The paper in general is clearly written and well structured. The newly implemented "concurrent radiation scheme" is adequately motivated and precisely explained, and its performance properties are comprehensively discussed. The new scheme is verified on the basis of a few specific, but realistic example ESM calculations.

While the "concurrent radiation scheme" has the potential to overcome a principle bottleneck in the overall ECHAM performance concerning scalability and parallel efficiency (this is convincingly described and demonstrated in the manuscript) the new concept introduces a fundamental load-balancing issue. The latter can be cured by tuning the fraction of the number of processors (resp. MPI ranks) which are dedicated to the radiation calculations and/or tuning the length of the radiation timesteps. While such tuning can in principle be used to achieve almost perfect load balancing, there are implications concerning the numerical accuracy and possibly robustness, because both, the number of processors used for radiation, and the (relative) length of the timesteps for radiation and atmospheric physics are directly linked to the numerical resolution in space and time, respectively. Moreover, the optimal ratios (spatial resolution and time stepping) depend on the performance characteristics (hardware and software environment) of the compute system used. Hence, I’m wondering about the general applicability of the new "concurrent radiation scheme" to a broad range of simulation scenarios. What is the overall strategy to achieve load balancing? Will adapting the spatial and temporal resolution of the radiation calculations always be feasible and sufficient to load-balance with the atmospheric physics? What will be the effect of the suggested "adopting a coarser domain decomposition" (cf. lines 255-258) on the accuracy of the models? Related to this, I'm missing a more systematic assessment in section 5 which currently is somewhat qualitative (e.g. "This implies that a further tuning for the concurrent radiation scheme may be needed", lines 321-322: what does this actually mean? how specific is the "tuning" for the specific model setup? Will it generalize to other ESM model setups and compute systems?). The new scheme also introduces a change in the operator splitting technique, the potential effects of which are not discussed nor systematically assessed in the paper: In the classical radiation scheme, the radiation timestep takes the most up-to-date quantities of the atmospheric physics as input, whereas in the concurrent radiation scheme the input quantities systematically lag behind by one ATM timestep (cf. Figs 2 and 6)

Fully addressing my concerns might be out of the scope of this paper. Some of the open
questions may well be left for future research, but in the present paper they should at least be clearly stated (e.g. in the conclusions/outlook) and possible weaknesses or todos should be identified and discussed.

Minor points:

- line 143; "receives feedback ... upon the request" what does this mean?
- line 178: specify type of InfiniBand (EDR, HDR, or alike)
- line 205: "... is adopted to ..." what does this mean?
- typo in line 408: scalability
- Figs 2, 6, 13-16: some of the labelling is hard to read, in particular red font on blue background
- It was not possible to anonymously (i.e. without registration of an account at DKRZ) download the modified ECHAM6 source code from https://doi.org/10.35089/WDCC/SC_PalMod_ECHAM6 for inspection and for an assessment of reproducibility aspects, but I'd consider this only a minor point, given that ECHAM is such a well-established code in the ESM community.