Reviewer comments are provided in blue. Author responses are given in black text, and changes made are provided with red for removals and blue for additions, where reasonable. Line numbers for the changes are given with respect to the line numbers in the actual manuscript.
Between creating the preprint and the newest manuscript, all of my Zotero citation keys changed, so all citations within the manuscript appear to be changes accoriding to latexdiff. Only a few instances have actually different citations between the preprint and the final manuscript.

Response to reviewer 1

Dear reviewer,
Thank you for thoroughly reading the manuscript and your enthusiastic response. The feedback has been valuable in making changes and improvements.

One common theme of the responses from both reviewers was about the figures having too small text. To this end, we have doubled the font size across all figures, and will work with the editors to make sure that the figures appear properly in any final copy that is produced. The figures have also been enlarged within the document, allowing more space for them to be viewed on the page.

Reviewer comments are provided in blue, with author responses given in black below.

Specific responses:

1. The text in figures would, ideally, be similar in size to the text within the captions. While those reading a PDF can zoom in, those of us foolish enough to print off papers to get some time away from a screen cannot read any of the labels in the figures of this paper. Is it possible to regenerate the diagrams with text scaled to their printed size?
   See above.
2. Though it isn't necessary, the paper may be improved by a qualitative description of some of the mathematical operations in S2.3 as most researchers in atmospheric science do not have much training in the formalism of mathematics. For example, "nearest neighbour distances between samples" would likely be interpreted to refer to the physical distance between observations rather than the distance in state space. The references given are accurate but not necessarily useful for the audience of this journal; a textbook or lecture course may be useful (as was done for the description of the copula).
   I have included the mention of "sample space" over state space to the description in Sec. 2.3. I have also added a reference to a book by James V. Stone (Information theory: a tutorial introduction) to Sect. 2.2 to direct readers to an easy to digest introduction to information theory.

L192 – L195
"Another commonly employed method is estimating the mutual information from nearest neighbour distances between samples in the sample space (e.g. Kraskov et al., 2004; Holmes and Nemenman, 2019). Regions of sample space with high probability density will

likely be sampled more frequently than low probability density regions with lower probability density, resulting in samples drawn with low separations between them, indicating structure in the distribution."

3. L68: It's not clear to me why optimising the correlation coefficient biases the results *high*. This may be a matter of definition, as I'm reading this line to say "we are preferentially selecting results where one is consistently larger than the other" as that is how 'bias' tends to be used in atmospheric science. You may have meant "we are preferentially selecting results that are large" (in magnitude) or "we are preferentially selecting results that resemble each other". My instinct is that the sample will be biased but with unpredictable sign. There is every chance this is a standard result in statistics of which I am unaware.

   The description in the introduction has been changed to reflect that it is the value of the correlation coefficient itself which is biased high (towards "better" values), not the bias between the values being compared.

L66 – L68

By selecting a co-location parametrisation that maximises the correlation coefficient, we are preferentially selecting results that are biased high with correlation larger which coefficients, which could arise from the variance in the estimation, rather than an actually better comparison between the measurements.

4. L161 and L163: The values given in the text differ slightly from those given in Fig. 2. Please check which is correct.

   I have updated the values referring to figures 2 and 3 since the figures have been regenerated.

L163 – L164

The value of IKSG =−0.001 IKSG =−0.002 nats can be negative as a result of it being empirically estimated (see Sect. 2.3).

L165 – L167

There is obvious structure in the joint probability distribution that can be learned, and as a result, the mutual information estimate IKSG = 3.167 IKSG = 1.965 nats is higher when X and Y are dependent compared to being independent.

5. L270: I agree that, in the 2-D presentation of Fig. 4, Cloudnet data is a point. However, in 3-D space, the satellite swath is a (mostly vertical) surface and the site is a vertical line. I'm not sure the document would be improved by making this point, but the pedant in me could not leave this distinction unmentioned.

   I have addressed this by adding "Projected onto the Earth's surface,...", which hopefully clarifies the point.

L284 – L285

The Projected onto the Earth's surface, the spatial co-location scheme treats each Cloudnet site as a 0-dimensional point-like source.

6. L401: It is not obvious to me that the number of profiles scales with \cosh(R). I would have expected this to come from the rule of cosines giving the length of a bisector of a circle to be $R\sqrt{2(1-\cos\gamma)}$; hence linear R dependence. Regardless, if you are correct, isn't $\lim_{R \to \infty} cosh(R) = \infty$ rather than R?
   I agree, and I've realised the mistake I made. The distance between a satellite and a fixed point (the Cloudnet site) follows a hyperbolic curve of the form $(\frac{r(t)}{r_{min}})^2 - (\frac{v(t-t_0)}{r_{min}})^2 = 1$, with t-t_0 being the temporal displacement of the satellite from being at the minimum separation of r_min. However, N_profiles scales as the length of the cord traced within the circle of radius R by the satellite ground path. This can instead be derived through Pythagoras, and gives a cord length $s = 2\sqrt{R^2 - r_{min}^2}$. This function N_profiles \propto s(R) is hyperbolic, but not cosh(R). In the limit of R >> r_min, this is approximately linear in R. The text has been changed.

L420 – L422

We expect Nprofiles to be proportional to R2, one power coming from a proportionality to Nevents, and the other deriving from the fact that for each given co-location event, the number of included vertical profiles scales as a function of cosh(R) sqrt{R2 −rmin2}, which is linearly proportional to approximately linear in R in the limit of large Rvalues R≫rmin.

7. These are two very good questions:
   Fig. 5 and L422: The discontinuity in significant solutions deserves some more comment to guide readers in how to communicate their confidence in the selection of optimal parameters. A continuous region of significance can sensibly be communicated through traditional uncertainty notation (even if not strictly appropriate), but this second solution is much harder to communicate.

   a. Would the authors expect this degeneracy to resolve as more data is added to the problem (i.e. wait for a longer dataset) or is this an unavoidable aspect of the problem?

      This is a very good question, and unfortunately, hard to answer due to the nature of the datasets being used. I would personally expect that the degeneracy would lift and the region of possible optimised co-location parametrisations would become simply connected. However, as you elude to, this could only really be tested in this scenario by waiting for more data. This would be an interesting target for a simulation to determine.

   b. What would you do if the difference in mutual information between the two solutions was negligible?

      In this scenario – where two co-location parametrisations provide the near-identical mutual information between them – I would choose the one with the smaller standard deviation estimate for the mutual information. If the two co-locations were similar in both magnitude and variance of the mutual information estimate, I would personally choose to use the co-location that maximises the available data for further analysis, although I would argue that this then becomes the researcher's choice, and that this framework provides

no explicit way to choose amongst the candidate optimised co-location parametrisations.

As for how to communicate the connfidence/uncertainty in the region of candidate optimised co-lcoation parametrisations, I agree that for simply connected regions, traditional uncertainty notation (although not stricty proper) would give an appropriate way to describe the region. For disconnected regions, this could be extended to a set of traditional uncertainty descriptions, with each element corresponding to one of the disconnected regions.

8. L425: I understand this section to argue that I(p) decreases as you move away from \hat{p}, but I'm unsure that 'unimodal' is the correct word for that. Unimodal means the surface is single-valued at each point – which is true by construction for I. Collapsing the 2-D surface into a function of distance from \hat{p} clearly *isn't* unimodal because the surfaces clearly aren't isotropic about that point. 'Largely monotonic' strikes me as more appropriate but several options are available.

   I believe that unimodal is the correct term here, refering to a function with a singular maximum value.

9. L778: If the authors think there is any general utility to this simulation, it may be worth mentioning within the main body that $N > 10^4$ achieves consistency in these estimators. Some guide on the number of observations necessary to apply this method would be useful to most readers.

   I included the simulations at the end to visually demonstrate the idea of estimating mutual information in data-limitted and data-contaminated regimes. I think the idea of a simulation like this being used in advance of an analysis to (for instance) bound the values of \vec{p} that are checked is very interesting. However, I think that the model in App. A is too simplified to provide a rule of thumb for how many data points analyses should use in general. Extending this work would be interesting, but I don't have the scope to do that at present.

10. App. C: This is probably just a difference between the language of a mathematician and a physicist, but I would call these three points approximations rather than assumptions that are broken.
    Re-reading App. C, I prefer the idea of calling them approximations. I have changed the language to reflect this.

L863:
In this appendix, we will derive a formula to determine the across-track density of satellite orbits, under the assumptions that withthe following approximations:
L868:
These assumptions are broken approximations are incorrect, but in most circumstances will lead to reasonable results.
L871:
Secondly, treating assuming subsequent orbits of a satellite on the tangent plane are parallel is a broken assumption.

11. Sec. 3.6.2: There is no way you could have known this, but section 4 of https://doi.org/10.1007/s10712-025-09898-4 published after your submission explores some of the issues with collocation you discuss at the end of this section. Further, in the aerosol community there is precedent in the publications of Nick Schutgens for the conclusion that a one-size-fits-all collocation criteria is sub-optimal, such as Fig. 1 of https://doi.org/10.5194/acp-20-12431-2020 or the conclusions of https://doi.org/10.5194/acp-17-9761-2017. Admittedly, Nick aims to minimise the bias, which is exactly what you argue against doing, but you might find the comparison could be useful.
    Thank you for bringing the mentioned papers to my attention, they are all very interesting reads. I have included references to Schutgens et al. 2017 and Langsdale et al. 2025 in the manuscript.

L40 – L42:
Finding a parametrisation for a co-location scheme that balances the need for sufficient data, whilst minimising the co-location mismatch between the compared data within a co-location event is the crux of the problem, and is as yet unsolved (Langsdale et al., 2025).
L496 – L498:
We have demonstrated that the value of p is influenced by local factors, such as mountainous orography near the surface-based observatories, and non-local factors such as the satellite sampling strategy(Schutgens et al., 2017).

Technical corrections:

- L16: validation and constraining

  corrected

L16:
validation an and constraining

- L62: expectation to the comparison

  corrected

L62:
a prior expectation to the is effectively applied to the comparison metric in the analysis.

- Tab. 1: There should be a space before N_{profiles}.

  corrected

- L598: these different quantities

  corrected

L623:
for the evaluation of thesea these different quantities.

- L710: There should be a space after the comma.

  corrected

- Eq. C10: Shouldn't it be \lambda_{21} here? If I'm correct, it might be clearer for C11 to show \cos\lambda_{21} and switch to \cos\lambda_{12} for C12.

  corrected

- Maybe C14 should come before C13 as the more natural equation to follow from C5?

  corrected

- L888: this bearing into an across-track

  corrected

L917:
In order to convert this bearing in to into an across-track density of orbits,

- I randomly tried a few of the DOI links in the references; Crameri 2023 and Palm 2021a were dead links for me.

  corrected

Response to reviewer 2

Dear reviewer,

Thank you for reading my manuscript and giving thorough feedback. I have appreciated reading through your responses, they have been very helpful for identifying areas in the manuscript that could be clarified and improved upon.

One common theme of the responses from both reviewers was about the figures having too small text. To this end, we have doubled the font size across all figures, and will work with the editors to make sure that the figures appear properly in any final copy that is produced. The figures have also been enlarged within the document, allowing more space for them to be viewed on the page.

Reviewer comments are provided in blue, with author responses given in black below.

General comments

1. The Introduction is very informative but I would consider it too long. I would recommend to make it a bit more compact while keeping all relevant references in.
   I believe the topic of the manuscript has a lot of nuance that may be unfamiliar to readers new to the field. I have tried to justify each part of the motivation for the framework, which has lead to the introduction being two pages long. I worry that shortening the introduction too much risks the loss of important details. I have changed some of the writing to try to make some sentences easier to understand.

2. The paper is lengthy, so I would recommend to try out ways to reduce its length as it could be a big discouraging for a potential reader to read it, even though its topic of discussion is of a great significance. Towards this aim, I would recommend to move some from the Appendices to a Supplement, which will come in a separate document. To my opinion, Appendix B is a good candidate to be moved to the Supplement. Moreover, Appendix C could become another section in the supplement. If you still decide to keep it in the manuscript, please try to make it shorter. I believe that Appendix A should still stay in the manuscript.
   I agree that the paper is long, and this could be discouraging. We did try to keep the manuscript shorter, but the nuances of the framework and results needed explaining and this increased the length of the manuscript. We submitted to AMT specifically because the journal allows for longer articles, which allowed us to go into sufficient detail to explain and justify the new framework.

3. The section 4 “Discussion” is also too lengthy. I would not create separate subsections for 4.1, 4.2 and 4.3. I would just be merging it to one while explaining the important parts of the frameworks “The choice of mutual information estimator”, “Physical interpretation of $\hat{\rho}$” and “Choice of co-location scheme”.
   Having considered this suggestion, I believe keeping the headings 4.1–4.3 in the manuscript is important, as they represent specific and overarching questions about

the use of the framework in future studies. Condensing them under a single subheading would make the information harder to find for a reader referring back to the paper.

4. Section 3.6.3 contains some weak statements as the comparison of the two extreme parameterizations p00 and p11 is not that strong. Try to make this section compact and more meaningful.
   I have made some changes to 3.6.3 to hopefully make the meaning of the comparison more clear.

5. I would recommend to completely remove Section 3.7 too. I do not clearly see any strong point to keep it. On the contrary, the manuscript will benefit from the change as it will become shorter.
   Section 3.7 was added as a response from co-authors that the manuscript needed to show a “real-world” comparison and the effects of using different co-location parametrisations on the outcomes of the comparison. It reinforces the point made in the introduction that the selection of a co-location parametrisation that optimises the value of a comparison metric (in this case the mean bias and variance of the bias distribution) would not necesarily agree with a choice of co-location parametrisation that optimises for the mutual information between the measurand distributions.

6. The Figures have some issues with the axis labels. The font size is way too small. When the reader prints the manuscript, the axes are not even readable. Even on a typical-size laptop screen, the axes are not visible and only with zooming in you can read them well. But this interrupts the flow of the reading. I would recommend to simply increase the font size of all axes in the figures.
   See above

Specific comments

1. The reference (Loew et al, 2017) appears too frequent in the Introduction; precisely four times and then again in the second sentence of the Section 2 “Framework”. It seems like a repetition.
   I agree that I have heavily used the reference Loew et al. (2017). It is an important and comprehensive review of satellite validation methodology, and was a primary source of inspiration for this paper. I have changed the fourth reference to the newly published Verhoelst et al. (2026), as it is also applicable.

L35 – L37:

Once data have been co-located and co-location events have been identified, formal uncertainty characterisation can be performed, or other comparison metrics such as the bias, RMSE and correlation coefficients can be calculated (Loew et al., 2017, Sect. 3.4) (e.g. Verhoelst et al., 2026).

2. Line 60 – I would not start the sentence “This should be avoided”. Please rephrase it.
   This has been rephrased.

L60:

This should be avoided, as the The comparison metric is an unknown quantity (hence the need for the analysis in the first place)

3. Introduce numbering items for the co-location scheme, criteria, parameterization and event. Their definition is good as it introduces early enough the terminology to the reader. It is very much appreciated.
   The definitions have been numbered.

4. Figure 1 is very good introductory of the potential schemes. Please add main references for each scheme to the figure legend. There are many references in the text but it is still nice to have the visualization of the scheme together with the main reference(s).
   I tried several ways to introduce the references into the Figure itself, and into the Figure caption. They all cluttered the figure and required additional space for the figure which would make the manuscript even larger. I have instead highlighted in the caption that references can be found in the text in Sect. 2.1, in case a reader would like to find the lists of references.

Figure 1 caption:
Refernces for each co-location scheme are provided in Sect. 2.1.

5. Line 105 – What “best results” mean? Earlier in the introduction (line 60), it was recommended that this should be avoided.
   I have changed this to be “better results”. This is in a different context to the comment on line 60. If we co-located data by matching purely based on the same time, instead of including the location of the satellite as co-location criteria, then data could be matched when the satellite is antipodal to the Cloudnet site. This is not physically motivated, and would clearly perform worse than the co-lcoation scheme that is more phsyically motivated and includes a spatial co-location criteria. This is what I mean by “better results”.

L103 – L105:
The framework requires that the co-location criteria for a given co-location scheme can be be described by a finite number of parameters, which is applicable for all realistic co-location schemes. The co-location scheme can be arbitrary, but it should be physically motivated to achieve the best better results.

6. Line 122-124 – The sentence “For example, … ancillary wind data” is disconnected with the previous text. It only makes sense with prior explanation why/how local advection might affect the colocation scheme. It should be further explained and ideally be moved to another location in the text.
   This comment isn’t to tell a reader exactly how a co-location scheme should be changed to include additional ancilliary data, only to highlight that it can be changed, and in a myriad of ways.
   To include expectations of the advection into the co-location, the user of the framework could offset the circle of radius R from the ground-based station by the vector (<u>, <v>) * \tau * k, where <q> is the average of q over the duration spanned by \tau; u and v are the local wind velocity components; and k is a scaling constant.

In this case, k is a new free parameter in the co-location scheme. This is only one way, the user could instead make the circle eccentric based on the mean wind vector and co-locate data within an ellipse instead of a circle, oriented along or perpendicular to the mean wind vector. There is no single correct co-location scheme, so I would not want to prescribe a particular scheme in the manuscript, in case people interpret this as the only way to do it.

L122 – L126:

For example, if co-locating atmospheric data between a satellita and ground-based station, the co-location scheme in Fig. 1a could be augmented to encode our expectations about how local advection may affect the spatiotemporal co-location of the data. For example, more complex schemes could allow us to encode our expectations of how local advection would affect the spatial distribution of (in)dependent samples between data sources, through the implementation of logical criteria on ancillary wind data.

7. Figure 3 is very nice. Please make clearer that the preferred situation is depicted in plot 3b.
   I have updated the figure caption to highlight this. I have also created a box around 3b to make this obvious.

Fig. 3 caption:

Panel (**b**) represents the best scenario for a co-location.

8. Line 221 – Remove the first bracket from Eq. 3. It should appear as…
   The outer brackets have been removed.

9. Line 233 – The vertical (range) resolution is on the order of decimeters (~10–30 cm), determined by timing precision. 30 m should referto the horizontal segment length, not the vertical resolution. Do you mean the following: “Photon returns are accumulated into 30 m along-track segments, which are further aggregated over 400 consecutive laser pulses to produce a product with an effective along-track resolution of approximately 240 m.”?
   Your comment is partially correct, in that the vertical resolution of individual photon returns is dictated by the timing precision of the ATLAS detector, and is on the order of centimeters to decimeters. However, in order to boost the signal to noise ratio and produce a gridded data product, ATL04 consists of 30 m vertical bins, with any photon return falling within a bin contributing to the photon count of that bin.
   The along-track resolution was however misquoted, being 280m instead of 240m. The 280 m along-track resolution comes from the pulse repetition frequency of 10 KHz, and the orbital speed of ICESat-2, yielding an along-track resolution of (400 pulses per profile / 10_000 pulses per second)*(~7_000 m/s) = 280 meters per profile.

L246 – L248:

ATL04 consists of photon returns that are vertically aggregated and summed over 400 consecutive laser pulses, giving to produce a data product with 30 m vertical resolution and 240 280 m along-track resolution.

10. Line 278 – There are some problems with Eq.5: (a) The vertical bar | is ambiguous here (it usually means “such that” in set notation or conditional probability), (b) The

prime notation j′ is unnecessary and makes it harder to read. The easier is to express it like…
This has been implemented.

11. Line 286-287 – “The observed clouds are optically thick enough to attenuate the ATLAS lidar beam throughout the co-location event, so lower-level cloud layers will be missed.”. This is an important information which can have an important effect on the comparisons against the ground-based VCF. However, I would expect that the opposite might happen for the Cloudnet data. The higherlevel cloud layers will be then missed? Or what is the situation exactly?
    This is a correct observation. The ATL09 retrieval only consists of lidar measurements, so is senstive to attenuation due to liquid water in cloud layers. And this will be part of the interest when comparing ATL09 VCF retrievals against Cloudnet retrievals – is there so much attenuation that the ATL09 VCF profiles lose significant information with respect to the Cloudnet profiles. Equally, the attenuation itself can be considered additional information, providing upper bounds on the VCF values at lower altitudes. The Cloudnet retrieval suffers less from this problem as it is synthesised from lidar, microwave radiometer and radar data. The radar in particular experiences less attenuation than the lidar signal, and liquid attenuation acts to degrade the quality of the Cloudnet microphysical classification, but less so the macrophysical classification of hydrometeor presence in the atmosphere. Thus, the Cloudnet retrieval suffers less from missing higher level clouds (although it will still happen) than ATL09 suffers from missing lower level clouds.

Fig. 4f: [I have included the line “+attenuation” representing ATL09 cloud+attenuation]

12.
    a. Fig 4c should indicate a minimum of 18.4km, if I understood well? Please add the information to the plot.
        You are correct, I tried adding this to the plot labels but I think it clutters the plot and isn’t necessary information.
    b. Also, it might be preferable to show the area which corresponds to the contributing pairs ATL09-Cloudnet at the co-location parameterization
        The area of contributing vertical profiles given the parametrisation is the unhatched regions of b-e, and the unshaded region in a. I have made this more explicit in the figure caption.
    c. Fig. 4f shows that the profiles have huge differences for altitude 4km and below. How do the uncertainties look like for the two measurements? Are the uncertainties taken into consideration at all? It would be highly advisable to include the error bars into the VCF profiles.
        Computing error bars of an aggregation over a binary mask, such as the arithmetic mean of the cloud mask at a given height is an ill-defined process. I did compute error bars at one point through bootstrapping, but the uncertainties are not used in further analysis and do not meaningfully add to the plot. A full uncertainty analysis of these cloudmasks is outside the scope of this work.
        I have added in a line to represent the ATL09 cloud-and-attenuation profile, which provides an upper bound for the amount of cloud detected by ATL09.

Regions that are attenuated in the ATL09 data product could be considered to vary between totally cloudy or entirely cloud free. Interestingly, we see that when the ATL09 cloud deviates from the Cloudnet VCF, the ATL09+attenuation profile still correlates with the Cloudnet VCF to much lower altitudes.

d. Moreover, it seems like there is an overlap issue in the Cloudnet profile. Can you please explain that? The VCF profile increases exponentially from the ground up to the first 1.0-1.5km where it reaches the value of 1.0 (i.e., fully cloudy). At the ATL09 profile the VCF is exactly zero, which could imply that the ATLAS laser beam is completely attenuated at those altitudes. But then, the vertical bins should be removed from the dataset in a pre-processing step. Is this performed? Can you also elaborate on the potential problem of incomplete overlap of the Cloudnet lidar? The incomplete overlap problem refers to the situation in which, at short ranges from a lidar system, the laser beam and the telescope field of view (FOV) do not fully overlap. Remaining incomplete overlap problems in the retrieved vertical profiles could introduce artifacts which should be removed in a kind of pre-processing step.

   We can see from Figure 4b that the ATL09 data is attenuated at ground level throughout the entire scene, which explains the 0 VCF value at ground level from ATL09. Removing these bins from the data would however make the analysis infeasible, as each co-location event would be considering ATL09 data in different-dimensional state spaces, which would be incomparable, and no mutual information would be computable between the ATL09 and Cloudnet data.

   Interestingly, the lidar overlap problem should not affect the mutual information computed between the Cloudnet and ATL09 data. The overlap problem can be corrected for by a height-dependent multiplicative correction to the measured lidar backscatter signal. This transformation is invertible/reversible/lossless, as it is simply a scaling of the lidar backscatter values at lower altitudes. One property of mutual information is that it is invariant under reversible transformations of the input data distributions. Thus, whether or not the overlap correction is applied to the data would not affect the mutual information analysis (although it would affect subsequent analysis of the data after determining the co-location parametrisation). I have added in a paragraph to explain this in Sect. 2.2.

   This does raise another good point though, in that the Cloudnet categorize data product uses a height grid measured from mean sea level. On top of that, the minimum detection range of the radar is ~100m above the radar, and the Cloudnet stations are all located above 0m msl. The vertical gridding from the homogenisation process has a bin that is constructed from -120m to +120m, which includes ATL09 data, but does not include Cloudnet data, so it is uniformly homogenised to contain 0 cloud fraction. Removing ATL09 data from height bins below the Cloudnet station could result in the loss of information from the ATL09 data if it passes over lower-lying surfaces.

   I have removed the lowest data point from the Cloudnet VCF profile in figure 4f, and have also included a line showing the ATL09 cloud + attenuation profile.

Figure 4 caption:

The co-location of ATL09 and Cloudnet data at Ny-Ålesund, with ATL09 data from the granule with reference ground track 115 on cycle 12 (dated 01 July 2021), using the co-location parametrisation p = (125 km, 6 hours). (a) The Cloudnet observatory (star), and the three ATLAS strong beams (lines). A circle of radius R= 125 km is also drawn around the Cloudnet site. (b) The feature mask generated from the ATL09 data associated with strong beam 1, showing where ATL09 retrieves clouds and is attenuated. (c) the The distance between the ATLAS ground track and Cloudnet observatory, showing the co-location criteria subsetting of the cloudmask. Hatched regions are rejected by The unhatched region contains the co-location, and vertical profiles contributing to the darkest regions show where the ATLAS lidar beam is attenuatedco-location. (d) and (e) are the same as in (b) and (c), but for the Cloudnet data contributing to the co-location event. (f) Vertical cloud fraction profiles for the ATL09 and Cloudnet cloudmasks feature masks, as well as the vertical cloud and attenuation profile for the ATL09 feature mask, subset by the co-location parametrisation p = (125 km, 6 hours).

13. Line 301-302 – "Above 5 km in height, both VCF profiles visually correlate with each other, indicating that the co-location … unobserved by ICESat-2.": Could it be that simply above 5 km, the ground-based lidar face laser beam attenuation issues? Or this is certainly not a case in such conditions of optically thick clouds present?
    If the Cloudnet retrieval faced attenuation issues, we would expect the decrease in the Cloudnet VCF profile with height. However, we would expect it to not agree with the decrease in the ATL09 VCF profile, as they would be making independent observations. Attenuation in the Cloudnet VCF values would then miss cloud layers that we would expect ICESat-2 to observe. As noted above (point 11), the Cloudnet retrieval synthesises both lidar and radar data, which should be able to detect the same higher altitude clouds that ICESat-2 is seeing. Thus, I believe that ICESat-2 and Cloudnet are observing the same cloud layers at higher altitudes.

L319:

The ATL09 vertical cloud and attenuation fraction profile correlates with the Cloudnet VCF down to an altitude of roughly 1.5 km.

14. Line 311 – Can you elaborate more on the selection of k=10?
    The selection of k=10 was chosen as a compromise between having a larger value of k (i.e. k=20), which is less subject to variance induced by the input-data, but is more computationally expensive; against a choice of smaller k (i.e. k=1,2,4) which is subject to larger variances dependent on the input-data, but is computationally cheaper. I tested small batches of parametrisations with k=1,2,4,10,20, and found that k=10 was the largest value tested that ran within a sensible timeframe. I have reworded this in the text to hopefully make this clearer.

L330:

We use the estimator parameter k= 10, as it balances the reduction of estimator variance as kincreases with the increased bias resulting from increased k. We use the estimator parameter k=10, as testing on our data showed that it balances the decrease in estimator variance as k increases against the increase in bias and computational cost as k increases.

15. Line 319-320 – Further, can you elaborate on the ni=10 and nrepeats=20 selection?

The numbers n_i=10 and n_{repeats}=20 we selected fairly arbitrarily. The time taken to compute each MI variance estimate scales linearly with n_{repeats}, so 20 is picked to obtain a good number of values for a maximum likelihood estimate whilst not taking too long. n_i=10 non-overlapping subsets of the VCF profiles are used in each of the computations as the values computed are expected to follow a chi squared distribution with n_i-1 degrees of freedom. Too large a value of n_i takes a long time to perform many mutual information computations, whilst a too small value of n_i requires more repeats to reliably obtain a maximum likelihood estimate for B.

16. Line 325 – And if null hypothesis is rejected, then p is not a good candidate, right? Probably refer to the t-test statistical test you make. Consider rephrasing to "For each p≠p^p \neq \hat{p}p=p^, we apply a Welch's t-test to evaluate the null hypothesis that the mean estimated mutual information I^KSG(p)\hat{I}_{\mathrm{KSG}}(p)I^KSG(p) does not differ significantly from I^KSG(p^)\hat{I}_{\mathrm{KSG}}(\hat{p})I^KSG(p^). Parametrizations for which the null hypothesis is not rejected at the 0.05 significance level are considered candidate optimized parametrizations.".
    I like your phrasing about the null hypothesis and when not it is not rejected. This has been implemented.

L339 – L343:

To do this, we perform an unequal variances (Welch's) t-test for each p≠^p ,with to test the null hypothesis being that that the mean estimated mutual information at p is equal to the mean estimated mutual information at ˆp – that is that ˆIKSG(p) = IKSG(ˆp). If Parametrisations p for which the null hypothesis is not cannot be rejected with a significance of 0.05, we consider p to be a possible candidate for an optimised parametrisation are considered candidate optimised parametrisations. Conversely, parametrisations for which the null hypothesis is rejected are not considered as candidates for the optimised parametrisation.

17. Line 331 – Could you explain how the measurement uncertainties are counted into the selection of partially cloudy category of very small VCF (e.g. VCF < 0.1)?
    As mentioned in my response 12c, the uncertainties of the VCF profiles aren't well defined quantities.
    If there were uncertainties on the VCF values, then the distribution would presumably not be degenerate at 0 and 1 (or could be made continuous through a kernel density estimate that used the values and their uncertainties to compute contributions to the joint probability density of the VCF values at arbitrary values). It is the degeneracies at VCF values of 0 and 1 which make the VCF distribution a mixed measure distribution, and prompted the need to classify cases into the nc, pc and tc categories defined.

18. Line 341 – Please revise Eq. 8. Option A (copula in terms of uniform variables): Option B (copula expressed with the original variables): Both are correct; which one you use depends on context and clarity.
    I revised this to use Option A.

19. Figure 5 – Plot 5c looks like there is contamination from independent data? It appears very similar to Figure 3d. Please clarify this point. Also, if there are no

unique parameterization and the hatching shows several possible good parameterization candidates, then some analysis should be included which shows and explains how much the impact could be.

Figures 3d and 5c are showing completely different quantities. Figure 3d shows a pobability distribution between two non-linearly related variables that is contaminated by independent data. As a result, the distribution has a low mutual information value of 0.018 nats encoded in the distribution.

Figure 5c instead shows a collection of mutual information values computed over a grid of co-location parametrisations. At each point on the mutual information surface, you can imagine that there is a plot similar to 3a-d (but in 100 dimensions), from which a mutual information value is computed, determining the height of the surface at that point in the parametrisation space.

A sentance has been added to make this distinction clearer.

Given the mutual information across the set of candidate parametrisations should be similar, it would imply that the joint probability distribution between the Cloudnet and ATL09 VCF distributions is likely similar across the set of candidate parametrisations too. Thus I would expect that the selection of co-location parametrisations from within the set of candidates should not alter the resulting analysis significantly, except from the volume of data available for the analysis as this changes as a function of ***p***.

L344 – L346:

In our analysis, we will use the parametrisation ˆp that maximises the mutual information between the ATL09 and Cloudnet VCF profiles, but other strategies could be employed to select which candidate optimised co-location parametrisation will be used (e.g. maximising the data volume permitted by the co-location).

L424 – L427:

For each parametrisation p = (R,τ), a joint distribution of VCFs between the ATL09 and Cloudnet data could be plotted, akin to Fig. 3 (albeit in 100 dimensions), from which a value of the mutual information between the ATL09 and Cloudnet VCFs is computed.

20. Line 397 – “As with Nevents, the smooth color gradient … and τ.”But, Nevents have not been expressed as a function of τ. Please clarify this discrepancy.

    We can describe N_events(R, tau) \propto R (i.e. there is no tau dependency, or tau^0). We can also describe N_profiles(R, tau) \propto R^2 tau. The reasons for this are given in the text. The smooth colour gradient described is a result of the variable being plotted (N_events or N_profiles) being polynomial in both of the plotting coordinates (R and tau). Either way, this in depth description of the dependency on N_events and N_profiles to the co-location parametrisations is to show that they are consistent with our expectations, and that N_events and N_profiles vary accross the parameter space in a way that is dissimilar to how the mutual information varies across the parameter space. I have changed the proportionality descriptions to emphasise that N_{events/profiles} are functions of the co-location parametrisations.

L409:

The result is Nevents being approximately proportional to porbitsR outcome is that Nevents(R,τ) ∝R.

L423:

The results are consistent with Nprofiles ∝R2τ Nprofiles(R,τ) ∝R2τ.

21. Line 418-419 – I am not sure if I interpreted correctly. Do you imply that the difference of 10 hours (form 8 hours to 18 hours) is significant or not?
    The significance described in the text is with respect to the differences in mutual information computed at different parametrisations, not with regards to the values of the parametrisations themselves. Parametrisations with tau<8 hours will include less data, making the analysis data-limitted. Parametrisations with tau>18 hours will permit data that is independent between Cloudnet and ATL09, enough so that it would degrade the inferred relationship between the VCF distributions from each data source.

22. Figure 6c – The dispersion of the possible optimal co-location parameterization is high. There is some explanation/interpretation in the text which associates this to the flat terrain around the ground-based station. Could you give an estimate of how much such a condition could impact the comparisons over Hyytiala?
    I don't fully understand the question being posed here, but I will attempt to answer what I believe it to be – How would selecting different co-location parametrisations for Hyytiala affect the comparison between ATL09 and Cloudnet data, and how does this relate to the terrain?
    It is my belief that selecting any of the parametrisations within the hatched region in Figure 6c should produce quantitatively similar final results in a comparison between the Cloudnet and ATL09 data. In theory, the 50-dimensional marginal distributions of the VCF profiles for both ATL09 and Cloudnet do not significantly vary across the set of candidate parametrisations. As such, the only difference in the mutual information between the ATL09 and Cloudnet distributions will come from changes in the relationship between the two distributions, their joint distribution. I struggle to see how over a small range in parameter space that the joint distribution relationship would change drastically in form, such that the mutual information remains similar but that the conclusions that would be drawn from the relationship would significantly change.
    The argument relating the terrain to the larger hat{R} values is that if there were nearby mountains, that could induce changes in the cloud regime in the vicinity of the mountains. So the clouds near the mountains would be affected by different physical processes than the clouds at the ground-based station. Thus, a parametrisation whose R value was big enough to include the mountains would be introducing observations between physically independent clouds at smaller values of R compared to a location without nearby mountains (or other large changes in orography).

23. Line 427-428 – It appears like there is contamination from independent samples like in Fig. 3d. Could you please clarify this point and probably rephrase the sentence "regions where the input data to the estimator is contaminated with independent samples"?
    Figure 6 shows mutual information surfaces, not individual distributions which can be contaminated with independent data (see my response to 19a).

24. Line 434 – The statement “are similar orders of magnitude” is a bit misleading here as the minimum of 4 hours is less than half the maximum of 10 hours. Please rephrase that.
    I think 4 and 10 are similar orders of magnitude. I‘m considering 10 to be an order of magnitude, in which case the numbers differ logarithmicly by ~0.4.

25. Line 435 – There is a statement that the temporal values τ are consistent with other studies but there is no statement about the spatial values R. Are also those consistent with available literature?
    I have looked for available literature, I haven’t been able to find much explicitly about the horizontal scales of cloud variability from satellite observations. The closest I could find is Feijt and Jonker, (2000), *Comparison of scaling parameters from spatial and temporal distributions of cloud properties* (DOI: 10.1029/2000JD900414) but this more addresses finer scale spatial variation within a given cloud regime as opposed to the scale of variation between cloud regimes.

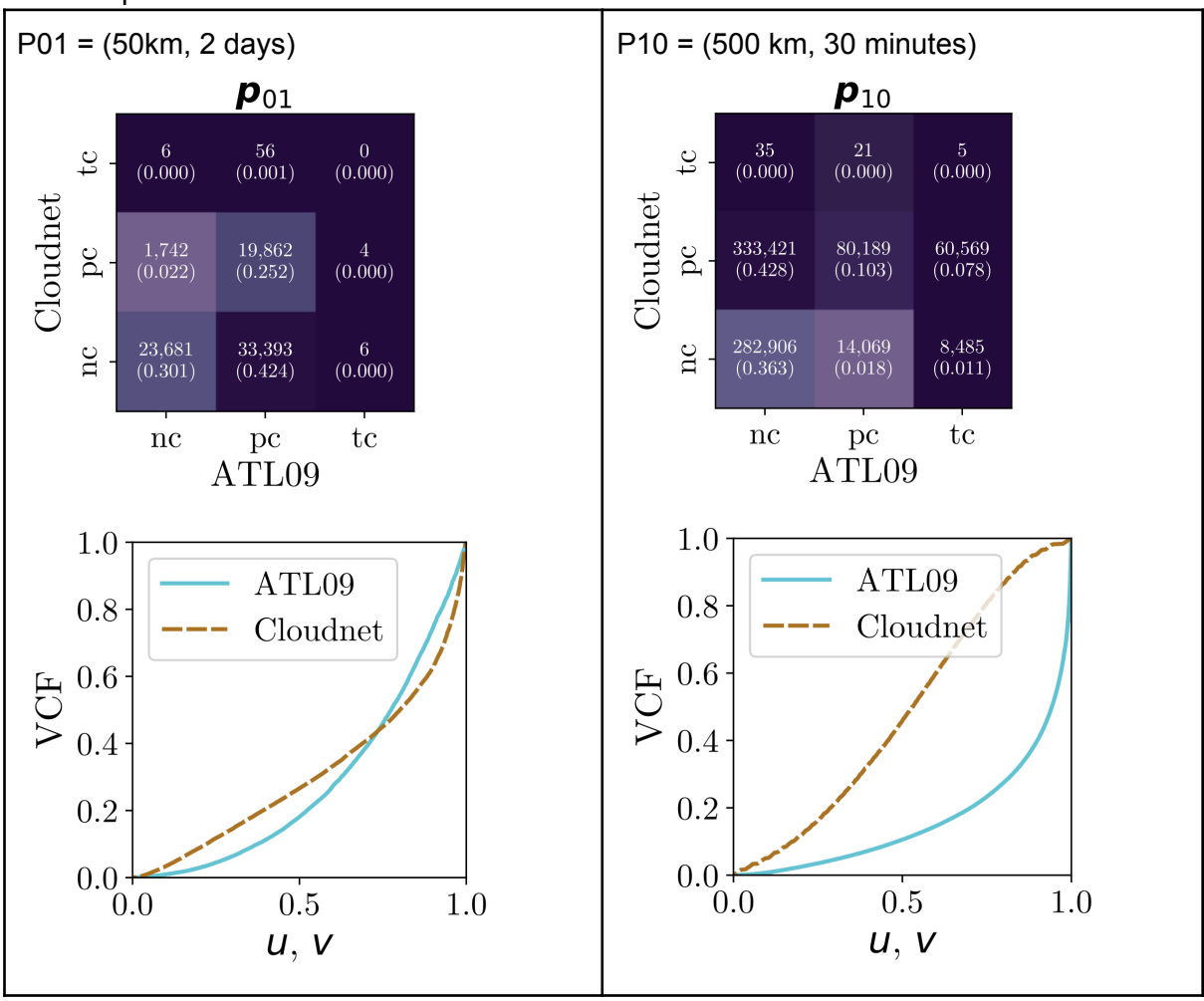
26. Line 455 – If I understood correctly, the selection of 100-dimensional joint probability distribution constrains the sampling in the other 3 locations. Probably for each location, depending on the latitude that the ground station lies, other selection should be done? What is the reasoning of having 100 dimensions?
    The dimensionality of the joint probability distribution is the sum of the dimensionalities of the individual measurements being compared. In this case, our VCF profiles consist of cloud fraction values recorded over a vertical profile containing 50 distinct vertical levels. As such, the distribution of each VCF measurement is 50 dimensional, and the joint probability distribution is 50+50=100 dimensional. If we were instead comparing integrated cloud geometric thickness from ICESat-2 (1 dimensional), and mean cloud geometric height and thickness from Cloudnet (2 dimensional), the resulting joint probability distribution would be 1+2=3 dimensional. The dimensionality depends purely on the measured values being compared, and not on where they are being compared.
    A 100 dimensional joint probability distribution is rather horrible. The dimensionality can be reduced by only considering coarser vertical grids for VCFs, but this reduces vertical resolution. If a different vertical resolution were used at each location to account for the different sample density at each location, the results of the subsequent comparisons between the ATL09 and Cloudnet VCFs would be incomparable between locations.

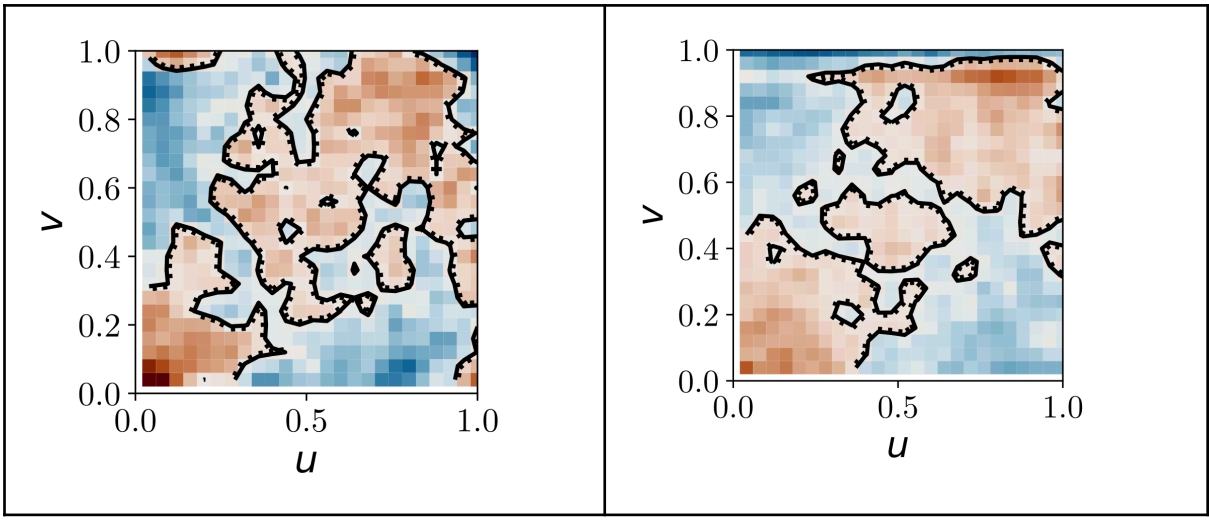
27. Line 476-480 – I cannot clearly see what it is hidden in those sentences. Could you proof that another set of co-location parameterization would not be suitable? Could you think of any validation of the optimal co-location parameterization?
    \hat{p}_{Ny-Ålesund} is provided as an example, but using any of the \hat{p} values from a specific site would induce problems at other Cloudnet sites (by selecting a parametrisation with reduced mutual information for whatever reason), hence the one-size-fits-all approach being flawed.
    Inherently, there is no proof that a different tested co-location parametrisation would be better as the parametrisation that produces the highest mutual information is the optimal choice by definition within the framework.

A validation of the optimised co-location parametrisation is possible however, by comparing its performance against other co-location parametrisations using independent metrics. This is demonstrated in Sect. 3.6.3.

28.

a. Instead of comparing with the two extremes p00 and p11, I would also compare the confusion matrices for other potential candidates of a co-location parameterization. Probably, I would pick the one with the smallest temporal window and the one with the largest spatial extend.
This was in fact done, and is the reason why p_{00} and p_{11} are not simply p_0 and p_1. Similarly to p_{11} and p_{00}, p_{10} (large R, small tau) and p_{01} (small R, large tau) underperform across all metrics in Table 2 when compared against p_{lit} and \hat{p}. They were removed to make the manuscript more concise.





parametrisation	R (km)	tau (hours)	ACC	c_min	c_max	c(1,1)	RMSD
p_{01}	50	48	0.553	0.35	2.08	0.35	0.16
p_{10}	500	0.5	0.466	0.50	1.48	0.83	0.13

b. Please also add in the legend what p00 and p11 mean. The parameterization is given in the text but the plots should also be self-informative without the need to read all the text.
I have directed the reader to Table 2 within the caption of Figure 7.

c. I also have a problem with the group of plots e-h: can they be biased due to the first 1.0-1.5 km that the VCF exponentially increases with height (w.r.t/ the incomplete overlap)? Please provide some more cases including also the other Cloudnet stations. If the VCF profiles have the same structure, then consider to apply some filtering/pre-processing prior to the co-location parameterization selection scheme.
The cumulative distribution functions shown in Fig. 7e–h account for all of the nonzero VCF values across all heights, across all considered VCF profiles from all sites, using the given co-location parametrisation. These are generated by flattening both the ATL09 and Cloudnet VCF profiles, and effectively generating a joint histogram (and then cumulative distribution) from all of the flattened VCF profiles together. Thus, all of the Cloudnet stations are already included in this analysis, and height information is ignored beyond the pairing of values between the Cloudnet and ATL09 VCFs.

Figure 7 caption:
The values of the co-locatio parametrisation vectors are given in Table 2.

29. Line 504-505 – From the statement I understand that plit is still a good choice. Therefore, the empirical choice of the parameters τ and R as taken from the literature appear as good as the one that you found from the maximum mutual information. Please clarify this point.

If we were only to consider the confusion matrix for the (nc, pc, tc) classifications, then yes, p_lit and hat{p} would appear to perform similarly. The rest of the section goes on to make the argument that if we look at the actual relationship between the ATL09 and Cloudnet VCF values through the copula density, that hat{p} has metrics that indicate a better relationship between the Cloudnet and ATL09 VCFs when using hat{p} over p_lit.

30. Line 508-509 – "The shapes of … at lower values." Could that be due to problems of incomplete overlap? Please clarify this point.
    The large distance between ICESat-2 and the highest altitude of the ATL09 data product, relative to the 15km height span of the ATL09 data product, means that incomplete overlap is unlikely to affect the distribution of VCF values determined from the ATL09 data. If there is a multiplicative bias in the ATL09 signal due to incomplete overlap, it would be fairly homogenous throughout the scene, which would have little impact on the shape of the cumulative distribution.

31. Figure 8 – There are essential problems with the visibility of the axis's labels in printed form. Please try to make them also friendlier for people with color vision deficiencies (CVD); certain color combinations are hard or impossible to distinguish and should be avoided.
    Thanks for highlighting this. I've been using the `cmcrameri` python package (Crameri, 2023) and the Coblis online CVD checking tool (https://www.color-blindness.com/coblis-color-blindness-simulator/). I have also made extensive efforts to make plots CVD friendly through other means, such as differentiating line markers and styles. For Figure 8, I'm using a colorblind friendly map (Acton) for the probability density, consistent with the rest of the paper. Plotting lines over the top of a colourblind friendly colourmap will almost necesarily have places with poor contrast (particularly under monochromaticity) when the line goes from lighter to darker parts of the underlying colourmap. I have tested Figure 8 against all given CVD conditions provided on Coblis. I agree that there are places in 8a–d where there is poor contrast between the underlying probability density and the line plotted over the top. As such, I have changed the colours of the lines to ensure that reasonable contrast is maintained between the lines and the probability distributions in 8a–d, and that there is some contrast in 8e–f, but distinction between the lines can be determined by their plotting styles.

32. Line 528-530 – I would not conclude the session with such a strong sentence. There is no validation which proves the statement, right?
    I have changed the end of the section to remove language like "better" and make the statements more objective.

L554 – L555:

These results show that using the optimised co-location parametrisation ˆp at each Cloudnet site yields a better This indicates that the dependency structure of the relationship between the VCF distributions than using other parametrisations, either by including fewer independent samples, or by including a larger number of dependent samples from Cloudnet and ATL09 data is stronger with ^p that with plit., and the other tested parametrisations.

33. Line 534-535 – Could this be related to signal attenuation from the ground-based lidar? Please comment on this.
    This is due to the different sensitivities of the retrievals to atmospheric scatterers at different heights. This is partially due to a decrease in sensitivity of the retrievals with separation between the scatterer and receiver (due to the inverse-square law) and it is partially due to attenuation effects. I have mentioned attenuation effects in the manuscript to clarify this point.

L560:
This could be explained by the viewing geometry geometries (i.e. ICESat-2 viewing clouds from above and Cloudnet viewing clouds from below) and the effects of signal attenuation on the retrievals, and is consistent with comparisons…

34. Line 824 – Equation B1 could be rephrased similarly to Eq.5 at Line 278
    This has been implemented.

35. Line 880 – Equations C10-C11 must have a wrong sign. Please check if the minus sign (-) must be present in the aforementioned equations.
    Thank you for spotting the sign error. I have changed the equations to rectify this, and have filled in the equations from (what was) C14–C16 (and what is now C13–C17) to show how the full derivation for how the final result is obtained.

36. Figure C1 – Why do you include in the plot also the ISS, EarthCARE and A-Train satellites? Relevant to that, there is some text in lines 901-902 about those satellites. Please provide a statement of why would this be relevant to this manuscript? Do you foreseen to apply this co-location parameterization to other missions?
    I include the other satellites in the plot, as they are satellites from which vertically resolved atmospheric measurements are made. The same framework and same co-location scheme described in the manuscript can be applied to retrievals from all of these satellites when comparing them against surface-based observations. I think the framework as a whole can be applied to any and all satellite retrievals, with the outcome being the identification of a co-location parametrisation that is location and data-source dependant.

Technical corrections:

- Line 16 – Typo of “validation an constraining” should be “validation and constraining”.
  corrected

L16:
validation an and constraining

- Line 62 – Remove the second “to the”.
  corrected

L62:
a prior expectation to the is effectively applied to the comparison metric in the analysis.

- Line 103 – Remove the second “be”
  corrected

L103:
co-location scheme can be be described by a finite number

- Line 118 – “Co-location” with capital letter.
  corrected

L117:
co-location Co-location events consist of paired homogenised data between

- Line 229 – Please correct the ATLAS acronym to Advanced Topographic Laser Altimeter System (ATLAS)
  corrected

L243:
The satellite has a single instrument payload, the Advanced Topographical Topographic Laser Altimeter System (ATLAS)

- Line 452 – There is an extra “a”. Please remove it.
  corrected

- Line 541-543 – This sentence reads like being replicated to the lines 535-537. Please remove or rephrase.
  corrected

L566:
This could be explained by the differences in viewing geometry between the platforms, and the general trend is consistent with results from other comparisons of vertically resolved satellite retrievals of cloud presence against surface based observations (McErlich et al., 2021).

- Line 562-563 – In case you do not remove the complete section 3.7, please remove the conclusive sentence.
  corrected

L587:
With no independent analysis on the choice of p, there is no way to distinguish between good and bad co-locations.

- Line 776 – “As 1/N ..” with capital A.
  corrected

- Line 784 – The sentence does not read well. Remove either the word “is” or add the word “that” after “we see”.
  corrected

- The doi of the reference “Herzfeld, U., Palm, S., and Hancock, D.: Ice, Cloud, and Land Elevation Satellite (ICESat-2) Project Algorithm Theoretical Basis Document for the Atmosphere, Part 2: Level 2 and 3 Data Products, version 4, https://doi.org/10.5067/JT3IX3YFW1RR, publisher: NASA National 960 Snow and Ice Data Center Distributed Active Archive Center, 2021b.” is wrong.
  corrected

- 48. The same for the doi of the reference “Palm, S., Yang, Y., Hertzfeld, U., and Hancock, D.: Ice, Cloud, and Land Elevation Satellite (ICESat-2) Project Algorithm Theoretical Basis Document for the Atmosphere, Part I: Level 2 and 3 Data Products, version 4, https://doi.org/10.5067/VZIDW16UA4S1, publisher: NASA National Snow and Ice Data Center Distributed Active Archive Center, 2021a.”
  corrected

- 49. The same for the doi of the reference “Vinjamuri, K. S., Vountas, M., Lelli, L., Stengel, M., Shupe, M. D., Ebell, K., and Burrows, J. P.: Validation of the Cloud_CCI (Cloud Climate Change Initiative) cloud products in the Arctic, Atmospheric Measurement Techniques, 16, 2903–2918, https://doi.org/10.5194/amt-16- 1095 2903-2023, publisher: Copernicus GmbH, 2023.”
  corrected