Reply to anonymous Referee #1

Remo Dietlicher

November 1, 2018

Thank you for carefully reading our manuscript. We are happy for your expertise regarding satellite data. In the following we answer the individual points you raise:

Main comments: In the model evaluation section, the authors do not describe the dataset at all and do not explain why they chose these specific datasets to evaluate their model. For example, three datasets are used for the fluxes and no reason whatsoever is given to justify this choice. The authors must define/introduce the datasets, even briefly, and explain why they use these. Also, the observed interannual STD can be used as an uncertainty estimates when nothing else is available. In addition, their model evaluation is more of a qualitative comparison than a quantitative one although it is possible to quantify the bias more precisely (see specific comments for further details). It also looks like (it is not specified in the manuscript) they didnt make a consistent comparison between the CALIPSO-GOCCP cloud phase dataset and their model outputs, i.e., they didnt use the simulator for the cloud phase diagnostics, which make the results difficult to interpret. In the second part of the manuscript, I feel like reducing the number of categories in the final part of the study would help to better determine the origin of the ice bias. Id like the author to either do that or better explain why they chose these categories and what is the added value of making this choice. Finally, the authors could easily check at least one of the mechanism that supposedly lead to the overestimation of the ice cloud occurrences (the overlap assumption).
(Introduction): The goal of the paper is a little bit confused and not clearly stated. Also, it is not clear to me how 'As has been eluded to above, the formation history of a cloud plays a decisive role, both for mixed-phase and cirrus clouds'. The authors should state clearly than one expect biases to come either from ice behavior with respect to liquid within the mixed-phase temperature range or from ice formation at temperature below -38C, and better explain the reasons.
We have restructured the introduction to better motivate the study. We now highlight that the cloud phase partitioning is governed by the ice phase parametrizations and that we want to figure out which process dominates in our model.

Define the acronyms (e.g., CALIPSO, GOCCP, COSP, CERES etc...)
Done.

P2 L18-19: The sentence could be re-phrased, K14 found that even... among a some (or number, I believe they used 6) GCMs was not reduced.
Done. Actually when checking the exact models, it was interesting to see so many models from the CAM family.

P2 L6: I believe ice-containing clouds would be more appropriate. Why not to compare the snow water content + IWC between the REF and 2M models?

We diagnose the snow mass flux in REF, assuming that all snow will reach the ground within one model timestep. Therefore we can quantify the column-integrated amount of snow but there is no way to assign snow contents to individual levels.

P8 L8: but may indirectly affect the cloud in the tropics, especially considering the large amount of high clouds removed

This is true, we adjusted the text.

P8 L10 version night time? The authors do not explain what is a simulator at all and why using cosp here. The sentence does not tell much. The SQRT(X2) of the bias and the correlation pattern number would help better assess the improvement of the new model version.

We use day and night. We extended this paragraph to motivate the usage of COSP better. We now also compute the Pearson correlation coefficient and RMSE between the models and CALIPSO to allow for a more quantitative discussion.

Fig. 3: There is no height in Fig3 Adding the contour of the difference in the original cloud cover on the bottom plot (i.e., the contour of the blue color, $-5\,\%$ in bottom right plot of Fig. 2) could help identifying areas of improvement. It seems like there is no change at all in middle cloud, which are lacking even in areas with no overlying high-cloud which could cause shielding effect of the lidar.

The height axis must have disappeared by mistake, its in there again. Thank you for noticing this. Adding the $-5\,\%$ contour line to highlight areas where the new cloud cover parameterization acts is a good idea. The new scheme has been designed to make the transition from mixed-phase to cirrus clouds continuous and consistent with the parameterization of the formation of cirrus clouds. Improving the cirrus cloud structure is a nice side-product. Improving the mid-level cloud structure has not been the focus of this study.

P8 L18-22: Im not sure I understand the sentence: The fact... The authors state that changing microphysics does not affect CRE, that is not true (e.g., Cesana et al., 2017; their Fig. 3). The authors might get similar CREs because they tune the TOA fluxes. Also in their Fig. 4, it is clear that there are regional differences in the GCMs CREs, i.e., over the Southern Ocean. This bias is worsened by the new GCMs, probably because of less supercooled liquid sustained in the mixed-phase clouds. The authors do not explain why they chose these particular observation datasets. For the fluxes, I believe CERES-EBAF is the most relevant dataset for model evaluation also the longest period of time available (therefore a better climatological estimate of the present-day mean state), which is not defined either. Same thing for the cloud cover, no reason for these specific datasets and while it is mentioned that the simulator is used before (although it is not mentioned why) here no information is given whatsoever. I would recommend using only simulator-derived model outputs against GCM-oriented observation datasets, e.g., ISCCP, simulator Klein and Jakob, 1999 and dataset: Pincus et al., 2012, MODIS, simulator and dataset Pincus et al., 2012, cloudsat simulator Marchand et al., 2008 and dataset Marchand et al., 2010, CALIPSO, simulator Chepfer et al 2008; Dataset Chepfer et al., 2010. The interannual STD may be used as an uncertainty...

We have rewritten this paragraph to make clear that CRE is a consequence of tuning

TOA fluxes. The fact that it is more negative than what observations suggest hints at a structural problem in the model that is not specific to the microphysics scheme.
Regarding the TOA fluxes in Fig. 4, we agree that the original Figure was confusing. We now only use the CERES-EBAF dataset as suggested and plot the interannual STD as a measure of uncertainty. Furthermore, we compute the correlation and RMSE of the full 2D fields. We agree that these statistics provide interesting and important information for a more quantitative assessment of these fundamental model variables. This analysis revealed that the statement about the new model correlating better with the observations was false, even though the zonal mean suggested that.

P8L26: I would suggest adding In the new scheme (i.e., 2M, 4M)... to avoid confusion.
We replaced this sentence with something more precise.

P8L31: Again, it is not quantified at all, so hard to say. With these 2D quantities (i.e., cloud cover), it is easy to compute means, biases and correlation, so please do so and compare to CERES-EBAF.
This has been addressed in a previous comment.

It is striking to see how little change there is between 2M and REF in terms of cloud cover whereas the vertical cloud fractions are tremendously different. Did the authors look at the high-cloud cover as well? Can they give a hint of why such a small difference in the cloud covers? The cloud overlap may explain this.
The new and reference models differ most in high-level clouds and are fairly similar for mid- and low-level clouds. Since both models tend to underestimate the cloud fraction, the overestimation of the high-level cloud fraction in the reference model improves the total cloud cover in areas where the cloud fraction would be small otherwise. We agree that the reason for the smaller difference among the models in terms of total cloud cover as compared to the vertical structure is due to vertical overlap. This is now mentioned at the end of Section 3.2.

P9L5: Again very little information is given about the observational dataset and its weaknesses/strengths.
We extended this paragraph to motivate the use of the Li et al, 2012 dataset better.

P10 Sec. 3.6: Is the simulator used in that comparison or do the authors compare CALIPSO-GOCCP to the direct outputs of their models?
We do not use a simulator for Fig. 7 but added a new Fig. 8 including output from the COSP simulator. More details follow below.

P11 Sec. 4: While I agree that the method used here to determine the origin of the overestimation of cloud ice is good, it is not new and it has been used in the past for different topics and referred to as tendency (i.e., Brient et al., 2016). It is usually not possible to do so when comparing multiple models unless a specific experiment is designed to tackle a problem and requires these such as in Brient et al. (2016) -, which is why it does not often appear in multimodel studies.
We changed the text to highlight the reason why this diagnostic is very helpful to answer the specific question at hand 'where does ice come from?' and better differentiate this method from analyzing model tendencies. In a nutshell, tendencies are a snapshot of

the strength of processes but provide no history, which source-tagged tracers do. However, implementing our method requires additional prognostic tracers, a substantial effort, which makes it even more unlikely to be used in model inter-comparisons.

I do not understand what justify the use of so many types of clouds. The question is where does this ice come from? The answer is threefold from what I understand. Therefore, there should be three categories: Fraction of ice from heterogenous processes Fhet, from homogenous processes Fhom and from nucleation Fnuc. The total would be 100 % and figures would be easier to understand.
The goal of the cloud types defined in this study is to differentiate clouds with fundamentally different properties. A priori, we did not know what to expect but wanted to include all information that is available in the model. For example, we were interested to know whether mixed-phase clouds are so rare because (1) mixed-phase freezing occurs infrequently or (2) whether the subsequent ice growth is slow. Distinguishing ice and liquid dominated mixed-phase clouds allows quantifying both aspects: (1) mixed-phase freezing is only important in a small fraction of clouds and (2) the effect it has on the cloud phase partitioning is even smaller because a majority of mixed-phase clouds do not (or slowly) glaciate. Similarly, including the vertical cloud structure allowed to diagnose sedimentation in vertically adjacent cloudy layers as the relevant pathway for the transport of ice into the mixed-phase regime. A priori, it could also have been the case that e.g. sublimation in clear-sky levels is underestimated such that ice can pass too many subsaturated layers. Finally, we differentiate warm and cold liquid clouds to quantify the cloud amount that could be affected by freezing.
The sum of all cloud types is exactly the total cloud cover.

P13L15: But how to define unrealistic pathways when no observations are available to compare to?
We have rewritten this paragraph to be more precise about the use of our method and refrain from the word 'unrealistic' which we agree cannot be assessed by observations.

P14: Again, a fraction compare to the total would make more sense.
We assume this comment is regarding Fig 12 which is discussed on page 14. Having both plots at hand, we believe there is no benefit of normalizing by total cloud cover. Qualitative statements about the relative contribution to the total cloudiness can still easily be made.

P14L20: If the simulator is used, then the same weaknesses should affect the model outputs. Also, in the mixed-phase temperature regimes, the undef-phase category can be considered as mixed-phase likely. By using ice/total cloud frequency you are considering these undef-phase clouds as being liquid clouds, which is true in the tropics at warm temperature but unlikely at freezing temperatures. Once again, this section raises the question of whether the lidar simulator was used in Fig. 7.
Since we are not using a simulator for Fig. 7, we added a new Fig. 8 which compares the phase ratios versus temperature lines from the satellite, the model and the model + simulator. This was actually an interesting and worthwhile exercise. The overestimation at $T > 15\,°C$ is much reduced when using a simulator, implying that attenuation is very important in this temperature regime. It makes sense, since the thick cloud type is also optically thick, the lidar signal will always be attenuated in the lower part of these clouds.

Unfortunately, this implies that the satellite cannot be used to constrain the cloud phase partitioning in this temperature regime.
Figure 7 was inspired by Cesana et al. 2015 where they conduct a comprehensive model inter-comparison of phase ratio vs. temperature histograms (their Fig. 10). For the satellite panel we reproduced their Fig. 10 by using daily night-time data from here: ftp://ftp.climserv.ipsl.polytechnique.fr/cfmip/GOCCP_v3/3D_CloudFraction/grid_2x2xL40/2008/night/daily for the years 2008-2014. These files contain a variable called *cltemp_phase* which is computed as ice/(ice+liquid), i.e. neglects pixels where no phase can be assigned (undef). This corresponds exactly to the computation within the COSP simulator.

One could also look at particular latitude bands to avoid the influence of these thick clouds and see whether it impacts the Phase-T relationship, e.g., in the Arctic where these clouds are less frequent.
Thick and cirrus clouds are the dominant cloud type almost everywhere on the globe (see Fig. 11).

L13: Did the authors mean sedimentation of ice at warmer temperature? i.e., the mixed-phase temperature range?
We meant to say from colder (below $-35\,°C$) to warmer temperatures (i.e. the mixed-phase regime). This is now stated more clearly.

L15 Im not sure simplifying ice category from ice crystals and snow ice to only ice can be called as an improvement, Id rather use the word 'difference'.
It is an improvement in terms of the physical realism of the scheme at the cost of increased computational demand because sedimentation of ice crystals needs to be resolved. So improvement might indeed be a little bit too general and we replaced it with 'difference'.

L18-20: No cloud bias below $-35\,°C$ is shown in this paper and the biases are not well quantitatively quantified. I dont understand the expression 'arguably more reasonable tuning parameters'. This should be clarified.
We show that the reference model overestimates the cloud fraction at temperatures below $-35\,°C$ in Fig. 3. In the updated version of the manuscript we now compute the correlation and RMSE and are confident that the new cloud cover parametrization is not just a conceptual improvement but also leads to a more realistic cloud fraction for these high clouds. We removed the sentence about tuning parameters. Even though we are more happy with scaling process rates by a factor of 5 rather than 1000, they are fundamentally unconstrained so there is not really a metric to assess good and bad.

P16 L5-6: Checking this out by changing the sedimentation overlap to random (or even minimum) overlap and running a short 1yr or even a few month simulation should be relatively easy to do and would strengthen the conclusions.
Given all the changes above, the part that is referenced here has been removed.