

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 $4 \times \text{CO}_2$. For a direct estimation, we select 4000 random samples from the test set and calculate the cloud radiative impact for the reference climate (CO₂ concentration of 2004) offline using pyRTE. Next, we increase CO₂ 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 $4 \times \text{CO}_2$. 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 T_{surf} 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

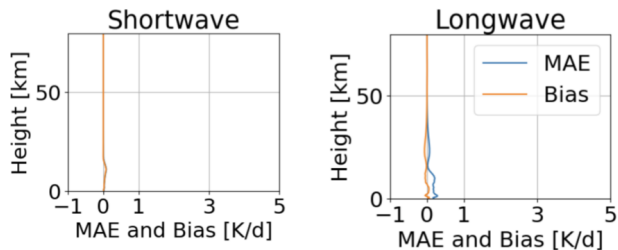


Figure 1: Mean absolute error (MAE) and bias induced by the linear decomposition assumption for $4 \times \text{CO}_2$. For comparison, the x-range is the same as in Figures 4, 5 and C1.

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 BiSTM 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.

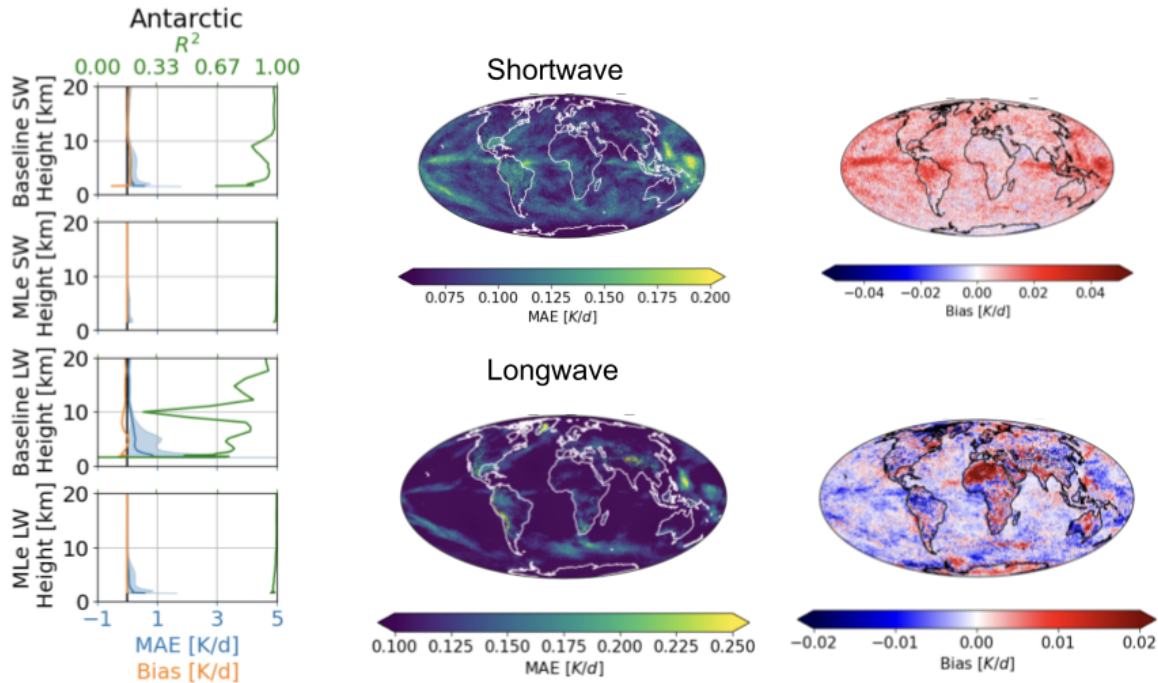


Figure 2: Left: similar to Figure 5 but for the Antarctic region. Top: MAE and bias map for SW. Bottom: MAE and bias map for LW.

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

"The corresponding R^2 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 R^2 , 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° . 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

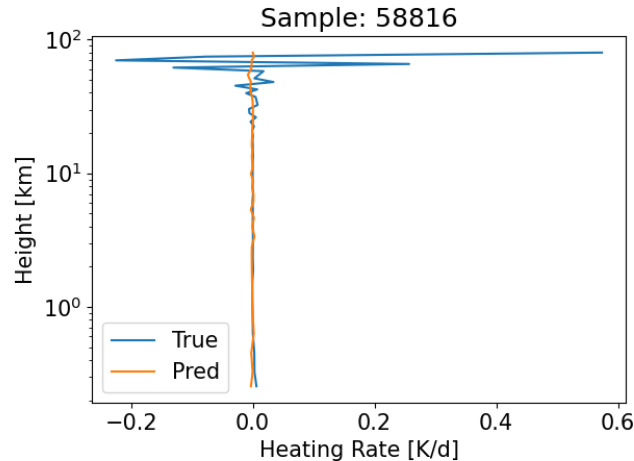


Figure 3: Example of cloud radiative effect on heating rate for SW radiation for a clear sky sample. Blue: coarse grained, orange: predicted by the neural network.

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 height. 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."

2 Response to Reviewer 2

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

2.1 Review Summary

This manuscript aims to improve the treatment of sub-grid cloud variability in coarse-scale simulations by using machine learning to predict cloud radiative effects (CRE). Specifically, the authors use the existing radiation scheme for clear-sky conditions, while enhancing all-sky calculations with CRE predictions from a machine-learning model trained on coarsened high-resolution simulations.

The idea is scientifically reasonable, and the introduction is well-written. However, I struggled to fully understand the results. The manuscript would benefit greatly from clearer and more systematic descriptions of what was exactly done in the datasets, especially in Sections 3 and 4.

We thank the reviewer for their comments and address the comments one-by-one below.

2.2 Major

2.2.1 Clarify the treatment of radiation in high-resolution and coarsened datasets

This is the most critical missing piece. Section 3 describes the QUBICC simulations and coarsening procedure, yet radiation is barely addressed. The only mention is that snow is considered in RTE+RRTMGP, which is relatively

minor to the overall methodology. Please clarify:

In the original QUBICC simulations (5 km):

- Was radiation computed using RTE+RRTMGP with maximum-random overlap?
- Or did you use an all-or-nothing cloud cover assumption? The manuscript loosely uses "high-resolution simulations," making it unclear.

We thank the reviewer for this comment and added more details to section 3 which is now: "Both simulations use the radiation scheme RTE+RRTMGP (Pincus et al. 2019). ICON-A uses McICA and maximum-random overlap to represent subgrid-scale cloud-radiative impacts. In the QUBICC simulation, the *all-or-nothing* scheme is used, assuming horizontal homogeneous distribution of clouds. The physics and radiation time step is 6 min for the ICON-A simulation, matching the radiation time step in QUBICC."

In the coarsening process: Are radiation fields simply averaged over coarse boxes? How is cloud fraction determined in the coarsened dataset? While Grundner et al. (2022) is cited for coarsening process, a more explicit description within this paper is essential for readers to understand the implications of coarsening, especially on radiation fields and cloud fraction.

We added more detailed descriptions to the text. Specifically, we added "All variables are horizontally and vertically coarse-grained from high-resolution simulations as in Grunder et al. (2022). For horizontal remapping, the high-resolution cells are weighted by their cell area that is contained in each coarse cell. High-resolution cells can be contained fully or partially in a coarse cell, which is represented by a smaller cell area. Similarly, the layer thickness was used for vertical coarse-graining, corresponding to a weighted average. This method is valid for all input and output variables used here, i.e., concentrations and fluxes. Cloud-related variables are expressed as mass fraction, ensuring consistent vertical coarse-graining. Absolute mass variables, such as air mass m_{air} in Equation 2, are coarse-grained by the (weighted) vertical sum that are partially or fully contained in a coarse layer."

2.2.2 What is the benchmark for "improvement"?

It is unclear what constitutes the "truth" when evaluating improvement. It appears that the authors compare ML-enhanced all-sky radiation against that computed by the existing radiation scheme using coarse-scale input. However, since neither inherently represents truth, it is difficult to claim "improvement." Rather, it is a comparison between two imperfect approximations.

We consider the (coarse-grained) high-resolution simulation as "truth". Coarse-scale radiation, which was calculated with pyRTE on coarse-grained fields for a one-by-one comparison, is considered baseline. We improved the description also with respect to the next comment.

2.2.3 Section 4 is very difficult to follow

The description of datasets and experiments in Section 4 is unclear. The repeated use of "with pyRTE" is ambiguous because presumably all simulations could involve pyRTE. I could not clearly understand which datasets were being evaluated, what the reference was, and what exactly was being shown in Figures 2-4. It would be helpful to provide a table summarizing all datasets. Also, please clearly address:

What is the ML model's performance on the coarsened high-resolution test dataset? Is Figure 4 meant to show this?

What is the difference between the datasets used in Figures 2&3, and those described in Section 4?

Much of the confusion could be resolved by more precise wording and descriptions and by consistently naming the datasets.

To resolve the confusion, we added a table summarizing all dataset (QUBICC, ICON-A, baseline, MLe-radiation), how they are computed and where they are used. We adjust the notation accordingly and added clear descriptions throughout the text. Specifically, we added the following paragraphs: "The output of pyRTE is used as baseline as

it represents the coarse-scale radiation scheme. This allows a sample-by-sample comparison to the coarse-grained heating rates obtained from QUBICC (first and third row in Figures 4 and 5). The ML-enhanced radiation scheme predicts the cloud radiative impact, which is added to clear sky heating. The resulting all-sky heating is compared to the coarse-grained QUBICC heating rates (second and third row in Figures 4 and 5). As evaluation metric, we use mean absolute error (MAE), bias, and the coefficient of determination (R^2). The improvement is measured by comparing MLe-radiation to the baseline."

2.3 Minor

Page 6: Printing uneven output intervals is useful. Have you investigated potential impacts on sampling global cloud distributions and cloud diurnal cycle?

So far, we only focused on uneven output intervals to increase the variability of different solar zenith angles at all locations. Considering a cloud-based sampling would be interesting aspect that could be used in future work.

Page 6: The phrase "In order to evaluate the high-resolution data" is unclear. Coarse-scale data cannot evaluate high-resolution data; please rephrase more precisely.

We changed the wording to "For comparison with the high-resolution data,"

Page 6: In addition to comparing variable ranges, it seems more important to analyze joint distributions, which reflect underlying physics and correlations that machine would try to learn.

We agree that analyzing the joint distribution is of interest, however the primary focus of this manuscript is the development of a new ML-enhanced radiation parameterization. The purpose of the range comparison was included as a diagnostic check for systematic differences. A detailed analysis of the joint distribution would shift the focus from the main objective and is therefore beyond the scope of this study.

Page 8: The statement "which may be due to the different spread in cloud water at 1 km (Figure 2b)" appears to be an educated guess. It would be more informative to provide a more insightful explanation, potentially linking this difference to specific physical processes or parameterization choices that could be responsible.

We have revised the wording to state this as an observation rather than a causal explanation. While the co-occurrence of a larger spread cloud water and LW heating is consistent, a definite attribution would require further analysis that is beyond the scope of this study. The new wording is now "However, the spread in heating rates is slightly different, which co-occurs with the different spread in cloud water at 1 km (Figure 2b)."

Page 9: The statement regarding "horizontally homogeneous input parameters" should explicitly specify the corresponding dataset, particularly in relation to the datasets described in the opening paragraph of Section 4.

We specified the statement to "While coarse-grained QUBICC heating rates reflect subgrid-scale variability of water vapor, coarse-scale radiation schemes assume a horizontal homogeneous distribution of water vapor. Therefore, the assumption of horizontally homogeneous input parameters introduces a small error of 0.367 K/d (SW) and 0.571 K/d (LW) for the baseline."

Page 9: The statement "exceeding 5 K/d" seems inconsistent with Figure 4 (SW max appears closer to 3 K/d).

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

Page 10: Figure 4 caption incorrectly labels "samples with partial cloudiness (right column)".

We thank the reviewer for spotting this mistake. The sentence was changed to "The results are shown separately for clear-sky samples (no clouds, left column), fully cloudy sky samples (second column), and samples with partial cloudiness (third column). Additionally, the results are separated by non-precipitating (fourth column) and precipitating clouds (fifth column)."

Page 10: The statement that cloudier scenes have larger errors is correct but it would be better to provide deeper insight.

The subsequent paragraphs further condition the error statistics on cloud-regime and regions, showing larger errors for deep convective regimes. A causal attribution of why the errors occur, would require additional analysis beyond the scope of this study.