Review of "Validation of cloud macrophysical properties from the ATLID Level 2a products A-TC, A-FM, A-CTH using airborne lidar observations during the HALO missions PERCUSION, ASCCI and NAWDIC"

by K. Krüger, M. Wirth, A. A. Floutsi, D. P. Donavan, G-J. van Zadelhoff, and S. Groß

https://egusphere.copernicus.org/preprints/2026/egusphere-2026-3051/

Response to Reviewer #1:

This manuscript presents a comprehensive validation of the EarthCARE ATLID Level 2 products - specifically the A-TC, A-FM, and A-CTH - using an extensive multi-campaign airborne dataset. The study is well-structured and demonstrates high scientific quality, providing a clear and methodical assessment of cloud macrophysical properties across diverse tropical and extratropical regimes. This research is important to the broader scientific community, especially for users of EarthCARE and airborne lidar data. Furthermore, the identification of systematic biases (e.g. the ice cloud overestimation in A-TC and the cloud top height discrepancies in A-CTH) provides critical feedback to the developers of the of the retrieval algorithms. I recommend that this paper be accepted for publication after addressing the (mainly minor) revisions discussed in the following comments/questions.

We are pleased that you gained a positive impression of our manuscript and recommend its publication in AMT after revisions. We sincerely thank the reviewer for taking the time to read our manuscript carefully and for the many detailed comments and suggestions, which have greatly helped us improve the manuscript.

We have considered all your comments and will give a point-by-point response. Answers are given in italics and bold. Text changes are marked in italics, bold, and blue.

1. "Spreading Effect": In the abstract (lines 25-26) and the Summary (line 820), the authors refer to a spreading effect as a primary cause for the overestimation of ice cloud occurrence. Does this spreading effect have to do with averaging or interpolation?

We consider the term "spreading effect" too vague in the Abstract and not sufficiently informative for the reader. The underlying reasons for the observed ice-cloud overestimation are discussed in detail in the Results and Discussion sections. We therefore decided to revise this sentence in the Abstract as follows:

"This overestimation is caused by ice clouds that are horizontally and vertically too large, but also, to a certain extent, by aerosol pixels incorrectly classified as ice clouds (at temperatures >0°C), and to a spurious occurrence of stratospheric aerosol features near the tropopause."

2. In Section 5.2 (lines 759–760) and the abstract (lines 28–30), the authors link the low-altitude cloud top height underestimations specifically to surface misidentification over the ocean. However, given that Figure 1 illustrates that the majority of the flight data were acquired over oceanic regions, it is unclear if this is an inherent ocean-specific retrieval issue or simply a result of the sampling distribution. Please clarify whether this would similarly happen over land surfaces if the sampling were equivalent, and consider

qualifying the statement to reflect that the observation may be a consequence of the mission's oceanic flight coverage.

***Indeed, most of our flights cover oceanic regions, which is why we state this so explicitly. However, our analysis of underpass segments above land masses provides indication that the issue is not exclusively associated with ocean surfaces. For example, the persistent fog situation (shallow stratus clouds) observed over Germany and Sweden during PERCUSION in November represent relatively "easy" targets for separating cloud and surface returns. Nevertheless, in broken cloud structures of these cases, extremely low cloud top heights occur in A-CTH.* As described in Sect. 2.4.3 A-CTH applies an altitude-dependent threshold-based WCT, so low-level CTH are especially challenging due to the low(er) signal-to-noise-ratio at that altitude range.**
Since we write "particularly over ocean surfaces," we see nothing wrong with this statement and have left it unchanged in the abstract. But we have revised the sentence LL 759–760 accordingly so as not to omit this statement of false detections over landmasses:

"These very low clouds are found in tropical and extratropical frames predominantly over the ocean, but also occasionally over landmasses (…)"

3. In lines 74–79, the authors discuss the limitations and advantages of validation methodologies. To provide a more complete picture of the current state of EarthCARE product validation, I suggest acknowledging and citing recent works that utilize ground-based networks (e.g., ACTRIS). Including references such as Baars et al. (2026) and Feuillard et al. (2026) would help readers better understand how airborne campaigns (like the one in this study) complement ongoing ground-based validation activities.

We agree that validation with airborne and ground-based validation are complementary approaches, and that both are required for the validation of EarthCARE products. To better reflect this, we included Baars et al. (2026) and revised ll. 74-77 accordingly (see below).

"Validation of spaceborne instrument products is commonly performed by comparing with observations made by ground-based or airborne instruments (Amiridis et al., 2025). Ground-based measurements, e.g., from the Polly^{NET} lidar network (Baars et al., 2016), enable continuous monitoring of EarthCARE products over extended time periods in particular regions (Baars et al., 2026). However (…)"

New reference used in the revised text: Baars, H., Kanitz, T., Engelmann, R., Althausen, D., Heese, B., Komppula, M., Preißler, J., Tesche, M., Ansmann, A., Wandinger, U., Lim, J.-H., Ahn, J. Y., Stachlewska, I. S., Amiridis, V., Marinou, E., Seifert, P., Hofer, J., Skupin, A., Schneider, F., Bohlmann, S., Foth, A., Bley, S., Pfüller, A., Gian nakaki, E., Lihavainen, H., Viisanen, Y., Hooda, R. K., Pereira, S. N., Bortoli, D., Wagner, F., Mattis, I., Janicka, L., Markowicz, K. M., Achtert, P., Artaxo, P., Pauliquevis, T., Souza, R. A. F., Sharma, V. P., van Zyl, P. G., Beukes, J. P., Sun, J., Rohwer, E. G., Deng, R., Mamouri, R.-E., and Zamorano, F.: An overview of the first decade of PollyNET: an emerging network of automated Raman-polarization lidars for continuous aerosol profiling, At mos. Chem. Phys., 16, 5111–5137, https://doi.org/10.5194/acp 16-5111-2016, 2016.

4. In line 116, the authors say that the A-CTH product is evaluated against a "postulated accuracy requirement of 300 m." Is this a specific mission requirement (e.g., the EarthCARE Mission Requirements Document or the cited Wandinger et al. (2023) paper) or another mission document?

The 300 m requirement originates from a pre-feasibility study conducted during Phase A and is first formulated in the EarthCARE Mission Requirements Document (Wehr, 2006) and taken up by Wandinger et al. (2023). We have decided to reference the EarthCARE Mission Requirements Document as the primary source and have included this reference in l. 118:

New reference: Wehr, T. (Ed.): EarthCARE Mission Requirements Document, Earth and Mission Science Division, European Space Agency, https://doi.org/10.5270/esa.earthcare-mrd.2006, 2006.

5. It is clear that WALES is a well-established instrument, but it would be beneficial to include a brief statement regarding the calibration and maintenance procedures, specifically how the system ensures long-term stability and accuracy across different campaigns. Additionally, if the WALES dataset has been utilized in previous satellite validation studies, I suggest acknowledging these contributions with relevant citations.

We would like to clarify that the quality of the WALES data is continuously assessed for each individual profile and measurement volume using a range of diagnostic indicators. These include, among others, detector saturation, signal-to-noise ratio, spectral purity of the emitted wavelengths, laser performance, and the correction for aircraft roll angle. These diagnostics are applied to flag affected measurement volumes and to provide an uncertainty assessment. Quantitative studies focusing on the quality of the WALES HSRL data are already included in Sect. Importantly, these procedures are applied consistently across campaigns, ensuring that the quality of the retrieved products remains comparable and stable over the long-term.

Secondly. we want to note, that individual CALIOP underflights were carried out during individual flights in the past decade, however we were not part of the CALIOP validation teams, and there is no study focusing on validation of the CALIOP data quality based on WALES HSRL.

6. In Table 1, the authors provide the "leg distance" for each flight, defined as the horizontal distance flown along the EarthCARE ground track. To assist reader's understand what is this leg distance, it would be helpful to include a brief explanation of how it is determined and its significance to the overall statistical power of the study.

To resolve confusion about the leg distance, and the underpass location we revised the sentence in ll.194-195 and added another sentence.

"For each underflight segment, the goal was to position the satellite's overpass (see diamonds in Table 1) above the aircraft near the midpoint of the flight leg in order to achieve the best possible spatial and temporal collocation of the airborne and space-based observations. Across the three field campaigns, the 40 conducted underpass segments provide a large data set, spanning a cumulative flight distance of 33,480 km (Fig. 1, Table 1)."

Furthermore, we improved the caption of Table 1 as follows:

"Summary of the 40 coordinated EarthCARE underflight segments carried out during 35 HALO research flights. The corresponding flight dates, associated campaigns, the distance with collocated WALES observations along the EarthCARE ground track with corresponding orbit frame, the number of WALES BSR profiles (with >1km coverage) and the baseline of the ATLID Level 2 products evaluated in this study are shown."

Furthermore, given the variability in leg distances across the campaigns, I suggest adding a sensitivity analysis (or a brief discussion) on how this horizontal extent affects the

collocation statistics. Specifically, would increasing the number of collocated profiles or refining the spatial matching criteria significantly alter the identified performance biases, or is the current sample size sufficient to represent the observed atmospheric variability?

We are actually not sure what "adding a sensitivity analysis (or a brief discussion)" refers to. In general, longer underpass segments yield more data and thus potentially more clouds. However, the absolute number of clouds measured also depends heavily on the cloud type and the features that were targeted for the validation (aerosols/ice clouds/convective clouds). Groß et al., 2026, provide more context on the validation strategies.

In our analysis, we were well aware that representativeness decreases with increasing distance from or time elapsed since the overpass. Precisely for this reason, two approaches were pursued in this paper: 1. the inclusion of all data points along the entire underpass, and 2. data points with good matching (within 10 minutes of the underpass). The results were presented in Section 4 and in the appendix, and the sensitivity of the results with respect to representativeness was discussed (Section 5.3). The comparable results obtained for both validation approaches show that the documented product performance is not heavily affected by the chosen approach i.e. reduced sample sizes.

7. Lines 243-244: Could the authors elaborate more on these "3 bins" window (e.g., how many meters vertically or how many seconds horizontally the window spans)? Specifically, does the window cover a fixed vertical altitude range, and how does the 3-bin horizontal width relate to the 1-second temporal resolution? Why was this specific window chosen?

As described in Sect. 2.1, the horizontal resolution of the used WALES data is 1 s, which corresponds to approx. 200–250 m given a typical HALO aircraft speed at high altitudes. The vertical resolution of the applied BSR data is 15 m. Thus, the applied 3 × 3 window corresponds to ~600–750 m in the horizontal and 45 m in the vertical.

The choice of a 3 × 3 window is not based on a fixed physical cloud scale, but represents a compromise between suppressing isolated spurious detections and still retaining small-scale cloud structures (that were also targeted for validation). Specifically, the 3 × 3 window makes the cloud detection more robust against isolated, one-pixel-wide enhancements in BSR. Such features, which generally occur only rarely, can result from measurement noise or from very small-scale cloud fragments (of 200x15m extent). For the latter, representativeness and comparability with the lower-resolution ATLID products can be challenging, as discussed in Sect. 5.2, and is not the focus of this study.

Previous studies using WALES BSR data (e.g., Urbanek et al.; Gutleben et al.; Dekoutsidis et al.) demonstrate that quantitative cloud detection from WALES BSR is also feasible without applying a 3 × 3 window. However, these studies addressed different scientific objectives, whereas the present approach was chosen specifically to ensure a robust and representative selection of cloud features for comparison with the ATLID products.

8. In the entire paper, there are no explicit units provided for the backscatter ratio (BSR). Could the authors please specify the units or define the normalization used? Additionally, is the BSR parameter consistent with the A-EBD BSR in Figures 3 and 4? Please define what does the parameter $R_{||}$ represent in these plots? Regarding the altitude axis in the plots: Is the altitude relative to the Earth's surface (above ground level), or is it referenced differently? Please clarify this convention in text as well, and if there is height correction implemented. It would also be beneficial to clarify the observational geometry of the WALES lidar during the coordinated underpasses: does the instrument only provide

observations of the atmospheric column above the aircraft, or does it also capture the column below the aircraft? This distinction is important for understanding the spatial matching relative to the EarthCARE ground track.

We thank the reviewer for pointing out these ambiguities.

The BSR is a dimensionless quantity. The term 'dimensionless' we included now in the text in Sect. 2.1 describing the WALES instrument.

The BSR values for WALES at 532 nm and ATLID at 355 nm are not directly equivalent due to their wavelength dependence. Theoretically, BSR values for cloud and aerosol features can be expected to be higher at 532 nm than at 355 nm, primarily because the molecular (Rayleigh) backscatter decreases strongly with wavelength ($\beta_{mol} \propto \lambda^{-4}$), while the cloud/aerosol backscatter generally exhibits a much weaker wavelength dependence.

Thus, we selected the colorbar of the A-EBD BSR (in Figs. 3e and 4e) such to enhance the visibility of the fine-scale structure at low BSR values. The point we aim to emphasize is that the patchy fine-scale structure of the ice clouds between 8–12 km is generally consistent between A-EBD and WALES, whereas A-TC represents this region as a homogeneous cloud layer across the entire cross-section.

We also updated the labelling (y-axis and colorbar labelling) in Figure 3 and 4 (see in the revised version) and also placed the panel labelling outside the window (as also requested by the other reviewer).

Regarding the WALES observational geometry, WALES is operated in a near-nadir, downward-looking configuration aboard HALO. It therefore measures the atmospheric column below the aircraft, from the surface up to approximately the aircraft altitude, rather than the column above the aircraft. Hence, we added 'downward-looking' to ll. 159:

"The instrument is near-nadir viewing (i.e. downward-looking) operated aboard the German research aircraft HALO (Krautstrunk and Giez, 2012) to provide profile data from the surface up to nearly aircraft altitude."

For clarification we furthermore also included the following sentence in the same paragraph describing the horizontal and vertical resolution of the WALES products.

"The WALES data are interpolated onto a vertically constant altitude grid above the EGM96 reference geoid, which corresponds to mean sea level and ensures the accuracy required for this study."

9. Did the authors use quality masking for the A-FM product?

We did apply quality masking to the A-FM product, as well as to A-TC and A-CTH. We excluded single pixels (for A-TC) and profiles (A-CTH, A-FM) carrying the highest-ranked quality flag (flag 4) for each individual product, following the definitions provided in the EarthCARE Handbook (https://earthcare handbook.earth.esa.int/catalogue/). This masking removed much less than 1% of the full data set.

10. Following the previous comment, in my opinion it's a confusing viewing geometry logic in Lines 375–377: The sentence regarding cloud top heights is ambiguous: "although fewer cloud top detections are visible compared to the WALES-derived cloud tops (Fig. 3b) as they occasionally are located above the flight level." Which instrument has fewer detections and why? If HALO is flying below the cloud top (e.g., at 10 km while clouds are at 13 km), a nadir-pointing WALES wouldn't see the cloud tops, but the EarthCARE satellite (looking from above) would. The authors need to rewrite this sentence to explicitly state

which instrument's view is obstructed and how that impacts the A-CTH vs. WALES comparison.

We hope the response of to comment 8 resolved the confusion about the viewing geometry. Due to the nadir-operated mode, it may be the case that clouds located above the aircraft not detectable by WALES, but are captured by the A-CTH. To improve clarity, we revised ll375-377:

"A-CTH shows cloud tops associated with high-, mid-, and low-level clouds, and occasionally detects cloud tops above HALO's flight altitude that are not detectable by WALES."

11. Lines 371 & 380: Regarding the Aerosol vs. Ice Cloud Misclassification, the text notes that A-TC classifies some aerosol pixels as ice clouds (often a known artifact with dust in lidar retrievals). The authors briefly mention Saharan dust at the very end. It would strengthen the analysis if they explicitly stated earlier that this A-TC misclassification is likely an artifact of dust depolarization mimicking ice crystals.

We agree the term "Saharan dust" should be mentioned earlier and adjusted the sentence in ll.353-355 accordingly:

"Another notable feature is an extended layer of intermediate BSR values (1.8 < BSR < 6) within the lowest 3 km, suggesting the presence of aerosols attributable to Saharan dust, within which denser cloud patches (BSR > 10) are embedded."

12. In lines 395 and 400, is it really intented to write Figure 3c and Figure 3d (extratropical case)? I understand that the authors compare to the previous case, but the text meaning would point the reader to figure 4.

We thank for pointing out this confusion. Corrected to Fig. 4c and 4d in this sentence in the revised manuscript.

13. Lines 396-398 lines: "Interestingly, the cloud formation between 6-10 km altitude in the A-TC product is subdivided near the approx. constant tropopause level into an ice cloud component below and a stratospheric aerosol layer above." If this is a cloud-to-aerosol transition at the tropopause, it is a scientifically interesting feature. However, calling it a "stratospheric aerosol layer" while discussing an altitude of 6–10 km is geographically sensitive. In the tropics, 6–10 km is mid-troposphere; in the Arctic 6–10 km can indeed be near the tropopause. I would recommend to explicitly state the altitude of the local tropopause for this flight to confirm why a 6–10 km feature is considered "stratospheric." Or maybe add a small description in the discussion.

We agree that the original wording could lead to confusion. We have therefore explicitly stated the tropopause altitude of 8 km for this flight. In addition, following the discussion in Comment #12, we now make clear that our interpretation refers to the extratropical tropopause.

Due to the nadir-viewing (downward-looking) geometry of WALES, observations are limited to altitudes below the maximum HALO flight altitude of approximately 15 km. This is well below the tropical tropopause, and therefore the observations do not allow us to draw conclusions about the tropical tropopause and to track a potential cloud–aerosol transition there. We have added this clarification to Sect. 5.1 of the revised manuscript (see also our response to Comment #25).

14. In many points in the paper flags of A-FM are mentioned (e.g. line 407) and it is a bit difficult to understand which flag is which mainly from the colorbar of figures 3+4. The reader may not have memorized the A-FM flag definitions; therefore, I would suggest to add a brief parenthetical description, e.g., (representing the most optically thick features) or include that to the appendix (if it is not already anywhere in the text) or point the reader to the hand book of that product (https://earthcarehandbook(...) /atl_fm__2a)

To facilitate the reader's understanding, we have added a link to the Data Handbook in the product description (Sect. 2.4.1). In addition, we have consistently replaced the term "flag" with "value" consistently throughout the manuscript to avoid potential confusion.

15. In the text (line 96), the authors state that the study comprises 35 dedicated research flights. However, Table 1 lists 40 flights. It appears that this is because a single flight may cover multiple EarthCARE orbits, with each orbit being treated as an individual entry in the table. Maybe this terminology could be conciled throughout the manuscript to avoid confusion. I suggest clarifying the distinction between a "research flight" and an "underflight segment" (the specific collocated orbit) early in the text so that the reader understands why 40 segments are derived from 35 flights.

To improve clarity, we revised the caption of Table 1 (see response to comment 6), and ll.181-183 of the manuscript which now reads as follows:

"As illustrated by the flight tracks in Fig. 1, two EarthCARE underpasses could occasionally be achieved for some research flights during PERCUSION-OP due to the increased satellite orbital coverage towards higher latitudes."

16. In Figure 6, where do the temperature data come from? While Figure 8 suggests the use of T_{IFS}, I suggest to state in the text or relevant figure captions whether the temperatures presented are derived from the EarthCARE L2 product files themselves or from an external model analysis. For the sake of reproducibility, it is good for the reader to know exactly which dataset (and maybe which version of the auxiliary data) is being utilized for these comparisons.

To clarify: We used operational hourly temperature data from the IFS analysis that were retrieved from the ECMWF Meteorological Archival and Retrieval System (MARS): https://www.ecmwf.int/en/forecasts/access-forecasts/access-archive-datasets, and then interpolated in space and time to the HALO flight track.

We added this information to the Data availability Section and provide the corresponding data access portal in the reference list as ECMWF (2026):

ECMWF, 2026: ECMWF operational analyses and forecasts, European Centre for Medium-Range Weather Forecasts [data set], https://www.ecmwf.int/en/forecasts/access-forecasts/access-archive-datasets, last access: 23 August 2026.

17. In the discussion regarding the overestimation of ice cloud occurrence (lines 437–440), the authors note higher pixel counts but leave the underlying physical mechanism open to interpretation. In the context of lidar remote sensing, such discrepancies often stem from

differences in instrument sensitivity thresholds, multiple scattering, or other algorithmic classification criteria. Could you clarify if these higher counts are a result of ATLID's sensitivity to thinner features that might be below the detection limit of the airborne HSRL, or if they are primarily artifacts of algorithmic misclassification (e.g., aerosol-as-cloud)?

We share your point of view, that selection of the BSR generally influences feature detection in the airborne data such that lowering the BSR threshold results in the detection of more "ice cloud features". However, we believe that these additional features are likely associated with aerosols rather than ice clouds. The shape of these aerosol features in the vicinity of ice clouds we can identified in both the ATLID L2 data and the WALES BSR data. We note that in Sect. 2.3, we provide a justification for the BSR threshold selected for the identification of ice clouds in the airborne data. The cloud-aerosol discrimination (focus on ice clouds) is work in progress and will be part of an upcoming paper.

This is precisely why we provide Fig. A1 in the Appendix, where we discuss and demonstrate this behavior. As shown there, there is no limitation in the airborne observations in detecting thinner features.

Given the mention of the "stratospheric aerosol flag" (line 443), please expand the discussion to explain how this flag interacts with the cloud classification logic and to what extent it quantitatively impacts the overall statistics.

We included the "stratospheric aerosol flags" here in A-TC as Fig.4 points already to potential aerosol-cloud misclassification near tropopause for extratropical flights (for Baseline BA), a finding that is discussed in detail in Sect. 5. To make this clearer we revised the corresponding sentence in ll. 443-444 which now reads as follows:

"These flights show an enhanced occurrence of the A-TC ice-cloud flag and, in particular, the stratospheric aerosol flag, pointing to potential cases of aerosol–cloud misclassification (cf. Fig. 4d) which will be discussed in Sect. 5.1."

18. Regarding the temperature thresholds discussed in the analysis of cloud pixel counts (lines 484–490), the physical role of the 236 K and 295 K peaks requires further clarification. Is the 236 K threshold a specific hard-coded limit in the A-TC classification algorithm, or does it correspond to a physical boundary in the underlying auxiliary data? I think that the local maximum at 295 K feels conceptually disconnected from the rest of the discussion. Please be more direct regarding what this peak represents. Are these counts identified as low-level clouds, or does this peak highlight a known challenge in the A-TC/A-FM product's ability to distinguish between boundary layer aerosols and near-surface cloud features?

In the description, we report two peaks at approximately 295 K and approximately 220 K (not at 236 K). The lines at 236 K and 273.15 K are intended solely to give the reader a rough idea of where to expect purely liquid or ice clouds. There is no temperature limit in the A-TC processing for detecting stratospheric aerosols – only for detecting polar stratospheric clouds (PSCs) which were not validated in this study. We believe that the number of stratospheric aerosol detections, which strongly decreases at temperatures >220 K and is hardly observed above 236 K, is due to the fact warmer temperatures simply correspond to data below the tropopause.

We have revised ll. 483–484 to clarify the context behind the ~295 K peak; the text now reads as follows:

"(…), exhibit another small local maximum at 295 K (1500 counts), which is related to the frequently captured (liquid) low-level clouds in the tropics."

19. The observation that no stratospheric aerosol flags were detected during the 2026 NAWDIC campaign (lines 444–446) is an important finding regarding algorithm robustness. This point is currently understated; I suggest moving it to the "Discussion" section. Please clarify if this absence reflects a genuine improvement in the BC baseline, a difference in atmospheric conditions, or a change in the classification logic.

We agree that this is an important finding that is related to the improved representation of the tropopause in Baseline BC. We discuss this in detail now in Sect. 5.1 (see response to Comment 25), but we have added a reference to this discussion in L444.

20. Figure 8 is packed with information. Please expand the caption to explicitly define the correspondence between the filled/unfilled circles and their respective violin plots, as well as the need to reference the secondary logarithmic y-axis. Furthermore, clarify if the shaded region (-300 to 300 m) represents the mission's formal accuracy requirement (in accordance to comment 3). If so, is this 300 m threshold tied to a specific physical constraint, such as the ATLID vertical sampling grid size? Please, if possible, also address whether this observed systematic bias is inherent to the retrieval architecture or potentially resolvable in future processing versions.

According to the reviewer suggestion we revised the caption of Figure 8, and included all the requested changes:

"Cloud top height distribution shown as violin plots for WALES observations colored by distinct defined cloud class (liquid clouds, extratropical and tropical mixed clouds, ice cloudsand for the full data set; see legend). For each class the solid-framed violins represent data for the full cross underpass segments, while the dashed-framed violins (the right violins of each pair) represent data within ±10 min. to the EarthCARE underpass. The counts per class (black filled for full data set, unfilled for ±10 min around underpass location) is shown with respect to the secondary, logarithmic-scale y-axis. (b) Differences in cloud top height between WALES and A-CTH (A-CTH − WALES-CTH), shown as violins for the same classes as in (a). The magenta shading in (b) depicts the ±300 m accuracy requirement according to Wehr, 2006 and Wandinger et al. (2023)."

21. I would suggest to include a regression line for Fig 7, it is intresting to see the correlation coefficient. Even if there are many CTH that WALES captured while A-CTH didn't, the agreement still look really good with most points around the 1-1 line.

We thank the reviewer for this suggestion. We considered including a linear regression and R2 in Fig. 7; however, after consideration, we decided not to include them which we want to explain in the following:

The number of detected clouds varies substantially across the 0–14 km altitude range. This uneven sampling means that a least-squares regression would be strongly influenced by the most densely populated CTH ranges, particularly the higher cloud tops (see Fig. 2 in the manuscript).

In addition, the differences between WALES and A-CTH can be larger for very low or very high cloud tops than for mid-level clouds. Such extreme deviations can disproportionately influence an ordinary least-squares fit, making the resulting regression less representative of the overall agreement between the two instruments. Hence, we consider a simple linear regression to be less suitable for this type of comparison. Instead think it is more helpful to provide an estimation of the percentage of A-CTH meeting the accuracy requirement (see Sect. 4 and 5). We therefore consider the 1:1 line and the ±2 km reference lines to provide a more direct and physically meaningful representation of the agreement between the two CTH

products.

22. Lines 556-558: “Interestingly the distribution of differences in this category is bimodal, with one maximum similar to the other categories at slightly above +300 m, and a broader secondary maximum ranging between –200 m and –600 m which may point to a potential systematic misclassification in this category.” The authors identified a bimodal distribution for liquid clouds (one peak at +300 m, one at -200 to -600 m). But potential systematic misclassification is a bit vague. Based on the physics, if A-CTH is underestimating liquid cloud tops, is it failing to detect the thin, wispy tops of boundary-layer clouds? Or is it incorrectly "locking on" to the cloud base in some cases?

We do not see a clear physical reason why A-CTH should fail to detect the tops of these boundary-layer clouds, as these clouds are predominantly optically thick, liquid-phase clouds, which are consistently detected by WALES and also the other products, A-TC and A-FM (see, e.g., Fig. 3). It is rather challenging to separate very shallow cloud tops from the surface return. We know that this issue is related to a known bug involving an incorrectly assigned surface return, which has been resolved in the Baseline BB revisions and thus in proto-CA (which we discussed then in Sect. 5). This issue is also documented in the A-LAY disclaimer, where surface returns are no longer misidentified as cloud tops. Thus, we avoid the term “misclassification” in this context, as these cases are better described as false detections resulting from the incorrectly assigned surface return. Hence, we changed ll. 556-558 accordingly in the revised manuscript:

“Interestingly (…), and a broader secondary maximum ranging between –200 m and –600 m which may point to a potential false cloud top detections in A-CTH near surface in this category.”

23. A general question for section 4.2: Since the authors are comparing WALES and ATLID cloud tops, it would be helpful for the reader to include the viewing geometry (see also comment 6). WALES is nadir-looking from inside the atmosphere and EarthCARE is looking from space (looking down through the atmosphere). Could the 300 m bias be related to the attenuation of the satellite signal as it travels through the atmosphere, vs. the much shorter path length for the aircraft?

We hope that the potential confusion regarding the viewing geometry of WALES has been addressed in our response to Comment 8.

In the manuscript, we compared the cloud-top heights retrieved by ATLID and WALES for different altitude and temperature intervals and find a rather persistent positive bias of approximately 300 m across all classes (apart from the surface bug at very low altitudes), indicating an overestimation of cloud-top heights by A-CTH. Furthermore, we only report a bias when both A-CTH and WALES simultaneously detect a cloud below the HALO aircraft altitude.

From a lidar perspective, the reported bias is therefore not expected to be substantially affected by the different viewing paths/attenuation of ATLID and WALES.

24. The text highlights that the WALES dataset is used for robust validation due to its spatiotemporal collocation. Is it possible that the high-altitude maxima observed during tropical campaigns (as noted) could introduce a geographic bias in the product validation if not properly accounted for?

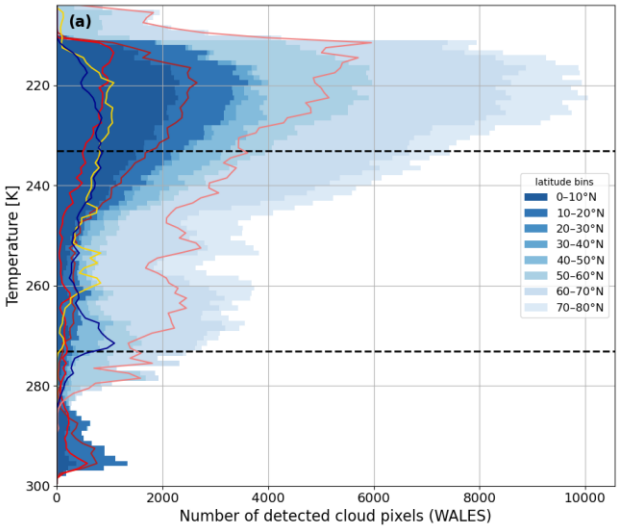
It is not fully clear to us to which product the comment "the high-altitude maxima observed during tropical campaigns (as noted) could introduce a geographic bias in the product validation" refers, we hope to clarify this point with the following discussion.

Fundamentally, for the validation of cloud properties, we find not only altitude but, in particular, temperature to be relevant, as it determines the cloud phase. To ensure that our validation is representative across different geographic latitudes and altitude ranges, we therefore analyzed the cloud-pixel validation additionally with respect to temperature (Fig. 6a–d). For the validation of cloud tops, we considered altitude (Fig. 7) as well as different temperature ranges and geographic latitudes (Fig. 8).

Regarding the validation of cloud pixels, we provide below an additional figure, analogous to Fig. 2, in which the contribution of cloud pixels within different temperature intervals is color-coded by latitude. It shows that the different latitude ranges and campaign segments all contribute across the full ice-cloud temperature regime. Fig.5 shows the overestimation of ice clouds by A-TC across all campaign segments, i.e., representing different latitude ranges. This is also clearly supported by the analysis of the individual flights shown in Fig. 5.

For the validation of cloud tops, we likewise performed our analyses for specific temperature regimes and considered tropical and extratropical observations separately. Figure 8 clearly shows a continuous overestimation of cloud-top heights by A-TC, independent of latitude and temperature, besides from the (in the manuscript) discussed "surface issue".

In our point of view this validation from different perspectives (individual flights, temperature, altitude) should support that the key findings are not caused by a "geographic bias".



25. In Section 5.1, the authors attribute improvements in the prototype CA version (e.g., elimination of stratospheric aerosol misclassifications) to both "improved tropopause determination" and the general evolution of the Baseline. Given that the EarthCARE handbook link is provided, I would suggest to disentangle the specific causes of the improvements: is it solely due to the new tropopause logic or other changes in the A-TC processor contributed?

The Data Disclaimer provides information about the most critical changes in the Level 2 processors and shows that various upgrades were implemented simultaneously. In addition, the quality of the Level 2 products is also influenced by changes in the Level 1 processors. Therefore, it is not always possible to attribute diagnosed improvements to specific, individual upgrades.

Nevertheless, the disclaimer provides an important indicator for certain improvements. We have revised Section 5.1 (also with regard to comment 13, and the general comment 26) to highlight this more clearly.

"Another feature observed in A-TC during mid- and high latitude flights is the occasional presence of stratospheric aerosol structures attached to ice cloud features, that are sharply separated at the tropopause. A-PRO uses the tropopause as a reference level to define backscatter thresholds for cloud detection (i.e., lower thresholds in the stratosphere, higher ones at the tropopause; see table 1 in Irbah et al., 2023). Cloud detection in A-PRO therefore also depends on the height of the tropopause. Although the tropopause is generally an important reference, it is not always clearly defined (e.g., it may exhibit discontinuities and large jumps in height near the jet stream, be poorly defined within the polar vortex, or be multi-valued; e.g., Schäfler et al., 2021; Krüger et al., 2022). Potentially incorrect tropopause heights can therefore influence feature detection in A-TC and A-CTH, and are likely responsible for the described sharp transition between ice clouds and "stratospheric aerosols" at the tropopause in the baseline BA (Fig. 4d). We emphasize that we were only able to diagnose this feature for extratropical frames, since the tropical tropopause lies far above the HALO flight level. We also want to stress that such misclassifications of ice cloud pixels as stratospheric aerosols are not observed in the more recent baseline BC (during the NAWDIC flights) and are much reduced for the midlatitude cases of the prototype-CA version (Fig. 9c). This is likely attributed through an improved tropopause determination as indicated in the data disclaimer (ESA, 2026a)."

Later in this Section we also added:

"The disclaimer indicates that this improved ice cloud-liquid cloud discrimination is related to the use of temperature from improved X-MET product that was already implemented for Baseline BC (ESA, 2026a)."

We also included the references of the data disclaimers to increase their visibility to product users:

ESA, 2026a: A-PRO data disclaimer: https://earthcarehandbook.earth.esa.int/documents/d/earthcare-data-handbook/earthcare-product-disclaimer-atlid-level-2a-a-pro; date of last access: 23.08.2026.

ESA, 2026b: A-LAY data disclaimer: https://earthcarehandbook.earth.esa.int/documents/d/earthcare-data-handbook/earthcare-product-disclaimer-atlid-level-2a-a-lay-a-cth-a-ald; date of last access: 23.08.2026.

26. A general comment: The technical depth of the analysis is commendable, but sections 3-5 feature some dense paragraphs that make it hard for the reader to digest the results effectively. To improve readability, I would suggest to break the massive blocks of text into smaller, and focus on the core findings for each product.

We agree with the reviewer that the presentation of the results and their discussion can be made more concise and accessible. This point was also raised by the other reviewer, and we have therefore addressed both comments together in our revision of Section 5. We have restructured and streamlined the section to break up dense passages, reduce repetition of findings, and place greater emphasis on the key results and their implications for each product. The corresponding changes can be found in the track-changes document and the revised manuscript.