1 Response to Reviewer 1

The reviewers comments are marked in blue and our answers follow the corresponding comment in black.

1.1 Major

The introduction effectively argues for separating cloud radiative impacts from all-sky radiation, suggesting this approach could benefit climate change simulations by avoiding the need for retraining. However, this potential benefit is not supported by evidence in the results. To strengthen this claim, the authors should either:

1. Provide verification, perhaps using a 4xCO2 high resolution simulation if one is available. (Or do some quick analysis on the cloud radiative effect with the coarser model with/without 4xCO2 as a proxy for the high resolution model.)

2. If evidence is not available, move the discussion of this benefit to the discussion section of the paper.

I believe the model would also work for all-sky heating rate as the target, performing as well as it does for the cloud radiative effect. In this case, we don't have to go through all the separation process.

We thank the author for this comment. To strengthen the claim, we did the following calculation:

1. calculate fluxes using pyRTE for a number of samples from our test set (4000 samples). This corresponds to the reference climate
2. calculate cloud effect (cloud_effect_reference)
3. calculate fluxes using pyRTE for the same samples with 4xCO2
4. calculate cloud effect (cloud_effect_4xCO2)

The relative mean absolute difference between cloud_effect_reference and cloud_effect_4xCO2 is 4% for LW radiation and around 2.5% for SW radiation in the troposphere and even smaller in the stratosphere. This is 10 times smaller than the error due to representations of the subgrid-scale cloud effects and in the order of errors of cloud effects of the ML-enhanced radiation scheme. We agree that a model for all-sky heating would perform similarly well but this requires much more training data that is currently not available. To not disrupt the flow of the manuscript, we added the following explanation with a supporting figure to the appendix "The linear decomposition assumption of clear-sky heating and cloud radiative impact may be questionable for different greenhouse gas concentrations. To provide some validity, we estimate the error induced by this assumption for $4 \times CO_2$. For a direct estimation, we select 4000 random samples from the test set and calculate the cloud radiative impact for the reference climate (CO_2 concentration of 2004) offline using pyRTE. Next, we increase CO_2 by a factor of 4 and repeat the calculation. The mean absolute difference between cloud_effect_reference and cloud_effect_4xCO2 gives an estimate of the error that is induced by the linear decomposition assumption for $4 \times CO_2$. The vertical resolved error and bias are shown in Figure D1. The x-range is the same as in Figures 4, 5 and C1 for comparison with the errors induced by subgrid-scale clouds and the remaining errors of MLe-radiation. In general, the error is smaller in the stratosphere than in the troposphere. For SW, the error is around 2.5% and for LW it is around 4% in the troposphere. This around 10 times smaller than the error from subgrid-scale clouds and around the same magnitude as the errors from MLe-radiation."

1.2 Minor

"O3, ρ, T, and Tsurf are normalized using their mean values μ and standard deviation σ": Please specify the dimension over which the mean and standard deviation are calculated. Are they computed over the whole dataset? Is there any height dependency?

The mean values and standard deviation are calculated over the whole globe from the first training time step. This yields good normalization factors as we don't have hard constraints on the input parameters. There is now

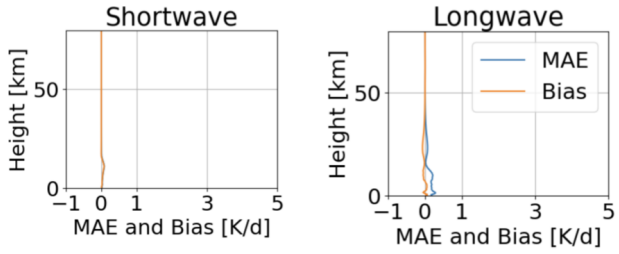

Figure 1: Mean absolute error (MAE) and bias induced by the linear decomposition assumption for $4 \times CO_2$. For comparison, the x-range is the same as in Figures 4, 5 and C1.

height dependency as we found that this leads to better results (not shown). We added the following text "The normalization factors are computed using all cells of one time step and are one-dimensional such that the vertical structure of the variables remain. The vertical structure is an important aspect that the BilSTM uses to make a vertically correlated prediction."

"We discarded a few coarse-grained cells, e.g., if the surface height of the coarse-grained cell deviated by more than 0.5m from the coarse-scale surface height." I don't get this part? Do you mean the variance of the fine-grained cell is larger than a certain threshold?

The coarse and fine grid are slightly rotated such that one coarse cell does not perfectly fit 256 fine cells. The horizontal coarse-graining was done using cdo and first order conservative remapping based on cell area. This lead to small differences in the surface height compared to the surface height from a coarse ICON simulation. If this height was off by 0.5m, we discarded that cell. We added more details to the description, specifically "We discarded a small number of coarse-grained cells if their surface height showed inconsistencies, which can occur over complex terrain. Specifically, the coarse-grained surface height was computed from the high-resolution grid file and compared to the surface height from the coarse-scale grid file, which is slightly rotated and shifted. Consequently, some high-resolution cells are only partially contained in a coarse cell, which can lead to a small mismatch in surface height. Cells with deviations of more than 0.5 m in surface height were discarded."

Figure 3. The difference between coarse-scale and coarse-grained cloud impact below 1km is quite obvious for both lw and sw. Is it concerning?

We agree that the difference is quite obvious, which was expected. The difference can also be observed for cloud liquid which directly affects heating rates. However, this is not concerning because the coarse-scale distribution is mostly contained in the coarse-grained distribution and the ML-model learned to react to different cloud states. So, the ML-model will effectively see a larger variability in states during training. The difference in spread that is larger for lower tropospheric LW cloud impact relates to only a fraction of samples.

Figure 4. The notation should be improved to avoid confusing. I assume the pyRTE results are meant to represent the coarse-scale radiation result, which is the baseline here. The ground truth is the saved results from QUBICC simulation. It would be less confusing if you can make this clear in both text and the figure/caption.

We changed the notation throughout the manuscript to make it more clear. We refer to QUBICC directly as ground truth and pyRTE is called baseline. We added a table to summarize all different dataset.

"The second column of Figure 4 shows results for fully cloudy samples (total cloud cover of 100%). For pyRTE, the MAE peaks near 10km, exceeding 5K/d for both SW and LW.": Is the pyRTE SW/LW MAE larger than 5K/d? The blue line is 0.5K/d for SW and 1K/d for LW.

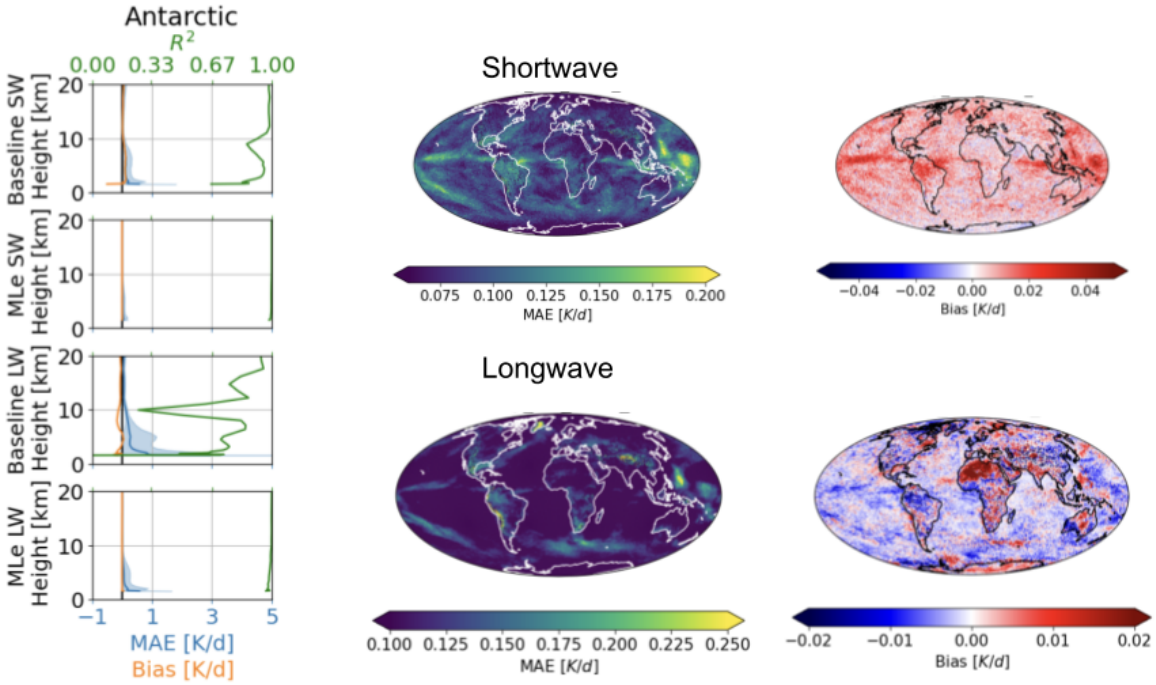

Figure 2: Left: similar to Figure 5 but for the Antarctic region. Top: MAE and bias map for SW. Bottom: MAE and bias map for LW.

We thank the reviewer for spotting this. We corrected it in the text.

"The corresponding R2 are low, with average values of 0.83 (SW) and 0.66 (LW), compared to 0.98 for the ML-enhanced scheme". How are the averaged values computed? Weighted by mass or simple average over values at different levels (how the levels are distributed)?

The average was computed as simple average over all levels. The levels are terrain following sigma coordinates. Considering the layer thickness or mass, when averaging the vertically resolved R2, won't change the result that the baseline (pyRTE) performs worse for this metric. A visual comparison of the green lines in Figure 4 and 5 shows the same result. We specified how the averages are calculated "The average R^2 values are computed by averaging over the vertical levels."

Figure 5. The breakdown of the different regions is informative. Is it possible to make a map of bias and MAE (if you have enough samples for the 80km resolution grid or even 200km)? It would provide more information for different audiences. For example, I am curious about the quality in the Antarctica region.

Yes, this is possible, and we provide the additional plots here. However, the sample size is limited and the maps should be interpreted with caution. Therefore, we don't add the plots to the main manuscript. Additionally, maps hide the vertical distribution and should be interpreted together with the vertically resolved results, e.g., Figure 4 and 5. For the Antarctic region, we filtered for samples with a latitude smaller than $-70\circ$. Here, the ML model performs similar to the Arctic region.

Figure C1. Could you comment on the large error in the stratosphere for both pyRTE and ML?

In the stratosphere, the heating rate is very sensitive to small differences in fluxes because the flux divergence

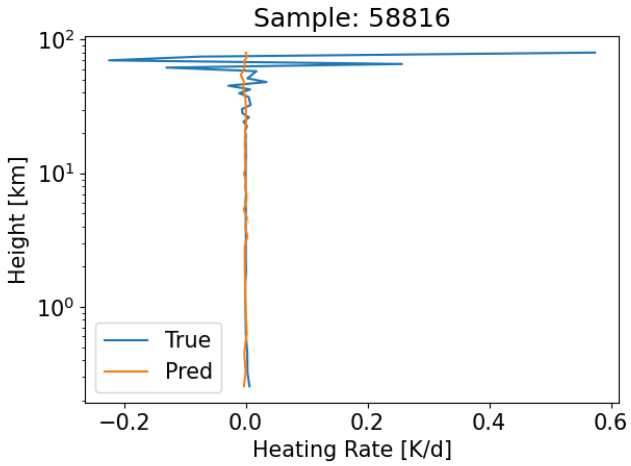

Figure 3: Example of cloud radiative effect on heating rate for SW radiation for a clear sky sample. Blue: coarse grained, orange: predicted by the neural network.

is divided by pressure difference (Eq. B1) or mass (Eq. 2), which is orders of magnitude smaller than in the troposphere. The sensitivity goes up to floating point precision, where coarse-graining induced numerical noise. Due to the large training size, the ML model was able to identify this as noise and therefore the bias is zero. Figure 3 shows one example of the cloud effect on SW heating rate. This is a clear-sky sample, therefore the cloud effect should be zero everywhere. In the troposphere, it is zero and in the stratosphere it oscillates around zero where the amplitude increases with hight. We added the following text to the description "The large errors in the upper stratosphere for the baseline are related to rounding errors increasing with height."