A framework for evaluating ice sheet altimetry uncertainty estimates

egusphere-2026-556

31st July 2026

Dear Prof. Machguth,

We thank you for considering our manuscript for publication in *The Cryosphere*. We are also grateful to both reviewers for their detailed, constructive feedback on our manuscript, and were pleased to read that they noted both its potential for high scientific impact and its novel contribution to the topic of uncertainty assessments, which is urgently needed. Below, we address in full each of the points raised by the reviewers. We have also made extensive revisions to the manuscript accordingly. As requested by Reviewer #2 (Martin Horwath), the most significant modification has been to update the uncertainty evaluation framework to include elements with a deeper mathematical grounding. Additionally, as requested by both reviewers, we have made significant changes to further clarify and refine the framework and methodology.

In the following response document, we first present a detailed summary of the major change described above (hereafter referred to as **Major Change**) and then provide point-by-point responses to the reviewers' detailed comments. We have included the reviewers' comments (*italicised in blue*), our responses (black text) and amendments to the original text (*italicised in grey*). Unless otherwise stated, line and figure numbers referred to below are in line with those in the original manuscript, available at https://doi.org/10.5194/egusphere-2026-556.

Kind regards,

Karla Boxall

(on behalf of the co-author team)

Author Response to Reviewers

Major Change

We thank both reviewers for their detailed and constructive reviews. We understand the primary concern, raised by Reviewer #2, to be the mathematical justification for our use of the similarity of the uncertainty and absolute unseen error distributions as a measure of uncertainty performance. First, we summarise our understanding of their principal concerns and proposed alternative approach, before outlining the implications for our work and the resulting revisions to the manuscript.

In their review, the reviewer illustrates their point, first, by plotting a set of measurements with a uniform (i.e. unique) uncertainty value, alongside a normally distributed set of random errors drawn from the associated distribution (Figure RC2.1). Here, we understand their argument that each (identical) uncertainty value *represents a single metric describing the distribution* of possible error values, and thus the distribution of uncertainty values cannot be expected to match that of the errors drawn randomly from that distribution. We agree entirely with the argument made by the reviewer here.

In the second thought experiment presented by the reviewer, an arbitrary distribution of uncertainty values is produced. For each uncertainty value, we understand that a corresponding error realisation is computed by first generating a corresponding Gaussian distribution under the assumption that the uncertainty characterises either the standard deviation or the median absolute deviation of that distribution, and then secondly drawing a random sample from that distribution, which is taken as the error estimate. The distribution of the resulting randomly drawn error values, alongside the distribution of the uncertainties from which they originated, is shown in Figure RC2.2. As such, our understanding is that the second thought experiment essentially adopts the same concepts of the first thought experiment, with the principal difference being that that it iterates over multiple non-unique uncertainty values. We therefore again agree with the reviewer's conclusion that the distribution of uncertainty values in this second thought experiment cannot be expected to match the distribution of errors exactly; given that a single summary metric of a distribution characterises one property of the distribution (through dimension reduction), and is therefore fundamentally distinct from the distribution itself.

Building upon the arguments above, the reviewer highlights an alternative measure of uncertainty performance, which has precedent in the literature in terms of its use for other satellite retrievals, and was not considered within our original study. Specifically, they suggest considering the approach described by Merchant et al. (2017), whereby the quotient of uncertainty and unseen error is computed, and the resulting distribution of the quotients is evaluated for normality. In their study, the authors conclude that a well performing uncertainty methodology will produce a normally distributed set of quotients with a standard deviation close to 1, and a full width at half maximum (FWHM) close to 2.35. Such a distribution is considered desirable, because it indicates that the discrepancy between uncertainty and error values is aligned with that expected for a standard uncertainty; that is, 68% of error values fall within ± 1 uncertainty.

The approach described by Merchant et al. (2017) is based upon the theoretical understanding that a level of discrepancy is to be expected between a stated uncertainty and any individual realisation of error, as the uncertainty corresponds to a metric that is used to describe the probability distribution from which the error is realised. The extent to which this discrepancy is normally distributed, with a standard deviation equal to unity, quantifies uncertainty performance as follows:

- **a)** FWHM > 2.35: a larger-than-expected proportion of error values exceed the stated uncertainty, which indicates that uncertainties are underestimated
- **b)** FWHM < 2.35: a smaller-than-expected proportion of error values exceed the stated uncertainty, which indicates that uncertainties are overestimated
- **c)** FWHM = 2.35: quotients normally distributed with a standard deviation equal to unity, which indicates that the stated uncertainty accurately reflects the error values

Having summarised our understanding of the reviewer's main analysis relating to the conceptualisation of our framework, we now provide a more detailed response to address the principal comments and points made. In summary, the additional work that we have performed to address this review comprises two new major elements:

1. ***Major Change 1:*** Substantial additional work to implement and analyse the alternative approach suggested by the reviewer, and to interpret the results within the context of altimetry observations over ice sheets.
2. ***Major Change 2:*** Additional edits and clarifications to better articulate the underlying motivation for the original approach that we had adopted, and to evaluate whether it provides complementary insight to the approach of Merchant et al. (2017), in terms of practical, interpretable and actionable information for downstream users of altimetry products.

In making these substantive changes, we believe that the revised manuscript – and the new analyses presented within it – is much clearer, more comprehensive, and has stronger mathematical justification. We are grateful for the reviewer's detailed comments and assessment in this regard.

Major Change 1

Having reflected carefully on the reviewer's comments, we propose to thoroughly revise the study so that our framework is fundamentally grounded according to the approach of Merchant et al. (2017), as suggested by the reviewer. Specifically, we conduct extensive new analysis – across all uncertainty metrics, study regions, and satellites – to assess uncertainty performance according to the extent to which the resulting quotients are normally distributed with a standard deviation equal to unity. This, we believe, will address the reviewer's request for a deeper mathematical grounding of the framework.

The updated framework is described in the updated Section 2.1, and the results of the new analysis are presented in the updated Section 3.

Major Change 2

We also make several important changes, additions, and clarifications to our existing assessment to address the points raised by the reviewer. Specifically, this new work evaluates the extent to which our original analysis is consistent, useful and complementary to the approach suggested by the reviewer.

First, we clarify that exactly matching distributions of uncertainty and error cannot be expected owing to their mathematical formulation, as was highlighted by the reviewer and described in the opening sections of this response; i.e. that individual uncertainty values, by definition, aim to characterise an *entire distribution* from which the error is drawn. Instead, we emphasise that the principal value of our original approach is that it provides a useful relative measure to compare different approaches, rather than implying that exactly matching distributions are an achievable goal.

The relevant text from the updated Sections 2.1 and 3, which addresses these points, is included below for reference.

> *"Although residuals are not expected to be exactly zero even for a well performing uncertainty, it is preferable for their distribution to have a relatively narrow peak centred close to zero, as this indicates a tendency towards a close correspondence between individual uncertainty and elevation difference values."*

> *"Although the distributions will not match exactly due to the fact that uncertainty values, by definition, aim to characterise an entire distribution of elevation difference values (Section 2.2), the relative agreement of the distributions' shapes and spreads provides an additional measure of the appropriateness of each uncertainty estimation method."*

> *"Overall, the findings drawn from the relative comparison of residual distributions mirror the conclusions drawn from the first element of the framework"*

> *"In this case, this manifests as a relative shift away from residuals distributions with a narrow peak, and that are centred around zero (Figure 3)."*

> *"The final element of the framework assesses the relative similarity between the overall distribution of uncertainties and the distribution of absolute elevation differences, across each of our uncertainty methodologies."*

> *"Specifically, for the POCA dataset, although distributions will not match exactly, the median metric results in an uncertainty distribution closest to the elevation difference distribution, according to the alignment of the medians and quartiles."*

> *"The final element of the framework supports these findings in Antarctica as relatively good agreement between the distribution of uncertainty and elevation difference is observed when the median is the chosen metric, with the agreement degraded significantly for other metrics (Figure 10)."*

Second, to further explore the value of our original approach in practice, within our revised manuscript we now compare the Merchant et al. (2017) approach to our original approach, to assess whether they arrive at consistent conclusions. We also include new discussion which critically reflects on these findings, in terms of the ability of both methods to differentiate the relative performance of different uncertainty methods.

Below, we include key elements of the revised Results (Section 3) relating to this:

> *"Overall, these findings mirror the conclusions drawn from the first element of the framework; namely, that using the median to populate the lookup table results in the most robust set of uncertainties for the POCA dataset, whilst the choice between the median and STD UB is less clear for the swath dataset."*

> *"Collectively, both elements of the framework therefore provide a consistent picture that increasing the number of co-variates with which to parameterise uncertainty enables a more accurate characterisation of the associated error."*

"The final element of the framework assesses the relative similarity between the overall distribution of uncertainties and the distribution of absolute elevation differences, across each of our uncertainty methodologies. <u>This analysis again supports similar high-level conclusions to those drawn from the first two elements of our framework</u>."

"<u>As was found for the quotient analysis,</u> Figure 6 demonstrates that the STD UB performs slightly better than the median for the swath dataset."

"Turning next to the analysis of the Sentinel-3 residuals (Figure 8<u>), as was the case for the assessment of the quotients</u>, we again find that for the Antarctic Ice Sheet, the choice of metric significantly influences the robustness of uncertainties generated."

Finally, we have added new text to reflect on the relative value and complementarity of the two strands of our revised framework, with the quotient-based approach quantifying the extent to which the discrepancy between uncertainty and error is as expected, and our original analysis contributing additional practical insight that will be informative for the reader and downstream user communities. Beyond identifying the relative improvement in performance between different uncertainty estimation methodologies, we believe that at a practical level our approach provides added value for the user in terms of ease of interpretation. Our approach shows the divergence in uncertainty and error magnitudes at both population and point scales, enabling immediate and intuitive comparisons to be made. For instance, the interquartile ranges (IQRs) of two residual distributions quantify, in each case, the range within which 50% of uncertainty estimates lie relative to the unseen error. Moreover, the potential to conduct further interrogation of the residuals, such as analysis of the cumulative distribution and spatial distribution, provides further, directly interpretable information for the user. While similar interpretations can be inferred from the Merchant et al. (2017) approach, the extent to which the quotient distribution is normally distributed is likely to be less directly interpretable to many downstream users.

The key revisions to the updated Sections 2.1 and 3, which address this point, are provided below.

"<u>This element complements the information derived by the quotient by providing the true magnitude of the bias, rather than the normalised bias (expressed as a ratio of elevation difference to uncertainty), while also enabling assessment of the proportion of over/under-estimation (via the cumulative distribution of residuals) and the spatial variation in bias magnitude (via a spatial plot of residuals).</u>"

"Although the distributions will not match exactly due to the fact that uncertainty values, by definition, aim to characterise an entire distribution of elevation difference values (Section 2.2), <u>the relative agreement of the distributions' shapes and spreads provides an additional measure of the appropriateness of each uncertainty estimation method</u>. <u>Specifically, this comparison can be used to identify the nature of over- or under-estimation, by distinguishing between methods in which the uncertainty is uniformly biased relative to the elevation difference, versus those in which the discrepancy is disproportionately concentrated in the tails of the distribution.</u>"

"This indicates a higher proportion of the dataset with a close correspondence between the uncertainty and the respective elevation difference, <u>with ~50% of uncertainty values falling within 1.03 – 1.22 m of the elevation difference</u>, depending on the choice of co-variates."

"This indicates that half of the swath uncertainties fall within 3 – 4 m of the elevation differences, regardless of the methodology chosen."

"Overall, these findings mirror the conclusions drawn from the first element of the framework; namely, that using the median to populate the lookup table results in the most robust set of uncertainties for the POCA dataset, whilst the choice between the median and STD UB is less clear for the swath dataset. Additionally, the analysis performed within this second element provides further complementary information relating to the absolute magnitude of uncertainty over- or under-estimation, which is useful at a practical level for downstream end-users."

"The second element of the framework can further aid the interpretation of uncertainty performance through the analysis of the cumulative distribution of residuals (Figure 4). The cumulative distribution shows the proportion of measurements meeting a given magnitude of uncertainty over- or under-estimation, with the point at which the distribution intersects zero indicating the percentage of negative residuals (where the uncertainty is greater than the recorded elevation difference)."

"Analysis of the spatial distribution of the point-to-point residuals provides additional useful diagnostic information relating to regions where under- or over-estimated uncertainties are found."

"Compared to the other two elements of the framework, this third component provides complementary population-level information, in terms of the extent to which the magnitude and spread of uncertainties mirrors that of the elevation differences. More broadly, this can help to diagnose specific sections of the elevation difference distribution that are inadequately represented by a given uncertainty method; for example, a failure to capture low- or high-magnitude errors."

"For the swath data, the magnitude of the residual biases demonstrates that the median results in a more desirable set of uncertainties. Specifically, for a four-dimensional swath lookup table, the distribution of the residuals produced by the median metric (median=0.29 m, IQR=2.12 m) are preferential to those produced using the STD UB (median=-0.64 m, IQR=2.48 m) (Figure B5). In this case, the optimal choice of metric is unclear based solely on the quotient analysis, with the standard deviation of the quotient distribution being 1.26 and 0.77 for the median and STD UB, respectively, indicating the same proportion of under-estimated uncertainties using the median, as over-estimated uncertainties using the STD UB (Figure B4)."

"This finding is less clear in the first element of the framework (Figure 7), which succeeds in demonstrating that the degree of uncertainty over-estimation is less in Greenland compared to Antarctica (as the FWHM values are larger and closer to 2.35), but fails to visually demonstrate that this results from an unrealistically concentrated distribution of uncertainty values (Figure 10), thus highlighting the complementary nature of the different elements of the framework."

Reviewer #1

General Comments

1. *This manuscript presents a framework for evaluating and improving uncertainty estimates in satellite radar altimetry measurements over ice sheets. The topic is timely and highly relevant, as robust and traceable uncertainty estimates are essential for combining multi-mission datasets and for reliable assessments of ice sheet mass balance and sea-level rise. In this respect, the study addresses an important gap in the current literature.*

We thank the reviewer for their assessment and were pleased to read that they believe the topic to be "timely and highly relevant" and that our manuscript "addresses an important gap in the current literature".

2. *Overall, the manuscript shows strong scientific potential. The work fills a gap in the scientific literature and could have a high scientific impact. At the same time, I believe it would benefit from further clarification and refinement before it can be fully assessed. In particular, clearer definitions of key concepts, along with a more structured presentation of the framework and methodology, would significantly improve readability and accessibility.*

We thank the reviewer for their suggestions to further improve readability of the manuscript. Please refer to our responses to Comments 4 and 5 (Reviewer 1) for the changes made to more clearly define key concepts, and to Comment 3 (Reviewer 1) for those improving the readability of the methodology.

3. *I found the section "Uncertainty Evaluation Framework" somewhat difficult to follow in its current form. The purpose of the section is not entirely clear, as it appears to serve partly as an introduction and partly as a method description, without fully achieving either. It might be helpful to either (i) integrate a concise overview of the framework into the Introduction (e.g., as a structured summary or bullet points), or (ii) expand this section within the Methods to provide a clearer step-by-step description, including definitions, inputs, and outputs.*

To address this point, we have made a number of changes to the text and structure of Section 1 (Introduction), Section 2 (Uncertainty Evaluation Framework) and Section 3 (Data and methods). We have now included a concise overview of the framework in the Introduction:

"In this study, we develop a new framework for evaluating the performance of uncertainties generated for radar altimetry-based ice sheet elevation measurements. More specifically, t<u>he aim of this framework is to assess uncertainty fidelity, by quantifying the extent to which uncertainty values adequately capture the deviation of a given measurement from a reference dataset.</u>"

We have also overhauled the structure of Section 2 (Uncertainty Evaluation Framework) and Section 3 (Data and methods). Section 2.1 now provides a clear description of the framework, with the aim, inputs and quantifiable measures defined for each element.

4. *In addition, it would strengthen the manuscript to clearly define the chosen metrics and covariates early on, as these form the backbone of the framework but are not always consistently introduced.*

The co-variates are now defined at their first mention in Section 2:

“For the Cryo-TEMPO POCA product, the elevation difference to a reference dataset is parameterised as a function of surface slope (gradient of the surface) and, optionally, roughness (irregularity of the surface defined as the standard deviation of the orthogonal residuals from a fitted plane), while for Cryo-TEMPO EOLIS, the elevation difference is parameterised as a function of across-track slope, along-track slope, roughness, power (strength of the backscattered signal), phase coherence (computed across the two antennas), and distance to POCA.”

The metrics are also defined at their first mention in Equations 4 and 5 in the updated Section 2.2.

5. *Also, as this is an uncertainty study, please clearly state what uncertainty, robustness, etc., means. Fx, does robustness refer to stability across configurations, agreement with residuals, or bias minimisation? You use the word, particular in Section 2, without describing what you mean by it (Comes later in Methods).*

A definition of uncertainty has been added to the Introduction.

“To capitalise fully on the extensive value of satellite altimetry, it is necessary for robust and traceable uncertainties to be provided alongside ice sheet elevation measurements, where uncertainty is defined to be a statistical measure characterising the dispersion of values that may reasonably represent the true surface height (JCGM, 2008).”

A definition of robustness has been added at the beginning of Section 2.1.

“Here, we define robustness to be the extent to which a set of stated measurement uncertainties (σ) adequately characterises the true deviations between a set of altimeter (E_{alt}) and reference (E_{ref}) measurements.”

6. *The decision to construct a look-up table based on a single reference year (2020) would benefit from further justification. Given known temporal and spatial variations in datasets (e.g., ICESat-2 sampling), it would be useful to discuss potential implications of this choice, including any seasonal or regional biases. Also, it is not clear to me that the resulting uncertainties are solely from 2022. Furthermore, I am missing your consideration in choosing exactly these years.*

The following text has been added to justify the choice of data for uncertainty generation.

“The choice to utilise all available data spanning the continent over an entire year was designed to ensure that the resultant uncertainties were not unduly affected by the influence of regional and seasonal biases, respectively.”

“ICESat-2 was chosen owing to its comprehensive ice-sheet-wide coverage with a 91-day repeat cycle at a high along-track spatial sampling of ~20 m, and its temporally and spatially overlapping availability with CryoSat-2 data.”

We have also clarified the choice to evaluate uncertainties from 2022 only.

“Following the lookup table creation, the table is then used to assign uncertainties to a distinct set of elevation measurements acquired in 2022. These uncertainties are subsequently evaluated using the framework. This was in place of using the uncertainties provided within the Cryo-TEMPO POCA and Cryo-TEMPO EOLIS products, which would have prevented the evaluation of multiple different uncertainty generation methods. The year 2022 was chosen to ensure that the evaluation was

performed using a temporally distinct dataset from that used to construct the look up table, and during which the reference data (ICESat-2) was fully operational and hence able to provide the required unseen reference data. In this regard, we note that any year with suitable unseen reference data (i.e. not used in the construction of the lookup table) could be evaluated using the framework."

7. *The manuscript aims to cover a broad range of analyses, including comparisons across multiple CryoSat-2 SARIn mode datasets and Sentinel-3 observations over both Greenland and Antarctica. While the motivation for this comprehensive approach is clear and appreciated, there are instances where the level of detail, particularly regarding how the uncertainties evolve in space and time, feels somewhat limited. At present, the scope appears quite ambitious for a single paper, and the material could potentially support multiple focused studies. I am not suggesting that the manuscript should be split, nor that its length should be substantially increased. Rather, I encourage the authors to consider whether certain aspects could be treated with slightly more depth or clarity, possibly by refining the focus or guiding the reader more explicitly through the key uncertainty-related findings.*

Please see our responses to Comments 29-35 (Reviewer 1) below that details the refinements made to the discussion of results, which we hope will fully address the comments raised here.

Specific Comments

8. ***Abstract***
   The abstract could be strengthened by including more specific results of your findings (e.g., comparisons between POCA vs swath, Sentinel-3 vs CryoSat-2, or regional differences between Antarctica and Greenland), and the years your analysis covers.

We have added the following sentence to the abstract:

"The framework also demonstrates improved uncertainty characterisation for point of closest approach elevation products relative to swath-processed products, for CryoSat-2-derived products compared to those derived from Sentinel-3, and highlights regional differences in uncertainty performance between Antarctica and Greenland."

For concision, we would prefer to use the abstract to introduce the framework in a generalised sense so as not to limit the application of the framework to other scenarios, such as additional datasets, years or regions. For this reason, we have not included in the abstract the specific years used to demonstrate the framework. This information is provided in Sections 2.3.1 and 2.4.1.

9. ***Introduction***
   Consider adding a short paragraph summarising existing approaches to uncertainty estimation in satellite altimetry (e.g., cross-over analysis, external validation with GNSS/airborne lidar/DEMs, inter-mission comparisons, statistical approaches).

We have added the following sentence and references to the introduction:

"Although satellite altimetry precision and accuracy have been estimated post hoc, using single-cycle cross-over analysis (Zwally et al., 2002; Brenner et al., 2007; McMillan et al., 2018; Schröder et al., 2019) and external validation with ICESat laser altimetry (Brenner et al., 2007, Rémy and Parouty,

2009), ICESat-2 laser altimetry (Huang et al., 2026), airborne lidar (Wang et al., 2015, McMillan et al., 2018, Schröder et al., 2019) or Global Navigation Satellite System (GNSS) data (Brunt et al., 2021), uncertainty estimation during product generation is often lacking, and uncertainty estimates are absent from the standard CryoSat-2 ground segment (ESA, 2023a; 2023b) and from existing Sentinel-3 ice sheet elevation products (ESA, 2025a)."

10. ***Uncertainty Evaluation Framework***
    L61–62: The statement regarding the absence of approaches to assess uncertainty may be somewhat strong; several studies have addressed this topic, even if not within a unified framework (see comment from the introduction).

We have removed this statement from Section 2.1. Please also see our response to Comment 9 (Reviewer 1).

11. *L66–71: The choice of 2020 (and 2022?) for the lookup table would benefit from additional explanation.*

Information regarding lookup table construction and the specific years evaluated have been removed from Section 2.1, which focuses on introducing the framework conceptually. Sections 2.3.1 and 2.4.1 detail and justify the year chosen for application of the framework (please see our response to Comment 6 (Reviewer 1) for further detail).

12. *L69: Please clarify which datasets are included in the look-up table.*

This information is now included in Section 2.3.2:

"For this reason, given that the purpose of this analysis is to use our framework to intercompare results from POCA and swath uncertainty generation, only the co-variates common to both datasets (i.e. slope, roughness, power and coherence) <u>are used to construct the lookup tables</u>."

And in Section 2.4.2.

"Specifically, <u>we evaluate lookup tables constructed using</u> all potential combinations of slope, roughness and power."

13. *L70: It would be helpful to clarify what is meant by "unseen," and to specify the reference dataset and how elevation differences are computed (e.g., crossovers, gridding, ect).*

We have clarified what is meant by 'unseen' at its first mention in Section 2.3.1, which now reads:

"In this regard, we note that any year with suitable unseen reference data (i.e. not used in the construction of the lookup table) could be evaluated using the framework."

We specify the reference dataset and how elevation differences are computed in Sections 2.3.1 and 2.2, respectively, but would prefer not to include this detail in Section 2.1 when introducing the framework so as not to limit its application to a specific reference dataset.

14. *L74: It is not entirely clear how the uncertainty values are derived (e.g., from the 2020 dataset).*

We now refer the reader to Section 2.2, where details of the uncertainty generation method are provided.

"We generate several different uncertainty lookup tables for each of the POCA and swath-processed datasets, corresponding to different co-variate combinations and metrics, using the methodology detailed in Section 2.2."

We also specify that the uncertainty lookup tables are produced using data from 2020.

"Uncertainty lookup tables are generated separately for Antarctica and Greenland using data spanning 2020."

15. *L77: Please define the lower and upper quartiles explicitly.*

The sentence now reads:

"As these distributions are typically non-Gaussian, distribution similarity is quantified through the comparison of their median and 25th/75th percentiles."

16. ***Data***
    Altimetry data section: Please specify which CryoSat-2 and Sentinel-3 elevations/ backscatter are used. Are the Sentinel-3 elevations from L2?

The elevation and backscatter variables used are now specified in the text.

"To generate the POCA lookup tables, we use ESA CryoSat-2 L2 in-depth (L2I) Baseline E elevation data (specifically the height_1_20_ku parameter, Figure B1a; ESA, 2023b)."

"We use ESA Sentinel-3 Level-2 Hydro-Cryo Thematic products BC-005 data (specifically the elevation_ice_sheet_20_ku parameter, Figure B1c; ESA, 2025a) to generate and evaluate ice sheet elevation uncertainties."

"The remaining auxiliary data, coherence (coherence_20_ku, Figure B1d) and backscattered power (sig0_1_20_ku, Figure B1e), are provided in the Level-2I data."

"The slope and roughness data are derived from REMA or ArcticDEM (Section 2.3.2), and backscattered power (sig0_ice_sheet_20_ku, Figure B1f) is provided in the Level-2 data"

17. *L128: "Binning" is introduced here; consider introducing this concept earlier in section*

This concept is now introduced in Section 2.2, which reads:

"This information is used to build a look-up table, for each ice sheet, containing estimates of elevation difference binned by each combination of the selected co-variates (Step 4)."

18. *L131: Please provide a reference for "these variables have been shown…".*

To avoid repetition, this sentence has been removed as this information, and its associated references, are already provided in Section 2.2.

"These approaches are based upon the understanding that radar altimeter measurement accuracy is sensitive to each of these variables, with surface slope providing a reasonable first order approximation (e.g. Schröder et al., 2019, McMillan et al., 2019)."

19. *It may be helpful to comment more explicitly on the differences in spatial coverage among ICESat-2, CryoSat-2, and Sentinel-3, as these differences complicate direct comparisons.*

We agree with the reviewer that there are important differences in spatial coverage between CryoSat-2 and Sentinel-3; namely (1) the difference in the latitudinal limit caused by the differing orbital inclinations of the two satellites, and (2) the denser spatial sampling achieved by CryoSat-2 owing to its longer repeat cycle.

Regarding the first point, we note that the latitudinal limit of the ICESat-2 reference dataset matches or exceeds that of both CryoSat-2 and Sentinel-3, which ensures that the full latitudinal range of both datasets is being evaluated. We have edited the following text in the manuscript to acknowledge this point.

"ICESat-2 was chosen owing to its comprehensive ice-sheet-wide coverage with a 91-day repeat cycle at a high along-track spatial sampling of ~20 m, and its temporally <u>and spatially</u> overlapping availability with CryoSat-2 data."

"The ICESat-2 ATL-06 v006 land-ice surface height measurements (Smith et al., 2023) are again used as the reference dataset <u>given their spatial and temporal overlap with Sentinel-3 observations</u>."

Second, owing to the differing spatial coverage of CryoSat-2 and Sentinel-3, we compare the overall adequacy with which uncertainties are computed for a given mission only. That is, the ability of CryoSat-2 uncertainties to characterise CryoSat-2-ICESat-2 elevation differences over CryoSat-2 regions is compared to the ability of Sentinel-3 uncertainties to characterise Sentinel-3-ICESat-2 elevation differences over Sentinel-3 regions. The difference in performance results from the combined effect of the differing spatial sampling and coverage provided by each of the satellites, in addition to the differing instruments and acquisition modes of CryoSat-2 and Sentinel-3. To clarify this point, the following text has been added to the manuscript.

"Overall, the application of a standardised framework across different radar altimeters (CryoSat-2 and Sentinel-3) provides a systematic process for making quantitative assessments of the relative accuracies of <u>mission-specific uncertainties that may arise from differences in instrumentation, acquisition mode, spatial sampling and coverage between satellites</u>."

20. ***Methods***
    Table 1: Additional clarification of the metrics (e.g., elevation differences, confidence intervals) would help the reader.

Following Comment 4 (Reviewer 1), we have defined the metrics and covariates used earlier in the text. In Table 1, we have also now referred the reader to the definition of elevation differences (Equation 3), the median absolute elevation difference (Equation 4) and the upper bound of confidence intervals of the standard deviation (Equation 5).

21. *L168–172: It might be worth discussing whether temporal sampling affects the results, and whether footprint differences (e.g., ICESat-2 vs. CryoSat-2) play a role.*

Agreed - owing to the differing temporal and spatial sampling of CryoSat-2 and ICESat-2 measurements, a corresponding ICESat-2 reference measurement does not exist for each CryoSat-2 elevation retrieval. For this reason, spatial (20 m) and temporal (1 month) co-location parameters are required when calculating the elevation difference to a reference dataset. We also agree that it is worth discussing the different footprint sizes of CryoSat-2 and ICESat-2. We have therefore added the following text to the manuscript to address both points.

"Owing to the different spatial and temporal sampling provided by the CryoSat-2 and ICESat-2 orbits, reference measurements are rarely acquired at the exact time or location as the radar altimetry retrievals. Similarly, differences in the spatial resolution of the CryoSat-2 and ICESat-2 footprints mean that co-located measurements do not necessarily sample the same area on the ground. […] Given that CryoSat-2 measurements are attributed to a point location, however, considering an ICESat-2 measurement within 20 m is a reasonable assumption."

22. *L179–181: This statement would benefit from further clarification.*

Upon reflection, this statement is not required and has been removed to aid readability.

23. *L203: Please elaborate on the motivation and implications of the different approach used here.*

We have addressed this as requested. Owing to the greater volume of data output from swath processing, there is less need for interpolation as bins are rarely unpopulated. The text has been edited to read:

"Such interpolation is not required for the Cryo-TEMPO EOLIS dataset, owing to the greater volume of data yielded by the swatch processing approach."

24. *L225, 249, L285, .., : Ensure consistent use of "median," particularly if used as shorthand for median absolute differences. I found the use of "median" in the sense of "median absolute difference" potentially confusing. A clearer distinction in terminology could help avoid ambiguity for the reader.*

As requested, we have ensured the introduction of the metric as 'median absolute difference' (hereafter median) and refer to it consistently as the 'median' for the remainder of the manuscript.

25. ***Results***
    L291–292: It is not entirely clear how the interpretation is derived from the figure; additional explanation would help.

We have edited the text for clarity, which now reads:

"As the quotient distribution varies more between metrics than between co-variate combinations, Figure 2 suggests that, for both datasets, the choice of metric has a greater influence on the distribution of uncertainties than the combination of co-variates utilised."

26. *L337: The term "true accuracy" is not well defined. Is there such a term? Consider replacing it with something measurable, e.g., agreement with reference data.*

Thank you. We have edited the sentence, which now reads:

"Thus, the framework illustrates that granting the uncertainty model more degrees of freedom in terms of the co-variates used to parameterise uncertainty enables the generation of a more robust uncertainty."

27. *L408: It would be useful to specify whether results refer to the entire ice sheet or specific regions (e.g., margins).*

We have specified in Section 2.4.1 the spatial coverage of Sentinel-3 data used.

"Uncertainty lookup tables are generated separately for Antarctica and Greenland using data spanning 2020. The choice to utilise all available data spanning the continent over an entire year was designed to ensure that the resultant uncertainties were not unduly affected by the influence of regional and seasonal biases, respectively."

The spatial coverage is also visualised in Figure B1.

28. *L425–427: A short explanation of why CryoSat-2 appears to perform better than Sentinel-3 would strengthen the discussion.*

Further detail on why CryoSat-2 is better suited than Sentinel-3 for elevation retrieval over ice sheets has been added to the text, which now reads:

"There, in most cases, the uncertainties do not adequately capture the tails of the elevation difference distribution, which are much larger than for CryoSat-2 (Figure 2) whose design is more optimised for the measurement of ice surfaces owing to its novel synthetic aperture radar interferometry mode, which provides more precise geolocation in areas of complex terrain."

29. ***Discussion***
    L511: Could the metrics be further refined for Sentinel-3?

Yes – we very much hope that the framework and initial results presented serve as the foundations and motivation for further refinements in the future. Specifically, as mentioned in the text, the framework succeeds in demonstrating that our initial Sentinel-3 uncertainties are less accurate than those derived for CryoSat-2, which can motivate and guide future work to further refine the methodology to generate Sentinel-3 uncertainties. The text has been edited to more explicitly acknowledge this point and highlight the potential for these results to guide such work in the future, and now reads:

"Ultimately, the framework suggests that these initial Sentinel-3 uncertainties are less accurate than those derived for CryoSat-2; a result that can motivate further improvements of Sentinel-3 uncertainties in the future, including using the framework to test additional metrics and/or co-variates."

30. *L514, L525: Grid size, spatial sampling, and repeat cycle may influence the results - a brief discussion would be helpful. A sensitivity analysis (e.g., grid size or parameter choices) could strengthen the conclusions.*

Neither the altimetry data nor the reference data are gridded, so we do not believe that grid resolution is an important parameter to consider. Grid extent is already discussed in Section 4.3 as a key factor explaining the differences in uncertainty performance between Antarctica and Greenland, as shown below:

"Due to its size and latitudinal extent, a smaller number of elevation retrievals are acquired over Greenland, compared to Antarctica, resulting in a lower number of elevation difference values being assigned to each bin during the uncertainty lookup table construction."

"Owing to both the smaller total number of data points and the lower quality of these data, the variability of elevation difference data within each bin is greater in Greenland, with a greater proportion of outlier values (Figure B8). The median metric is relatively insensitive to such pervasive outliers, implying that its performance may not be as robust in regions, such as Greenland, where the distribution of elevation differences is less Gaussian in nature."

A sensitivity analysis is a good suggestion for a future use of the framework to determine whether uncertainty performance varies based on input data volume. Initial analysis to this effect is provided in Figure B8, but we believe a full sensitivity analysis is out of scope for this manuscript, which aims to provide an initial demonstration of the potential applications of this new framework.

31. *The Sentinel-3 case study over Greenland is interesting; however, a slightly more comprehensive evaluation or discussion of the results would strengthen this section.*

As requested, we have added additional discussion of Sentinel-3 uncertainty performance to the Discussion in Section 4.3:

"Application of the uncertainty evaluation framework to both CryoSat-2 and Sentinel-3 data demonstrated that the performance of uncertainties differs between Antarctica and Greenland. In Antarctica, the use of the median metric produces a quotient distribution with a standard deviation closest to 1 <u>for both satellites (Figures 2 and 7)</u> indicating that the expected proportion of elevation differences fall within the stated uncertainty. Similarly, the median results in a more zero-centred and narrower distribution of residuals <u>in both cases (Figures 3 and 8)</u>, representing a closer agreement between uncertainty values and the unseen elevation differences. In contrast, the STD UB metric typically results in an overly conservative set of uncertainties. For Greenland, however, the advantages of using the median are generally less clear. For CryoSat-2 POCA data, for example, the STD UB produces a less biased residual distribution (median=0.01 m, IQR=1.43 m) than the median when all available co-variates are used (Figure B4). <u>Similarly, for Sentinel-3 data, the performance of the median and the STD UB is similar according to the first two elements of the framework (Figures 7 and 8), particularly when all three co-variates are used. According to the final element, however, STD UB results in an unrealistically concentrated distribution of uncertainty values (Figure 10).</u>"

32. *It would be interesting to examine whether the choice of covariates varies spatially.*

While we agree that it would be interesting to determine whether the optimal co-variate combination varies spatially, this would involve calculating regional lookup tables, which would complicate any subsequent operationalisation and roll out, and may not be as effective as ice-sheet-wide lookup tables. Specifically, we might expect a similar degradation in performance from continental-scale to regional-scale lookup tables as observed between Antarctica and Greenland, owing to the smaller volume of data used in their construction. Although beyond the scope of this study, we have added this suggestion to the text as a potential future use of the framework. The text now reads:

"Future work could consider evaluating the generation of uncertainties at both global and regional scales, to better understand any trade-offs associated with spatial scale, alongside the specificity or transferability of uncertainty generation methods between regions."

33. *Have the results been tested for different years? This could help assess robustness.*

As discussed in our response to Comment 6 (Reviewer 1), we have evaluated the uncertainties using data from 2022. Whilst this is sufficient to demonstrate the application of the framework, we have suggested in the text that other years could also be evaluated using the framework.

"The year 2022 was chosen to ensure that the evaluation was performed using a temporally distinct dataset from that used to construct the look up table, and during which the reference data (ICESat-2) was fully operational and hence able to provide the required unseen reference data. In this regard, we note that any year with suitable unseen reference data (i.e. not used in the construction of the lookup table) could be evaluated using the framework."

34. *Differences in altimeter modes (e.g., SARIn, LRM) are not discussed; it would be interesting to comment on whether these could influence the observed differences between missions. Why not compare CryoSat-2 LRM with Sentinel-3 as well?*

Our CryoSat-2 analysis in the manuscript was restricted to SARIn regions to enable comparisons between POCA and swath-processed datasets, as is stated in the text:

"Analysis was conducted over regions where CryoSat-2 operates in SAR Interferometric (SARIn) mode only, as swath-processed data is unavailable elsewhere and both datasets were required for inter-comparison purposes."

We agree, however, that the framework could also be used to evaluate uncertainties generated for LRM altimetry measurements. We have added the following sentence to acknowledge the possibility of evaluating LRM data in the future, and its comparison to Sentinel-3 data.

"Future work comparing these uncertainties to CryoSat-2 LRM uncertainties using an equivalent method may also shed light on the magnitude of uncertainties that can be expected from non-SAR altimetry measurements."

35. *In the abstract, you mention the absence of uncertainties in the Level-2 product. It would be helpful to briefly comment on the framework's readiness for operational implementation (e.g., at agency Level-2 processing).*

Thank you for this suggestion. The text has now been updated to specify the ways in which this framework has already be used in operational scenarios.

"By using this framework to compare a range of uncertainty generation approaches for POCA and swath-processed CryoSat-2 ice sheet elevation data, we have demonstrated how uncertainty calculation can be assessed and standardised between products. Indeed, the findings have contributed to the evolution of the Cryo-TEMPO POCA product (Huang et al., 2026), which will replace the two-dimensional lookup table in Baseline D with a four-dimensional lookup table in Baseline E. It is also hoped that this Cryo-TEMPO implementation will ultimately be taken up within a future baseline of the standard, non-thematic ESA Level-2 product.

"More broadly, this new framework can also play a key role in optimising uncertainties associated with upcoming polar radar altimetry missions, such as the Copernicus Polar Ice and Snow Topography Altimeter (CRISTAL) and the Sentinel-3 Next Generation Topography mission. Indeed, the framework has already been employed to select the uncertainty generation method used in the CRISTAL Land Ice Level-2 Ground Processing Prototype."

Technical Corrections

36. *L55–59: Consider adding a clearer introductory sentence describing the framework.*

We have made this change. Please see our response to Comment 3 (Reviewer 1) for further details.

37. *L96: Possibly add a reference to Section 2.*

To enhanced readability, this text has been removed in the updated manuscript.

38. *L261 (and elsewhere): Consider explicitly naming the reference dataset (e.g., ICESat-2) rather than using a generic term.*

Please see our response to Comment 13 (Reviewer 1).

39. *Ensure consistent terminology throughout (e.g., median vs. mean; definition of quartiles).*

We have made changes to ensure consistent terminology throughout. Please see our responses to Comments 24 and 15 (Reviewer 1) for details.

40. *Avoid repeated definitions of the same parameters; median and IQR are defined several places.*

Repeated definitions of median and IQR have been removed.

Reviewer #2

General comments

1. *Satellite altimetry is a primary source for observing the state, the temporal changes, and the mass balance of ice sheets. However, the quantitative characterization of uncertainties in data products from satellite altimetry has been challenging and incomplete, due to the complexity of the measurement systems and processing systems involved and due to the sparsity of validation data. Uncertainty measures provided with different datasets have been difficult to evaluate and to compare. At the same time, uncertainty characterization is crucial for evaluating the reliability of altimetry-based results all the way to ice-sheet mass balance and for the combination of altimetry results with results from complementary observations and modeling.*

   In this highly relevant context, the manuscript specifically addresses uncertainties of Level-2 ice sheet elevation products (ice surface heights assigned to positions). The manuscript proposes and realizes a framework to address two principal tasks:
   (A) 'uncertainty estimates' (also called 'uncertainty assessment', 'uncertainty generation' etc.), that is, the quantification of uncertainty by certain metrics;
   (B) 'evaluation of uncertainty estimates', that is, an evaluation of how realistic and robust the results from (A) are.

   The proposed framework is applied to two L2 products from CryoSat-2 SARIN Mode measurements, namely the CryoTEMPO POCA and with Cryo-TEMP EOLIS, as well as to data from the Sentinel-3 HydroCryo Thematic product BC-005. (For brevity, I refer to these products as radar altimetry (RA).) For the first two RA products, methods for uncertainty estimates have been developed previously, and these methods are evaluated by the present study. For the Sentinel-3 product, uncertainty estimates are generated for the first time and are subsequently evaluated. The framework is first elaborated for the Antarctic Ice Sheet and subsequently also applied to the Greenland Ice Sheet.

   The basic idea for both (A) and (B) is to compare the RA data with independent measurements from ICESat-2 laser altimetry (LA), where the comparison is restricted to pairs of RA measurements and LA measurements that are closely co-located in space and time.

   The proposed framework for (A) is as follows: define a metric for uncertainty. (Three alternative metrics are addressed). Realize that the uncertainty depends on covariates like surface slope, backscattered power etc. Use RA-LA differences to establish the relation between the covariates and the uncertainty metric in a lookup table approach. Use this lookup table to assign uncertainty to all RA measurements. The study finds that while surface slope appears to be the most important determinant of the uncertainty metrics, the use of more covariates improves the prediction of the uncertainty metrics.

   The proposed framework for (B) is as follows: use RA-LA differences that were not used in step (A) ("unseen differences"). Compare, in a statistical way, these RA-LA differences with the *uncertainty estimates assigned to the RA measurements. Specifically, statistics of the absolute value of elevation differences (|RA – LA|) are employed.*

The manuscript is a novel contribution to the developments and discussions on uncertainty assessments that are urgently needed. The methods and results are presented in detail in an overall well-structured way. Thereby, the manuscript provides a wealth of material for further discussion.

We thank the reviewer, Martin Horwath, for his detailed summary and were particularly pleased to read his acknowledgement of the potential novelty and importance of the work.

2. *However, I have concerns about details of the methodology. They don't appear to be grounded in clear concepts and justified by mathematics. These issues call some of the conclusions into question. I suggest that the methodology has to be revised [or] justified more clearly.*

Thank you for this comment, which is addressed in the **Major Change** above.

3. *In addition, I suggest some improvements in the clarity of the presentation of the overall framework and how it is placed in the wider context of uncertainty assessments for altimetry results.*

Thank you. We have addressed this comment below in response to Comment 5 (Reviewer 2).

Specific comments

4. *The work would benefit from a clearer and more thorough concept of uncertainty. My following comment is guided by the "Guide to the expression of uncertainty in measurement" (JCGM 2008). You may follow a different concept. So, consider my comment as a contribution to the discussion.*

   The error (defined as 'measured value minus true value') is assumed to be a random variable. By 'uncertainty characterization' we mean some quantification of the probability distribution of the error. The most common way is through the variance (and covariances in case of multi-variate data) where uncertainty is quantified in terms of standard uncertainty.

   Two of the three metrics considered in the manuscript are oriented towards the concept of standard uncertainty: the STD UB and the RMSE. The third metric, the median of the absolute value, is different.

   The manuscript's basic idea for the uncertainty evaluation is outlined in Section 2: "In the first stage of the framework, the overall distribution of all elevation differences [|RA - LA|] [...] is compared to the corresponding distribution of uncertainty values. Closely matching distributions indicate that the overall distribution of uncertainties is realistic [...], which is a desirable characteristic." The same idea is further elaborated in the second stage of uncertainty evaluation, where it is stated that the distribution of |RA – LA| minus uncertainty *metric' should ideally be narrowly centered around zero to indicate realism of the uncertainty metric.*

I am missing a mathematical justification for these criteria of uncertainty assessment.

To explain and illustrate my point, let's do two thought experiments (accompanied by numerical experiments) where I assume that the error has a Gaussian distribution. (Of course, in reality, the distribution may be non-Gaussian. However, any evaluation framework should work, at least, for the simple case of Gaussian distributions.)

For a first thought experiment, let's assume a set of measurements of equal quality. That is, the errors of all measurements follow a Gaussian distribution with an identical standard deviation at, say, 1.0. That is, all measurements have a standard uncertainty at 1.0. The median absolute deviation (MAD) of this Gaussian distribution is 0.67449 (e.g. https://en.wikipedia.org/wiki/Median_absolute_deviation)

I have done calculations with pseudo-random realizations of this situation. Figure RC2.1b (red histogram) illustrates the distribution of the standard uncertainties, which is concentrated at the value 1.0 for this first thought experiment. The grey histogram illustrates the distribution of the absolute values of the random variable itself (that is, the absolute deviations). Likewise, Figure RC2.1a compares the (one-valued) distribution of the MAD with the distribution of the absolute deviations.

The figures are shown in a style analogous to that of Figure 2 of the manuscript. The standard uncertainty is conceptually similar to the manuscript's RMSE metric, while the MAD is conceptually similar to the manuscript's Median metric. In this thought experiment, the two metrics are two alternative, and perfectly correct, quantifications of the uncertainty. That is, the distribution of the absolute errors (grey histograms) perfectly matches the distribution indicated by the uncertainty metrics. However, it obviously does not match the distribution of the uncertainty metrics (red and yellow histograms).

The criterion for uncertainty performance used by the manuscript is the match between the distribution of absolute errors and the distribution of uncertainty metrics. According to this criterion, the two metrics would perform quite badly in our thought experiment, with the standard uncertainty metric performing worse than the MAD metric, according to the comparison of median values. This interpretation is difficult to justify, and so is the underlying performance criterion.

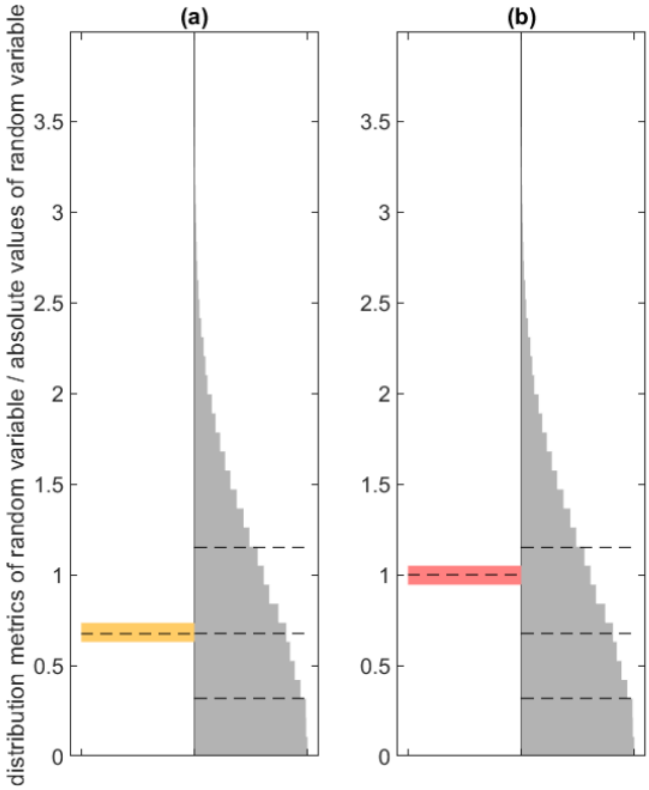

Figure RC2.1: Results from the first thought experiment, presented in a way similar to Figure 2 of the manuscript. Grey histograms: Distribution of absolute values of random variables according to a Gaussian distribution with a standard deviation at 1.0. Red and yellow histograms: Medium absolute deviation (a, yellow) and standard deviation (b, red) underlying the pseudo-random generation of the distribution. Dashed lines show the 25%, 50% (median), and 75% quantiles. Histograms are normalized to their maximum value.

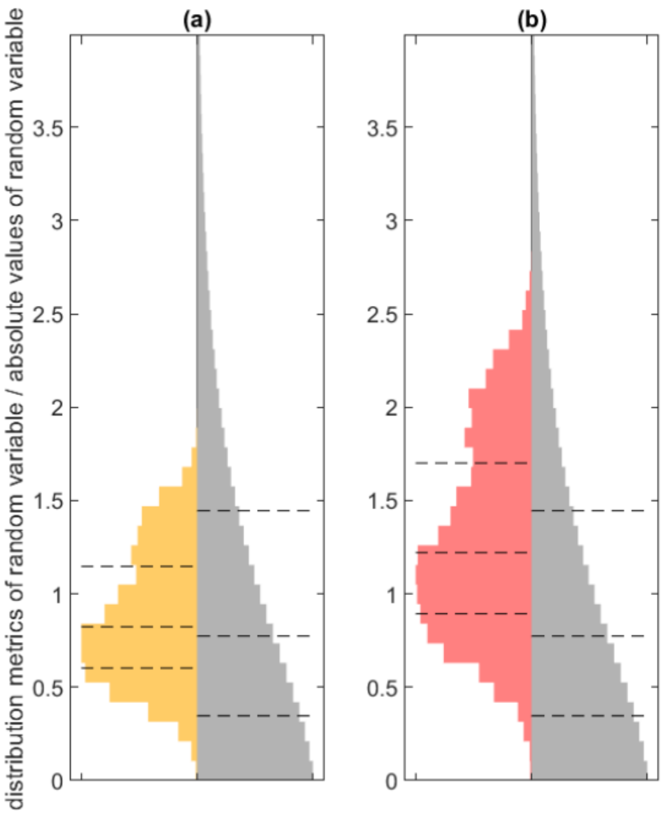

Figure RC2.2: Similar to previous figure, but now for the second thought experiment. A set of Gaussian random variables with different standard deviations, or likewise, different mean absolute deviations (MAD), is assumed. Red and yellow histograms show the distribution of the standard deviations and the MADs, respectively. Grey histograms show the distribution of absolute values of the set of random variables.

For a second thought experiment, which comes closer to the real situation, I assumed that different measurements have different uncertainties, hence different values of the uncertainty metrics. I designed an arbitrary distribution of standard uncertainties, shown in Figure RC2.2b (red). I generated a corresponding pseudo-random distribution of absolute errors (Figure RC2.2b, grey). Likewise, in Figure RC2.2a I compare the distribution of MADs with the distribution of the absolute errors.

In this second experiment, again, the distribution of the errors perfectly matches the distribution indicated by the uncertainty metrics. But it does not match the distribution of the uncertainty metrics. In particular, the median of the standard uncertainty is larger (by a factor of ~1.48) than the median of the absolute error, for good mathematical reason.

In the manuscript, the match between the distribution of the absolute error and the distribution of the uncertainty metrics is taking as the performance criterion. I conclude from the above that this approach has limited mathematical justification.

It is interesting to note that some qualitative features of how the distributions compare in the second experiment resemble features observable in Figure 2 of the manuscript.

The conceptual limitations propagates to the second stage of the uncertainty evaluation, where the manuscript states that "It is preferable for the distribution [of |RA – LA| – uncertainty value] to have a narrow peak centered close to zero", and, "A zero-centered distribution suggests the uncertainty estimates are largely unbiased".

In light of the above, this approach needs to be questioned. If we take the RMSE as our metric, then it is clear that the metric is larger than the median of |RA – LA| (cf. my Figure RC2.1b and RC2.2b), and this does not mean that the metric overestimates uncertainty.

As a related remark, evaluating different uncertainty metrics (as done in the manuscript) likely requires different methods of evaluation. However, the manuscript uses one and the same method for the three metrics. This chosen method (using statistics of |RA – LA |, with emphasis on the median rather than, for example, the RMS) is more suited to evaluate the Median metric than to evaluate the two other metrics. This is, at least partly, the reason why the Median metric performs best in the evaluation.

More reflection is needed on the overall mathematical concepts underlying the framework, and a literature review may help. Merchant et al. (2017, https://doi.org/10.5194/essd-9-511-2017) provides a review in the context of ESA Climate Change Initiative data products. In the course of a revision, the authors might consider whether the quotient between |RA – LA| and the uncertainty value are more informative about uncertainty performance than the magnitudes considered so far.

We thank the reviewer for their detailed and constructive comment, which we have addressed in **Major Change** above.

5. *The overarching motivation outlined in the abstract and in Section 1 is ice mass balance and applications like model assimilation. Meanwhile, the study is restricted to Level-2 data. That's perfectly fine. However, some discussion could be added how the presented work relates to the ultimate goal of providing uncertainties that are useful to evaluate altimetry-based mass balance (to mention a prominent example.) I guess, propagation of the L2 uncertainties through the analysis of temporal variations, through gridding and maybe through integration over an entire ice sheet, will require considerable additional work, including the analysis of spatial and temporal error covariances in addition to point-wise uncertainties. I would be curious whether the presented framework offers potential towards such extensions.*

Thank you for this comment – we agree completely with the points raised. Indeed, one of the principal motivations for this work is that we need robust and consistent uncertainties, computed across multiple missions, to integrate into a new Kalman Smoothing Framework that aims to compute spatially aggregated estimates of ice sheet elevation and volume change. Ultimately, these uncertainties will then be propagated through to mass balance estimates, to provide improved uncertainties associated with this important parameter. We hope to have a manuscript prepared on this topic later this year.

As we believe this framework has potential towards such extensions, and indeed this serves as one of our main motivators – to better constrain uncertainty in altimetry ice sheet mass balance estimates, we have added the following text to the manuscript to explicitly refer to the applications for downstream users, such as Kalman Filtering.

"A consistent and well-validated approach to uncertainty estimation is also an essential component in utilising newly developed methods, such as a Kalman Smoothing Framework, to combine these data with observations from other coincident missions, such as the existing Copernicus Sentinel-3 constellation. Robust and consistent uncertainties will be integrated into such frameworks to compute spatially aggregated estimates of ice sheet elevation and associated uncertainties, which will in turn enable more refined estimations of ice sheet mass balance and global sea-level rise."

Technical comments

6. *Here I note a few editorial comments, even though some of them will likely become obsolete during revision*
   As a reader I will appreciate additional insights (additional to Figure B7) into the number and distributions of RA-LA differences that were available for the analysis.

We have added maps, histograms and summary statistics of the elevation differences in 2020 used in the creation of the uncertainty lookup tables as Figure B2, which is now referred to in the text.

7. *When reading the abstract and the introduction I was not sure I understood what this manuscript is about, because the distinction between (A) uncertainty assessment and (B) uncertainty evaluation is not so clear. It took me two rounds of careful reading the manuscript to arrive at my own summary I have given above in my general comment. The authors might scrutinize the wording and the manuscript structure to make this distinction clearer.*

Thank you for this comment. In response, we have overhauled the structure of Section 2 (Uncertainty Evaluation Framework) and Section 3 (Data and methods) to produce an updated Section 2 (Data and Methodology), which clearly distinguishes between the evaluation and generation of uncertainties.

The updated structure comprises:

- Section 2.1 (Uncertainty evaluation framework) – describes the conceptual framework used to evaluate the robustness of generated uncertainties (point B above).
- Section 2.2 (Uncertainty estimation approach) – describes the approach taken to generate the uncertainties (point A above) that are later evaluated using the framework.
- Section 2.3 (Inter-comparison and harmonisation of existing methods) – describes the first application of the framework, which ascertains whether a standardised uncertainty solution for POCA and swath-processed CryoSat-2 products is suitable.
- Section 2.4 (Creation of new uncertainties) – describes the second application of the framework, which evaluates a set of uncertainties generated for the Sentinel-3 thematic product for the first time.

We have included a summary of the updated structure in the text.

"In this section, we first introduce our conceptual framework used to evaluate uncertainties (Section 2.1). Then, we present the approach used to generate the uncertainty values themselves (Section 2.2). The first application of the framework (Section 2.3) ascertains whether a standardised uncertainty solution for POCA and swath-processed CryoSat-2 products is suitable. The second application of the framework (Section 2.4) evaluates uncertainties generated for a Sentinel-3 product for the first time."

8. *For example, Section 2 explains (B) the uncertainty evaluation. For better clarity, it doesn't need to mention how uncertainties have been generated prior to their evaluation. So all the explanation related to the lookup table (which confused me) could be omitted here.*

Thank you for this suggestion. We have removed details concerning lookup table construction from the section that introduces the uncertainty evaluation framework (Section 2.1), as suggested.

9. *The abstract could be a bit more specific about the results. Currently, it mentions no results of the uncertainty evaluation.*

We have added mention of the results to the abstract. Please see our response to Comment 8 (Reviewer 1) for further details.

10. *Line 26-28: Since only a small selection of the many milestone references can be given to support the statement on altimetry for monitoring elevation changes, I wonder why the Suryawanshi et al. 2025 reference is included here, which is not on monitoring elevation changes.*

We have removed this reference.

11. *line 44: Reference "Hebeler et al. 2008" is missing in the list of references.*

Thank you. We have added this reference to the reference list.

12. *Line 66. There is some redundancy here, and a more straight formulation is possible. For example, you might replace "Here, we present our framework for uncertainty assessment that evaluates..." by "Our framework evaluates".*

Thank you for this suggestion. We have reworded the sentence so that it now reads:

"Our framework aims to provide a standardised mechanism for quantifying the robustness of any given uncertainty estimation method."

13. *In Section 2, you could give the reader a better grasp by stating that the "reference dataset" consists in ICESat-2 measurements.*

Please see our response to Comment 13 (Reviewer 1).

14. *Since you introduce mathematical symbols anyway a bit later in Section 2, it could be useful to introduce them earlier, and use them in conjunction with the verbal explanation. In short: E_alti,i is the measurement, sigma_i is its assigned uncertainty, E_ref,i is the co-located validation measurement (which is from ICESat-2 in our specific case), |E_alti,i – E_ref,i| is the "unseen difference". In the first stage we compare the distribution of |E_alti,i – E_ref,i| and sigma_i. In the second stage we analyze resid = |E_alti,i – E_ref,i| – sigma_i.*

Thank you for this suggestion. We have introduced the mathematical symbols at the beginning of Section 2.1.

"Here, we define robustness to be the extent to which a set of stated measurement uncertainties (σ) adequately characterises the true deviations between a set of altimeter (E_{alt}) and reference (E_{ref}) measurements. Hereafter, the quantity $E_{alt} - E_{ref}$ is referred to as the elevation differences."

15. *Line 100: State more explicitly that you do not use the uncertainties that are provided with CryoTEMPO POCA and with Cryo-TEMP EOLIS products. Instead, you perform your own generation of uncertainties for these products, reproducing their approach and variants of this approach.*

We have added the following statement to clarify that we do not use the existing Cryo-TEMPO uncertainties.

"Following the lookup table creation, the table is then used to assign uncertainties to a distinct set of elevation measurements acquired in 2022. These uncertainties are subsequently evaluated using the framework. This was in place of using the uncertainties provided within the Cryo-TEMPO POCA and Cryo-TEMPO EOLIS products, which would have prevented the evaluation of multiple different uncertainty generation methods."

16. *line 124 typo: ALT06 should read ATL06.*

Thank you. We have corrected this typo.

17. *line 131 maybe write "backscattered power" instead of just "power" to be even clearer.*

Please see our response to Comment 4 (Reviewer 1). We have also edited this text to read:

"The remaining auxiliary data, coherence (coherence_20_ku, Figure B1d) and backscattered power (sig0_1_20_ku, Figure B1e), are provided in the Level-2I data."

18. *line 146 It would ease the flow of reading if you didn't repeat the full names of the products, the acronyms of which were already introduced*

Thanks. We have edited the text to use the acronyms here.

"Two CryoSat-2-derived Level-2 products currently include uncertainty estimates alongside ice elevation measurements: the Cryo-TEMPO POCA thematic product (ESA, 2025b), and the Cryo-TEMPO EOLIS thematic product (ESA, 2019)."

19. *By the way, the EOLIS dataset is sometimes referred to as EOLIS and sometimes referred to as "Swath". You might consider being consistent.*

We have clarified that EOLIS is a swath-processed dataset (as demonstrated in the text below), and thereafter consistently referred to swath-processed data only.

"Here, the framework is used to compare the approaches to uncertainty generation adopted in the Cryo-TEMPO (POCA) and Cryo-TEMPO EOLIS (swath-processed) products, respectively."

20. *Table 1: last line: The entry appears to apply for both POCA and EOLIS. At the same time, the entry mentions "POCA elevation measurements". Is this correct?*

Thank you. This has been corrected to remove mention of POCA in the last line of Table 1.

21. *line 174: "Once a set of i elevation difference pairs are identified": Here, i seems to denote a number of elevation pairs, while in the equation itself, i seems to denote an index.*

Thank you. This has been corrected and now reads:

"For each pair of CryoSat-2 – ICESat-2 measurements identified, a corresponding elevation differences is computed:

$$\Delta E_i = E_{alt,i} - E_{ref,i} \qquad (3)$$

where $E_{alt,i}$ and $E_{ref,i}$ are the ith pair of CryoSat-2 and ICESat-2 elevations, respectively."

22. *line 190 maybe replace "slope bin j" by "bin j" as the bins refer to various covariates.*

Thank you. We have removed the word slope.

23. *Reference CryoTEMPO-EOLIS (2024): The given weblink does not work*

Thank you. The link has been corrected.

24. *line 198 "where σ_i is standard deviation of the sample" should read "where s is the standard deviation of the sample", I guess.*

Thank you. We have corrected this typo.

25. *Line 225f. I find the terminology "median absolute difference" versus "root mean square error" a bit inconsistent, as the two metrics are based on the same "differences".*

Thank you for this comment. We have replaced reference to 'root mean square error' with 'root mean square difference' throughout the manuscript.

26. *line 225: maybe refer to Eq. 3 and 4 for the "median" and the "STD UB".*

As suggested, we now refer to the equations in the text.

"The metrics selected for assessment are the median absolute difference (hereafter median) (Equation 4), the upper bound of the confidence interval of the standard deviation (STD UB) (Equation 5) and the root mean square difference (RMSD)."

27. *Section 3.2. The title could be more specific, e.g. "Uncertainty estimation". Not all "methods" are addressed in this section. An essential part of the methodology, namely the uncertainty evaluation framework, is already covered by Section 2*

We have restructured Sections 2 and 3. Please see our response to Comment 7 (Reviewer 2) for details.

28. *Line 254ff: I found the two introductory sentences hard to grasp, while reading the titles of subsection 4.1.1 and 4.1.2 and their first sentences made it clear quite immediately what this section was about.*

These introductory sentences have been removed to aid readability.

29. *Figures 2 and B4. The grey histograms of "absolute elevation difference" show low frequencies for the bin containing zero, as compared to the next bins. This doesn't look plausible. Why*

should the absolute elevation difference be less likely between 0 and 1 than between 1 and 2? This effect might be an artifact of the bin discretization, which could be repaired.

Thank you. The visualisation of the grey histograms in these figures have been altered to minimise this artefact.

References

JCGM (2008): Evaluation of measurement data – Guide to the expression of uncertainty in measurement, Int. Organ. Stand. Geneva, available at https://www.bipm.org/documents/20126/2071204/JCGM_100_2008_E.pdf/

References

Merchant, C.J., Paul, F., Popp, T., Ablain, M., Bontemps, S., Defourny, P., Hollmann, R., Lavergne, T., Laeng, A., De Leeuw, G. and Mittaz, J. (2017) Uncertainty information in climate data records from Earth observation. Earth System Science Data, 9(2), pp.511-527. https://doi.org/10.5194/essd-9-511-2017