Response to reviewer 1

We would like to thank the reviewer for his/her important comments and suggestions.

The manuscript presents a neural network-based approach to above-liquid-cloud aerosol retrievals from the multi-angle polarimetric measurements by PARASOL. The method utilizes 3 separate NNs: one to determine if there is a liquid cloud layer, one to perform the ACA retrieval, and one for an approximate forward model for goodness-of-fit evaluation. Seasonal distributions of above-liquid-cloud aerosol optical thickness appear to generally agree with past studies. However, the manuscript lacks crucial details and tests that preclude its publication in the present state. Detailed comments are provided below.

☐ 0. A high-level comment: there are some acronyms that are defined in multiple places in the manuscript (for example, see comment #3), as well as places where the defined acronyms are not used, e.g., lines 14-15 spell out "above cloud aerosol" and "direct radiative effect" rather than using the ACA and DRE acronyms that were defined in the sentences before it. It would be good to define these acronyms only at their first usage and consistently use them throughout the manuscript.
We have a thorough check among the whole text and revised them

☐ 1. Line 7: "ACAOT (above cloud aerosol optical depth)" should be changed for consistency, either to "ACAOT (above cloud aerosol optical thickness)" or "ACAOD (above cloud aerosol optical depth)".
We changed ACAOD to ACAOT and use it consistently throughout the revised manuscript

☐ 2. Line 64-65: "The level 1 data are provided on a common sinusoidally grid of approximately with ground pixels of approximately 6 × 6 km2." - This sentence phrasing is confusing and I am not entirely sure of its intended meaning. Please rephrase it to more clearly convey the intended meaning.
We rephrased it as “The level 1 data are provided on a 6x6 km2 sinusoidally grid.”(line 69 of the revised paper)

☐ 3. Line 89: The AOT acronym was defined earlier in the text on line 45, it doesn't need to be re-defined here.
We removed this definition.

4. Section 3.2: While this section does a great job explaining how the training data were produced, it lacks sufficient detail for other aspects relevant to NN training:

☐ a) Lines 154-155: Real measurements would have noises that vary across different measurements. Why is the noise assumed to be a constant relative noise of 0.02 for the intensity and a constant absolute noise of 0.012 for DoLP? These seem arbitrarily chosen, based on the provided information. This also suggests that the model will not be as accurate when noise levels deviate from these assumed values. Were other noise levels considered? See also comment #5d below.
This noise setting is only used in the calculation of chi2 (equation 2 in the preprint), but for the training, a variable intensity noise of 1-3% is considered while the noise of DoLP is constant at 0.012. We investigated different noise settings on intensity and DoLP (fixed value and variable range) in the training, both in this work and previous work (Yuan et al 2024) and the current setting produces the best results for cloud fraction retrievals (Yuan et al, 2024) as well as above-cloud aerosol retrieval (this paper), although the effect of the noise setting is minor. We would like to note that also the two main aerosol retrieval algorithms for PARASOL (GRASP and RemoTAP) assume a constant noise for evaluation of goodness-of-fit and during the inversion process. See Hasekamp et al. (2024) and references therein.

☐ b) Lines 162-163: How was the ensemble size for each model component determined?
This was empirically determined using synthetic experiments. We added this information to the revised manuscript (line 184)

☐ c) Lines 159-161: Regarding the description of NN ensembles, the current phrasing suggests that the described approach ("the whole training set is equally and randomly divided into several parts") is the only way to perform this, but in reality there are many other methods to achieve NN ensembles that do not follow this procedure (see, e.g., Dietterich, 2000). Please rephrase this sentence so that it is clear that this is your elected methodology to achieve an ensemble of NNs, not that this is the only methodology to achieve it.
We have rephrased it to “in our approach, the whole training set is equally and randomly divided into several parts (line 179 of the revised paper)

☐ d) Lines 165-166: For the cloud mask and retrieval NN, how was the measurement noise model determined? These values seem arbitrarily chosen based on the provided information.

The noise setting is inherited from Yuan et al 2024, where different noise settings are investigated for cloud mask from PARASOL measurements (see also our response above).

- ☐ e) Line 169: How was the batch size of 12,000 selected for this problem, and were other batch sizes considered? This is an unusual value, as typical batch sizes are chosen as 2 to some power, for efficiency when using a GPU (see, e.g., Kandel & Castelli, 2020). This is also unusual given the magnitude, as batch sizes are often significantly smaller than this; existing literature suggests that small batch sizes perform better (e.g., Bengio, 2012; Masters & Luschi, 2018; Kandel & Castelli, 2020).
  Indeed, a smaller batch size can help in increasing the generalization and decreasing the memory used in the training process. Nevertheless, a large batch size benefits the convergence rate (Soham De, et al, 2017). Compared to applications such as image processing, one training sample in this paper consumes less memory, which makes it possible to use a larger batch size. We did several tests for different batch sizes (from 512 to 20000) and didn't find significant differences over the NN's performance. We added a statement on it in the revised paper (line 190)
- ☐ f) Line 170: "mean root square error (RMSE)" should be fixed to "root mean square error (RMSE)".
  Corrected.
- ☐ g) Lines 170-172: How were the model architectures determined? It is surprising that a given model has the same number of neurons in each layer, and that all 3 models have the same number of hidden layers.
  1) We don't find an obvious increase of performance from using different number of neurons in each layer in this application. However, using different number of neurons in different layers will lead to a non-square weight matrix, which has a potential risk of rank diminishing in forward and backward propagation. 2) In all the NN retrieval experiments (from PARASOL measurements), we found a worse performance (from the holdout set) when the number of layers increased or decreased.
- ☐ h) What activation functions were used, and how were they determined?
  We use ReLU as activation function. but we have tried different activation functions (ReLU, leaky ReLU, sigmoid, tanh, etc) and no obvious performance difference is observed. Therefore, for computational efficiency, we chose the "simplest" ReLU as activation function. We mentioned this in the revised paper (line 193)
- ☐ i) What learning rates were used to train these models? How were those learning rates determined?
  Initial learning rate is 0.01 which is default and suggested for the Adam optimizer. The Adam optimizer changes the learning rate based on the convergence situation at each

iteration step and is not very sensitive to the initial value of the learning rate. We added a description on it in the revised manuscript.

☐ j) Line 119: How were the 8 million samples split into subsets for training, validation, and testing?
Ninety percent of the samples are in the training set and the rest are in the holdout (test) set. The holdout set is only used for assessing the training process. The validation/evaluation of the NN approach is based on an individual data set that is independent from the training procedure. We added the information to the revised script.(line 180)

☐ k) Line 144: How did you determine the number of leading PCs to use for each variable? How much explained variance do these PCs comprise? Did you first attempt this without using PCA and find poor results?
We tried different values to find the best settings (from the performance of synthetic experiments), and we also tried without PCA, which performs not as good as the current setting. For intensity, 33 PCs comprise 99.99% explained variance. For DoLP, 25 PCs comprise 99.14% explained variance. We included this in the revised paper (line 157).

5. Section 4.1 is missing some key synthetic tests/results:

☐ a) What is the confusion matrix for the NN liquid cloud mask model? This is a crucial table to provide for any classification NN to better understand the rate of true/false positives/negatives and thereby the reliability of the NN for this task.
We added a confusion matrix (as shown below) to the supporting information:

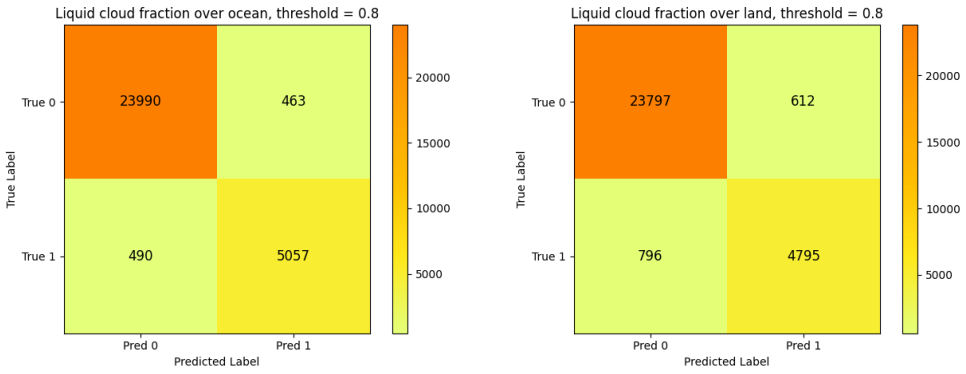


Fig 1. Confusion matrix of liquid cloud mask, left is over ocean and right is over land.

☐ b) What is the accuracy of the goodness-of-fit determination by the NN forward model? For a given state vector retrieved by the NN retrieval model, a forward model can be computed by either the NN forward model or the physics-based forward model. Doing

both for all cases considered in this synthetic test will enable the determination of the reliability of the NN forward model for this task based on metrics like the correlation coefficient, coefficient of determination, RMSE, etc.

Below we show the comparison of intensity and degree of linear polarization (DoLP) between NN forward model and RemoTAP forward model, at 565nm. The rstd (relative standard deviation) of intensity is 0.7% and the std (standard deviation) of DoLP is 0.0025, both of which are below the instrument measurement noise. This suggests that the NN forward model is good enough to replace the full physical model (RemoTAP) in estimation goodness-of-fit. We added the figures in the revised manuscript (Figure 2 of the revised manuscript).

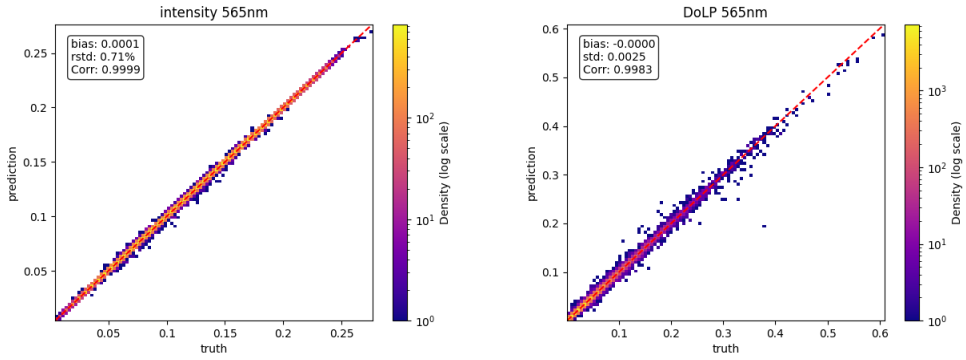


Fig 2. Intensity (left) and degree of linear polarization (DoLP, right) from NN forward model (prediction) and RemoTAP forward model (truth) at 565nm.

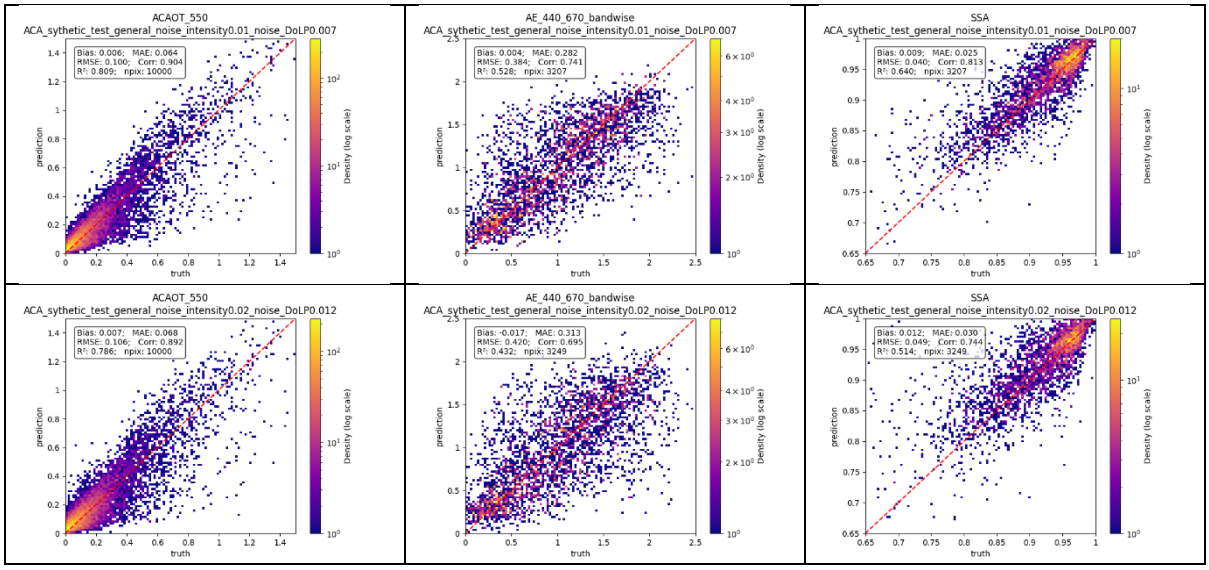
☐ c) The correlations (I assume the R correlation coefficient? Please be explicit) reported here suggest that the NN model poorly captures AE behavior, especially under fine- or dust-dominated conditions. This is also the case for SSA under dust-dominated conditions. This seems to suggest that the retrieval method is only weakly sensitive to these parameters. How does this compare with physics-based retrievals of these parameters under these conditions? If it similarly struggles, it would be good to discuss this and clearly point out that this is not a limitation of the NN approach. However, if the physics-based retrievals do not have this issue, then it suggests that the presented NN-based approach is not optimal.

R indeed indicates the correlation coefficient. We have better clarified that in the revised manuscript. The relatively low correlation coefficients for the fine- and dust-mode dominated cases can largely be explained by the small range in AE, where most values are between 1.5-2,0 for for fine- dominated cases and between 0-0.5 for dust-dominated cases, where this range is close to the retrieval accuracy itself (RMSE of ~0.4). So, it means little capability to distinguish size within these categories, but clear capability to distinguish between fine-mode dominated and dust-mode dominated

cases (as shown in the top panel). A similar explanation holds for the dust SSA. Compared to clear-sky full-physics retrievals (Hasekamp et al, 2024), the performance for SSA of our above-cloud retrievals is similar. The performance for AE is worse than for clear-sky full physics retrievals, but this is most likely because of smaller information content for AE for above-cloud aerosol retrievals than for clear sky retrievals. Namely, performing clear sky aerosol retrievals with an NN gives similar results to full physics (this is current research ongoing in our group).

☐ d) Related to comment #4a: How does the model perform when applied to synthetic data with a different noise level than that assumed in Sec. 3.2? It would be helpful to understand how this impacts the results, given that real measurements will not exactly follow the noise model assumed when generating the training data.

Below we show three different noise level on the synthetic dataset with both fine and dust mode aerosols. It can be seen that increasing the measurements noise leads to the increase of retrieval error on all the three properties, but the change is small in this noise range (intensity noise 1-3% and DoLP 0.007-0.017). We include those figures in the supporting information (Fig 6 of SI).



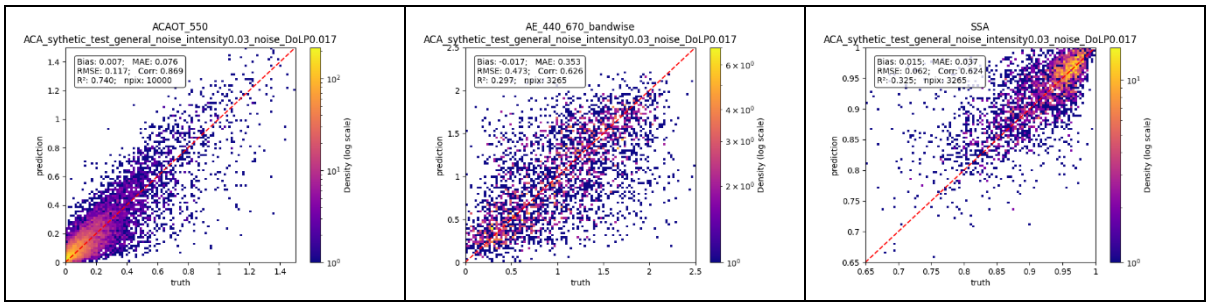


Fig 3. ACAOT 550nm, AE 440-670nm and SSA 550nm on the same synthetic dataset with different noise levels. The first line's noise settings are 1% to intensity and 0.007 to DoLP. The second line is 2% to intensity and 0.012 to DoLP. The third line is 3% to intensity and 0.017 to DoLP.

6. Section 4.2 / Figure 3:

☐ a) For Fig 3's a-c panels, it looks as if only a single retrieval was performed at the 4 chosen optical depths. I assume this is really showing the **mean** RMSE over all 10,000 retrievals at a given optical depth, rather than the RMSE of a single retrieval as the labeling and caption currently indicates? For clarity it would be good to include error bars showing the standard error of the mean, which will better show whether the observed trends are statistically significant. Please also update the caption to clarify that these RMSE values are averaged over the 10,000 cases considered at each COT value.

The RMSE is calculated over all the 10000 retrievals for a single wavelength (550 nm). Below shows the ACAOT 550, AE 440-670 and SSA 550 error (prediction - truth) as a function of COT, the error bars stand for the standard deviation of all the retrievals in a COT value. The figure indicates that the retrieval error on the dataset is dominated by the standard deviation. We included this information in the caption of Figure 4 in the revised manuscript, but prefer to keep the plot with RMSE instead of the plot below.

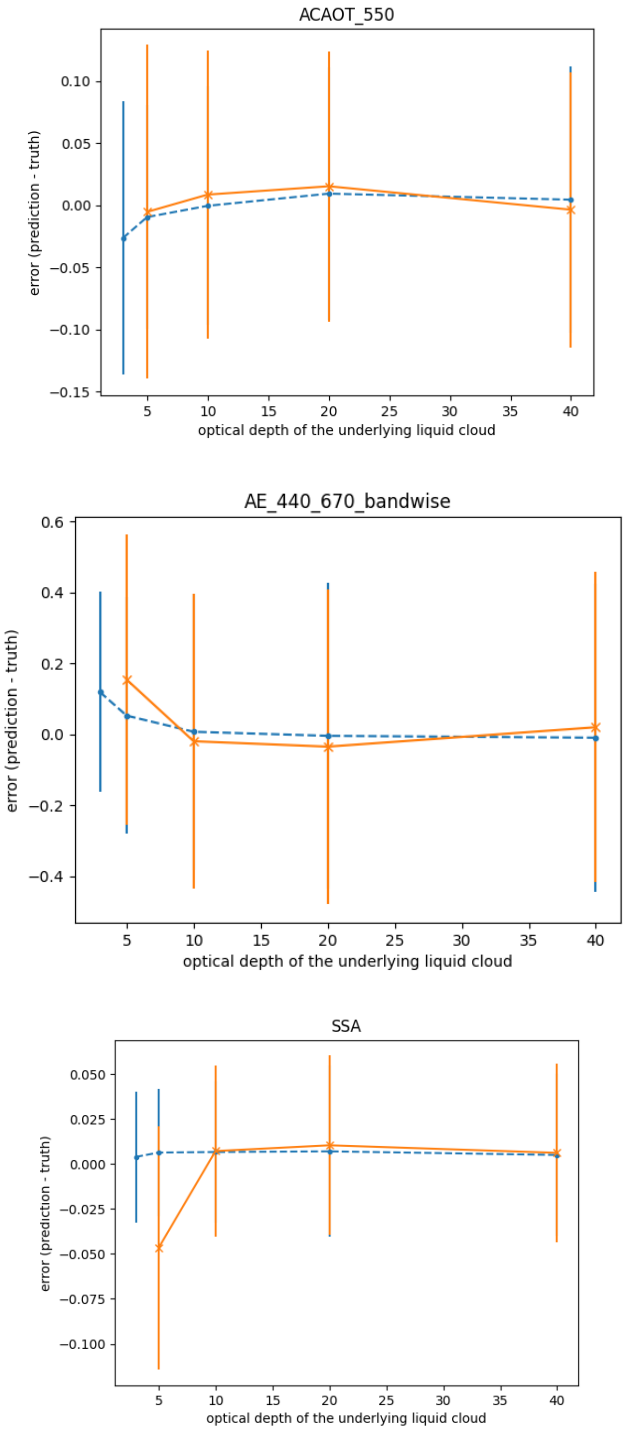


Fig 4. Error and standard deviation of ACAOT, AE and SSA retrievals in different COT bins. Blue lines are over ocean and orange lines are over land.

☐ b) Am I correctly inferring that the "number of remaining pixels" means the number of cases that were not screened out by the retrieved cloud fractions or goodness-of-fit

metric? Assuming this is the case, rather than reporting the absolute number of pixels in Fig 3's d-f panels, it would be clearer to report the fraction of pixels where retrievals weren't screened out or, conversely, the fraction of pixels that were screened out.
We changed it to “fraction of successful retrievals” which is the fraction of pixels not screened out.

☐ 7. Lines 219-221: I don't necessarily agree with this statement. Looking at SSA, the synthetic test found a correlation of ~0.76, with the correlation lowest for the dust-dominated cases at ~0.56. Meanwhile, in Fig 4, the SSA correlation is even lower at ~0.37. Fig 4 also shows a >27% increase in RMSD for SSA when compared to the RMSE of the synthetic tests. From this, I would conclude that the AE is more comparable for above-cloud and adjacent clear-sky retrievals than AOT or SSA.
Indeed the AE looks most comparable between ACA retrievals and clear-sky retrievals among the three properties (ACAOT, AE and SSA). We rephrased it to “the intrinsic aerosol properties (AE and SSA) are more comparable for above-cloud and adjacent clear-sky retrievals than the AOT, although the correlation of SSA is low (0.37)” (line 250)

☐ 8. Line 245: "The high AE and low SSA is an expected feature of the smoke in mid-Africa." Can you please provide a reference for this statement?
Mallet et al 2024 (https://doi.org/10.5194/acp-24-12509-2024) clearly stated this feature. We included this in the revised manuscript. (line 304)

☐ 9. Lines 245-247: Such a large disagreement in AE suggests some fundamental or systematic difference between the method considered here vs. that considered by Waquet et al. (2013), and it would be good to understand the origin of this difference. Is there a reasonable explanation why the reported AEs are less than 1/2 that reported by Waquet et al. in these midlatitudes? Does Waquet et al. also only consider above-liquid-cloud AE? If not, given that above-ice-cloud AE is disregarded here, is that perhaps a common occurrence and the results of Waquet et al. are elevated due to that above-ice-cloud AE? Does the Waquet et al. approach overestimate ACAOT, or does the RemoTAP method considered here underestimate AE? Did Waquet et al. similarly consider the differences between above-cloud AE with adjacent clear-sky AE? Hopefully these questions can help to better elucidate the origin of this stark difference.

In the revised manuscript, we performed a more detailed comparison with AERO-AC (setction 5.2, see also our response to reviewer-2)

☐ 10. Lines 255-257: I don't agree that the synthetic experiments indicate the NNs have the ability to retrieve AE from fine- and dust-mode-dominated aerosol. Figure 3 shows that the performance is poor in these regimes: the NN underestimates the AE for fine-dominated cases, and it overestimates the AE for dust-dominated cases.
See our response above. We rephrased it to: ... the NNs have to ability to retrieve AE that allows separation between fine-mode and dust dominated cases ..." (line 319)

11. Lines 273-275: "... the NN-based surrogate forward model, just like the full-physical model, can provide goodness-of-fit mask to filter unphysical retrievals, which may due to imperfect cloud mask or some challenging aerosol/cloud/surface combination.":

☐ a) I don't see where this is substantiated in the manuscript; no comparisons are performed between the NN's goodness-of-fit calculation and a physics-based model's corresponding goodness-of-fit metric. Please perform the test in comment #5b to substantiate this claim.
We added the evaluation of NN forward model to the revised manuscript (see above).

☐ b) The last clause of this sentence is phrased awkwardly, and I cannot discern the intended meaning in the context of the rest of the sentence. Please rephrase this so that it clearly conveys the intended meaning.
We rephrased it as "... unphysical retrievals. Those unphysical retrievals are caused by reasons such as imperfect cloud screening or challenging aerosol/cloud/surface combinations" (line 342)