M. Coulon-Decorzens, F. Hourdin, N. Villefranque

1 Reviewer #1 (Formal Review for Authors (shown to authors)):

~~Thank you very much for your work and your suggestion. The manuscript has been consequently rewritten and your suggestion lead to us to run new experiments. We decided not to include new figures in the main manuscript, and prefer to add a new figure in the supplementary information (see Figure S1). We hope having improved the quality of the manuscript thanks to your useful questions and remarks and having well answered to your review. In the following, the overview, comments and grammatical corrections from your review are reported in blue, our answers are in black italics and citations from the text in green. Line numbers in citations correspond to the new manuscript.~~ General comment: In this study, the authors investigated an important question, that when using an inaccurate radiative transfer model, if structural error in the radiative transfer model might be compensated by errors in cloud properties due to targeting top-of-atmosphere radiative fluxes in the tuning process. They mainly focused on the structural error due to lack of 3D radiative effects of clouds. Their results suggested that when 3D effects are neglected, accurate fluxes are obtained only at the expense of overestimated cloud fractions, compensating underestimated cloud reflectivity at low sun. Finally, the authors proposed to see tuning as a powerful way to investigate compensating errors and guide model development. The manuscript is generally well organized and well written. I have some minor comments/questions, mostly related to the designment of experiments and implication of the results.

Thank you very much for your work and your suggestion. The manuscript has been consequently rewritten and your suggestion lead to us to run new experiments. We decided not to include new figures in the main manuscript, and prefer to add a new figure in the supplementary information (see Figure S1). We hope having improved the quality of the manuscript thanks to your useful questions and remarks and having well answered to your review. In the following, the overview, comments and grammatical corrections from your review are reported in blue, our answers are in black italics and citations from the text in green. Line numbers in citations correspond to the new manuscript.

1. During the SCM tuning process, parameters (i.e., decorrelation length and fractional standard deviation) in Tripleclouds were fixed, while parameters in boundary-layer and cloud schemes were calibrated targeting 3D MC fluxes. In so doing, the structural error is entirely due to lack of 3D effects. However, I would assume that in practice, before applying Tripleclouds in GCMs (especially at the stage when Spartacus was not available yet), Tripleclouds has been tuned against 3D MC fluxes in offline simulations (i.e., forced by perfect cloud profiles) or concurrently tuned with other processes in SCM simulations. If so, the structural error due to lack of 3D effects has been partly compensated by errors of parameters within Tripleclouds. Then, the question needs to be addressed becomes how the remaining structural error (after tuning Tripleclouds parameters targeting 3D MC fluxes) will be affected or compensated by parameters in boundary-layer and cloud schemes. My question is, if parameters in Tripleclouds was tuned targeting 3D MC fluxes in the "perfect cloud" framework, what the cloud-radiation error compensations would be in the SCM simulations?

We could indeed have chosen to tune parameters in Tripleclouds targeting 3D MC fluxes in the "perfect cloud" framework, and look at the cloud-radiation error compensations in SCM simulations. However we choose not to do that intentionally, because we wanted to disentangle radiative code errors from that of clouds, and quantify the possible error compensations between them in a well constrained framework. The motivations for our

choices were not explained well enough probably in the original manuscript and we spent much efforts when revising the manuscript to be more explicit on our motivations. To test however the suggestion of the reviewer, we ran a new tuning experiment, where cloud geometry parameters of Tripleclouds are tuned targeting 3D MC reference in this perfect clouds framework, following the same protocol. Results of this experiment are shown in Figure S1 of the supplementary information and are analysed line ?? to ?? and line ?? to ?? reported below [MCD : mettre les lignes et références]. Because the very best Tripleclouds configuration at simulating 1D MC fluxes is also among bests ones at simulating 3D MC fluxes, we expect that runing tuning experiment with fixed 1D-RT parameters that best simulate 1D MC fluxes or that best simulate 3D MC fluxes would lead to the same results. Therefore, cloud-radiation error compensations would be the same to first order in the SCM simulations as those shown in brown Figure S2. In practice, if we choose to use 1D-RT in the GCM, we would prefer targeting the average flux metric of Section 4.3 together with cloud-oriented metric of (?, ?).

2. The authors suggested that we can see tuning as a useful tool to investigate compensating errors, which is important for model development. In this study, we know exactly that the structural error is due to lack of 3D effects and can quantify it's impacts via adding or removing such effects in the experiments. However, in real GCM tuning exercises, the source and manifestation of structural errors are usually not well recognized. Hence, how to use tuning to guide model development is still challenging. I suggest adding some discussion regarding this issue. Qian et al. (2018) and Yang et al. (2019) can be added in the discussion.

Thank your very much for the bibliography suggestion. We agree that in a real GCM, the sources and manifestation of structural errors are not always known, and in fact it is one of the main motivations to investigate this issues in idealized situations as done here. We add some discussion regarding this issue in the last section of the article, [MCD : Est-ce qu'on cite les articles -j'aurais tendance à dire oui- et si oui où ça ?]
▷ L. : L'endroit où l'on aura citer les articles
▷ L. : Paragraphe de la CCL

Specific questions:

1. Line 23, "When 3D effects are neglected, accurate fluxes are obtained ...". "accurate"? Did the authors mean the structural error can be entirely compensated?

Thank you for pointing out this lack of precision. We would not indeed mean that structural can be entirely compensated. The sentence has been modified.
▷ L. 23-25: When 3D effects are neglected, reasonable fluxes are obtained only at the expense of overestimated cloud fractions, partly compensating underestimated cloud reflectivity at low sun.

2. Line 182, "Large scale dynamics, radiative heating and surface conditions are imposed". Please elaborate on this.

Some precisions had been added in the paragraph describing LES
▷ L. : Large scale dynamics, radiative heating cooling and surface conditions are imposed throughout the simulation for each case. The effects of large-scale advection and radiative cooling are represented by prescribed source terms in the heat and moisture evolution equations, applied in each column of the domain. These sources/sinks of heat and moisture are functions of height and time and replace an explicit resolution of the large-scale dynamics and of radiation. In the ARMCu case, surface fluxes are prescribed as

a function of time, following a typical diurnal cycle of turbulent fluxes over land, while in RICO, the sea surface temperature is set to a constant during the whole simulation and turbulent fluxes are calculated by the model

3. Line 286-291, "The next iteration starts by sampling a set of parameter vectors in the NROY space of the previous iteration, ..., mostly because emulators uncertainty decreases, which is due to denser information being collected for training". Have the samples that were generated during previous iterations (but still within the NROY space) been reused when training the emulator?

No, samples that were generated during previous iterations are not reused to train emulators of other waves

4. Line 344, "each consisting in 101 simulations." Which sampling algorithm was applied? e.g., LHS? QMC?

For the first wave, a latin hypercube sampling algorthme is applied to select the 101 vectors over the parameter hypercube. For the next wave, a latin hypercube sampling algorithm of 10^5 vectors is performed over the full parameters hypercube, and the first 101 members that satisfy all (across previous waves plus the current one) implausibility conditions are selected to run the simulation of the next waves. We add the following sentences in the section that describes the HighTune:Explorer tuning tool.
▷ L. 255-259: In practice, the parameter vectors used to run the first ensemble of simulations are sampled randomly from the hypercube using a Latin hypercube sampling algorithm. For subsequent iterations, the first $10 \times N$ parameter vectors that satisfy the implausibility condition are selected from a larger random sample of the hypercube comprising 10^5 samples, also obtained using a Latin hypercube sampling algorithm.

5. Figures 3c-3f, it seems the experiments were run at more than three solar angels, but I didn't see any description about this configuration in the method section.

The tuning experiments use metrics at three solar zenith angles (0°, 44° and 77°). For diagnosis purpose, RT was also computed at other solar zentih angles (11°,22°,33°,55° and 66°)
▷ L. ?-?: For SZA=0°, 44° and 77°, fluxes plotted are directly those used as metrics. For the other one (SZA=11°,22°,33°,55°,66°), they are further computed following the same framework, for diagnosis only.

6. Line 429, "Yet, Spartacus fluxes are globally accurate." How to derive this conclusion?

We agree that the sentance as such was somewhat meaning less. What we had in mind saying this was the following. Figure 3 shows that the 10 best simulations of the Spartacus PPE simulate radiative fluxes fairly close to 3D MC ones, and simulate significantly better fluxes than Tripleclouds, considering the fact that Tripleclouds can't do better that fluxes shown in this figure, even when targeting 3D MC fluxes in the tuning process. The sentence hower despeared in the revised version of the paper.

7. Line 438, "as clearly seen in the second line of panels in Figure 3." What is "the second line"?

Changed to:
▷ L. 428: as clearly seen in Figure 3 (c-f).

8. Line 439, "The Tripleclouds is able to well simulate 1D radiative reference computations but it will always underestimate the reflected radiation at high solar angle compared to zenith". Compared to zenith or compared to Spartacus? In addition, Tripleclouds was calibrated targeting 1D MC flux. Can we improve the angular dependency of Tripleclouds if targeting 3D MC flux?

Thanks for having higlited this mistake; it is in comparison with 3D MC of course. To answer the last the question, we ran the same tuning experiment targeting 3D MC fluxes rather than 1D MC. It happens fact the 10 best Tripleclouds configurations that best match 3D MC fluxes show results very similar to that of the 10 best Tripleclouds configurations that best match 1D MC fluxes (red points of Figure 3 (g-l)). Targetting 3D MC fluxes instead of 1D MC ones thus does not improve the angular dependency of Tripleclouds. We changed the sentences to the following ones :
▷ L. 429-433: The 1D-RT always underestimates (resp. overestimates) cloud reflectivity at high SZA (resp. at low SZA) compared to 3D radiation. This is not an artifact of the calibration experiment: tuning the 1D-RT scheme against 3D-MC fluxes leads to the same result (see Figure S2 in Supplementary Information).

2.9. Line 451, "two hours from ARMCU et two hours from RICO"?

This sentence has been replaced to the following ones :
▷ L. ?-?: At each iteration, 130 free-parameter vectors are sampled in the NROY space and ARMCU and RICO are simulated using these 130 configurations of LMDZ. One RT configuration is then run offline for the two chosen hours of the two cumulus cases and each of the 130 SCM configurations, to compute solar reflected fluxes at three solar zenith angles each

2.10. Figure 4, What are the yellow marks?

Yellow marks indicate the cloud cover of the 10 best simulations of the Hybrid experiment. Unlike in Figure 3 where there is two colors, one for the 1D-RT (`Tripleclouds`) experiment and the other of the 3D-RT (`Spartacus`) one, here there is only one color because 1D-RT and 3D-RT share the same decorrelation length in this experiment so they share the same cloud cover. We add this sentence to add clarity in the legend of Figure 4.
▷ L. ?-?: in a-b, decorrelation length parameters used to compute cloud cover (yellow) are those of the 10 best simulations from the Hybrid experiment that are the same using 3D-RT (`Spartacus`) or 1D-RT (`Tripleclouds`)

2.11. Line 540, "layer-wise maximum cloud fraction". Maximum cloud fraction should be defined in an earlier place, so the readers can understand the results related to Figures 3 and 4.

We choose to define this maximum cloud fraction when introducing Figure 3:
▷ L. : We start by analysing the time evolution of the cloud cover obtained directly in the LES with the evolution computed with the exponential-random model of Ecrad, using the decorrelation lengths obtained in the best experiments using 1D-RT (red) and 3D-RT (blue) scheme. For comparison, we show on the same graph the time evolution of the maximum cloud fraction on the vertical, which is also the cloud cover that would be computed with a maximum overlap hypothesis or an infinite decorrelation length for the exponential-random overlap scheme.

2 Reviewer #2 (Formal Review for Authors (shown to authors)):

~~Thank you very much~~ General comment: Error compensation is an obstacle to the model development and improvement. Particularly in the model/parameter tuning processes, the energy budget balance at the Top of the Atmosphere (TOA) inevitably includes the compensation error. This study investigated the cloud-radiation compensation errors through PPEs simulations by SCM, with and without 3D radiative effects of clouds. Results indicate that the accurate fluxes result from an overestimated cloud fraction and an underestimated cloud reflectivity, and that 3D radiative effects are crucial for flux simulations. The insight of this study is novel and the result is interesting. However, the manuscript is not well-written or clearly presented. The current version seems a technical report for the model teams, making it difficult to follow the logical flow for me. I think the manuscript requires substantial revision or rewriting in accordance with the writing standards for academic articles. Since I cannot completely understand the current manuscript, I will only offer some detailed suggestions.

Thank you for having reviewed the paper. We ~~significatly~~ understand that the paper may appear as too technical although we believe it is difficult to discuss such modeling and tuning aspects in deep without giving a significant place to technical issues. We agree nevertheless that it is of first importance that the manuscript appears not as a technical document but as a piece of research on modeling, that we believe it is. Therefore, following your comment and some comments from the two other reviewers, we decided to significantly re-work the paper to add clarity, and better highlight the scientific results of our study~~.~~ ~~We have~~ , and the motivation of the various protocols used. Following your advices as well, we also revised certain word choices, which may have been confusing or too technical. We hope that this new version will be of a higher quality and will be easier to follow. We also hope having well answer to your review. In the following, the overview, comments and grammatical corrections from your review are reported in blue, our answers are in black italics and citations from the text in green. Line numbers in citations correspond to the new manuscript.

1)Words such as "Spartacus," "Tripleclouds," and "History Matching" should preferably be introduced after their first occurrences. They are important to understand the results.

In the revised version, we introduce the words "`Spartacus`" and "`Tripleclouds`" later on in the paper that was the case in the submitted manuscript, i.e. at the beginning of section 2.2 [MCD : Mettre lee numéro de ligne]. We made the choice then to refer to `Spartacus` and `Tripleclouds` as 3D-RT and 1D-RT in the remaining of the manuscript.
▷ L. : In the rest of this article, the tripleclouds solver is referred to as 1D-RT and the spartacus solver as 3D-RT, referring to the presence or absence of 3D radiative effects.

History Matching is still first used in the introduction of the paper, as this method is at the heart of our study. We change the sentence of its first use to the following:
▷ L. : Following the History Matching approach proposed by Williamson et al. (2013), [...]

2)Some references are not cited correctly (e.g., L73 - 75). Please re - cite the references according to the citation format of JAMES.

This reference is now L298 and appears to follow the JAMES citation format.

3)The full names (e.g., PPEs at Lines 670 - 671) and their abbreviations (e.g., on Line 151, the abbreviation SZA is missing its full name) are misused throughout the manuscript.

Thank you for pointing out those mistakes. We carefully checked the abbreviations and their full names are correctly used in the revised manuscript.

4)It is advisable to move the experiment designs in result analysis sections 3 and 4 to section 2.

We decided not to follow your recommendation on this point, because it did not seem to make the article any clearer. We present in section 2 the general approach, but we think that presenting the specific choices made on the experiments in section 3 and 4 helps clarifying the link between the protocol chosen and the scientific question to be addressed. Nevertheless, we put the description of experimental design of section 3 in a subsection, to improve the whole structure of section 3
▷ L. : 3.1 Experimental desgin

5)As my understanding, the experiments and conclusions in section 3 are the idealized bases for those in section 4. However, the logical relationship between them is not mentioned. Another interesting thing in section 3 is that the results in both spartacus and tripleclouds are close when the SZA is 44degree. What is the underlying mechanism?

As said above, We made significant work to better explain the link between sections in this revised version. Please refer to the manuscript with enlightened differences to have a global view of this changes.

Results in both spartacus and tripleclouds are close when the SZA is 44 degree is explain by the fact that 3D radiative effects are close to zero for this SZA. We add the folling lign at the beginning of section 3 to clarify this point:
▷ L. 307-312: The shift in the sign of 3D radiative effects as a function of SZA is explained by the compe- tition between two processes: side leakage, which reduces cloud reflectivity and predominates at low SZA, and side illumination, which increases cloud reflectivity and predominates at high SZA. This also explains why 3D radiative effects are null for intermediate SZA, around 40-50° for the cumulus cloud scenes under study.

6)Line410 "betwen"->"between" *Done*

7)Figure 3c, the lines for MC ICA and MC 3D should be black (or changing the figure caption), MC ICA better changes as MC 1D as in figure caption *Thank to point this out,*

[MCD : A faire]

8)Fig. 4C same as Fig. 3c [MCD : A faire]

9)For the conclusions, the compensation errors in most CMIP models are the underestimated cloud fraction (especially total cloud fraction) and the overestimated cloud reflectivity. Are the different errors in this study from its LMDZ model itself? or from the methods?

~~[MCD : Mince, j'avais oublié cette remarque !!!!!! Pas sur de l'avoir comprise je voulais en parler avec vous p~~
It is true. We added the following sentence to the concelusion:

▷ L. LXXX: This demonstration is made in an idealized configuration, and our results should not be directly extrapolated to 3D coupled climate models. And indeed, the compensation errors obtained here differ from the generally admitted although not clearly demonstrated result that GCM systematically overestimate cloud reflectivity to compensate for an underestimated occurrence of clouds (?, ?). Many reasons can be mentioned to explain this difference. Most CMIP RT models were using so far a max-random overlap approximation, which results in an underestimated cloud cover, even for a well simulated vertical profile of cloud radiation. Also, some models may have difficulties to simulate cloud fraction large enough, which may induce an overestimation of cloud reflectivity to compensate for this lack of clouds. It is not the case of the LMDZ cloud parameterizations which easily simulate cloud fractions which are large enough, or even too large depending on the values of some free parameters, for the cumulus clouds considered here. Finally, the tuning of global 3D GCM is done usually against observed TOA climatological fluxes, for which compensation are at work between various zenith angles, at least as concerns the diurnal cycle. By requiring that the model works reasonably well at high SZA without accounting for 3D effects in the RT computation, we induce compensation effects that selects large cloud fractions, that produce at the end a large effective cloud covers which can compensate for the absence of 3D effects in the RT computation.

[FH : Ajouter le paragraphe ci-dessus et enlever: Indeed, in the SCM setup considered here, only shallow convection and cloud parameterizations can compensate structural radiative errors, whereas much more processes are at work in a 3D GCM, which can result in other compensating errors.]

3 Reviewer #3 (Formal Review for Authors (shown to authors)):

Thank you very much for your very clear questions and remarks, that really helped us revise the paper. We decided to limit the changes in the organization of the sections and concentrated our efforts on transitions between sections and on the explanation of the motivations and rationale behind the experimental protocol chosen, following your suggestions, for a better guidance of the reader. We therefore significantly modified the introduction and end of section 3 and 4, as well as the general introduction. We also moved some discussion that were initially along the analysis of the experiment in the last section, to better emphasis the results of the experiments. We hope that we succeeded to improve the manuscript, having well taken into account your suggestions. In the following, the overview, comments and grammatical corrections from your review are reported in blue, our answers are in black italics and citations from the text in green. Line numbers in citations correspond to the new manuscript.

Summary : This work studies potential compensating errors that may appear in a GCM during tuning. All GCM contain compensating errors, and this work is a good example of how a tuning protocol can be used to identify this type of error. First, the radiative transfer scheme, ecRad, is tuned using two different solvers, Spartacus, that considers clouds 3D effects, and Tripleclouds, that does not consider 3D effects. The target radiative fluxes are computed using, respectively, a 3D and a 1D Monte Carlo simulator ran on the outputs of a LES simulation. This first analysis shows that considering 3D cloud effects allows the radiative transfer scheme to be more precise than Tripleclouds at high solar angles. Second, a single column atmospheric model (SCM) is tuned to match the radiative fluxes computed with the reference LES model and the 3D Monte Carlo radiative transfer simulator. The SCM radiative fluxes are computed using ecRad scheme with the configuration found during the first part of the paper. Uin this section, it is first shown that using a solver that ignores 3D effects can reach the right radiative fluxes by introducing errors in the cloud fraction. It is also shown that the model can be well tuned with Tripleclouds solver if the target metrics are averaged over multiple solar angles. The paper is well written, and the presented analysis is detailed and seems very accurate scientifically. The importance of this study is well justified by the rising importance of machine learning in the tuning of models.

Overall comments:

However, it is sometimes hard to understand the motivation behind specific choices and methods used in this study. I wonder if this issue could be easily solved be reorganizing some sections or by including more text to guide the reader. Here is a list of questions and confusions that emerged my reading:

Section 3: After reading all the paper, I understood that ecRad with Tripleclouds is not tuned using the Monte Carlo 3D fluxes as the target, to make sure it ignores the 3D effects of clouds. Hence, when it is used during tuning with the SCM outputs, it can be demonstrated that ignoring 3D cloud effects in the radiative transfer scheme can introduce compensating errors in the clouds. It would have helped my understanding to read this explanation at lines 333-338.

We add those sentences in the experimental design subsection of section 3 to clarify this point.
▷ L. 329-333: Reference radiative fluxes used as targets are 3D-MC fluxes for the 3D-RT tuning experiments, and 1D-MC fluxes for the 1D-RT experiments. By requiring the 1D-RT model to match 1D MC fluxes rather than 3D MC ones, we intentionally avoid 1D-RT to compensate for a structural error that we know is present in the solver, namely the absence of 3D radiative effects.

Section 3.2: It is not clear to me why the authors try to find a paramete set that works well for the two different models simultaneously. Is it to reach the objectives of section 4? Please clarify how at the beginning of this section.

We add a specific paragraph to clarify this point at the beginning of this section (that is now section 3.3).
▷ L. 435-438: The best 3D-RT configurations were found to yield the right reflectivity for the wrong cloud geometry, which we interpret as a sign of internal compensating errors. Our main hypothesis here is that a "good" 3D-RT model should be able to simulate 3D fluxes as well as 1D fluxes when 3D effects are removed from the simulation.

4.3: I read the first part of this section multiple times without completely understanding the objective. The title of this section is not clear either. A few more minor comments follow.

[MCD : repondre]

3.0.1 Minor comments :

2.1. L. 57 : In so doing -> In doing so

Thank you

2.2 L. 150-153: This sentence is confusing. Would it be possible to add a few words more precise then "more appropriate ecRad parameters"? I was confused by the words "(instead of best Spartacus)". It seems trivial to me that using the parameters tuned with Spartacus would not work well with Tripleclouds.

The introduction has been consequently rewritten, so this sentence doesn't appear anymore in the revised version of the manuscript.

2.3 L. 218: Effective radius of cloud "droplets".

Done L. 179

2.4 L.232-236: Does it account for 10 W/m2 for high and low sun?

[MCD : J'ai pas compris la question]

2.5 L. 362: remove "a".

[MCD : done]

2.6 L.375-376: There is a part of this sentence missing.

▷ L. 380-382: The optical depth at a given point along the horizontal axis corresponds to the sum, along the vertical axis, of the optical depths of the cloud layers above it.

2.7 L. 439-441: Tripleclouds seems to underestimate reflectivity at SZA=77degree sign but not at SZA=44degree sign, could you comment why?

We add thes sentences to clarify this point:
▷ L. 307-312: The shift in the sign of 3D radiative effects as a function of SZA is explained by the compe- tition between two processes: side leakage, which reduces cloud reflectivity and predominates at low SZA, and side illumination, which increases cloud reflectivity and predominates at high SZA. This also explains why 3D radiative effects are null for intermediate SZA, around 40-50° for the cumulus cloud scenes under study.

2.8 L.444-446: Could you clarify here why you want to satisfy both solvers at the same time?

Yes, as previously said, we add a specific paragraph to clarify this point at the beginning of the section 3.3.
▷ L. 435-438: The best 3D-RT configurations were found to yield the right reflectivity for the wrong cloud geometry, which we interpret as a sign of internal compensating errors. Our main hypothesis here is that a "good" 3D-RT model should be able to simulate 3D fluxes as well as 1D fluxes when 3D effects are removed from the simulation.

2.9 L. 500-502: Could you include the radiative transfer parameters in this experiment as well? Could it be a way to mitigate the compensating errors?

[MCD : Fred si tu veux répondre, je crois qu'au final on a pas discuté de cette idée dans l'article]

2.10 L. 580-581: Reading this, it seems that the "best Tripleclouds" was not used in sections 4.2.

[MCD : Pour le coup c'est une question à laquelle je peux répondre mais vas-y si tu veux]