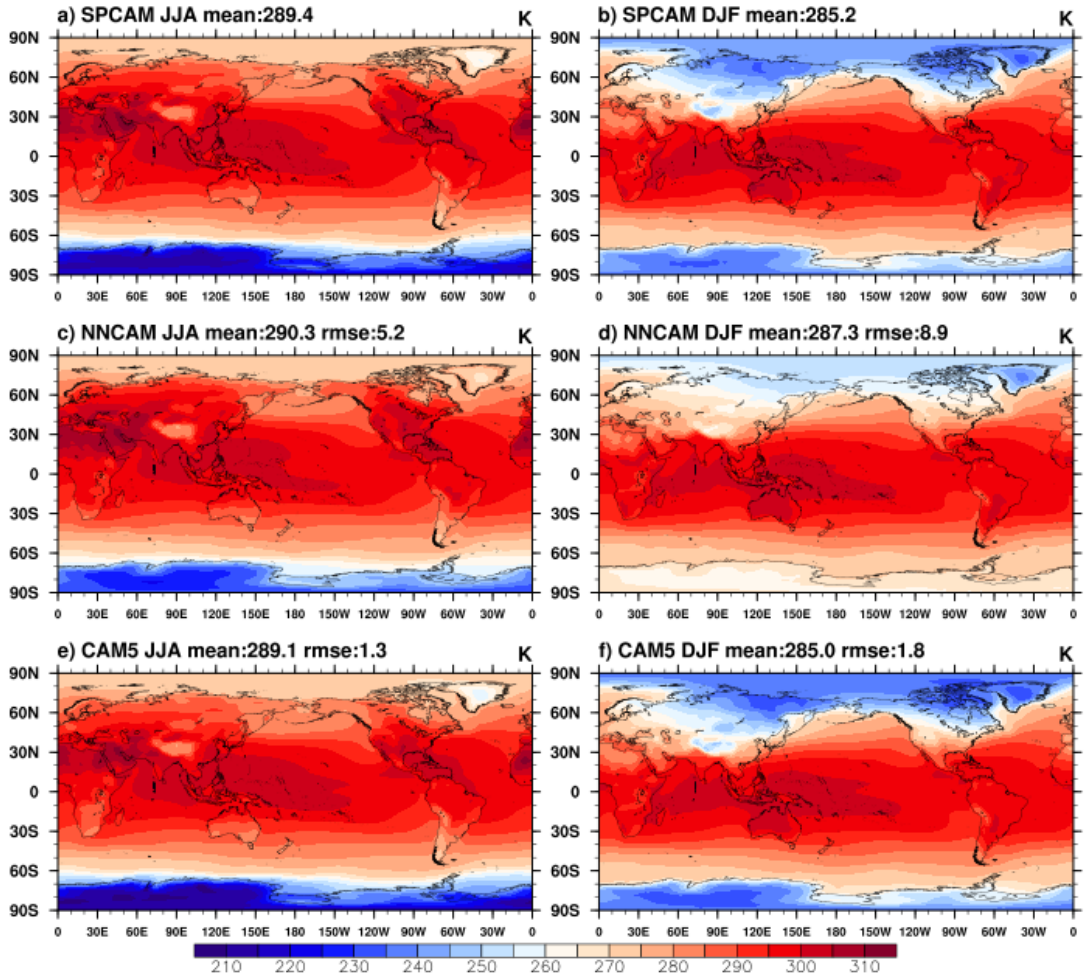

Figure R3. The mean 2m temperature of June-July-August (left panels) and December-January-February (right panels) for (a, b) SPCAM (1999−2003), (c, d) NNCAM (1999−2003), and (e, f) CAM5 (1999−2003)

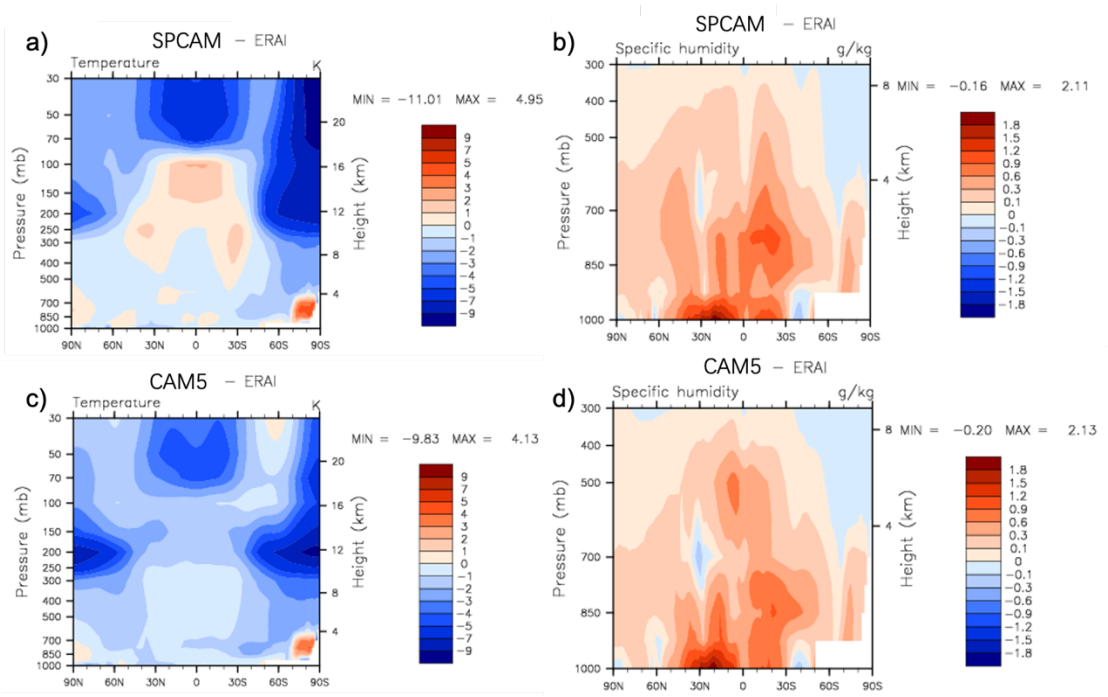

Figure R4. Latitude-pressure cross sections of annual and zonal mean differences for temperature (left panels) and humidity (right panels) between SPCAM and ERA-Interim (a & b) and between CAM5 and ERA-Interim (c & d).

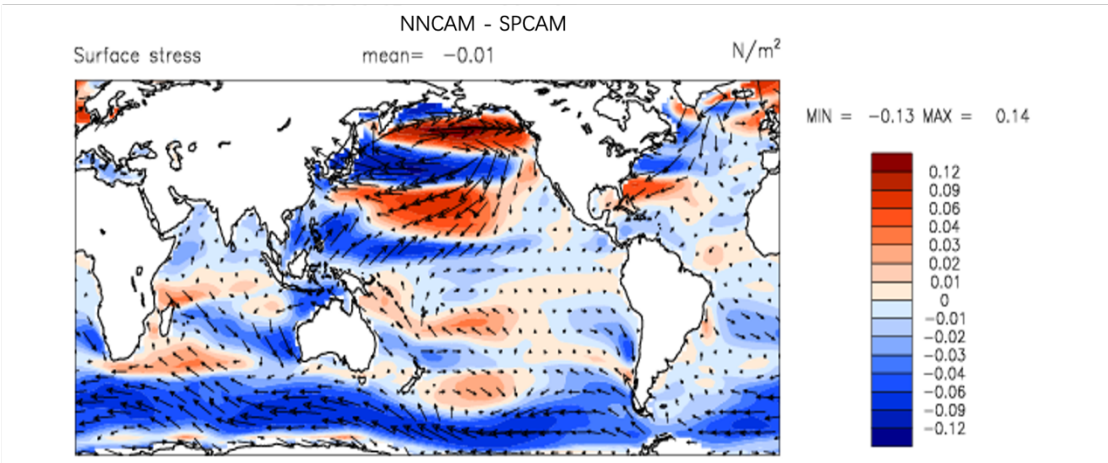

Figure R5. Global distribution of DJF surface wind stress differences between NNCAM and SPCAM.

Figure 1 – 14

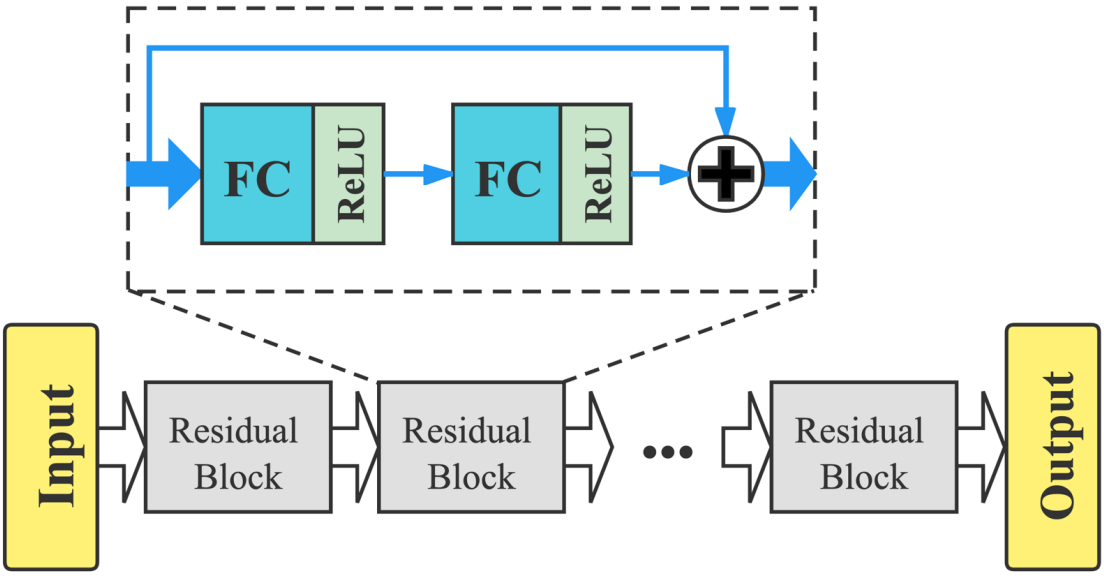

Figure 1. Schematic showing the structure of ResDNN. It consists of 7 residual blocks, each of which (shown in dashed box) contains two 512 node-wide dense (fully-connected) layers with a ReLU as activation, and a layer jump. The input and output are discussed in section 2.2.2.

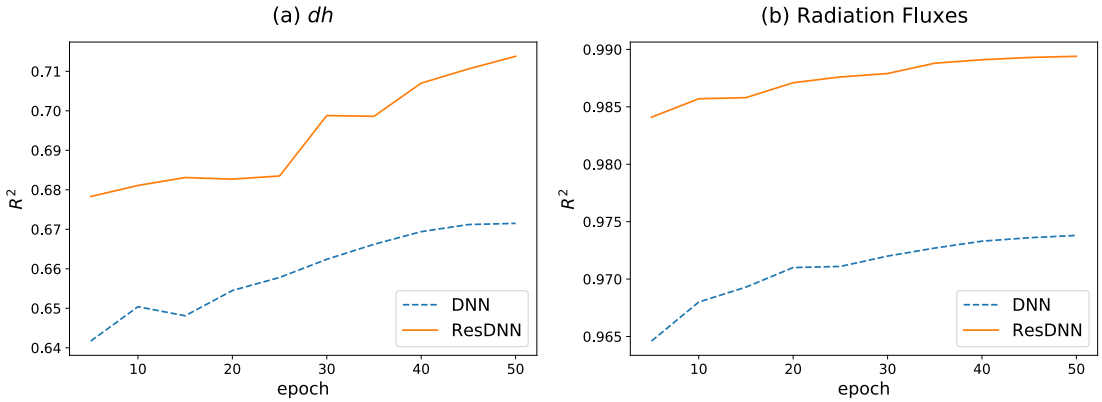

Figure 2. Fitting accuracies (R^2) of both the proposed ResDNN (orange solid lines) and DNN (blue dashed lines) for different targets. (a) shows the R2 of moist static energy changing rate (dh) versus training epochs and (b) shows the fitting accuracy of the average R^2 over the 8 radiation fluxes. Note: Spatial averaging of MSE is performed before calculating R^2.

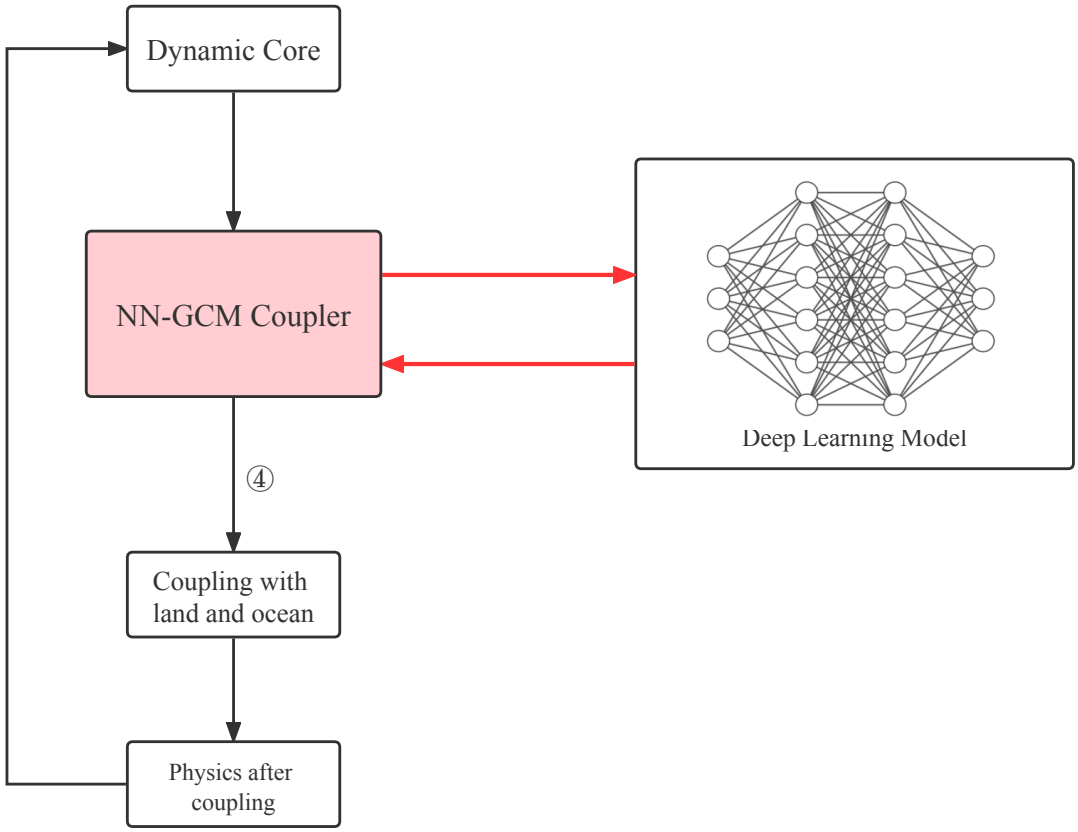

Figure 3. A flow chart of NNCAM including NN-GCM Coupler. NNCAM runs in the direction of the arrow, and each box represents a module. Among them, NN-GCM Coupler is indicated by light red. NN-Parameterization is shown in the sub-figure on the right. Note: ① represents the dynamic core transmits data to NN-GCM Coupler; ② and ③ represent the data communication between DNN-GCM Coupler and NN-Parameterization; ④ represents the host GCM accepts the result from NN-Parameterization.

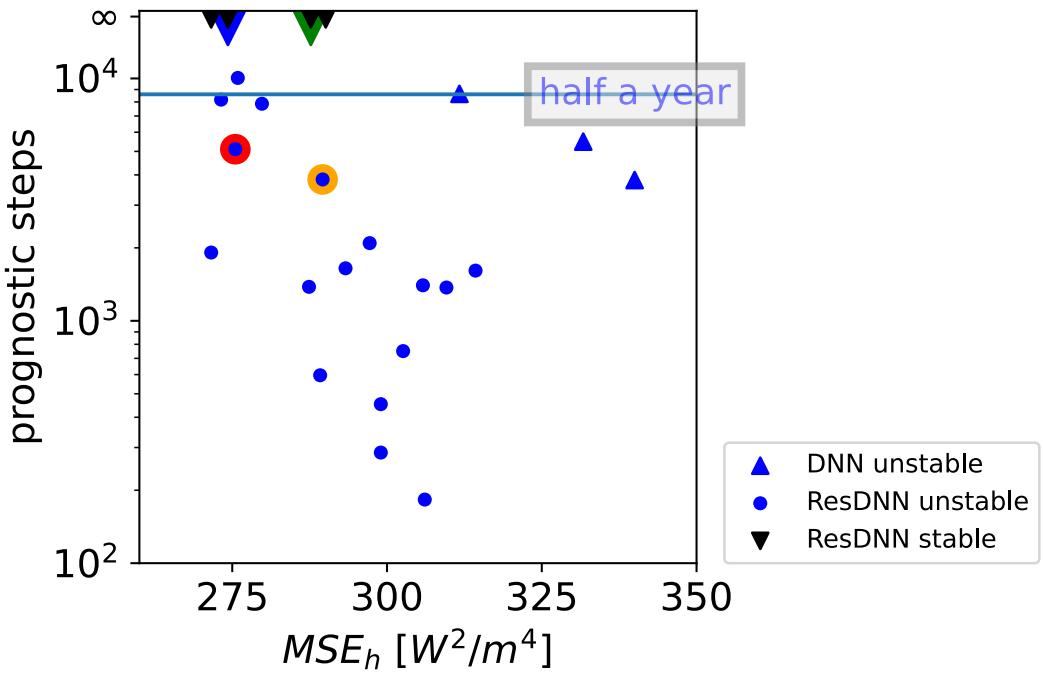

Figure 4. The offline moist static energy mean square error vs. prognostic steps. The black reversed triangles are stable NN coupled prognostic simulations lasting more than 10 years, blue ones are unstable simulations, and the blue triangles are for DNNs. The marked dots with colored outline are later exhibited in Figure 5 for time evolution of global averaged energy.

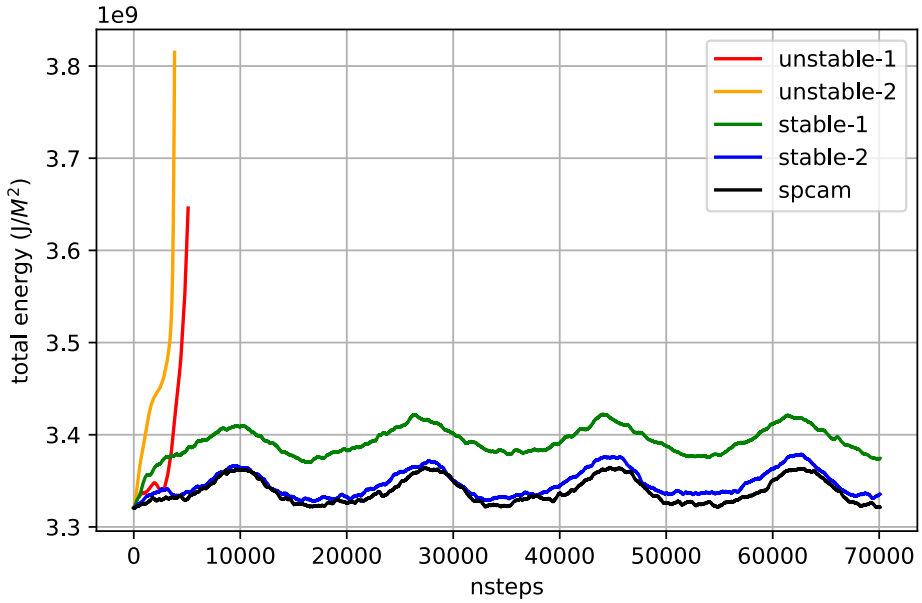

Figure 5. Time evolution of global averaged column integral total energy of NNCAM with different ResDNN parameterizations (marked with the same colors in Figure 4) and SPCAM target (the black line): Blue for stable and accurate ResDNN, green for a stable but deviated ResDNN, orange and red lines for unstable ResDNN.

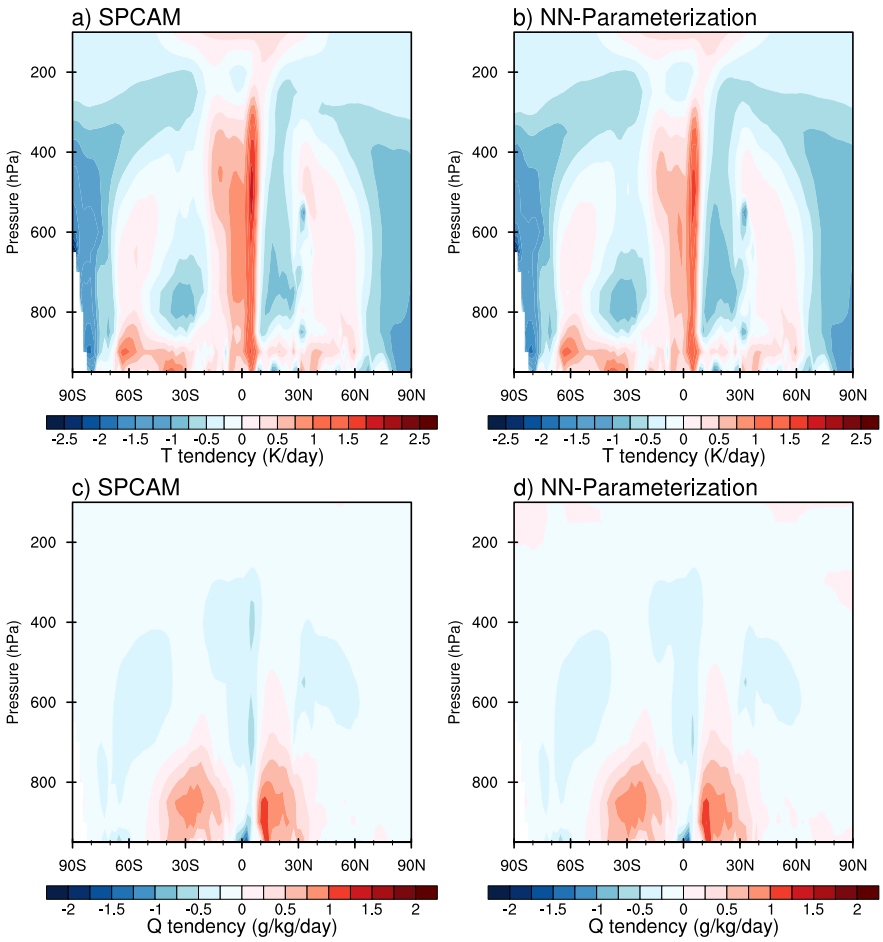

Figure 6. Latitude-pressure cross sections of annual and zonal mean heating (top) and moistening (bottom) from moist physics during the year 2000 for (a, c) SPCAM simulations, and (b, d) offline test by the NN-Parameterizations.

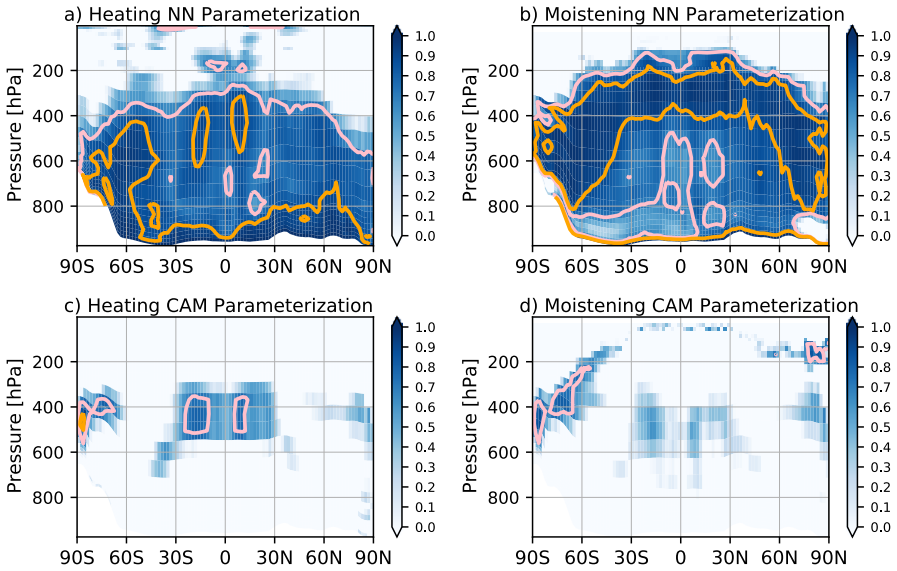

Figure 7. Latitude-pressure cross sections of coefficient of determination (R^2) for zonal averaged heating (left panels) and moistening (right panels). They are predicted by (a & b) NN-Parameterization in the offline one-year SPCAM run, and (c & d) by offline CAM5 parameterizations. Both are evaluated at 30-min timestep interval. Note: areas where R^2 is greater than 0.7 are contoured in pink and those greater than 0.9 are contoured in orange.

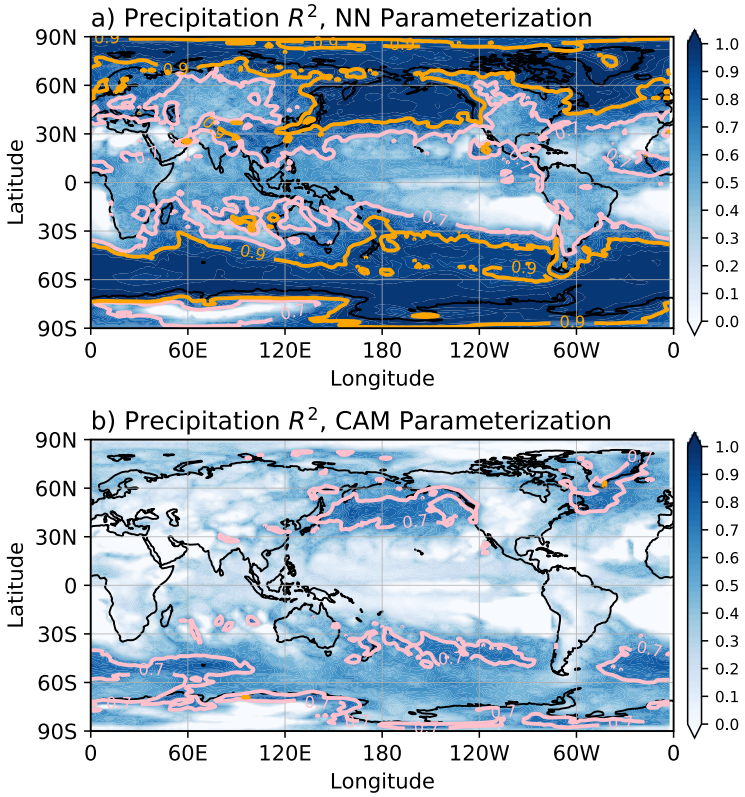

Figure 8. Latitude-pressure cross sections of coefficient of determination (R^2) for the derived precipitation predicted by NN-parameterization (a) and total precipitation from CAM5 parameterization (b) in the offline one-year SPCAM run. The predictions and SPCAM targets are in 30min timestep interval. Note: areas where R^2 is greater than 0.7 are contoured in pink and those greater than 0.9 are contoured in orange.

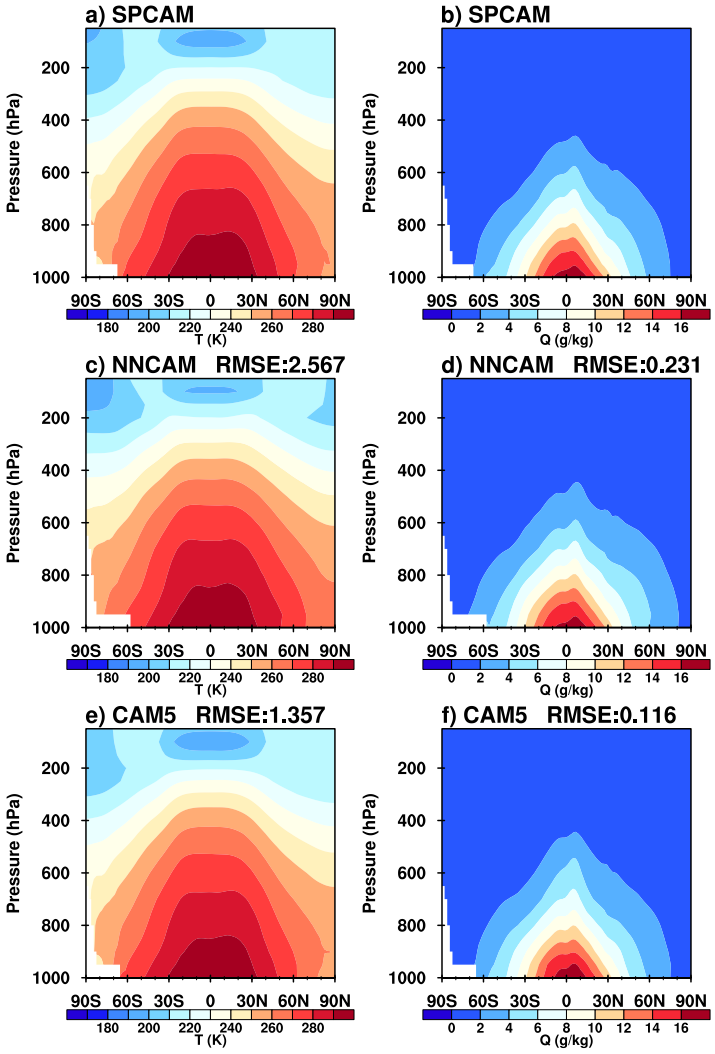

Figure 9. Latitude-pressure cross sections of annual and zonal mean temperature (left panels) and specific humidity (right panels) from (a, b) SPCAM (1999−2003), (c, d) NNCAM (1999−2003), and (e, f) CAM5 (1999−2003).

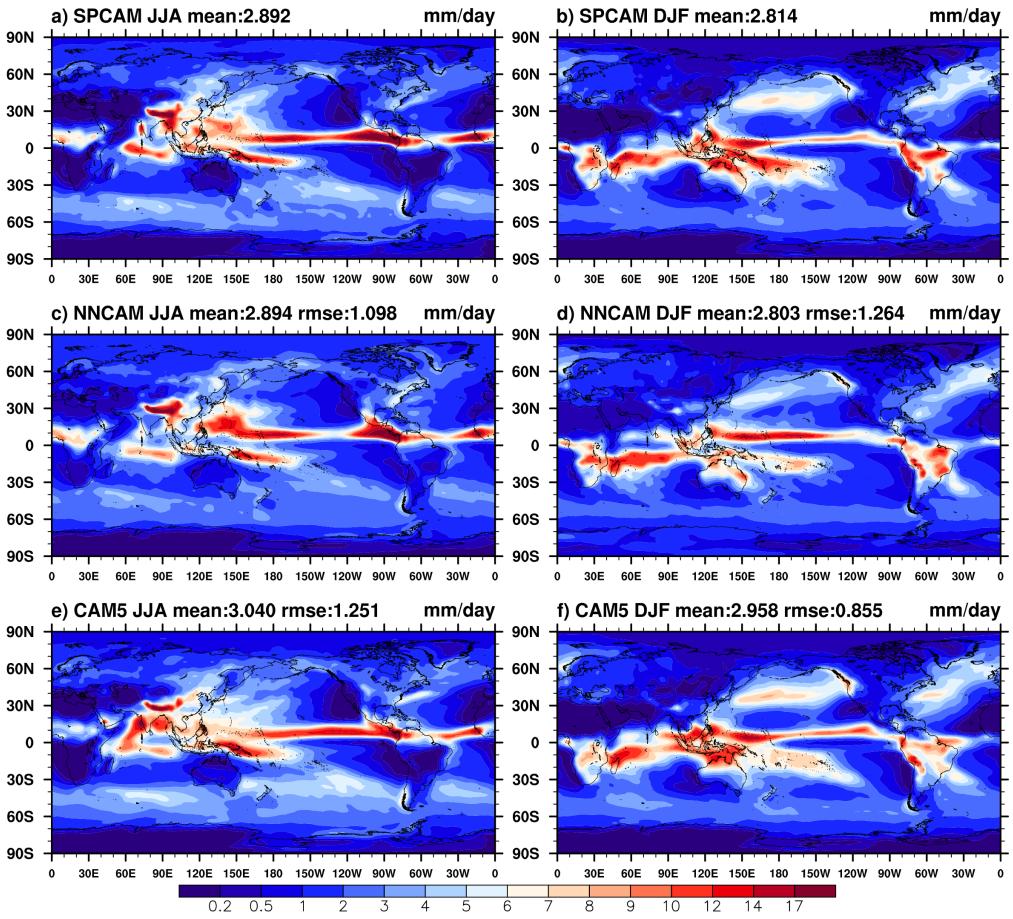

Figure 10. The mean precipitation rate (mm day^{-1}) of June-July-August (left panels) and December-January-February (right panels) for (a, b) SPCAM (1999−2003), (c, d) NNCAM (1999−2003), and (e, f) CAM5 (1999−2003).

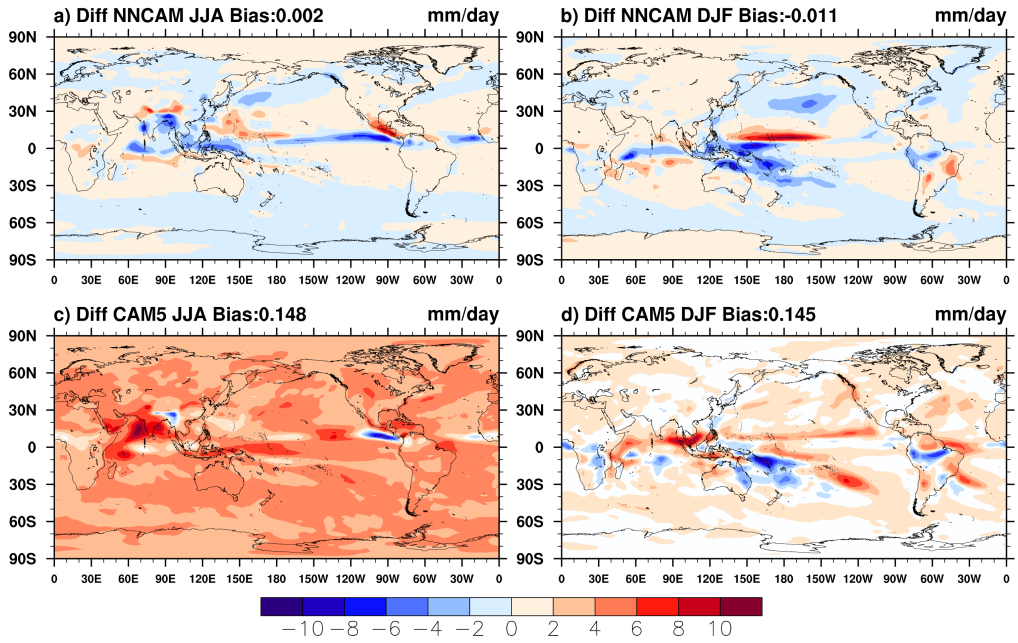

Figure 11. Global distribution of precipitation difference averged over boreal summer (left panels) and winter (right panels) between NNCAM and SPCAM (a & b) and between CAM5 and SPCAM (c & d).

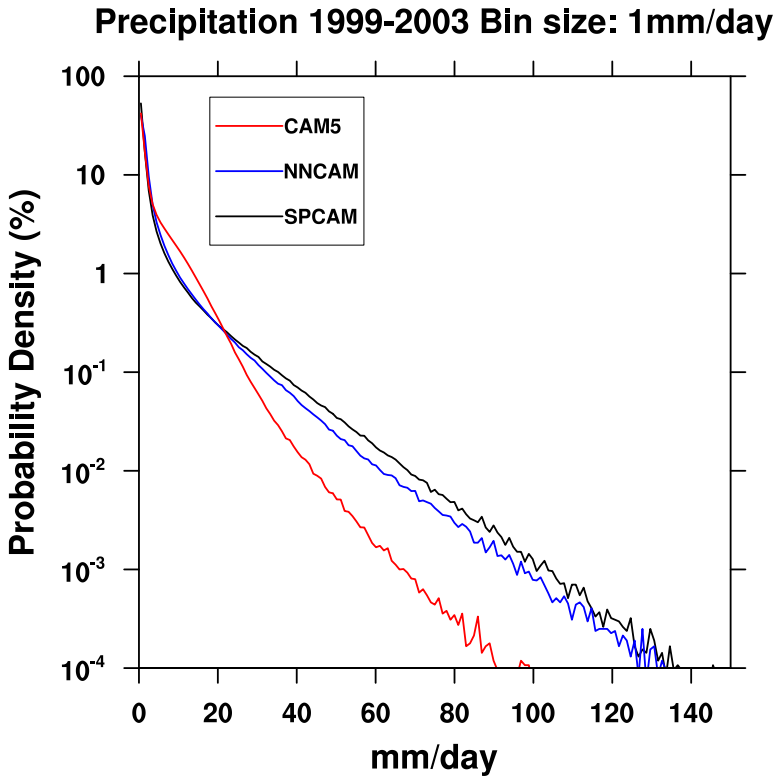

Figure 12. Probability densities of daily mean precipitation in the tropics (30°S−30°N) from the three model simulations. Black, blue and red solid lines denote SPCAM, NNCAM and CAM5, respectively.

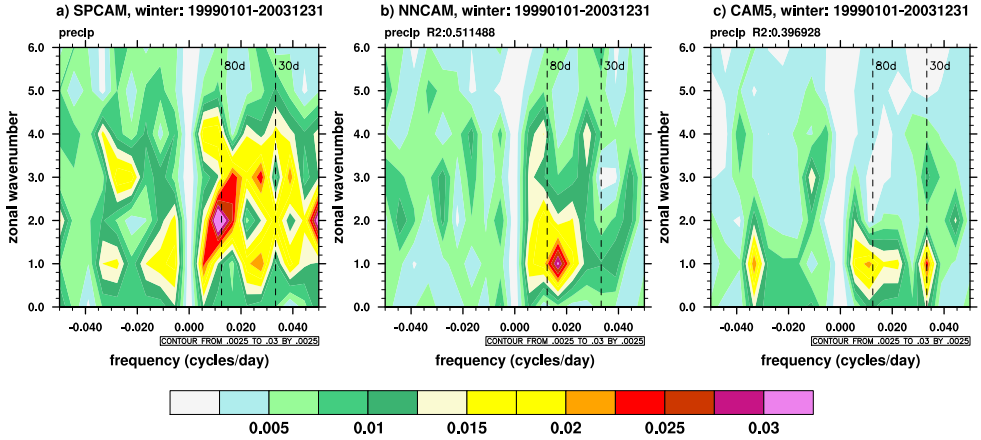

Figure 13. The wavenumber-frequency spectra of 10°S−10°N daily precipitation anomalies for (a, b) SPCAM, (c, d) NNCAM, and (e, f) CAM5 simulations for boreal winter.

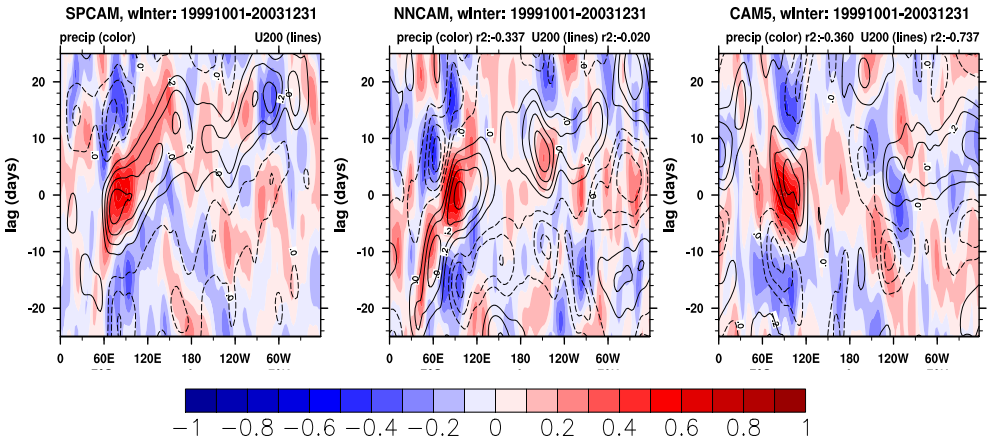

Figure 14. Longitude-time evolution of lagged correlation coefficient for the 20-100 day band-pass-filtered precipitation anomaly (averaged over 10°S−10°N) against regionally averaged precipitation (shaded) and zonal wind at 200hPa (contoured) over the equatorial eastern Indian Ocean (80E−100°E, 10°S−10°N). Dashed lines in each panel denote the 5 m s^{-1} eastward propagation speed.

Figure S1 – S4

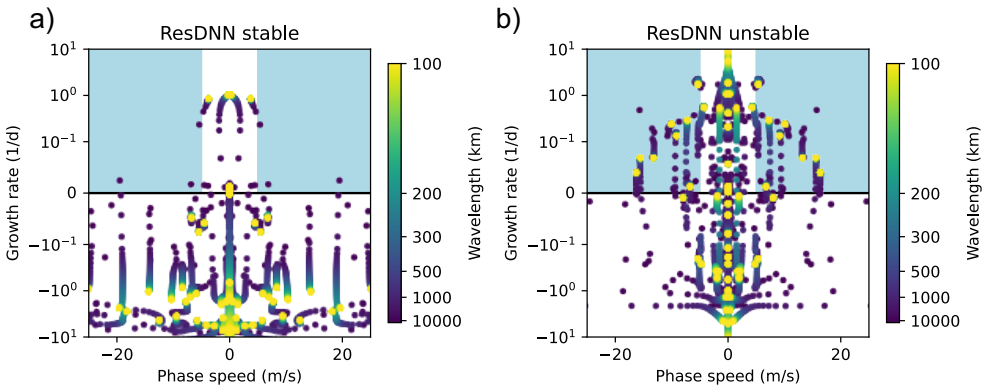

Figure S1. Wave spectra of a) a stable NN parameterization and b) an unstable parameterization. The light blue background indicates where the phase speed is above 5m/s and the growth rate is positive. The stability diagrams are obtained by coupling linear responses of NN parametrizations to simplified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

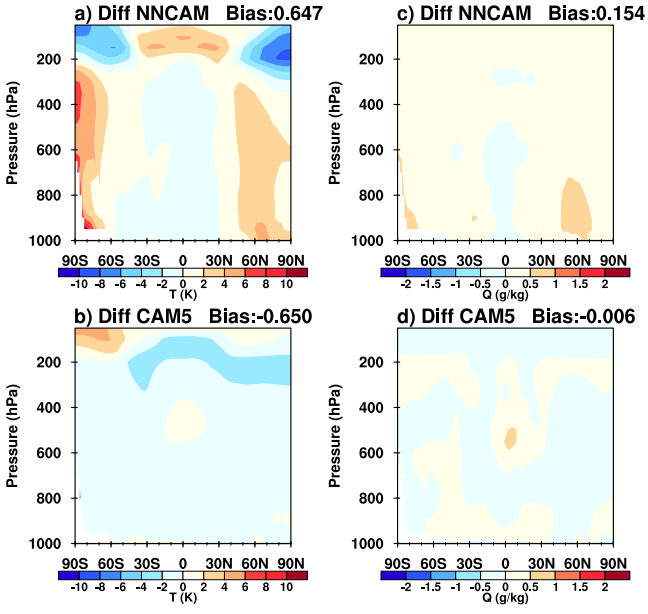

Figure S2. Latitude-pressure cross-section zonal and annual mean differences for temperature (top row) and specific humidity (bottom row) between (a & c) NNCAM and SPCAM and (b & d) CAM5 and SPCAM. The simulation period for all model is from 1999 to 2003.

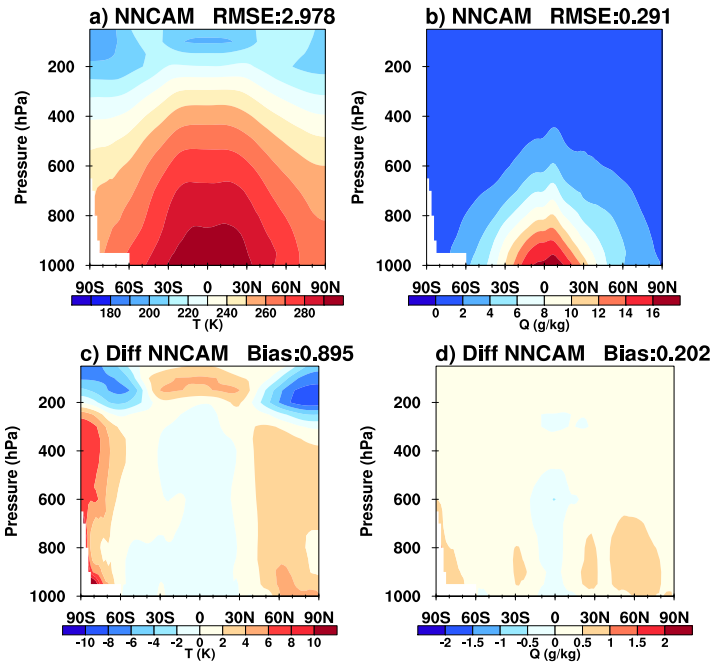

Figure S3. Latitude-pressure cross-section zonal and annual mean for a) temperature and b) specific humidity in NNCAM simulated from 2004 to 2008 with their differences with the SPCAM simulation from 1999 to 2003.

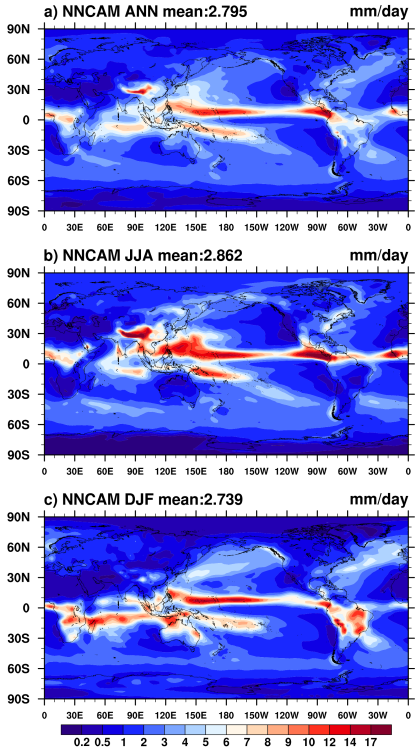

Figure S4. Global distribution of temporal mean precipitation predicted by NNCAM from January 1st 2004 to December 31st 2008 for a) annual, b) boreal summer (JJA), and c) boreal winter (DJF).