Review of Boxall et al. „A framework for evaluating ice sheet altimetry uncertainty estimates“ by Martin Horwath

General comments

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 Hydro-Cryo 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.

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 of justified more clearly.

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.

Specific comments

(1)

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.

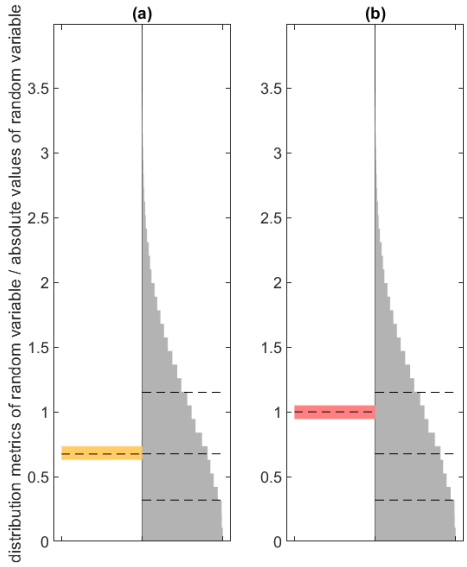

Figure RC2.1: Results from the first thought experiment, presented in a way similar to Figure 2 of the manuscript. Grey histograms: Distribution of absolute values of random variables according to a Gaussian distribution with a standard deviation at 1.0. Red and yellow histograms: Medium absolute deviation (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.

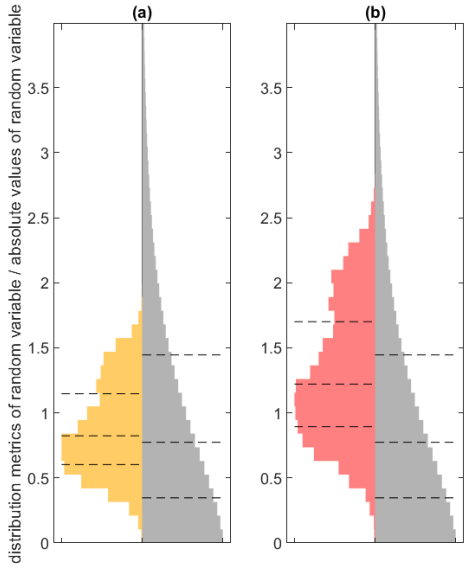

Figure RC2.2: Similar to previous figure, but now for the second thought experiment. A set of Gaussian random variables with different standard deviations, or likewise, different mean absolute deviations (MAD), is assumed. Red and yellow histograms show the distribution of the standard deviations and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

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.

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.

(2)

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.

Technical comments

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.

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.
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.

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

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.

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

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".

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

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, σ_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 σ_i. In the second stage we analyze $resid = |E_{alti,i} - E_{ref,i}| - \sigma_i$.

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.

line 124 typo: ALT06 should read ATL06.

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

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

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

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?

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.

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

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

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

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".

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

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

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.

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.

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/