Response to referee

ilo and hsk

January 2025

We thank the reviewer, Alek Petty, for the time, effort, and expertise spent reviewing our paper. Please find below reviewer comments provided in black, and our comments in red.

1 Answer to review by Alek Petty - General review

The paper by Olsen et al., introduces a compiled dataset of sea ice thickness related 'reference' measurements (freeboard, ice draft, snow depth, sea ice thickness) from various sources towards the goal of validating satellite-derived (radar) products across both poles through an ESA Climate Change Initiative project. They aim to align the various data with monthly gridded (25/50 km) satellite grid-scales to more easily enable evaluations. The authors make the claim in the abstract (and similar statements in the main manuscript) that this is "the first published comprehensive collection of sea ice reference observations including freeboard, thickness, draft and snow depth from sea ice-covered regions in the Northern Hemisphere (NH) and the Southern Hemisphere (SH)".

Overall, I think this was a decent effort to compile various sea ice datasets of interest, but I was ultimately disappointed with how basic the methodology was for processing the different datasets and accounting for the different uncertainties and significant differences in spatial scales (representation errors) that I remain unsure how useful this 'reference' catalog will really be. It also didn't include a lot of the more recently available data I was expecting to see. Our community hasn't produced an agreed upon 'reference' data collection as it's very hard to do this and be consistent with the uncertainties and include a full accounting of things like representation/sampling error, and it often depends on the exact goal of the validation effort.

If your primary goal is to bring in datasets that measure sea ice at vastly different spatial/temporal scales to convert these into 'reference' measurements to validate (gridded) satellite products, then you really need to consider how best to do that. I know a lot of studies just bin data into a grid-cell (myself included), but if this paper is focused on creating a reliable/useable reference processed dataset, then I think you need to acknowledge when this works and when it doesn't and ideally explore better ways of doing that through more sophisticated statistical means.

The reference dataset produced in this study was designed for quality control of the ESA Climate Change Initiative (CCI) sea ice thickness (SIT) climate data record (CDR), prioritizing simplicity due to the many unknowns in the field. Rather than assuming that more sophisticated statistical methods would resolve these uncertainties, we believe that a simple approach is desirable given the complexity of the uncertainty budget, including sampling bias and thickness conversions. The discussion of how to treat the uncertainty budget is peaking, and it is a highly complex topic without an agreed-upon consensus. As the reviewer himself mentions, the community has not yet established a consensus on a reference dataset or standardized guidelines for the processing of reference observations, further reinforcing the need for a straightforward methodology at this stage. We are in a new era where the community is focusing more on defining the framework of Fiducial Reference Measurements (FRMs). The authors are involved in several ESA projects, that are in the early stages of developing a more thorough and sophisticated approach to define and handle FRMs in terms of best practices, including uncertainty diagrams, time-space averaging, representativeness, and error propagation. This is ongoing work and we are yet to know the outcome of these projects. Given the uncertainties, such as the representation error, which you highlight, we believe that it is advisable to use a simple approach (which the reviewer also notes that he utilises) until more ideal methods are possibly discovered in these projects. However, there is yet no proof that "more sophisticated statistical means" will address the underlying issues of the complex uncertainty budget of sampling bias and thickness conversions.

To our knowledge, there is currently only one published paper that discusses initial thoughts on how to treat FRMs for satellite altimetry over sea ice i.e., Da Silva et al. 2023, which we reference in the discussion section. Additionally, some dedicated data products, e.g., AWI IceBird, are now beginning to incorporate and provide uncertainty estimates in their newer products (available in data from 2017 and 2019 campaigns used in Jutila et al. 2022), even though they, currently, still rely on the constant single values for AEM, which are those we refer to in Table 4 in our paper.

Thus, we believe that waiting to publish until all uncertainties have been resolved, all error propagations and averaging protocols have been identified, and best practices have been defined or resolved, would not be in the community's best interest. Our method represents a first approach - but not the only approach - currently relatively widely used (as suggested by the reviewer himself) by the broader community. To allude to this aspect, we propose to change the title to "A first approach to dual-hemisphere reference measurements from multiple data sources for evaluation and product inter-comparison of satellite altimetry over sea ice".

Furthermore, by making this dataset publicly available, we contribute not only with the data processed to a format widely used by the community, but also provide a processing pipeline that can use the originally published reference observations in their native format and performs all the pre-processing steps as described in Section 3 (this step needs to be done by all users, and thus every user shall track down the documentation and apply the necessary steps), add uncertainty estimates, flags, and finally time-space average to a level which is comparable to satellite scales. By providing the code for the processing pipeline (given in the ”Code availability section” available on GitHub (Ida Olsen and Henriette Skourup 2025)), the user can easily accommodate and integrate alternative time-space averaging or apply new methods for uncertainty estimates. To our knowledge, this is the first time such a dataset has been published with added uncertainty estimates and a preliminary method for handling flags.

It is worth mentioning that the Sea Ice-thickness product iNter-comparison eXerciSe (SIN'XS) project (`https://sinxs-tools.noveltis.fr/`), which is much larger in scope than ours, also incorporates reference measurements and data comparisons using monthly gridded data – showcasing, that this methodology is currently the go-to in the community, and while newer and optimal solutions could be sought, we are not aware of many studies that have conducted such sensitivity studies and proposed new approaches, definitively. We also do not claim our approach to be conclusive; we recognize its limitations, as you have pointed out. To address the concerns raised by the reviewer, we propose to create a second version of the dataset employing a more restrictive approach to data inclusion. For example, we suggest re-evaluating the flags to be able to exclude data that we do not recommend to be used e.g., the ship observations, where the number of measurements is below a certain threshold, or where the standard deviation remains below a pre-defined threshold (whenever reasonable and applicable). We would further provide the statistics by including all the observations, and compare to statistics by excluding those with flags of different categories. In this way, we provide the statistics to support our recommendations for which classes based on the flags to include and based on this the user can easily use these flags to identify the most optimal data for their purpose while maintaining all the data in the RRDP.

As for exploring more ”sophisticated” methods, this is unfortunately beyond the scope of this study, although we are keen to know which particular methods you have in mind and would encourage you to provide references for us to include in the discussion of the paper. It would be a natural extension of this work to investigate different averaging approaches, such as e.g., the Lagrangian approach proposed in Section 7.6.

We suggest extending the already included reference observation time-series upon availability. Additionally, we propose including available reference observations of similar types to those already included, such as HEM from MOSAiC, Nansen, and N-ICE. However, new types of reference observations such as drone measurements, will not be included, as we aim to maintain long time-series of consistent measurement types for quality control of the satellite CDR. We would greatly appreciate if you could share any other datasets you have in mind.

In a lot of your results example cases, you compare one of the ‘reference’ datasets with a satellite product, observe differences between the two, then say well they are maybe different because the reference dataset has issues (e.g. related to spatial scales and how they were aggregated) . . . so why produce this reference dataset and use it in the first place? What's the value of a bad reference dataset that we don't really trust?

As the reviewer is surely aware of, we have limited available reference data in the polar regions to support validation of altimetry observations over sea ice (especially for earlier periods, and for specific regions e.g. the Laptev Sea). Therefore, the alternative to many of these cases will be to have no reference data. Furthermore, this depends a lot on the requirements of the users. If one is for example looking to verify thin ice, then the ship data might be appropriate, as the bias towards choosing a route with thinner ice is expected to be smaller if the ice thickness in the region is generally thinner. Understanding and evaluating the limitations of different parts of the reference data is a central element in this study.

Similarly, you treat airborne data as a 'reference' dataset, but I think that is very dangerous. NASA's Operation IceBridge is great for coverage and the multi-sensor nature of the mission, but it still has a lot of issues that are frustratingly yet to be resolved, e.g. the big uncertainties in snow depth from different algorithms applied to the snow radar (King et al., 2015, Kwok et al., 2017) and significant biases between the quick-look and final snow depths (Petty et al., 2023, Fig. S3) which needs to be acknowledged. I was quite surprised this wasn't mentioned at all really.

Thank you for pointing this out. We intentionally use the terms *reference measurements* or *reference observations* instead of *validation data* to acknowledge that reference observations are not necessarily the absolute truth and each comes with its limitations, including the airborne data. We acknowledge that the caveats and limitations of the OIB data are not sufficiently addressed. Additionally, reviewer #2 has notified us that helicopter EM measurements tend to preferentially sample thicker ice, avoiding thin ice and open water for safety reasons. We will adjust the text to highlight these limitations and include the suggested references.

I also think for this study to work, you should try to actually characterize the uncertainties and/or errors in a consistent way. Your effort to summarize how the uncertainties are described in the product is a decent one and I appreciated the effort you put into this. But take IceBridge for example, you neglect all the algorithm differences I point to above, so how useful really are those individual product uncertainties?

We appreciate your acknowledgement of the effort put into summarizing the uncertainties. Uncertainty quantification in individual validation/reference products is an important and pertinent topic being also one of the subjects of lasting discussions on how to proceed with defining this 'reference' data collection and what the reviewer himself wrote earlier. We therefore consider questioning the provided uncertainties to lie beyond the scope of the presented study. Uncertainty characterisation for every single reference measurement used in this study is a significant undertaking, which requires a study in itself. It is also the focus of several FRM projects (e.g., SIN'XS, St3TART/St3TART-FO) and a complete traceable uncertainty characterisation (through either error propagation or Monte Carlo simulations) from initial measurement to the provided product used in the RRDP requires a thorough evaluation of the different assumptions and methodological steps taken in the initial processing conducted, as well as consideration for correlation, covariance and distributions. For now, we will have to rely on the uncertainties provided in the products, while noting the limitations mentioned by the reviewer (such as representation errors). However, as already mentioned, several projects are currently looking more into this topic (e.g., FDR4ALT, SIN'XS, St3TART-FO, FRM4ALT). However, we will acknowledge the uncertainty introduced by using the different OIB algorithms as also stated in the previous comment.

You state that the reference data should be 'used with care' a few times, but to me this is the job of this study! Decide which data to remove as it is just not a trust-worthy reference dataset for satellite validation for whatever reason. Seems like a cop-out to just say use it with care.

We see your point, however, different reference data is useful for different purposes and removing some data, when the availability is so sparse, also does not seem like the right approach. As suggested above we propose creating a second version of the dataset where we aim to classify the observations with an additional flag to suggest the most useful ones. However, ultimately, "what data is the most useful" is still up for debate when it comes to which data is "best used" to validate satellite observations (e.g., we still need to smooth up to 25 km or more to get reasonable comparisons to radar altimetry, even with airborne data).

Finally, the datasets listed as future work (IceBird, MOSAIC, Nansen Legacy) would have been great to see in this study! Again I think this paper was neither exhaustive of all available data nor thorough in the methodology, so I encourage the authors to decide on a better strategy based on my comments above.

New data become available continuously. The ESA CCI SIT CDR, for which this reference dataset was intended for quality control, is currently only available until 2020. We already agreed to extend the datasets in the reference observation dataset to cover the period of 1993 to the present, including snow depths measurements from the Icebird campaign since 2017. We will consider including reference measurements similar to those already included in the reference dataset e.g., helicopter EM measurements from MOSAiC, Nansen Legacy, and N-ICE upon availability. However, we will not include new methods, such as drone measurements, in the current version, as we would like to keep relatively long consistent time series as these are used for quality control of CDR. Please, let us know if there are specific reference observations you had in mind.

2 Answer to review by Alek Petty - Specific comments

I thought it was strange how much the intro talked about radar issues. Why not make it more about the science of why we want to measure basin-scale sea ice thickness? Then if your focus is radar, make that clear from the start, laser creeping in sometimes was confusing. Probably also easier to reference the papers that discuss the various issues in more detail, keep your focus on the reference datasets.

We agree that the introduction has a strong focus on satellite altimetry methods, which are important for understanding how to handle reference measurements – particularly since the reference dataset was produced as part of the ESA CCI project with the specific aim of being compared with the ESA CCI SIT CDR for consistency. We will update the introduction to place greater emphasis on reference observations. However, we would like to retain the most relevant satellite altimeter considerations in the paper, as they support the discussion section and contribute to understanding the comparability of sea ice variables from reference observations and satellite altimetry. Please also see our responses to the more detailed comments below.

L39 – I think that's still very much TBD and depends on the approach/freeboard used etc!

We agree. Indeed, snow density and snow depth are not the only main contributors depending on instruments, e.g., surface roughness has recently been stated as a significant contributor for the radar-derived altimetry aspects by Landy et al. 2020. We will rephrase this statement for clarity.

L41 – this is mixing up actual errors and theoretical uncertainties propagation which I think is confusing. We agree. We will correct these statements.

L45 – this seems like a bit of a stretch for an introduction! Do we really know that with confidence? Is that true for all types of freeboard and ice regime? None of the statements of this line is stated with certain confidence (e.g., it states that snow load **may** be most important over thin ice, whereas sea ice density *may* be the largest for thicker ice. Furthermore, we also state that it depends on the snow and ice conditions). However, we will revise this section and adapt accordingly.

L47 – well this is really 'a lack of uncertainty quantification data' rather than uncertainties directly I think.

We are not sure what you are referring to here. Can you, please, elaborate?

L80 onwards – ok so your aim is to reconcile radar thickness measurements. I think it would thus help to start with what you interested in then provide the uncertainty discussion to back that up, as before it was confusing how little you talked about laser.

The CDR is SIT, so shouldn't thickness be the main validation target?

The reference dataset was created by CCI to aid and assess the radar altimeter-based SIT CDR. The estimated sea ice thicknesses provided in the CCI SIT CDR are derived from the satellite radar altimetry freeboard measurements. Therefore both freeboard and thickness reference data are necessary. From a metrological point of view, the most accurate comparisons are made by introducing the least uncertainties, e.g., by comparing freeboard reference measurements (whether radar or laser, depending on the sensor used) with freeboards from the satellite CDR. Thus, we aim to use the measurands in their most native form when comparing the reference observations to the satellite altimetry observations, albeit considering thicknesses or freeboard. This is also highlighted in the first paragraph of Section 6.

L135 – “this data package and the methodologies applied herein have the potential of becoming the reference for future comparisons of current and future SIT products.” This is a big claim and I don’t think you have demonstrated this potential considering all the caveats and issues, and the basic methodology (aggregation) discussed here and even in your results.

We acknowledge that the wording is too strong in this sentence. We propose to change the formulation to:
“This data package and the methodologies applied herein provide initial efforts in collecting, unifying and comparing SIT reference measurements from different reference data sources. The data package and this paper provide a starting point for future work in assessing the uncertainty and reliability of reference measurements.”

L407: How is accuracy qualitative? A little confused by that statement. I think it’s basically the same as error, no? So it requires a known truth? Whereas uncertainty can be more theoretical.

It is true, that both accuracy and error require a known truth. However, accuracy is qualitative—it describes how close a measurement is to the true value in broad terms (e.g., excellent, good, poor). In contrast, error is a quantitative measure that specifies the deviation between the true value and the measured value (e.g., 10 cm). That said, in practice, the two terms are often used interchangeably despite their distinct statistical meanings. The uncertainty is a parameter characterizing the spread of the quantity values attributed to a measurand. Uncertainty can in plain language be seen as doubt, whereas error can be seen as a mistake as stated by Harris et al. 2017.

L505 ok so maybe stick with the higher number of 10 cm then?

The number of 10 cm refers to an upper limit of the bias, as stated in Lee et al. 2015, and does not necessarily reflect the uncertainty associated with representation error, which, as we have previously stated, is yet to be resolved. Since we acknowledge that we have yet to quantify this uncertainty term, we will consider alternative values such as the 10 cm value proposed here. However, we believe a more appropriate approach would be to further emphasize the issue of representation error in the text and to implement a quality flag, as previously proposed, to indicate data affected by known representation error issues.

L598: “Collocation is performed by finding all satellite data points obtained within ± 15 days from the date of the reference data, and within the 25 km (50 km for SH) grid cell of the reference coordinates. The average (arithmetic mean) of these satellite points are subsequently allocated to the reference data.” Ok so what uncertainties do we think this introduces? I think you need to provide some educated guesses at the very least.

We agree, that we somehow need to address this pertinent issue. This also relates to representativeness and error propagation. As we currently are working on defining the framework of error propagation in other projects (see general comments), we will not be able to fully implement this in the current study. However, we will suggest to investigate the representativeness; for example, if there is only one available flight, the average over a month would not be as representative as e.g., for the moorings, where both the reference and satellite measurement are expected to have the same temporal averaging. However, here we still assume that the sea ice covering the mooring due to the sea ice drift over a month is equivalent to the sea ice covered by a satellite within a grid cell, even though the mooring is permanently fixed to the same location. We suggest making a sensitivity study where we change the time averaging to include e.g., ±7 days or even ±3.5 days of the satellite CDR around the time-stamp of the reference measurements to see the impact of these for the different types of reference measurements. This would allow us to provide the study with a statistical basis for temporal representativeness. However, by using this approach we do not take into account that the satellite data is not equally distributed over a month either, which would also introduce another uncertainty. The above-mentioned approach could go along with a more extensive discussion of the representation and the uncertainties introduced by using our method in sections 7.5 and 7.6

L660 – why bother comparing if you then say it's not right to compare them? Would you have stated the same if the stats were better? Much better to state from the off which data are appropriate to compare against and why, then show how to use those..!

We see your point. However, polar reference observations are very sparse and we do not have data that match the spatial and temporal scales of satellite altimetry. Therefore, we prefer to retain the reference observations currently included in the dataset. If we do not perform some form of statistical comparison, even with data that have limited use for satellite CDR quality control, we cannot determine which reference measurements are appropriate to use. We here propose to clarify how and where different reference data sources are most appropriately used. This will be supported by our proposed dataset update, which introduces flags to categorize reference observations based on our interpretation of the statistical results according to their suitability for satellite CDR quality control. For example, ship-based observations and IMBs would rank low, and could be filtered out by users through these flags. To support our discussions we believe it is important to retain all currently included data, even those with representativeness issues, since it is undeniable used in scientific work. We will make a critical assessment of the reference data, cite relevant literature using these data for validation, and align the manuscript with this new approach i.e., implementing flags to categorize data according to their usability.

IMB discussion – ok so there's two things – you're underestimating the actual uncertainties AND also not really dealing with the representation error.

We agree and acknowledge that we need to further consider how to address these two issues. Since IMBs measure localized thermodynamic growth, we believe the best solution, as previously proposed, is to implement flagging to highlight the representativeness issue of these measurements, which makes them unsuitable for comparisons with data on a 25/50 km scale. We are, however, very keen to know if you have any suggestions for us on how to implement this. Specifically, in terms of the uncertainty contribution due to representation error (which to our knowledge is not well described in papers on satellite altimetry sea ice thickness or in comparisons to reference measurements) and in terms of quantifying the actual uncertainty. Do you have a study in mind?

"Additionally, no specific uncertainty for SD versus SIT is provided, resulting in the acoustic rangefinder sounders' accuracy used as the uncertainty for both SD and SIT." Why? I think you should be attempting to figure out what that should be, even if you have to make some assumptions.

We agree to look more into this and will make a dedicated section or paragraph in the manuscript, which discusses this in more detail. The uncertainty of the sensor retrieval, which includes the initial snow depth, sensor tilt and undetected snow-ice formation, is only one thing. We lack studies investigating the typical uncertainty of measurements of initial snow depth, which according to Nicolaus and Katlein 2017 is the primary source of uncertainty in snow depth measurements itself. The much bigger issue is the representativeness of the target variable. A realistic uncertainty assumption likely needs to be in the order of at least 20 to 30 cm. We cannot think of a single altimetry thickness study that also takes into account the representation issue (whether it be inequal thickness distribution from localised buoy measurements compared to satellites, different coverage of grid cell by both satellite and buoys etc.), and we will make sure to discuss this in the manuscript. Are you aware of any such studies or do you have an alternative idea of how we can address this uncertainty?

References

Lee, J. E. et al. (2015). "Uncertainty Analysis for Evaluating the Accuracy of Snow Depth Measurements". In: *Hydrology and Earth System Sciences Discussions* 12.4, pp. 4157–4190. DOI: 10.5194/hessd-12-4157-2015.

Harris, I. et al. (2017). "Updated high-resolution grids of monthly climatic observations – the CRU TS3.10 Dataset". In: *Earth System Science Data* 9.1, pp. 511–527. DOI: 10.5194/essd-9-511-2017. URL: https://essd.copernicus.org/articles/9/511/2017/.

Nicolaus, M. and C. Katlein (2017). "Observations of the Snow Depth on Arctic Sea Ice". In: *Journal of Geophysical Research: Oceans* 122.9, pp. 7167–7183. DOI: 10.1002/2017JC012838.

Landy, Jack C. et al. (2020). "Sea Ice Roughness Overlooked as a Key Source of Uncertainty in CryoSat-2 Ice Freeboard Retrievals". In: *Journal of Geophysical Research: Oceans* 125.5. e2019JC015820 2019JC015820, e2019JC015820. DOI: https://doi.org/10.1029/2019JC015820. eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2019JC015820. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019JC015820.

Jutila, Arttu et al. (2022). "High-Resolution Snow Depth on Arctic Sea Ice From Low-Altitude Airborne Microwave Radar Data". In: *IEEE Transactions on Geoscience and Remote Sensing* 60, pp. 1–16. DOI: 10.1109/TGRS.2021.3063756.

Da Silva, Elodie et al. (2023). "Towards Operational Fiducial Reference Measurement (FRM) Data for the Calibration and Validation of the Sentinel-3 Surface Topography Mission over Inland Waters, Sea Ice, and Land Ice". In: *Remote Sensing* 15.19. ISSN: 2072-4292. DOI: `10.3390/rs15194826`. URL: `https://www.mdpi.com/2072-4292/15/19/4826`.

Ida Olsen and Henriette Skourup (Feb. 2025). *ESA-CCI-RRDP-code*. Version v1.0. DOI: `https://doi.org/10.5281/zenodo.14808969`. URL: `https://github.com/Ida2750/ESA-CCI-RRDP-code`.