Response to RC2 of egusphere-2023-2235

First, we want to thank the second referee for reviewing our manuscript. In the following, we will respond to the given suggestions and questions and describe the changes made within the manuscript corresponding to the comments in detail.

General Comments:

1. The authors make the claim that they are "validating" retrieval results in several places in the introduction. Given the way they are comparing "model" and "retrieval" results I am not sure that is entirely the case, as there is no evaluation of whether polarized reflectance misfits are related to "errors" in retrieved properties. There is also no evaluation of the effects of noise on the retrievals. Nonetheless the use of a simple model sampling scheme to relate observations to model variables is of considerable value for understanding how to use polarized observations for model evaluation and should probably be more emphasized.

   Thank you for noting the perhaps misleading use of the term "validate" in the submitted paper. We changed the affected formulations to weaker terms, for example "evaluation of the accuracy of the retrievals". The arguments of the referee that there is no full validation of the retrievals are understandable. But in fact, measurement noise was considered in our evaluation: As explained in Sec. 3, the noise of the Monte Carlo simulations was set to 6%. This is slightly higher but on the order of the recently found total radiometric uncertainty for the two polarization cameras of specMACS of 3.8% to 5.8% (see Weber et al, https://egusphere.copernicus.org/preprints/2023/egusphere-2023-2209/egusphere-2023-2209.pdf) and could of course lead to higher variations in the aggregated polarized radiance signal. We also simulated a subset of the image for the full overflight with a smaller noise of 3% previously to the submission of this manuscript, and did not find significantly different results for the retrievals. In addition, the comparison between the real measurements and the simulations in Sec. 4 shows that the simulations performed are realistic and thus, they were treated as if they were real measurements. We emphasized the point mentioned in this comment at the end of the Introduction (l. 118f.).

2. The reverse Monte-Carlo sampling of singly scattered photons described at the beginning of Section 5 is the crux of the paper and presents a reasonable and simple way of sampling model output, in order to evaluate it against polarimetric observations. However, the rationale for the use of "the average of all scattering event locations" (line 199) in sampling the model fields and its consequences should be discussed. For example, a "model" simple average, unweighted by the probability of the path will tend to give a sample that is biased deeper into the cloud than that of the optical signal if singly scattered light dominates. In contrast, if multiply scattered light

(e.g. multiple forward scattering events caused by the diffraction peak in the phase function) dominates the signal the "model" simple average will give a sample that is biased higher in the cloud than the optical signal. This difference between the "model" sampling and what one expects to be the source of the optical signal is one potential explanation for the compensating biases in the stereo cloud top height results. Some discussion of the single scattering "model" versus full Monte Carlo sampling would therefore be helpful.

The authors agree with the argument given by the referee that taking an unweighted average of the scatter event locations for the determination of the "expected" model quantity results in biases with respect to the actually observed quantities. From the backward Monte Carlo simulations of singly scattered photons, we get the locations of the last scatter events of photons on the way from the sun to the detector weighted with $\exp(-\tau)$ along the line of sight. Hence, the "average of all scattering event locations" is not an unweighted average. Nevertheless, one can discuss about the influence of multiple scattering on the optical signal detected at the sensor and thus seen by the retrieval algorithms and how its neglect will bias the expected model quantities. To begin with, the polarized signal of the cloudbow is generated by both single and multiple scattering, the latter, however, has no significant influence on the angular structure of the cloudbow and hence the droplet size distribution retrieval (see Sec. 3 and Alexandrov et al. (2012)). Therefore, it is reasonable to consider the distributions seen by singly scattered photons. In contrast, the stereographic retrieval uses simple intensity measurements, such that multiple scattering might dominate the signal. However, it is based on the identification of contrast gradients. Those will decrease significantly, when multiple scattering becomes more important as the signal will be smoothed. We plan to study this in future by addressing the influence of the different scattering orders on the contrasts of the image. We noted this as a potential bias in the manuscript in Sec. 5 (l. 241f.) and discussed it further in l. 298f. Moreover, the stereographic cloud top heights are used for the geolocalization of the cloud targets evaluated by the cloudbow retrieval. Hereby, a small error in the cloud top height can already affect the aggregation of the cloudbow signal significantly as explained by Pörtge et al. (2023) as well as in the Introduction (l.70f.) and in Sec. 5. Thus, the comparison of the stereo heights to the expected ones from the single scattering simulations gives valuable information about the uncertainty of the height information used for the cloudbow retrieval. We pointed this out in line 299 onward.

3. Figures 4f & l, 5f and l and 6f and l are not useful. The statistics of the regressions are informative but for the reader presenting histograms of differences would be a more effective use of the graphic.

   The suggestion made by the reviewer to present the difference histograms was implemented in the manuscript including the mean and the 1σ-interval.

4. As noted in point 2 there will be a difference in the vertical weighting of the "model" sample and that which is expected from the polarimetric retrievals. It is particularly important to note this when making a comparison between the effective variance

retrievals and the model since the effective variance itself is constant. This means that there is no underlying signal, and the comparison is primarily a comparison of differences in sampling. An additional point regarding the effective variance comparison is that the retrievals have clearly failed when veff=0.32. Some additional comment on this and ideally examining whether there are a particular range of effective radii where this failure occurs would be desirable.

Correct, because of the constant effective variance throughout the model domain, there is no natural variation in the effective variance and we have not tested the retrieval on variations in the effective variance itself for now. For that, a parametrization to derive the effective variance from the outcome of the LES model would be needed. We plan to include this in future studies. As explained in l. 408f now, the variation in the effective variance which is currently expected from the model input is only due to variations in the effective radius within the sampling volume.

Concerning the cases where veff=0.32, we performed a slightly deeper analysis, however, as pointed out from l. 423 onward, no significant correlation between retrieved effective variances of 0.32 and other parameters was found. Concerning the question of the range of effective radii, it can be said that veff=0.32 occurs for nearly all effective radii retrieved as can be seen from Fig. 7a.

5. While I do not think that additional simulations are in order the authors should note that an effective variance of 0.1 is quite large for a cloud top size distribution and this should be born in mind when planning future work.

   Thank you for pointing this out, we will certainly keep that in mind. For the EUREC4A campaign, effective variances on the order of 0.1 were found. Since we are continuing these model-based evaluations, we will use observed veff distributions in future studies.

Editorial Suggestions

Rewrite Eq.(6) as v_eff_tot= veff + (reff_4*reff_2/reff_3^2-1)*(1+veff) where reff_4 is the 4th moment of the effective radius etc. I suggest this because it is then clear that sampling variability in reff can only increase the apparent effective variance.

Where the Marshak et al. (1998) paper is cited at line 121 it should be noted that there conclusions are for overcast clouds. E.g. insert the word overcast between "for" and "marine" on that line.

Both editorial suggestions were applied to the manuscript.