1 Response to Reviewer 2

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

1.1 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.

1.2 Major

1.2.1 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."

1.2.2 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.

1.2.3 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 discriptions 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."

1.3 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.