Response to Reviewer 1

We sincerely appreciate the reviewer for the thorough evaluation of our manuscript and for the constructive and insightful comments. These comments have helped us identify several aspects of the manuscript that require clarification, particularly the representation of neural network inputs and outputs, the physical interpretation of radiative transfer processes, and the presentation and interpretation of forecast verification results. We have carefully addressed all the major and minor comments point-by-point, and revised the manuscript accordingly. The detailed responses are as follows:

Major Suggestions

Lines 173-175. The statement is hard for me to understand. What does "multi-channel one dimension vector" mean? As documented in Table 1, input cloud parameters could have dimensions of (ngptlw, ncol, nlay), (ncol, nlay), and (ngptsw, ncol, nlay). What did the authors specifically do to merge the cloud parameters at different g-points (i.e., quadrature points)? What does "6-channel one-dimensional vector" mean? It would be better to be specific.

Response: We thank the reviewer for this thoughtful and precise comment.

In our implementation, each training sample corresponds to a single vertical atmospheric column. As shown in Table 1, the input variables have different original dimensions. Variables without a spectral g-point dimension, such as pressure and temperature, are directly represented as vertical profiles. Cloud-related variables with an additional g-point dimension, such as those with shapes (ngptlw, ncol, nlay) or (ngptsw, ncol, nlay), require further preprocessing.

For these cloud-related variables, we considered two preprocessing approaches. The first approach treats the values at individual g-points as separate input channels, preserving the complete g-point information but substantially increasing the number of input channels. The second approach aggregates the values by summation along the g-point dimension, thereby reducing the number of input channels and improving computational efficiency. We adopted the second approach in this study. Specifically, for each atmospheric column, the values of each cloud-related variable are summed over all corresponding longwave or shortwave g-points at each vertical layer. Consequently, variables originally having dimensions of (ngptlw, ncol, nlay) or (ngptsw, ncol, nlay) are reduced to vertical profiles of length nlay. Although this approach reduces the input dimensionality, the detailed distribution of information across individual g-points is no longer retained.

After preprocessing, all input profiles are padded where necessary to obtain a consistent vertical dimension of 89 and stacked along the channel dimension. The resulting input tensor for each atmospheric column has the shape (N_channels, 89), where N_channels denotes the total number of input features after preprocessing.

Therefore, the term "multi-channel one-dimensional vector" refers to a set of vertical profiles, with each channel representing a distinct input feature and the one-dimensional axis corresponding to the vertical coordinate.

Regarding the network output, the longwave and shortwave radiation networks are trained separately, and each network predicts six radiative quantities: total sky upward flux, total sky downward flux, total sky radiative heating rate, clear sky upward flux, clear sky downward flux, and clear sky radiative heating rate. These six quantities constitute the "6-channel one-dimensional vector" described in the original manuscript. Since radiative fluxes are defined at the 89 model levels, whereas heating rates are defined within the 88 model layers, padding is applied where necessary to obtain a unified vertical dimension of 89. Consequently, the output tensor for each atmospheric column has the shape (6, 89).

We have revised the corresponding manuscript text to clarify the preprocessing procedure, the treatment of the g-point dimension, and the physical meanings and dimensions of the input and output channels.

The revised manuscript text reads as follows:

"Input and output variables were processed separately for each vertical atmospheric column. For cloud-related input variables containing an additional spectral g-point dimension (ngptlw or ngptsw), two preprocessing approaches were considered: treating individual g-points as separate input channels or aggregating them by summation along the g-point dimension. The latter approach was adopted to reduce the number of input channels and improve computational efficiency. Variables without a g-point dimension were retained as individual vertical profiles. All input profiles were padded where necessary to obtain a unified vertical dimension of 89 and stacked along the channel dimension, resulting in an input tensor of shape (N_channels, 89), where N_channels represents the total number of input features after preprocessing. The longwave and shortwave radiation networks were trained separately, with each network producing six output profiles: total sky upward and downward fluxes, total sky radiative heating rate, clear sky upward and downward fluxes, and clear sky radiative heating rate. Since radiative fluxes are defined at the 89 model levels and heating rates at the 88 model layers, padding was applied where necessary to obtain a consistent output tensor shape of (6, 89) for each atmospheric column."

Line 332. What do you mean by "converging to highly fragile local minima"? Would it be helpful to explain more using mathematical language?

Response: We thank the reviewer for this insightful comment. We agree that the expression "converging to highly fragile local minima" was not sufficiently precise and requires further clarification.

Our original intention was to describe neural network solutions that achieve low training errors but exhibit high sensitivity to small perturbations in atmospheric input states. Such sensitivity may lead to large prediction errors when the network encounters atmospheric conditions that are not adequately represented in the training

dataset, potentially affecting the numerical stability of the coupled hybrid model. However, this behavior is more appropriately characterized in terms of input sensitivity rather than the geometric properties of local minima in the parameter optimization landscape.

Mathematically, for a neural network $f_\theta(x)$, where x denotes the atmospheric input variables and θ represents the network parameters, the response to a small input perturbation ϵ can be approximated using a first-order Taylor expansion:

$$f_\theta(x+\epsilon) \approx f_\theta(x) + J_\theta(x)\epsilon$$

where $J_\theta(x) = \partial f_\theta(x)/\partial x$ is the Jacobian matrix of the network outputs with respect to the inputs. Accordingly, the magnitude of the prediction change satisfies approximately

$$\|f_\theta(x+\epsilon) - f_\theta(x)\| \lesssim \|J_\theta(x)\|_2 \|\epsilon\|.$$

This relationship indicates that a large input Jacobian norm can amplify small perturbations in atmospheric states, potentially resulting in substantial changes in the predicted radiative quantities.

To improve robustness against such perturbations, random noise can be introduced into the training inputs. For a mean squared error objective, the corresponding noise-augmented training loss can be expressed as

$$L_{noise}(\theta) = \mathbb{E}_{(x,y),\epsilon}[\|f_\theta(x+\epsilon) - y\|^2]$$

where y denotes the reference radiative quantities and ϵ represents the random input perturbation.

Assuming that the perturbations are sufficiently small, have zero mean, and satisfy $\mathrm{Cov}(\epsilon) = \sigma^2 I$, and approximating the neural network locally by its first-order Taylor expansion, the noise-augmented loss can be expressed approximately as

$$L_{noise}(\theta) \approx L_{original}(\theta) + \sigma^2 \mathbb{E}_x[\|J_\theta(x)\|_F^2]$$

where $\|\cdot\|_F$ denotes the Frobenius norm. Under this local linear approximation and the assumed noise distribution, noise injection can be interpreted as introducing a Jacobian-related regularization effect, which may encourage reduced sensitivity to small input perturbations.

This provides a mathematical explanation for how noise-augmented training may improve the robustness of the radiation emulator. Nevertheless, noise injection does not guarantee convergence to a particular type of local minimum or necessarily produce a smoother optimization landscape. We have therefore removed the expression "highly fragile local minima" and revised the corresponding description to focus on input sensitivity and robustness.

The revised manuscript text reads as follows:

"In addition, random noise is injected into the training data to improve the model's robustness to input perturbations. By exposing the neural network to small variations in atmospheric states during training, this approach acts as a form of regularization that can reduce the sensitivity of network predictions to input perturbations. This may help mitigate large prediction errors and improve the numerical stability of the coupled hybrid model."

Line 340. Figure 5. Would it be better to add an introduction of what the curves and bars represent in the caption of this figure? It appears that the curves in each panel above y = 0 show anomaly correlation, but what do the bars in the lower portion of each panel show?

Response: Thank you for this helpful suggestion. We agree that the original caption of Figure 5 lacked explicit descriptions for different graphical elements. The upper black and blue curves represent the anomaly correlation coefficients (ACCs) of the operational CMA-GFS model and the hybrid model, respectively. The lower blue curve denotes the ACC difference calculated as hybrid model ACC minus operational model ACC. The blue shaded bars mark the critical bounds for statistical significance at the 5 % level. We have revised the figure caption in the revised manuscript as follows:
"**Figure 5:** Comparison of 500-hPa geopotential height anomaly correlation coefficients (ACCs) between the hybrid model and the operational CMA-GFS model for the Northern Hemisphere, Southern Hemisphere, and East Asia domains. Results are shown for January (top row) and July (bottom row) 2024. The black and blue solid curves represent ACCs of the operational and hybrid models, respectively. The horizontal dashed line marks the ACC threshold of 0.6. The lower blue curve shows the ACC difference (hybrid minus operational). The blue shaded bands indicate the critical bounds for statistical significance at the 5 % level. The x-axis corresponds to forecast lead time in hours."

Line 360. Figure 6. Are the threat scores shown in the figure global or regional results? According to what shown in Figure 8, it seems the scores are computed based on the precipitation forecasts over China? I am not a weather forecaster. This may be commonly known in the weather forecast community, but I am curious why boreal summer (July) precipitation forecasts have higher threat scores than boreal winter (January) precipitation forecasts, particularly for heavy and torrential rain forecasts?

Response: We appreciate this thoughtful question. The threat scores (TSs) shown in Figure 6 are calculated over land areas of China rather than globally. We have explicitly clarified the verification region in the revised figure caption as follows:
"**Figure 6:** Threat scores (TSs) of precipitation forecasts from the hybrid model and the operational CMA-GFS model for different rainfall thresholds, calculated over land areas of China. Results are shown for January 2024 (upper panels) and July 2024 (lower panels). The x-axis denotes forecast lead time in hours."

Regarding the seasonal differences, precipitation over China exhibits a pronounced annual cycle strongly influenced by the East Asian summer monsoon. In July, heavy and torrential rainfall events occur much more frequently, whereas such intense precipitation is relatively rare in January. The higher TS values in July, particularly for heavy and torrential rainfall, may therefore be partly associated with the greater occurrence of these events during the summer rainy season.

In addition, TS is sensitive to event frequency, as it is calculated from hits, misses, and false alarms without considering correct negatives. The relatively small number of heavy precipitation events in winter may contribute to greater sampling variability and make the TS more sensitive to individual misses and false alarms. Nevertheless, event frequency alone does not determine TS, and the seasonal differences also reflect variations in forecast performance.

We have also added a brief explanation of these seasonal differences in the revised manuscript:

"The generally higher TS values in July, particularly for heavy and torrential rainfall, may be partly associated with the greater frequency of intense precipitation events during the East Asian summer monsoon season. In contrast, heavy precipitation events are relatively rare in January, which may contribute to greater variability in the corresponding TS values."

Lines 360-365. Figure 7. Do the biases shown in this figure have units of mm? It appears the thick horizontal red line in each panel marks y = 1 (mm?). Would it be better to adjust the y limits of the relevant panels to avoid the overlapping of the red lines with x labels or tick-labels?

Response: We thank the reviewer for this careful comment. The precipitation bias shown in Figure 7 is a frequency bias, defined as the ratio of the number of forecast precipitation events to the number of observed precipitation events. It is therefore dimensionless rather than expressed in mm. The thick horizontal red line indicates the unbiased reference value of 1, corresponding to equal frequencies of forecast and observed precipitation events.

We acknowledge that the red reference line overlaps with the x-axis tick labels in several panels. In the revised manuscript, we have adjusted the y-axis limits of the relevant panels to eliminate this overlap. We have also revised the caption of Figure 7 to clarify the definition of the bias metric and the meaning of the red reference line, as follows:

"**Figure 7:** Frequency bias (dimensionless) of precipitation forecasts from the hybrid model and the operational CMA-GFS model for different rainfall thresholds, calculated over land areas of China. Results are shown for January 2024 (upper panels) and July 2024 (lower panels). The thick horizontal red line indicates the unbiased reference value of 1. The x-axis denotes forecast lead time in hours."

Lines 353-354. I assume the authors must have verified the statement about the forecast of precipitation extremes using some metric not presented here. It seems the authors used the purple curves to show the major rivers and lakes in China in the upper panels of Figures 8 and 9. However, according to the colormaps of the two figures, purple also means the regions with accumulated precipitation > 100 mm, making it hard to see that the hybrid model shows a slightly increased consistency with the observation.

Response: Thank you for your valuable comments. We agree that the original statement regarding the forecasting of precipitation extremes was not sufficiently supported by the presented results, and that the purple lines representing rivers and lakes in Figures 8 and 9 could be confused with the colours indicating extreme precipitation.

First, we would like to clarify that the overall precipitation forecast performance is quantitatively assessed by the Threat Score (TS) and multiplicative bias shown in Figures 6 and 7. Those metrics show that the hybrid model outperforms the operational model for certain precipitation thresholds and forecast lead times, yet such improvement is not uniform across all conditions. By contrast, Figures 8 and 9 serve only as qualitative case study illustrations of spatial precipitation patterns and are not used for quantitative performance evaluation.

We acknowledge that our original wording may have overstated the improvement in precipitation forecasting. We have therefore revised the corresponding description to highlight the general similarity between the two models, while describing case-to-case differences in precipitation spatial distribution and intensity, without claiming systematic improvement for extreme precipitation forecasts.

Second, we agree that the original colour scheme in Figures 8 and 9 was potentially misleading. We have revised both figures by changing the colours of rivers and lakes to neutral gray, making them clearly distinguishable from the precipitation colour scale. The original precipitation colour scale is kept unchanged to maintain consistency between observations and model forecasts.

The corresponding revised manuscript text reads as follows:
"For daily accumulated precipitation below 10 mm, the hybrid model exhibits broadly similar spatial patterns to those of the operational model. For precipitation above 10 mm, the two models exhibit differences in the spatial distribution and intensity of precipitation, as illustrated in Figures 8 and 9. Although the hybrid model shows stronger precipitation signals in some regions, these selected case examples do not provide sufficient evidence for a systematic improvement in extreme precipitation forecasts."

Lines 390-395. Figures 10 and 11. What do the two contractions "MAE" and "ME" in the two figures stand for? Are the results shown in the two figures also the forecasts over China?

Response: Thanks for this comment. MAE stands for mean absolute error, which measures the average magnitude of forecast errors, while ME stands for mean error, which indicates systematic forecast bias. The verification results shown in Figures 10 and 11 are calculated over land areas of China. We have clarified these definitions and the verification domain in the revised figure captions, as follows:
"**Figure 10:** Mean absolute error (MAE) and mean error (ME) of 2-meter temperature forecasts from the hybrid model and the operational CMA-GFS model, calculated over land areas of China. Results are shown for January 2024 (upper panels) and July 2024 (lower panels). The x-axis denotes forecast lead time in hours."

"**Figure 11:** Mean absolute error (MAE) and mean error (ME) of 10-meter wind forecasts from the hybrid model and the operational CMA-GFS model, calculated over land areas of China. Results are shown for January 2024 (upper panels) and July 2024 (lower panels). The x-axis denotes forecast lead time in hours."

Lines 406-408. I assume the authors must have verified the statement about the forecast of the hurricane precipitation using some metric not presented here. By eyeballing Figure 14a, it looks that the hybrid model succeeded in forecasting the outer rain belt of the hurricane while the operational model did not?

Response: We thank the reviewer for this careful observation regarding the precipitation forecasts for Typhoon Kaemi in Figure 14a.

The apparent difference in the outer rainband between the two models was caused by an inconsistency in the plotting domains. Specifically, the operational-model panel was inadvertently plotted over a domain covering only mainland China, which excluded part of the offshore outer rainband, whereas the hybrid model panel covered a larger domain. This inconsistency created the impression that the hybrid model captured the outer rainband more successfully than the operational model.

We have corrected this issue in the revised Figure 14a by using identical plotting domains for both models. The revised figure provides a more consistent comparison and shows that the two models exhibit broadly comparable precipitation distributions. We would also like to clarify that the assessment of Typhoon Kaemi precipitation in this section is based on a qualitative comparison of the spatial distributions, rather than additional quantitative verification metrics.

Accordingly, the original sentence "the hybrid model showed no significant improvement in precipitation forecasting, though its predicted precipitation intensity was slightly stronger" has been revised as follows:
"The hybrid model and the operational model exhibited broadly comparable precipitation distributions, while the hybrid model predicted slightly stronger precipitation in some regions."

Lines 415-416 and 430-431. The statements are vague. What and how can satellite observation data be used to improve the performance of the hybrid model in forecasting synoptic fields and 10-meter wind speed? How can LBLRTM be used to achieve this goal? Depending upon how fine the selected wavenumber interval is, LBLRTM could be several orders of magnitude more computational expensive than RRTMG.

Response: We thank the reviewer for this important comment. We agree that our original statements were overly general and did not adequately explain how satellite observations and LBLRTM could contribute to the further development of the hybrid model.

Satellite radiance observations could potentially provide additional constraints for evaluating radiation-related processes through suitable observation operators. Such

observations may help identify biases in simulated radiative quantities, although their use would require careful consideration of observational uncertainties and differences between observed radiances and model-calculated radiative fluxes. Importantly, improvements in radiative transfer calculations do not necessarily lead to better forecasts of synoptic fields or near-surface wind speed. Any potential benefits to forecast performance would need to be evaluated through dedicated coupled forecast experiments.

Regarding LBLRTM, we agree that its computational cost is substantially higher than that of RRTMG, making direct integration into operational global forecast simulations impractical. Our intention was to consider LBLRTM as an offline high-accuracy reference model for assessing radiative transfer calculations and potentially generating supplementary training data for future ML emulators. Since LBLRTM would be used only for offline calculations rather than during operational forecasts, its high computational cost would not directly affect the runtime efficiency of the hybrid model.

To avoid implying that these approaches would necessarily improve forecast accuracy, we have revised the relevant manuscript text to clarify their potential roles and acknowledge the need for further investigation. First, the paragraph beginning with "Overall, the current hybrid model achieves forecast performance …" has been revised as follows:

"Overall, the current hybrid model achieves forecast performance comparable to that of the operational model. However, slightly degraded performance is observed for some synoptic fields and 10-meter wind speed. The causes of these differences require further investigation, and additional experiments are needed to determine whether refinements to the ML-based radiation scheme could improve forecast performance."

Second, the final part of the "Discussion and future work" section has been revised as follows:

"Future work will focus on further optimizing and refining the neural network model. In particular, high-accuracy radiative transfer calculations from LBLRTM could be used offline to assess the accuracy of radiative transfer calculations and potentially provide supplementary training data for future ML emulators. Satellite radiance observations may also provide additional constraints for evaluating radiation-related processes through suitable observation operators. The effectiveness of these approaches and their potential impacts on forecast performance remain to be investigated through dedicated numerical experiments."

Lines 416-418. Yes, you did a great job. I have a suggestion that I do not expect the authors to address in this study. This study already has abundant results. Now, the authors have successfully speeded up the radiation computation in the weather forecast model by a factor of 7. What if the authors increase the frequency of calling the radiation scheme in running the hybrid model? Will weather forecast improve if the radiation scheme is called more frequently? Experiments for answering the questions may be a future study.

Response: We sincerely thank the reviewer for the encouraging assessment of our work and for this valuable suggestion. We agree that the improved computational efficiency of the radiation emulator offers an opportunity to investigate more frequent radiation calculations. More frequent updates may improve the representation of rapidly evolving cloud–radiation interactions and the diurnal cycle, although the actual impact on forecast accuracy remains to be evaluated. We consider this an interesting direction for future research and will explore the impact of radiation scheme calling frequency on forecast performance in our future work.

Minor Suggestions

Lines 62-64. The SPeedy Algorithm for Radiative TrAnsfer through CloUd Sides (SPARTACUS) solver in ecRad is a 3D solver that incorporates 3D cloud radiative effects on grid-box-mean radiative fluxes. The statement here as written reads like the ecRad had not have a 3D radiative transfer computational capability until Alves et al. (2023) added one. It would be better to make it clearer.

Response: Thank you for pointing out this misleading phrasing. We agree that the original wording could give the incorrect impression that three-dimensional cloud radiative effects were first incorporated into ecRad by Alves et al. (2023). In fact, ecRad already included the SPARTACUS solver for representing three-dimensional cloud radiative effects before that study. We have revised the sentence to clarify this distinction. The revised sentence reads:
"Alves et al. (2023) investigated ML approaches related to three-dimensional cloud radiative effects within the ecRad framework (Hogan and Bozzo, 2016), which already included the SPARTACUS solver for accounting for these effects on grid-box-mean radiative fluxes."

Line 67. What do you mean by "non-cooling effect of cloud reflection"? Cloud sunlight reflection results in a heating of the cloud top, i.e., a heating effect locally. For example, see the offline radiation simulation study of Ren et al. (2021).

Response: Thank you for this insightful comment and for recommending the study by Ren et al. (2021). We agree that our original wording was ambiguous and could be interpreted as implying that cloud reflection directly produces local radiative cooling. Our intention was to describe the non-local influence of clouds on shortwave heating rates, as demonstrated by Hafner et al. (2025). Their interpretability analysis showed that clouds can reduce shortwave heating in atmospheric layers below the cloud through the reflection and attenuation of incoming solar radiation. This does not contradict the local shortwave heating within clouds, particularly in their upper portions, as discussed by Ren et al. (2021).

To avoid confusion, we have revised the sentence to explicitly distinguish the non-local reduction of shortwave heating below clouds from local cloud radiative heating:

"Hafner et al. (2025) developed a Bidirectional Long Short-Term Memory (BiLSTM) emulator for the ICON model and employed Shapley Additive exPlanations (SHAP) to interpret the learned relationships. Their analysis demonstrated that the BiLSTM captured physically meaningful radiative processes, including the non-local reduction of shortwave heating below clouds due to the reflection of incoming solar radiation and the local temperature dependence of longwave emission. This non-local effect is distinct from the local shortwave radiative heating within clouds, as also discussed by Ren et al. (2021). "

We have also added the reference of Ren et al. (2021) to the reference list of the revised manuscript.

Ren, T., Li, D., Muller, J., and Yang, P.: Sensitivity of radiative flux simulations to ice cloud parameterization over the equatorial western Pacific Ocean region, Journal of the Atmospheric Sciences, 78(8), 2549-2571, 2021.

Line 100. Clouds and aerosols in the atmosphere are particles, not gases, but how electromagnetic waves interact with them is important and handled in the atmospheric radiative transfer computation. Hence, would it be more accurate to say "an absorption and scattering medium" rather than "a gas medium"?

Response: Thank you for this suggestion. We agree that atmospheric radiative transfer involves absorption and scattering by gases, clouds, and aerosols. We have replaced "a gas medium" with "an absorbing and scattering medium" to provide a more accurate description. The original sentence is replaced with:

"Radiative transfer refers to the transport of electromagnetic radiation through an absorbing and scattering medium, including atmospheric gases, clouds, and aerosols."

Lines 115 and 120-121. Can the authors be more specific? What is "subgrid-scale cloud variability" is considered in RRTMG but not in RRTM? Correct me if I am wrong, the impact of cloud mass horizontal variability in a cloud-containing model layer on grid-box-mean radiative fluxes is not handled in either RRTM or RRTMG. In other words, cloud is assumed to be homogenous in each cloud-containing model layer in both RRTM and RRTMG. In addition, correct me if I am wrong, random and maximum-random cloud overlapping treatments are available in both RRTM and RRTMG?

Response: We thank the reviewer for this careful comment. We agree that our original description of subgrid-scale cloud variability was overly general and could be misinterpreted.

The subgrid-scale cloud variability referred to in our manuscript primarily concerns fractional cloud cover and vertical cloud overlap, rather than the horizontal variability of cloud condensate within cloudy regions. We agree that the standard

implementations of both RRTM and RRTMG assume horizontally homogeneous cloud optical properties within the cloudy fraction of each model layer, unless extra subgrid variability is explicitly prescribed.

Regarding cloud overlap, the reviewer is correct that RRTM_LW already supports random and maximum-random overlap assumptions. The key distinction is that RRTMG is designed to work with the Monte Carlo Independent Column Approximation (McICA), which enables efficient representation of fractional cloud cover and cloud overlap through stochastic subcolumn sampling, rather than introducing cloud overlap assumptions for the first time.

We have revised the relevant text to clarify these distinctions and avoid overstating the differences between RRTM and RRTMG. The sentence "While RRTMG shares the same fundamental physics and absorption coefficients as RRTM, it incorporates several modifications to improve computational efficiency, and to represent subgrid-scale cloud variability." is replaced with "While RRTMG shares the same fundamental physics and absorption coefficients as RRTM, it incorporates several modifications to improve computational efficiency and facilitate the treatment of fractional cloud cover and cloud overlap."

In addition, the incorrect original statement "Subgrid-scale cloud variability is not considered in RRTM" has been removed entirely from the manuscript.

Lines 121-122. The McICA method is developed for computational efficiency. What aspects of cloud-radiation interactions does this method improve?

Response: We thank the reviewer for this question. We agree that the primary purpose of McICA is to improve computational efficiency rather than to introduce new cloud physics. Its main advantage is that it decouples the specification of subgrid cloud structure from the radiative transfer calculation, enabling different cloud overlap assumptions and, in principle, horizontal variability of cloud condensate to be represented within a unified framework.

For a prescribed subgrid cloud distribution and consistent radiative transfer assumptions, McICA provides a stochastic estimate of the corresponding ICA radiative fluxes that is unbiased in expectation, although individual realizations are subject to sampling noise (Pincus et al., 2003). Its primary advantage is computational efficiency rather than an intrinsic improvement in cloud radiative physics.

RRTMG is commonly coupled with McICA to efficiently represent fractional cloud cover and vertical cloud overlap via stochastic subcolumn sampling (Pincus et al., 2003; Iacono et al., 2008). As noted in our previous response, RRTM_LW already supports random and maximum-random cloud overlap but does not use McICA. The standard RRTM_SW, by contrast, is restricted to clear sky or fully overcast conditions. This decoupling of subgrid cloud sampling and radiative transfer also offers flexibility to accommodate more complex cloud distributions, including condensate inhomogeneity when the relevant information is available.

We have revised the corresponding manuscript text accordingly, as follows:
"RRTMG is commonly coupled with the Monte Carlo Independent Column

Approximation (McICA), which efficiently represents fractional cloud cover and vertical cloud overlap through stochastic subcolumn sampling (Pincus et al., 2003; Iacono et al., 2008). Although RRTM_LW already supports random and maximum-random cloud overlap, it does not employ McICA, while the standard RRTM_SW is limited to clear sky or fully overcast conditions. By separating the specification of subgrid cloud structure from radiative transfer calculations, McICA also provides flexibility to incorporate more complex cloud distributions, including cloud condensate inhomogeneity when such information is available. "

We have also added Pincus et al. (2003) to the reference list and removed Barker et al. (2008) in the revised manuscript.

Pincus, R., Barker, H. W., and Morcrette, J.-J.: A fast, flexible, approximate technique for computing radiative transfer in inhomogeneous cloud fields. Journal of Geophysical Research: Atmospheres, 108(D13), 4376, 2003.

Lines 123-124. Can the authors provide references for this statement? The largest longwave wavelength band of RRTM/RRTMG has a wavenumber range of 10–350 cm-1 or a wavelength range of 28.57–1000 μm. What and how have measured sub-millimeter radiance spectra been used to validate RRTMG?

Response: We thank the reviewer for identifying this misleading statement. We agree that our original text incorrectly attributed spectral-range validation originally performed for LBLRTM to RRTMG.

The validation against measured atmospheric radiance spectra spanning the sub-millimeter to ultraviolet wavelengths applies to LBLRTM rather than RRTMG (Clough et al., 2005). As a broadband radiative transfer scheme, RRTMG is mainly validated by comparing its radiative fluxes and heating-rate outputs against reference calculations from LBLRTM and RRTM, instead of direct comparisons against observed sub-millimeter radiance spectra (Iacono et al., 2008).

Given that the primary purpose of this paragraph is to illustrate the broad practical usage of RRTMG within numerical weather prediction and climate models, we have removed the problematic sentence about spectral validation and revised the opening of this paragraph accordingly. The revised paragraph starts with:

"RRTMG has been widely adopted in numerical weather prediction and climate models, including the IFS system at ECMWF, the Community Earth System Model and WRF regional forecast model at NCAR, and the CMA-GFS model at the China Meteorological Administration Earth System Modeling and Prediction Centre."

References

Clough, S. A., Shephard, M. W., Mlawer, E. J., Delamere, J. S., Iacono, M. J., Cady-Pereira, K., Boukabara, S., and Brown, P. D.: Atmospheric radiative transfer modeling: A summary of the AER codes, Journal of Quantitative Spectroscopy and Radiative Transfer, 91(2), 233-244, 2005.

Iacono, M. J., Delamere, J. S., Mlawer, E. J., Shephard, M. W., Clough, S. A., and Collins, W. D.: Radiative forcing by long lived greenhouse gases: calculations with

the AER radiative transfer models, Journal of Geophysical Research Atmospheres, vol. 113, no. 13, D13103, 2008.

Lines 165-170. Table 1. I can guess out the meanings of all the dimension name abbreviations except for this one naerec. Does it refer to the number of aerosol species? Would it be better to spell out the full names of all the abbreviations somewhere? The two variables reicmcl and relqmcl in the Inputs table each shows up two times. Would it be better to combine the two rows of the same variable?

Response: We appreciate the reviewer's careful reading and constructive comments. As the reviewer correctly inferred, the abbreviation naerec stands for the number of aerosol species (aerosol components for shortwave calculations). Following this suggestion, we have integrated explanations for all dimension-related abbreviations from Table 1 into the main manuscript text. The inserted text reads:
"In the dimension labels used in Table 1, ncol denotes the number of horizontal atmospheric columns per computational process; note that radiation calculations and model training operate on individual atmospheric columns. nlay is the total number of vertical model layers, ngptlw denotes the total number of g-points for longwave radiation, ngptsw corresponds to the total number of reduced g-intervals for shortwave radiation, nbndlw refers to the number of longwave spectral bands, and naerec indicates the number of aerosol species considered for shortwave calculation."
We apologize for the duplicated rows of reicmcl and relqmcl in the table, which stemmed from a copy-paste error during table preparation. In the revised Table 1, these redundant entries have been removed, with only one row retained for each variable.

Line 175. It is hard to understand "full model layers" and "half layers". Would it be clearer if 88 model layers and 89 model levels were used instead?

Response: Thank you for the suggestion. We agree that using "88 model layers and 89 model levels" delivers a clearer and more precise description of the vertical discretization. We have revised the sentence to read:
"Since some input variables are defined on the 88 model layers and others are defined on the 89 model levels, the corresponding data were padded to obtain a consistent vertical dimension of 89."

Line 176. "RseCNN" or "ResCNN"?

Response: Thank you for identifying this typographical error. We have corrected "RseCNN" to "ResCNN" in the revised manuscript.

Line 194. The phrase "machine learning" appears in the main text for the first time in Line 21. Would it be better if the contraction "ML" were introduced at the very beginning?

Response: Thank you for this suggestion. We have introduced the abbreviation "ML" at the first occurrence of "machine learning" in the main text and use it consistently thereafter.

Line 195. "panel compares" or "panels compare"? In the left panel of each of the two figures, I find only 3 variables, not 6 variables.

Response: Thank you for pointing out this inconsistency. Each radiation component includes six predicted quantities: total sky and clear sky radiative heating rates, as well as upward and downward fluxes. We have revised the text to clarify that these six quantities are displayed across the subpanels rather than in a single panel. The original sentence is replaced with:
"As shown in Figures 2 and 3, the left-hand subpanels compare the neural network predictions with the corresponding RRTMG reference values for six radiative quantities, while the right-hand subpanels present the mean absolute error (MAE) of these six variables and their biases across different percentiles."

Line 313. "becomes" or "become"?

Response: We appreciate this careful grammatical check. We have revised "becomes" to "become" in the revised manuscript.

Line 320. Would it be better to use "in" or "of" rather than "at" right be "layer i"? The atmospheric heating rate of a layer is computed with the variables at its upper and lower levels (or interfaces) next to the adjacent layers.

Response: We appreciate this valuable suggestion. We have adopted "in layer i" in the revised manuscript.

Line 373. There is an unwanted backtick symbol "`" before the word "model's".

Response: Thank you for spotting this typographical artifact. It is an unintended formatting error introduced during manuscript typesetting. We have removed the redundant backtick symbol before "model's" in the revised manuscript.

Line 402. "on the afternoon" or "in the afternoon"?

Response: We appreciate this careful proof-reading. We have corrected "on the afternoon" to "in the afternoon" in the revised manuscript.