Authors' reply to reviewer Hartwig Deneke of the paper: "Towards an improved treatment of cloud-radiation interaction in weather and climate models: exploring the potential of the Tripleclouds method for various cloud types using libRadtran 2.0.4", submitted for publication in GMD by Nina Črnivec and Bernhard Mayer.

We thank the reviewer Hartwig Deneke for good assessment of our work and valuable comments, which helped improving the quality of our original manuscript. Below please find reviewer's comments in blue and our reply in black. Changes in the revised paper are marked with quotation marks and additional indent.

The paper studies the ability of the TripleClouds method to reduce biases in radiative fluxes due to unresolved inhomogeneities in coarse-resolution models. Overall, I find this study very interesting, with sound methodology and a good presentation. The only critical comment I'd like to raise that in some parts, the language could be improved, in particular some long sentences are too long/hard to follow. As such, I recommend the paper for publication after some minor technical corrections.

* As one of several very long sentences, I refer to an example sentence from the conclusions , which addresses several independent points simultaneously (physical understanding/cloud type dependence/parameter tuning/operational use). I strongly recommend to shorten such sentences, and address points eparately: "The acquired physical understanding of radiative biases, in particular those stemming from neglected cloud horizontal heterogeneity, for three fundamentally contrasting cloud case studies as highlighted in this work, is a necessary first step for properly setting the TC parameters in its possible future operational usage."

Thank you for this remark, we agree that several long sentences should be shortened.

Sentence:
"The acquired physical understanding of radiative errors, in particular those stemming from neglected cloud horizontal heterogeneity, for three fundamentally contrasting cloud case studies as highlighted in this work, is a necessary first step for properly setting the TC parameters in its possible future operational usage."
was reformulated as follows:

> "This work provides the physical understanding of radiative errors, in particular those stemming from neglected cloud horizontal heterogeneity, for three fundamentally contrasting cloud cases. This is a necessary first step for properly setting the TC parameters in its possible future operational usage."

We have additionally shortened the following sentences:

Sentence:
"Although the representation of unresolved clouds in radiation schemes of coarse-resolution weather and climate models has progressed noticeably over the past years, a lot of room remains for improvement, as the current picture is by no means complete."
was reformulated as follows:

> "The representation of unresolved clouds in radiation schemes of coarse-resolution weather and climate models has progressed noticeably over the past years. Nevertheless, a lot of room remains for improvement, as the current picture is by no means complete."

Sentence:
"These case studies were deliberately chosen in a way, that cloud vertical arrangement tends towards the assumed maximally-overlapped scenario, thus focusing on radiative effects associated with cloud horizontal inhomogeneity and eliminating the error arising from the misrepresentation of assumed vertical overlap as would be expected to occur in conditions with strong vertical wind shear (Naud et al., 2008; DiGiuseppe and Tompkins, 2015)."

was reformulated as follows:

> “These case studies were deliberately chosen in a way, that cloud vertical arrangement tends towards the assumed maximally-overlapped scenario. This enables us to focus on radiative effects associated with cloud horizontal inhomogeneity, while eliminating the error arising from the misrepresentation of assumed vertical overlap as would be expected to occur in conditions with strong vertical wind shear (Naud et al., 2008; DiGiuseppe and Tompkins, 2015). ”

Sentence:
“To obtain the pair (LWC^{cn}, LWC^{ck}) from $\overline{\mathrm{LWC}}$, which is indeed available in a GCM, we introduce the so-called $\overline{\mathrm{LWC}}$-scaling factors for the optically thin and thick cloudy region, termed s^{cn} and s^{ck} respectively, fulfilling the following relationships:”
was reformulated as follows:

> “To obtain the pair (LWC^{cn}, LWC^{ck}) from $\overline{\mathrm{LWC}}$, which is indeed available in a GCM, we introduce the so-called $\overline{\mathrm{LWC}}$-scaling factors for the optically thin and thick cloudy region. These are termed s^{cn} and s^{ck} and fulfill the following relationships:”

Sentence:
“Based on a comprehensive review of numerous observational studies encompassing diverse cloud data sets, Shonk et al. (2010) converted various variability measures into a single globally applicable FSD parameter, whose mean value and uncertainty are:”
was reformulated as follows:

> “Based on a comprehensive review of numerous observational studies encompassing diverse cloud data sets, Shonk et al. (2010) converted various variability measures into a single globally applicable FSD parameter. Its mean value and uncertainty are:”

Sentence:
“Thus persistent cloud-top radiative cooling, a typical feature of marine stratocumulus-topped boundary layers (STBLs; Wood, 2012), drives convective instability and controls turbulence within the underlying mixed layer (Randall, 1980; Deardorff, 1981; Stevens et al., 1999), when adequately coupled to a dynamical model.”
was reformulated as follows:

> “This persistent cloud-top radiative cooling is a typical feature of marine stratocumulus-topped boundary layers (STBLs; Wood, 2012). It drives convective instability and controls turbulence within the underlying mixed layer (Randall, 1980; Deardorff, 1981; Stevens et al., 1999), when adequately coupled to a dynamical model.”

Sentence:
“Thus the GCM boosts radiatively driven destabilization of the stratocumulus layer during daytime and nighttime, by overestimating cooling at the uppermost region of the layer (by -14 K day^{-1}) and overestimating warming in the region underneath (error up to 9 K day^{-1}).”
was reformulated as follows:

> “Thus the GCM boosts radiatively driven destabilization of the stratocumulus layer during daytime and nighttime. It overestimates cooling at the uppermost region of the layer by -14 K day^{-1} and it overestimates warming in the region underneath by up to 9 K day^{-1}.”

Sentence:
“It was found that in the majority of applications, the ICA is significantly more accurate than the GCM experiment, indicating a large potential for Tripleclouds, which reduces the error related to unresolved cloud structure, but not to horizontal photon transport.”
was reformulated as follows:

> “It was found that in the majority of applications, the ICA is significantly more accurate than the GCM experiment. This indicates a large potential for Tripleclouds, which reduces the error related to unresolved cloud structure, but not to horizontal photon transport.”

Please note also that because the paper was thoroughly reconstructed (as suggested by the second reviewer), some other sentences were reformulated and shortened as well (see marked changes within the pdf file).

Some additional comments on the text:
* L90: “which has received considerably less attention in the previous debates”: less than what?

Thank you for pointing this out. We changed the sentence to:

> “Furthermore, we aim to consistently analyze the atmospheric heating rate and net surface flux, which has received little attention in the previous debates.”

* L 99: “...which requires an upgrade of vertical overlap rules” … “the maximum-random overlap was thus retained” => these phrases seem contradictory, maybe this can be written differently/more clearly.

We changed the word “upgrade” to “extension” to make the vocabulary consistent with that in a later sentence. This hopefully clarifies the paragraph (crucial considerations are marked bold):

> “… which requires **an extension of vertical overlap rules**”

“the maximum-random overlap was thus **retained** for the entire fractional cloudiness and **additonally applied** for the optically thicker segment. **This extended vertical overlap formulation**...”

* L138: “Diverse models employed to generate these cloud fields … ensure that the three selected cases comprise a wide range of inhomogeneity.” => This could be interpreted that the differences in models is mainly responsible for inhomogeneity. Isn’t it the main goal to reproduce inhomogeneity observed in nature? All in all, I think model and resolution-induced differences are something which cannot be avoided, rather than something desirable.

Thank you for this comment – you have certainly raised an important remark. We have modified the paragraph accordingly:

> “Input data for radiative transfer experiments is a set of 3-D highly-resolved inhomogeneous cloud fields, defined in terms of LWC and IWC distributions. These differing cloud cases comprise a wide range of inhomogeneity observed in nature. In the following, each cloud type is characterized briefly.”

* L262: “The difference between the ICA and 3-D was”: output? results? Word missing!

We changed the sentence to:

> “The difference between the ICA and 3-D results...”

L528: “These findings are in support of cloud regime dependent approaches, which ought to be further boosted to be used in radiation schemes of next-generation atmospheric models.” Is “boosted” the right word here? At least to me, the meaning of this sentence is somewhat unclear (caveat: I am not a native English speaker)

Thank you for this remark. The word “enhanced” would be a better choice than “boosted”. Nevertheless, we have additionally simplified the sentence to:

> “These findings are in support of cloud regime dependent approaches, which ought to be used in radiation schemes of next-generation atmospheric models.

Authors' reply to reviewer Robin Hogan of the paper: "Towards an improved treatment of cloud-radiation interaction in weather and climate models: exploring the potential of the Tripleclouds method for various cloud types using libRadtran 2.0.4", submitted for publication in GMD by Nina Črnivec and Bernhard Mayer.

We thank the reviewer Robin Hogan for good assessment of our work and many valuable comments and suggestions, which helped improving the quality of our original manuscript. Below please find reviewer's comments in blue and our reply in black. Changes in the revised paper are marked with quotation marks and additional indent. The line numbers refer to those in the original manuscript.

This paper presents an interesting evaluation of the Tripleclouds (TC) method for representing cloud structure in a radiation scheme using benchmark Monte Carlo calculations and three contrasting, realistic 3D cloud scenes. While the McICA method is a more common way to treat cloud structure in operational models, Tripleclouds is attractive because it is free from stochastic noise. The paper goes into some depth in testing and optimizing the various ways in which Tripleclouds can be configured to reduce the errors. However, I have some major concerns that should be addressed before this paper can be published. The paper is generally well written otherwise.

MAJOR COMMENTS

1. The use of maximum-random overlap is well behind the state-of-the-art and will have led to appreciable biases in your GCM and TC schemes. The cirrus case is the most obvious: Hogan and Kew (2005, Table 2) found that for this exact case, a GCM-type radiation scheme with maximum-random overlap underestimated the TOA cloud radiative forcing by 42% in the longwave and 48% in the shortwave, compared to the same GCM-type scheme but with the "true" overlap. I expect significant overlap biases to be present in the other two scenes, and indeed I can only explain your TC flux biases with respect to ICA as being due to overlap being too maximal. To test this, simply compute the total cloud cover predicted from the cloud-fraction profile and your maximum-random overlap assumption, and compare it to the actual total cloud cover of the scene. Note that even the overcast stratocumulus case can have an overlap bias because of the overlapping of sub-grid cloud structures, which ought to be represented by non-maximum overlap of the high-LWC and low-LWC cloudy regions (some schemes use a vertical decorrelation scale for cloud heterogeneities of half that used for cloud boundaries, i.e. around 1 km). I don't believe it would be too complicated to incorporate realistic overlap: this can be done using overlap matrices introduced by Shonk et al. (2010), and an example implementation is in the open-source ecRad package - see the radiation_overlap.F90 source file on GitHub. The capability has been in older TC implementations for longer - for example, Shonk and Hogan (2010) evaluated the impact of horizontal heterogeneity and vertical overlap using an implementation of Tripleclouds incorporating exponential-random overlap in the Met Office radiation scheme, and stressed the importance of making both horizontal heterogeneity and vertical overlap as realistic as possible.

We agree that the maximum-random overlap is not the state-of-the-art approach, thank you for pointing this out. We have thus removed the term "state-of-the-art" when describing our Tripleclouds radiative solver (line 5, caption of Fig. 1, line 91). Unfortunately, we did not manage to incorporate more realistic overlap within the scope of this study, but we plan to generalize the overlap rules in the future. We have extended the final section "Summary and conclusions" as follows:

> "Furthermore, in order to properly consider clouds in sheared environments, the vertical overlap rules in the present TC implementation have to be generalized."

In order to eliminate the effects of vertical wind shear in the present study, however, we have intentionally selected cloud cases, where the realistic overlap should be close to the assumed maximum-random overlap.

Firstly, we would like to emphasize that we did not use the exact same cirrus as Hogan and Kew (2005), although it is true that both cloud fields relate to the same case study (June, 24th, 1999). In particular, Hogan and Kew (2005) used coarser horizontal grid spacing of 1.56 km. We adopted the cirrus data presented in doctoral dissertation of Sophia Schäfer, where the simulation of Hogan and Kew (2005) was rerun with a higher resolution (50 m horizontal grid spacing) to enable proficient geometry analysis. Finally, we have additionally smeared the latter case onto grid with 100 m horizontal grid spacing to facilitate the radiation simulations. We have additionally clarified these issues in Section 2.1.2 (footnote):

> “It should be noted, however, that the studies of Hogan and Kew (2005) and Zhong et al. (2008) use coarser horizontal grid spacing (1.56 km). We adopted the cirrus data from Schäfer (2016), where the simulation of Hogan and Kew (2005) had been rerun with higher resolution (horizontal grid spacing of 50 m), whereby we eventually smeared the data onto the grid with horizontal grid spacing of 100 m to facilitate the Monte Carlo radiative transfer simulations.”

Furthermore, we probably also use different vertical resolution, which should also affect the difference in vertical overlap errors. We use constant vertical grid spacing of 109 m, whereas Hogan and Kew (2005) state: “This yields a horizontal resolution of 1.56 km and between 45 and 110 m in the vertical (depending on the case)...”(?). Finally, the extent of the calculation domains in Hogan and Kew (2005) and Schäfer (2016) differs as well. In summary, the two cirrus cases have different 3-D geometry, so that the overlap errors do not necessarily match.

In order to gain further insight into vertical overlap issues (as you advised), we have computed the actual total cloud cover of the 3-D scene and compared it to the maximum of the cloud-fraction vertical profile. The latter namely implies the total cloud cover corresponding to the maximum overlap assumption (if the cloud layer is vertically continuous without cloud-free regions, as is the case throughout our study, then the maximum-random overlap simplifies to the maximum overlap). For our cirrus case, for example, we found the following: There is indeed discrepancy between the actual total cloud cover of the 3-D scene (which is about 76 %) and maximum layer cloud fraction (which is about 47 %; see Fig. 3). However, it should be emphasized that in our maximum-random overlap implementation, we employ pairwise overlap (in this manner the matrix problem is faster to solve). This means that the total cloud cover predicted by our overlap scheme is generally larger than the maximal layer cloud fraction and is therefore closer to the actual total cloud cover! In other words, the pairwise overlap “luckily” relaxes the stiffness of the maximum assumption, acting to reduce the error due to true overlap not being maximal.

Regarding your comment: "..., and indeed I can only explain your TC flux biases with respect to ICA as being due to overlap being too maximal." - We don't see this so clearly (?). In general, the TC biases with respect to the ICA are due to several reasons:

- Overlap being too maximal in the TC compared to realistic overlap, which is taken into account in the ICA (but hopefully the error is small in our case);
- Inhomogeneity being misrepresented in the TC (either underestimated or overestimated) compared to the full realistic inhomogeneity, which is taken into account in the ICA;
- Effective (partial) treatment of 3-D radiative effects in the TC, which are entirely neglected in the ICA.

Please see also our answer to comment 18, which further elucidates these issues.

2. When comparing TC results to the benchmark, a number of errors are compounded and there is insufficient effort to separate them out for the reader: (1) the overlap parameterization is biased (see above); (2) the split percentile of 50% cannot adequately represent large FSDs; (3) a value of FSD=0.75 is used at various points which is different (often much lower) than the "truth"; and (4) TC is benchmarked against at 3D model when it makes no attempt to represent 3D effects. This leads to tuning of one parameterization to fix a problem caused by another (see comments 18 and 19). Wouldn't it be more satisfactory to evaluate TC against ICA when all the inputs are correct, in order to identify the intrinsic error in TC, then the impact of not knowing the exact overlap or FSD, or not representing 3D effects, can be quantified separately? Indeed, this is the approach taken by Hogan et al. (2019): their Fig. 7a-c shows that when the correct overlap and FSD is used, the shortwave bias against ICA over 65 scenes is less than 3 W m-2. This was then a firm foundation for them

to look at the representation of 3D effects, whereas in your case it would be a firm foundation to look primarily at errors due to parameterizing FSD (although you also have some interesting 3D results).

We agree that when we compare the TC results to the benchmark, a number of errors are compounded. The ecRad package (and the study of Hogan et al., 2019) contains the Tripleclouds radiative solver (Shonk and Hogan, 2008) merged with SPARTACUS (Schäfer et al., 2016), where you have the possibility to consider: horizontal cloud inhomogeneity + general (true) vertical overlap + proper 3-D effects (SPARTACUS). Our TC scheme, on the other hand, can only account for horizontal inhomogeneity (and to a minor extent for 3-D effects), therefore we can not easily separate all different effects out for the reader. We indeed tried to discuss the contribution of the various effects to the extent that we could (horizontal inhomogeneity, 3-D effects). Please see our answer to comments 18 and 19, where we explain how we have extended the discussion by mentioning the vertical overlap issues, the problem of compensating errors etc. (in order to better explain the various error sources).

Nevertheless, our primary aim was to explore various TC configurations, which can in practice be used in GCMs. To that end, the 3-D experiment is the proper benchmark, because it is the best proximity to the real world (and not the ICA). Please see also our answer to comment 13, which discusses the issues related to FSD parameterization.

3. The proposed solution to the high-FSD problem in section 4 should be compared to the solution to the same problem presented by Hogan et al. (2019, appendix). Their solution is derived from theoretical distributions, rather than by trying to minimize an error against a benchmark calculation in which several other errors (notably in overlap) are also present. The structure of the present manuscript is a little frustrating - section 3 contains some puzzling errors that are only addressed, or even properly mentioned, when the reader gets to section 4. Why not flag up the need to address the problem sooner in the paper?

Thank you for this suggestion. We have reconstructed the manuscript as you advised. We have thus introduced the various Tripleclouds experiments (baseline as well as optimizations for both overcast and highly heterogeneous cloudiness) already in Section 2. We have simultaneously reconstructed the “Results and discussion” section. Unfortunately we can not test the full solution to the high-FSD problem presented in Hogan et al. (2019), because in our Tripleclouds solver the cloud fraction scaling factor (which determines the geometrical splitting of cloud fraction in two parts) is implemented as a constant (i.e., not height-dependent). Hogan et al. (2019) presents the solution, where the “cloud fraction scaling factor” is a function of FSD (with the true high FSD being height-dependent). Once we manage to generalize the vertical overlap rules as you suggested in the first place (for arbitrary height-dependent cloud fraction scaling factors), we will be able to test the solution of Hogan et al. (2019) for high FSDs on our cloud data (which would be interesting to do).

SPECIFIC COMMENTS

4. Abstract line 6: while the optically thicker part could be used to represent convection, most clouds are not convection and the use of a thicker and thinner parts are simply a first-order approximation to the horizontal distribution of optical depth that is found in stratiform clouds.

We agree. We have thus shortened the sentence to:

> “This subject is addressed with the Tripleclouds radiative solver, the fundamental feature of which is the inclusion of the optically thicker and thinner cloud fraction.”

5. Figure 1: Caption should stress that this is a schematic; the vertical resolution shown here is coarser than any operational model.

Thank you for pointing this out. We have stressed these issues:

> “Note that the schematics are illustrative and that operational models employ finer vertical resolution.”

6. Introduction: This seems unnecessarily long. The primary purpose of the paper concerns testing the Tripleclouds scheme for representing horizontal cloud heterogeneity in a 1D radiative transfer context, for which the appropriate benchmark is the Independent Column Approximation. Many of the references and discussion concern 3D radiative effects, which seems not so central to the topic of this paper; a little shortening would therefore seem in order. The introduction should mention the McICA scheme of Pincus et al. (2003), which is much more commonly used to represent cloud structure than Tripleclouds in current global models. Figure 1, panel 2, could just as easily be used to illustrate the McICA scheme as a cloud-resolving model.

We have shortened the Introduction as you advised. We have omitted a few sentences, redundant words and a few references (concerning 3-D radiative effects and some other). Please see the manuscript file with marked changes. We have also mentioned the McICA scheme of Pincus et al. (2003) – thank you for pointing this out. The added sentence is as follows:

> “While an alternative technique known as the McICA (Pincus et al., 2003) is currently operationally employed in the majority of coarse-resolution models, the TC scheme is attractive because it is free from stochastic noise.”

In addition, we would like to let you know that we have thoroughly described/compared the McICA and the Tripleclouds scheme (in terms of computational efficiency, stochastic noise and related issues) in the Introduction of the parent paper Črnivec and Mayer (2020). Please see also Fig. 1(e) of Črnivec and Mayer (2020), which nicely illustrates the McICA algorithm. We didn’t want to exactly repeat all these contents in the present paper. Finally, we have retained the 3-D experiment as the benchmark, since it is the best proximity to the real world.

7. Line 120: Define FSD here, saying particularly that it is the standard deviation divided by the mean, in both cases considering only the non-zero water content values in the horizontal LWC distribution.

Thank you for pointing this out. The FSD parameter was originally defined in Section 2, but we agree that it is better to introduce the definition earlier within Introduction. New formulation:

> “The latter is conveniently defined by the fractional standard deviation (FSD) of cloud condensate as well as the shape of condensate distribution. The parameter FSD (e.g., of LWC) is defined as the standard deviation (σ_{LWC}) divided by the mean ($\overline{\mathrm{LWC}}$), whereby only the non-zero values in the horizontal LWC distribution are considered.”

8. Figure 2: the linear colour scale is not really suitable for the cirrus cloud since it is entirely white for optical depths up to 5, yet a significant fraction of the radiative impact of this cloud in the longwave will be from optical depths less than 5.

Thank you for this remark. We have improved the figure by applying the logarithmic colour scale.

9. Line 155: Please say how the distributions were fitted (least-squares or fitting three of the moments of the distribution?) and I'm not sure what use a Gaussian distribution is, except as an excuse to use the 16th percentile, since it is unbounded on the lower end and so negative water contents are predicted. I'm also curious as to whether you can say that either lognormal or gamma are really better; Hogan and Illingworth (JAS 2003, Figs. 4-5) found that there was little to choose between them when comparing to real data, and often the gamma and lognormal were much closer to each other than either were to the noisy distributions of individual scenes.

We have extended the sentence as follows:

> "To gain further insight about the subgrid cloud variability, the theoretical distributions (Gaussian, gamma, lognormal) were fitted to the actual LWC distribution in each vertical layer, so that they have the same mean and standard deviation as the actual data."

Thank you for pointing out some interesting results by Hogan and Illingworth (2003). It is true also in our case that lognormal and gamma distributions performed similarly well when fitted to the actual data in several layers. We have extended the discussion as follows:

> "The investigation revealed that the actual LWC distribution throughout the majority of the upper portion of the cloud, where radiative effect is maximal, is best approximated with the lognormal distribution (best fit in 5/8 of top layers), followed by the gamma distribution which performs similarly well."

Nevertheless, there are also layers where one distribution clearly outperforms the other: see for example Fig. 14 of Črnivec and Mayer (2020) where the same analysis was applied for the cumulus cloud field (and the gamma distribution is clearly closer to the real distribution compared to the lognormal distribution).

The reason why we chose the three distributions (including the Gaussian distribution) is given later in the text (lines 223-225), where we also state that the lognormal and gamma distributions are more realistic. In addition, the comparison of the TC(FSD) experiment assuming Gaussian distribution with the TC(LP) experiment gives insight into the error due to FSD parameterization (e.g., global constant).

10. Line 202: \bar{LWC} as defined here is not the layer-mean LWC, but the in-cloud mean LWC, i.e. ignoring the clear region.

Yes, we are aware of this (we thought it is clear that only cloudy pixels in the layer are considered when computing the average LWC). Nevertheless, we agree that the following sentence is better:

> "… is determined under conservation constraints of the in-cloud mean LWC (denoted as $\overline{\mathrm{LWC}}$):"

11. Line 234: a short further discussion is required - like the bottom row of your Fig. 3, Hogan et al. (2019, Fig. 10a-d) also found much larger FSD values than reported by Shonk et al. (2010). This is not possible to represent if the two cloudy regions are assumed to have the same area, but in the appendix to that paper they showed how an improved representation of large-FSD distributions could be achieved by making the optically thinner region occupy a larger area.

Thank you for this remark. We agree that the baseline TC configuration (where the two cloudy regions have the same area) is not best suitable for cloud scenes with large FSD. We have thus extended Section 2.2, by adding an extra subsection 2.2.3 entitled "Optimization for highly heterogeneous cloud scenarios" (dealing with large FSDs in the way that you described: making optically thinner region occupy a larger area). In our initial paper these issues were addressed later in Section 4 ("Parameter optimizations"), where we have also pointed out the study of Hogan et al. (2019). Nevertheless, we agree that it is better to raise these issues already in the Methodology section (Section 2).

12. Eq. 8: surely to be a bias, x and y should be averaged over more than one event? Otherwise it is an instantaneous error. Also it should be clarified whether x and y represent horizontal averages (e.g. of heating rate)... but are they also vertical averages of heating rate through the cloud layer? "Cloud-layer RMSE" is ambiguous - does the "layer" refer to model layer or the entire layer of cloud?

Correct, it is an instantaneous error. We have changed the terms "absolute bias" and "relative bias" to "absolute error" and "relative error" throughout the entire paper as well as on the figures.

In the case of 3-D and ICA experiments, x and y indeed represent horizontal averages. This is already stated in the text: “The 3-D and ICA experiment were both performed on the high-resolution cloud grid, with the result subsequently horizontally averaged across the domain.” The GCM and TC experiments, on the other hand, are single-column experiments. We always show vertical profiles of heating rate, therefore we did not perform any vertical averaging.

The “cloud-layer RMSE” refers to the entire layer of cloud, as it is already explained in the text (line 280): “The cloud-layer RMSE denotes the RMSE evaluated throughout the vertical extent of the cloud layer of each particular cloud field case study.”

13. Table 1 and text: I understand that the TC(FSD) method uses FSD=0.75, but it is unclear from either this table or the text whether the TC(LP) method uses the 16th percentile of the *true* in-cloud LWC distribution for each cases, or for an idealized (e.g. lognormal/gamma) distribution with an FSD of 0.75. If the former, then surely the main difference between the two methods is whether the true distribution is used or not, information that is not stated (at least not clearly). It would seem much more satisfactory in all cases to use the observed FSD in order that we are evaluating the intrinsic TC method, not the rather old and simple parameterization of Shonk et al. (2010).

The TC(LP) method uses the 16th percentile of the true in-cloud LWC distribution. This is written at the beginning of Section 2.2 (lines 198-204), where we have summarized the original TC method of Shonk and Hogan (2008) which we apply in our study (term the “observed distribution” is used to denote the “true distribution”). In order to clarify that we are employing this particular LP methodology as well as to clarify the difference between the LP and the FSD method (that you pointed out), we have reformulated the text in Section 2.3:

> “The Tripleclouds solver was employed in conjunction with the LP method based on the observed condensate distribution and the FSD method utilizing the distribution assumption in various configurations as outlined in Sect. 2.2.”

To explain the TC(FSD) method once more: The TC(FSD) method in the baseline configuration utilizes an idealized distribution assumption (Gaussian/gamma/lognormal) together with FSD of 0.75. We intentionally didn’t use the observed (actual) FSD based on high-resolution 3-D cloud data, because we assume that this exact value is not known in a GCM (in fact, this is our main motivation to test various FSD parameterizations). In other words - our primary aim was to answer the question: How to properly configure Tripleclouds for use in GCMs, where FSD parameterizations have to be used: e.g., global FSD estimate of Shonk et al. (2010), which is certainly rather old and simple; but we have also tested more recent and more sophisticated FSD parameterizations of Boutle et al. (2014) and Hill et al. (2015). The latter were initially introduced later in Section 4, whereas in a revised paper we have introduced them already in Section 2.

We could extend the discussion by comparing TC(FSD=parametrized) and TC(FSD=actual) in order to evaluate the TC error related to the FSD parameterization. We have performed these experiments (not shown in the paper), although the configuration of these experiments is actually not so straightforward – e.g., for the cumulonimbus. In the mixed-phase region of the cumulonimbus, namely, the actual FSD of liquid phase is very large (implying that simultaneously the asymmetrical cloud fraction splitting should be applied), whereas the actual FSD of ice is close to the global estimate (implying that the split percentile of 50 is adequate) – how can one properly deal with these issues (i.e., which split percentile should be used? - because the cloud fraction of liquid and ice phase are not separate parameters for TC solver...) is left for the future study.

The figure below shows the results of the “test experiments” for cumulonimbus, where the split percentile was simply held at 50: shown is the heating rate error in the baseline FSD experiment using global constant (“G”) and using the actual FSD (“A”) with various distribution assumptions. Whereas the heating rate error within the anvil is reduced when the actual FSD is used (and one could indeed estimate the error related to the FSD parameterization), the heating rate within the stratiform mixed-phase region is only slightly changed (but this is not a fair evaluation of the error related to the FSD parameterization due to SP of 50).

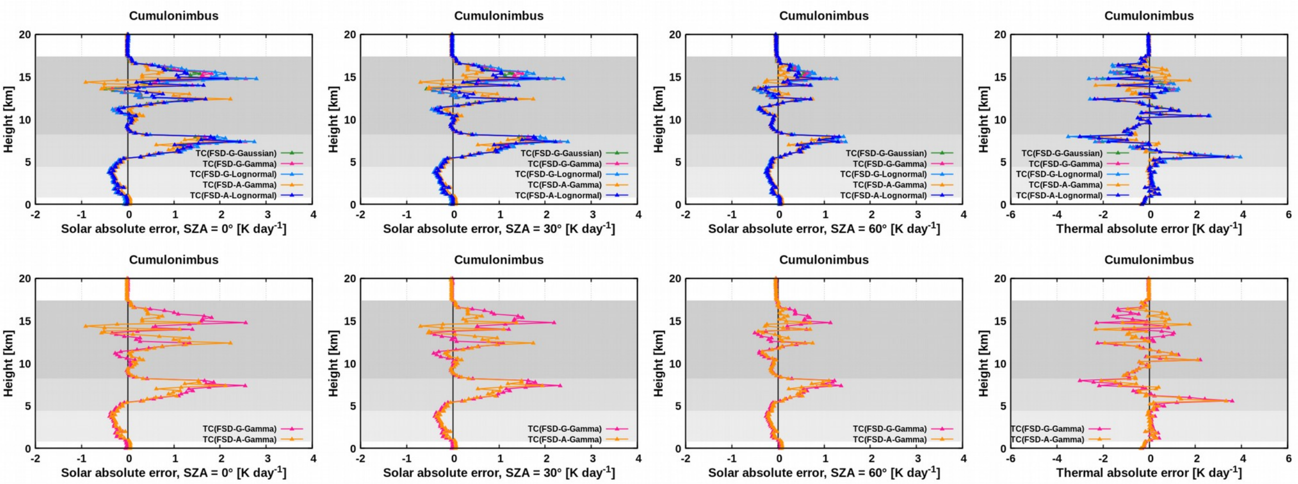

Figure: The heating rate error of various TC experiments for cumulonimbus (evaluated against a 3-D benchmark). Top row: The three baseline TC(FSD) experiments with global FSD estimate ("G") and additional experiments with the actual FSD ("A") together with various distribution assumptions. The bottom row shows only the two experiments using the gamma distribution with different FSDs (for clarity). For simplicity, the split percentile of 50 is used.

Nevertheless, we are primarily interested in the error of TC quantities if the parameterized FSD is used, since this is the experiment, which is actually performed within a GCM.

Finally, we have added the following paragraph at the end of Section 3, exposing general difficulties regarding the TC treatment of mixed-phase region:

> "In overall summary, the baseline Tripleclouds setup performed well for the apparently most complex deep convective scenario. Nevertheless, improved configurations should be further sought in the future. It would be especially intriguing to contemplate how to better treat the mixed-phase region, where the actual FSD of liquid phase is extremely large. Thus similar optimizations as for the cirrus case study could be introduced, although in the mixed-phase region of the present cumulonimbus, where the actual FSD of ice is close to the global estimate, caution needs to be taken when asymmetrically splitting the cloud fraction."

14. Table 1: "Conventional GCM" is not an appropriate label since many/most operational GCMs now use the McICA method. You could call it "Homogeneous cloud"? For this table to be most easily understood, the acronyms LP and FSD should be defined in the table.

Thank you for this remark, we agree that the "conventional GCM" is not an appropriate label. We have changed the description of this experiment in Table 1 to: "GCM radiation scheme utilizing homogeneous cloudiness". We have also removed the word "conventional" everywhere in the text. We have defined the acronyms LP and FSD in the table.

15. Line 310: Say *why* the "GCM" scheme has his bias: by homogenizing the cloud, the probability of radiation being scattered or emitted is greater. This is even more clearly evident in the cirrus case.

We provided a detailed explanation for the GCM biases (due to homogeneous cloudiness) in the parent study Črnivec and Mayer (2020) - see their Section 5.1. We have thus extended the paragraph as follows:

> “The physical explanation for the GCM bias arising from homogeneous cloudiness is given by Črnivec and Mayer (2020).”

16. Line 350 and lower panels of Fig. 6: In a fractional sense both the "GCM" and TC errors are very large, especially in the infrared. However, this is not a fair evaluation of the TC method because the fractional standard deviation is extreme in this case (around 3) whereas you are feeding it with a value of 0.75, or using a split percentile of 50% which cannot capture large FSDs. It would be better to discuss your solution to the problem earler.

Although the true FSD of the present cirrus is very large, we aimed to evaluate the TC with existing FSD parameterizations (e.g., global constant of 0.75), since this is an experiment which could be performed in a weather or climate model (where the “true FSD” - i.e., the value based on high-resolution 3-D cloud data - is not known). We therefore think it is fine to evaluate the TC with parametrized FSD (even though the values might be much lower than the “true FSD”). We showed in a later section (originally Section 4: “Parameter optimizations”) that even using more sophisticated FSD parameterization for ice inhomogeneity of Hill et al., (2015) (which would hopefully result in FSD being closer to reality and thus larger) does not bring any improvements compared to the FSD of 0.75. Simultaneously we have shown that it is crucial to modify the split percentile. In the revised paper, we have exposed these issues earlier (Section 2).

17. Figure 8, left panels (and also discussion at lines 431, 437 and elsewhere in section 3.2): It is not the net surface flux that should be shown here, but the cloud radiative effect. This way the true fractional error of the various methods can be worked out. For example, the cirrus case at 60 degrees SZA shows the "GCM" method has a solar bias of -25 W m-2, but the net flux is around 300 W m-2, implying that this is less than a 10% bias. However, it should really be compared to the cloud radiative forcing which Hogan and Kew (2005) estimated to be -39 W m-2 for this case. Thus the error is more like 64%.

Our aim was to consistently analyze the atmospheric heating rate and net surface flux throughout the entire study, because these are the two quantities that are actually being computed in a weather or climate model. In addition, we think that a net surface flux error of 10 % is not negligible and therefore worth showing.

We therefore did not perform the analysis of the top-of-the-atmosphere radiative fluxes (cloud radiative forcing), where the error could perhaps be larger as you suggested. We would like to emphasize again, however, that the cases analyzed in Hogan and Kew (2005) and in the present study are not the same, as they employ different resolution (Hogan and Kew, 2005 used horizontal grid spacing of 1.56 km, whereas we use horizontal grid spacing of 100 m; please see our answer to the first major comment for further details). In summary, the analysis of the low- and high-resolution cirrus might not necessarily bring the same conclusions regarding the magnitude of errors.

18. Section 3.2: I find the discussion of the biases in Fig. 8 rather unsatisfactory because there is inadequate discussion of the role of heterogeneity and overlap, or the problem of compensating errors (although the 3D effect is discussed at length). My interpretation would be as follows:

18a. Stratocumulus: GCM is biased low because it overly homogenizes the cloud, but then the question is why TC is biased high even though the homogeneity is about right. This raises the question as to whether the assumption of maximum overlap of the in-cloud heteorogeneities explains the excessive transmission to the surface. The vertical decorrelation length of in-cloud heterogeneities is typically assumed to be half that of cloud boundaries, so about 1 km, and if this was implemented it would block more of the solar radiation and reduce the positive surface-flux bias. Just an idea.

It is true that the maximum overlap of optically thicker cloudy regions should result in an increased transmission of radiation towards the surface (compared to more realistic overlap considering vertical decorrelation length) – thank you for this idea, which we also included in the discussion. We however later showed (in Section 4.1 of the original manuscript) that the TC is biased high mainly because the inhomogeneity in the baseline experiment is not about right: when using better inhomogeneity estimate of Boutle et al. (2014) the net surface flux bias is significantly reduced. We have extended discussion about the TC net surface flux for the stratocumulus as follows:

> "When the TC(LP) is applied, the net flux error is mostly slightly reduced, whereas in TC(FSD) baseline experiments the error is increased compared to the GCM. The latter finding is consistent with previous considerations, where it was pointed out that global FSD introduces excessive inhomogeneity to the radiatively important part of the stratocumulus, unrealistically reducing the absorption of solar radiation within the cloud layer. The corresponding increased cloud-layer transmittance, as we demonstrated herein, has important implications for the surface budget, therefore proposed optimization is highlighted in the next section.
> It should finally be noted that also the assumed maximum overlap of optically thicker cloudy regions could result in a somewhat excessive transmission towards the surface (compared to the situation in a GCM). The vertical decorrelation length of in-cloud heterogeneities is typically assumed to be half that of cloud boundaries (Shonk et al., 2010). If this phenomenon was implemented in the TC scheme it would block more of the solar radiation and reduce the positive surface net flux error."

18b. Cirrus: GCM/TC models all significantly underestimate surface transmission because they all fail to capture the strong cloud heterogeneity. Only later does the reader find Fig. 10 in which a "fix" is presented, but it would be simpler to present the fix at the same time as the original problem.

We agree. In the revised paper we have presented the methodology to fix the treatment of highly heterogeneous scenes already in Section 2.

18c. Cumulonimbus: There is a strong 3D effect for this case so if TC is closer to the 3D benchmark than it is to ICA then it is for the wrong reason: surely you should aim for TC to agree with ICA, then use some other scheme to try to capture the 3D effects on top of TC? The fact that all the TC calculations have a positive bias with respect to ICA is probably due to the incorrect maximum-random overlap assumption in this deep cloud system.

Yes, ideally we would use some other scheme to properly capture the 3-D effects on top of TC (as you do with SPARTACUS in ecRad), but this is unfortunately out of the scope of the present study (we don't yet have the scheme that would take subgrid 3-D cloud-radiative effects into account). So we aim for TC to agree with 3-D rather than with ICA (because the 3-D is the realistic benchmark and we can even capture part of the 3-D effects in our present Tripleclouds implementation; which ICA can not capture at all). See also for example Section 5.1 of Črnivec and Mayer (2020), where we have also shown for the cumulus cloud field that at low Sun the TC bias is smaller than the ICA bias (when both were evaluated against the same 3-D benchmark). This was attributed to the effective treatment of 3-D radiative effects in the TC scheme (which we don't consider as being "wrong" in our case, but rather as "desired"). It would of course be better if we had a more advanced 3-D scheme. We have extended the final section "Summary and conclusions" as follows:

> "Furthermore, in order to properly consider clouds in sheared environments, the vertical overlap rules in the present TC implementation have to be generalized. Finally, if the subgrid horizontal photon transport is to be accounted for in a proficient manner, the two-stream equations need to be extended to include terms representing the in-layer horizontal radiative energy exchange between the cloud and the cloud-free part of the grid box as well as that between the optically thicker and thinner

> part of the cloud. We currently investigate some of these topics, which will be addressed in a forthcoming study.”

The fact that all TC calculations have a positive bias with respect to ICA (true in the solar spectral range) can indeed be partly due to overlap being too maximal. This leads to an increased amount of direct (and hence net) solar radiation reaching the ground in the TC calculation compared to that in the ICA. In addition, the aforementioned effective treatment of 3-D effects in the TC scheme (the relevant one in this case: “cloud side escape effect”) simultaneously leads to an increased amount of diffuse (and hence net) surface radiation compared to its counterpart in the ICA. We have included these issues in the discussion as follows:

> “The fact that all TC solar calculations have a positive bias with respect to the ICA can be partly also due to the assumed overlap being too maximal. This leads to an increased amount of direct solar radiation reaching the ground in the TC calculation compared to that in the ICA. It should however be kept in mind that the partial effective treatment of 3-D radiative effects (e.g., cloud side escape) in the TC scheme simultaneously leads to an increased diffuse surface radiation compared to its counterpart in the ICA. Future studies should try to disentangle and quantify these effects.”

19. Line 493 and bottom-left panel of Fig. 10: I don't think you can conclude much from this analysis because you have a large error from the maximum-random overlap assumption, so tuning the treatment of cloud horizontal heterogeneity is probably leading to the wrong conclusions.

We don’t necessarily have a large error from the maximum-random overlap assumption (see our answer to comment 1). We have however concluded the paragraph with the following remark:

> “Finally, in more detailed future studies additionally considering vertical overlap issues, it should be kept in mind that the fixing of cloud horizontal inhomogeneity and vertical overlap should be addressed concurrently to avoid the problem of compensating errors (Shonk and Hogan, 2010).”