Referee Report on "Comparison of Antarctic polar stratospheric clouds observations by ground- and satellite based lidars and relevance for Chemistry Climate Models" by Snels et al.

General comments

This paper presents a statistical comparison of Polar Stratospheric Clouds (PSCs) occurrences for different Antarctic PSC fractions including NAT mixtures, STS, ice, and enhanced NAT mixtures between ground-based measurements and CALIPSO data. In a second step, the CALIPSO data are compared with 5 different Chemistry-Climate Models (CCMs) using several diagnostic methods, based respectively on the vertical extend of PSCs for all PSC fractions, the total PSC frequency for NAT mixtures and ice, the histogram of SAD values for all PSC fractions, and the evolution of the NAT and ice fractions as a function the difference between the temperature and the NAT equilibrium temperature.

I am not convinced that the way the authors process the CALIPSO and ground-based lidar data is always rigourous and adequate, and this might be a source of many biases and difficulties.

Further, the way to evaluate the agreement between the CALIOP and ground-based datasets, but also the agreement between the different models and CALIOP, look subjective in some cases (e.g. comparison CALIOP-ground-based lidar based on Figure 1, distinction between "rather good agreement above 15 km" and "biased below 15 km" on Figure 2, general rejection of "outlier" LMDZrepro model although this model scores not so bad following some specific criteria).

Concerning the comparison between CCM's and CALIPSO, I find striking that the "best model" giving the best agreement with CALIPSO is highly depending on the methodology used: Based on total PSC frequencies (Table 2), LMDZrepro and WACCM-ccmi are performing the best; based on the SAD histogram, LMDZrepro shows the best agreement based on the range of Log_{10}(SAD); WACCM and CAM3.5 give the closest evolution of the NAT and ice fraction as a function of $T\text{-}T_{NAT}$. Hence, CCSRNIES is the only one of the 5 models considered here that cannot pretend to the status of "best model" following any diagnostic method, although the authors reject overall another model, namely LMDZrepro, and outlier. Overall, I don't see any clear conclusion from this work, and my general feeling is mainly that the way the CALIPSO data ground-based lidar data are processed might present biases or be inadequate, and that the implementation of the different diagnostic methods should be improved.

Finally, the text is often lacking in rigour or written in a language punctuated by approximate expressions and mistakes, making the reading sometimes very difficult. This should be improved.

Detailed comments

Abstract

- L. 3-5, p.1: This sentence is particularly difficult to read. Please reword in a more fluent way.
- L. 1 and 6, p.1: The authors repeat partly the same idea. The text could be written more efficiently, or in another way to put the emphasis on the main focus of the sentence.

1. Introduction

- L. 7-8, p.2: "Many different schemes…": Do the authors mean that the different schemes use different thresholds for detection and classification ?
- L.11-12, p.2: "Ground-based lidar observatories… from the early nineties to today": The authors might be only interested by the period from the early nineties until today, or by a specific location (probably McMurdo), but there exist ground-based lidar time series spanning at least 2 decades more ! (See for instance Jäger, J. Geophys.Res., 2005). Hence, they should be more specific.
- L.12-13, P.2:" A clear issue …": Do the authors mean that the ground-based time series above Antarctica are not representative enough for climatological studies and model evaluation above Antarctica ? This should require a reference.

2. Comparison of PSC observations by ground-based and satellite based lidars

2.1 CALIPSO observations

2.2 Ground_based PSC observations at McMurdo

- L.20, p.3: "Klett algorithm": This requires a reference.
- L.2-3, p.4: What do the authors mean by "facilitate" ? Is it about reducing the dataset ? Or having a regular time base ? Or something else ?

2.3 PSC detection and classification

- L. 24, p.4-l. 8, p.5: The authors are restarting an overview of the literature, citing the same works as in the overview literature in the introduction. This cares for unnecessary repetitions. The authors should focus on the message needed at this point of the discussion, without repeating what was said before.
- L.1-2, p.5: These lines include 2 almost similar sentences about the same work ! Please remove what is not necessary.
- L. 1-6, p.5: The same reference is cited 3 times during the description of this work. Please remove two of them !

2.4 PSC detection and classification criteria for the CALIPSO V2.0 data

- L. 10-12, p.5: Here again, the authors repeat what has been written in the introduction (on ll. 8-10, p.2).
- L.13, p.5: "below" is actually immediately after the sentence. "As follows" might be more appropriate.
- L. 14, 16, p.5: The use of "now" brings some confusion: do the authors mean "in Version 2" or "in the present work" ? Using "In Version 2" (if this is what is meant) might clarify this point.
- L.17-19, p.5: These two sentences are difficult to read. Do the authors mean that there are two criteria, and that a PSC occurrence is assumed if at least one of the criteria are fulfilled ? Writing that two threshold for background aerosols, respectively for the perpendicular backscatter and the scattering ratio, are defined as their median value plus one median deviation, might already clarify the text. Using formulas might also make it more clear. It is also not clear for me what is the relationship between the median deviation and the "unc" quantity. I understand from the text that, in both cases, the effective threshold is the median value+median deviation+ uncertainty. Is it what the authors mean ? Again, an expression using an equation may remove any ambiguity.
- L.2, p4; l.17, p.5; l.30, p.6: the time references are confusing. In l.2, p.4, it is indicated that about 1 data point estimated from 30 minute observation is considered every 6h at most; In l.30, P.6, this becomes "1 or 2 measurements occurring per day". And in l. 17, p.5, the authors consider a "daily median". On which sampling do they compute the median ? And does the explanation in p.5 mean that a different threshold is considered every day ? An hence that the "background value" is changing every day ? This seems a strange concept of "background value" !
- L. 20-31, p.5: Again, all this long description of PSC types would be much more easy to read if they were included in a table and supported by some equations in the text. Also, if the authors find necessary to repeat the change of criteria performed in the CALIPSO dataset, they should at least explain why all these changes are made. Is it a response to the conclusions of the work by (Pitts et al., 2018) explained in ll. 3-6, p.5 ? If yes, the conclusions of (Pitts et al., 2018) might be moved to here.
- L. 26-29, p.5: I understand that MLS is used to select the PSC type observed by CALIPSO, and that CALIPSO is used to determine the selection criteria. Is there here any problem of snake biting its own tail ? How effective is then this selection ?
- L. 32, p.5: "the PSC classified grid": What does it mean ?
- L. 32, p5: Which optical parameters ?

2.5 PSC detection and classification criteria for the ground-based data

- L. 5-9, p.6: Here, the threshold for PSC detection are clearly constant. In which extend are these criteria consistent with the criteria used in ll. 17-19, p.5 ?

- L. 11-13, p.6: I am not sure if this selection occurs in the same way as for the CALIPSO data (See L. 25-26, p.5). Which is the criteria used in that case and how consistent are the selection criteria for the CALIPSO data and the ground-based data ?
- L.13, p.6: Why do the authors consider here monthly averages while they consider daily averages before ? Isn't there a lack of coherence in their choices?
- L. 4-15, p.6: Again, using a table for all the selection criteria could be more readable and make the comparison with equivalent selection criteria applied to CALIPSO more readable.

2.6 Comparison of coincident PSC observations at McMurdo from the ground and from CALIPSO during the 5-year observation period

- L. 19, p.6: What do the authors mean by "unique definitions" ? Here, the criteria used for ground-based and CALIPSO measurements are different !? This sentence sounds also not very fluent.
- L. 3-4, p.7: Does it means that the criteria provided in §2.4, specifically for CALIPSO, are actually not the ones that are really used ? This is quite confusing !
- L. 8, p.7 – l.11, p.9 and Table 1: It is extremely difficult to conclude that the agreement between both plots is good. When focusing on very limited periods showing a clear pattern related to a specific PSC type on one of the plots, the other plot often doesn't show a similar pattern at the same time and same altitude range. Hence, I cannot agree with the statement in l.6, p.8, that "the overall agreement is rather good". The authors try to confirm the agreement by providing a statistical comparison over 5 year: this is quite a long time, and I don't think that the relatively good agreement found between ground-based and CALIPSO for STS, NAT mixtures and ice may provide any real evidence of the agreement between both datasets. I guess it rather gives an overall probability to find a specific PSC type above Mc Murdo, which is something quite different. For the enhanced NAT mixtures, the situation is even worse since there is about a factor of 2 between the statistics, despite the long time period. Results presented in Figures 2 and 3 are also calculated as averages over a five-year time period, so that they don't bring more evidence on the agreement between ground-based and CALIOP measurements. Hence, as suggested by the authors higher in the text, the difference in measurement rate and coverage, different geometry and measurement protocols may induce significant biases in the PSC classification. Did the authors compare directly coincident measurements at specific very limited periods ? Even if, as explained by the authors in l.5-6, p.7, a point-to-point profile comparison may be unsatisfactory, we should expect that a comparison within a short period shows similar patterns in both plots.
- L. 3, p.8: "at the core of the PSC winter season": it might be useful to mention the corresponding period in terms of months.

- L. 1-5, p.11: I don't see how the different geometries could justify the differences in the results, since Figure 2 presents PSC fractions, and not absolute values. It can be argued that CALIPSO will be more sensitive at high altitude and the ground-based lidars at lower altitude, but I guess this applies to all kinds of PSC. Hence, it is conceivable that the total number of observed events could be affected, but probably not the PSC fractions. Concerning the differences in statistics, how do the authors expect them to influence the agreement between datasets ?
- L. 3-4, p.12, Figures 2 and 3: What can explain that the the temperature dependence of the NAT fraction max agree quite well between CALIPSO and ground-based measurements (Figure 3), while the same NAT fraction are so different at some altitudes, e.g. around 20-22 km (Figure 2) ? It is unlikely that the number of events is too small at these altitudes to make the estimated fractions statistically not significant.
- L. 8-10, p. 12: I don't understand this conclusion: the differences are manifest on Figure 2.

3. Comparison of CALIOP PSC observations in the Southern Hemisphere with CCM simulations

- L. 17-31, p.12: The resolution should be mentioned for the different models and datasets. Resolution aspects play most probably a crucial role in the comparison between models, and with CALIPSO (See also comments on L.4, p.17 and Figure 7).
- L.14-15, p.13: Which kind of threshold do the authors apply to the SAD when applying the observation operator ? Do the authors mean that they use a mask recording the amount of lidar measurements in every grid cell and putting to zero all grid points that are not covered by any lidar presence ?
- L. 16, p.13: The formulation is confusing: is "the sum of all layers" an amount of layers or a distance in km (= amount of layers x 1.5 km) ?
- Caption Figure 4: "the number of km": Please be more specific: does it concern the altitude range ?
- L. 6, p.14: What do the authors mean by "NAT-like" ? The ensemble NAT mixtures + enhanced NAT mixture ?
- L.1, p.17: Are there no reasons to think that it is the CALIPSO PSC frequencies that are underestimated with respect to the reality ? I have in mind the way the statistics are processed, the use of monthly means, and the characteristics of the CALIPSO/ground-base station coverage.
- L.4, p.17 and Figure 7: "a very large underestimation": with respect to what ? In July, it is very similar to WACCM-cmmi, and very similar to WACCM in August. In September, LMDZrepro is much larger than WACCM. The "very large underestimation" is certainly not general when considering the total PSC frequency. However, it is true when considering the SAD criteria (Figure 7). It has to be noted that LMDZrepro gives overall the closest to CALIPSO in both cases (Total PSC frequency and SAD). Would the similarity with

CALIPSO and the outlier character with respect to the other models in the case of the SAD diagnostic be related to the coarser grid resolution of the LMDZrepro model with respect to the other models ?

- L.5, p.17: "The largest biases are found for ice PSCs that tend to be significantly overestimated": Do the authors mean: "underestimated" ? I guess they are still considering the LMDZ model ?
- L. 7-8, p.17: Taking into account the difference in assumptions, what is the reliability and the robustness of such diagnostic method ? A sensivity study might be needed.
- L. 6, p.18: "This in turn would give less irreversible denitrification processes than in the case of simulation by the models with larger NAT SAD" ?
- L.4, p.19: occurences of what ? Please be more specific.
- L. 6, p. 19: How is the averaging performed ? As a simple mean of all numbers ? Or by weighting by the grid cell area ? Concerning CALIPSO, how do the authors use the monthly means ? By making a mean of means ? Averaging yet averaged values may affect significantly the results.
- L. 10-12, p.19: "Too slow", "too fast": with respect to CALIPSO ? This should be specified. What do the authors mean by "progression for ice/NAT" ?
- L. 1, p.20: "The fraction of data with different PSC": Please revise the formulation.
- L. 3, p.20: the fraction of what ? Please be specific ! "an increase of ice with T-TNAT < -5K": Please revise the formulation: increase with decreasing temperature.
- L.5, p.20: "a sharper increase of the fraction": fraction of what ?
- L. 7, p.20: "while for the other models, the ice…".

4. Conclusions

- L. 12, p.20: A point-to-point comparison is always feasible ! The issue is to know if it is valid and reliable.
- L. 14, p. 20: "very similar": Based of the results presented in Figure 1, I don't agree. (See comment above). At least, a statistical indicator and quantitative estimates of the uncertainty should provided.
- L. 16, p.20: As already mentioned, I don't understand the emphasis on "below15 km". Is it based on Figure 2; If well, this seems very subjective to me.
- L. 16-17, p.20: "rather good above 15 km": this looks particularly subjective. At least, the "rather good" should be quantified in some way !
- L. 20, p.20: "Models fail to reproduce realistic geographical distributions of PSCs": I am really not convinced by the demonstration made in this paper. A significant part of the problem might come from the way the authors implement their different methodologies, and more particularly from the comparison of things that are not really comparable.
- L. 22, p.20: The more recent WACCMI-ccmi model compared better with CALIOP only for one specific diagnostic method (based on the total PSC

frequency). The issues is to understand why: in view of all my previous criticisms, it might be fortuitous.

Technical corrections

- L. 32, p.2: “The most recent…”: “CCM” could be introduced at the first occurrence of the expression “Chemistry Climate Models”, in L. 15.
- L.11, p.3: the acronym CALIOP has already been expanded above.
- L.2, p.4: “acquisition”.
- L.18, p.4: remove one “of”.
- L. 14-20, p.5: Putting “Data pre-processing”, “PSC detection”, and “PSC composition” in subtitles, or putting the ensemble in a table might simplify the layout and make the reading easier.
- L. 32, p.5: “The dataset provides” ?
- L.4, p. 6: I am not sure that “reelaborated” is correct English.
- L. 9, p.6: “ground-based”.
- L. 13, p.6: “corresponding”.
- L. 16, p.6: “5-year”.
- L. 18, p.6: “the differing measurement procedures used for each of them” might be more correct.
- L.19, p.6: [the procedures] “induce”.
- L. 2 and 6, p.7: “signal-to-noise ratio”.
- L. 8, p.11: I guess the sentence is incorrect. What is “the it” ?
- L.17, P.11: I am not sure that T_{NAT} is defined yet.
- L.1, p.12: “a similar behaviour”.
- L.6, p.13: I guess the latitude range mentioned concern the southern hemisphere. Latitude values have thus to be indicated with a “S” or with negative numbers.
- Caption Figure 4: Missing point at the end of the sentence.
- L. 1, p.14: “90°-0°”:Please specify: W or E ?
- L. 6, p.14: “0°90°”: Please adapt the notation (cf. remark on L. 1, p.14 + use “-“).
- L. 1, p14-l.7, p.15: It could ease the reading to use separated paragraphs for each PSC type.