Responses to Reviewer #1 (Robbie Mallett)

We thank the reviewer for the thorough review of our manuscript and for the valuable comments and suggestions. We have carefully addressed each point and believe that the revisions have strengthened the manuscript. Our responses are provided below, in blue below each comment in black. When relevant, we have added quotes from the manuscript in orange.

Comments in the synopsis

The first is that the method used to assess the skill of the products is dominated by the seasonal cycle of ice thickening at the BGEP moorings. This is an issue because many product users (including the authors when approaching their time-trend analysis) are interested in interannual variability and trend detection. The seasonal time scale that dominates the evaluation does not say much about this, so is of limited use to product users. The metric's narrow scope may be masking real development-driven improvements in product skill over time, and its ubiquity may be masking inter-product diversity via Goodhart's Law.

We agree that the dominance of the seasonal cycle is an important limitation when assessing the ability of products to capture interannual variability and long-term trends. As discussed in our response to Reviewer 2, our primary objective was to evaluate the first-order consistency of the available sea-ice thickness products rather than their skill in reproducing anomalies or temporal trends. The seasonal cycle clearly is a key metric. We have clarified this objective throughout the manuscript and acknowledge that an anomaly-based evaluation represents an important extension of the SIN'XS assessment framework that will be investigated in future work.

We have included a paragraph explaining the problem and giving due credit to the studies of Nab et al. (2024) and Landy et al. (2022). However, as the SIN'XS project funding has ended, we are unable to extent the analysis in this regard.

Regarding the "broken" product, the actual satellite products do a better job when looking at the thickness extremes because their thickness range is closer to the ULS range. And there is also the fact that at least for the CryoSat-2 period there is hardly any trend in thicknesses, thus it is not too surprising that the climatology is a fairly good representation for the individual years. And all products owe parts of their seasonal cycles to the use of snow products that do of course include snow thickness cycles that strongly impact the common thickness retrievals.

1) Secondly, given the presence of a companion data-description paper submitted to Scientific Data, I think this paper should be more comprehensive if it is to be published in TC.
2) In terms of the inputs, it omits the sea ice thickness data produced by the UK Centre for Polar Observation and Modelling. This has historically been widely used, so I think will be of interest to many readers.
3) But the paper also uses only a small fraction of hard-won in-situ and airborne SIT evaluation data. As a community we constantly request more evaluation data. On that basis we should at least use the data that are available in papers like this. The explanation that these data weren't analysed because nobody put them into the SIN'XS database surprised me, given that the authors were and presumably are still in charge of the database, and are also free to download and analyse additional data themselves in this manuscript. If this paper is not the right place for a full analysis, where is?

We agree but the products included in the SIN'XS database were voluntarily provided by their respective data producers. We therefore did not independently download or preprocess products that had not been submitted, including the CPOM dataset. We have clarified this point in the manuscript.

Regarding the reference datasets, we fully agree that additional in situ and airborne observations would strengthen the analysis. However, the present study was intended to demonstrate the capabilities of the SIN'XS database using the currently available reference datasets. The integration of additional ULS, airborne and field measurements is already planned as part of the continued development of the database.

Finally, I'm curious about the Author Contributions statement, which suggests a major data curation effort was necessary for this paper, in addition to curation of the dataset already submitted to Scientific Data.

The Author Contributions statement has been revised by replacing "data curation" with "providing data in the SIN'XS format", which more accurately reflects the contribution of the corresponding authors.

Broad comments

Seasonal Bias in Assessment Method I think the assessment method used in the paper (shown well in its Figure 1) leads to a flattering and quite narrow evaluation of sea ice thickness products. This is because when calculated in this way, the correlation and RMSD statistics are dominated by the fidelity with which a product captures the seasonal cycle of ice thickening rather than any interannual variability or trends. We studied this phenomenon in Nab et al. (2024), and concluded that the SIN'XS project should carefully consider its rubric for assessing a product's skill (and avoid the approach that has been taken here).

I won't recap the paper in full, but essentially the seasonal cycle of sea ice thickening is much bigger than interannual variability or decadal trends. So the top right of the manuscript's Figure 1a is occupied by climatologically thick sea ice in April, and the bottom left is occupied by climatologically thin sea ice in October, and the skill metrics are therefore sensitive to how well a product represents this climatology. If the top right of the plot were dominated by years where the sea ice were thicker than usual, then high scores would represent a good representation of interannual variability. If the top right were full of datapoints from one end of the timeseries, then high scores would represent good capture of temporal trends (to the extent that they exist). But since the plot is currently dominated by the seasonal cycle, the scores are dominated by the skill in this quite narrow regard.

The limitations that come from the dominance of the seasonal cycle in the assessment are neatly illustrated by construction and evaluation of a "broken" sea ice thickness product. I made this by imagining CryoSat-2/Sentinel-3 were simply returning radar freeboards of zero all year round. I converted these radar freeboards to sea ice thickness with the unmodified Warren snow climatology and compared all EASE grid pixels within 75 km of the BGEP moorings to match the authors' analysis. This hopefully represents the most trivial sea ice thickness product one can imagine, using no data actually measured by a satellite. It has no year-to-year variability or trend. My Figure 1 (below) suggests that it's very comparable in skill to the AWI, GSFC and CPOM products when R2 and RMSE are calculated with the authors' method.

The reason the broken product performs well is because a seasonal climatology is encoded into it by the input of external snow data which reliably thickens over winter and thus reliably increases the derived thickness, even in the case where the radar data doesn't seasonally evolve at all.

Given the performance of my 'broken' sea ice thickness product in the manuscript's evaluation framework, I recommend further investigation of how the skill of a product on longer (perhaps more meaningful) timescales can be measured. Obviously I would say this having been part of Nab et al. (2024), but I do think one should consider that a

product's skill is best assessed by its ability to capture anomalies from its climatology. When assessed in this way (i.e. with the seasonal cycle removed), my "totally broken" product rightly per forms poorly, is meaningfully outperformed by the 'real' products. This methodology isn't new, having been used in Landy et al. (2022): '*Furthermore, after removing the climatological mean seasonal cycles of ice draft from the three long time series in the Beaufort Sea, the anomaly correlation coefficients between ULS and CryoSat-2 observations are 0.45, 0.51 and 0.37 for moorings A, B and D, respectively. This suggests that CryoSat-2 summer observations can capture a significant portion of the interannual variability in SIT recorded by moored ULS sensors*."

But why should we care about this seasonal bias in the assessment method? Some product users will indeed value the faithful capture of the seasonal cycle, but many others want to know how much the sea ice in a region or on a route is thinner or thicker than usual (operational ice charters for instance). They therefore will want to know which products have higher or lower variability, and whether this variability is real and matches the "true" variability seen at the moorings. Other users (including the authors in this manuscript) want to assess temporal trends in thickness over a decade or more. The assessment structure in Figure 1 is inadequate for that because it's so heavily weighted towards capturing skill on seasonal timescales. To summarise this point: even if the products produce a convincing climatology, we can't trust the trends unless the products actually know whether the sea ice is thicker or thinner than average in a given year. But also the ubiquity of this fairly narrow assessment rubric may actually be masking real improvements (or deterioration) in IAV and trends that come out of product development.

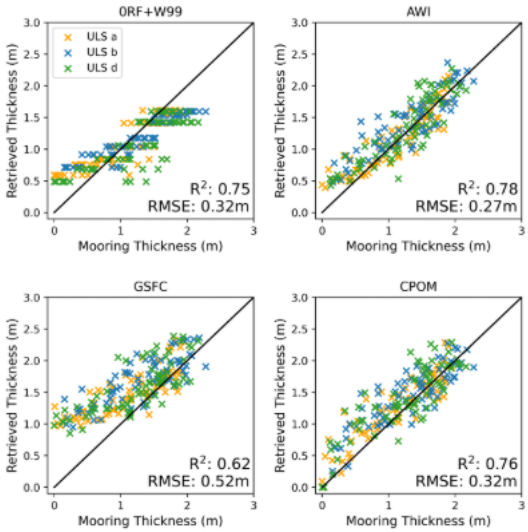

Figure 1: Broken sea ice thickness product vs three other CryoSat-2 products. The broken one (top left- constructed with zero-radar freeboards and W99) has comparable R2 and RMSE scores. The fact that it performs so well despite not showing no

interannual variability or trend suggests that these metrics of intercomparison are of limited use when used in this way.

I will finally add (and this is a minor point) that the ubiquity of this rubric in the literature raises the question of whether these metrics have become targets for product developers over time. This should incentivise other methods of intercomparison (and use of different evaluation data), because the statistics used here have already been quite extensively 'tuned to' by product developers. Analysing the seasonally-dominated statistics at BGEP moorings may therefore mask the true diversity between products because they're often tuned to score well in this particular analysis in this particular place.

We really appreciate this very interesting proposition as the proposed anomaly-based assessment would provide better evaluation on the capability of products to reproduce interannual variability. However, implementing this approach consistently across a diverse ensemble of products requires additional methodological developments, including the definition of product-specific climatologies.

The objective of the present study was to assess the overall consistency between products in terms of biases, agreement and long-term behaviour rather than to evaluate their ability to reproduce interannual anomalies. We have revised the manuscript to make this objective clearer and we have removed all the conclusions that seem to refer to the sea-ice (in general) to refer explicitly to the trends exhibited by the products. We also acknowledge that extending the evaluation framework to include anomaly-based metrics will be an important direction for future developments of the SIN'XS tools.

The concluding paragraph states: 'We would like to extend [the analysis] to other reference measurements, particularly airborne measurements (Icebird, Cryo-TEMPO, CryoVEx, OIB) and more ULS data (Lindsay, 2010) to account for a broader range of sea-ice types and regions'.

If the authors would like to do this, I'm curious about why it didn't happen? I can't think of a better team to do it; some authors on this paper are the leading experts in these in-situ and airborne observations (with the data often collected by them, or their own institutions). It would massively add value to the analysis, and without this analysis it's hard to interpret the results because by their own admission, the authors are limited by sea ice type and region. Noting that there are thirty authors on this paper and the distributed SIT products were pre-prepared and published anyway for another paper, this seems like a reasonable ask.

To give a concrete example, why not include the ULS moorings from the Fram Strait Arctic Outflow Observatory? These are publicly available at the appropriate time

resolution, and the analysis method is basically identical to the BGEP moorings but on the other side of the Arctic. There are also similar moorings in the Laptev Sea that I believe authors on this paper will have access to (see Landy et al. (2022) for a previous evaluation of this type). On line 258 reads: 'Since no more ULS reference data are available in the SIN'XS database, we cannot extend our analysis to other regions'. I would gently ask the authors: Why not just download these reference data onto your local machine? And were you not (at least partially) in charge of what went into the database?

The satellite sea ice thickness community often calls for more in-situ data for evaluation, and emphasizes their critical importance to funders. To be consistent with these calls, I think community assessments like this should use the full range of publicly available data or at least multiple sources. If the BGEP moorings are enough for us to properly evaluate our products, why do we spend so much money, carbon and time proposing and gathering more?

We fully agree that additional in situ and airborne reference datasets should be incorporated into future analyses. The primary purpose of this study, however, was to present the products currently available in the SIN'XS database and demonstrate the range of analyses that can be performed to evaluate their consistency.

The manuscript already represents a substantial body of work, and it was therefore necessary to limit the scope of the present analysis. We have nevertheless clarified that the integration of additional reference datasets, including further ULS measurements and airborne observations, is already planned.

We would like to thank the reviewers for the quality and professionalism of their comments. Many of these comments emphasize that, in order to quantify sea ice trends, it would be necessary to perform validations using other reference measurements than BGEP only (other moorings, airborne measurements), to study the uncertainties, and to correct anomalies for seasonal variability.

We fully agree and are taking these points into account for future developments of SINXS data and tools (some of which have already been implemented or are currently being implemented).

However, the objective of this paper was not to characterize sea ice trends themselves, but to characterize SINXS products. This ambiguity is our fault, as we have on several occasions referred to "sea ice trends" rather than "trends in sea ice products". We have therefore corrected all such ambiguities to ensure that we are referring specifically to the products.

Narrow/Technical comments

Going back to the concluding paragraph: *'We would like to extend [the analysis] to other reference measurements … to account for a broader range of sea-ice types'*.

If the authors strictly mean 'type' as in MYI and FYI, I think they can already split the analysis at the BGEP moorings into months when they're dominated by FYI cover vs MYI cover? See Figure 3 of Nab et al. (2024) for how this can be done. I totally agree that this is a desirable piece of analysis, particularly for some product users who are more sensitive to ice thickness biases over certain ice types. There's also clearly an argument that skill over FYI will become increasingly important for SIV estimates. But it would also be useful for product developers to identify routes to improvement, as well as for a discussion about why these products might be better or worse at the moorings in given years.

We thank the reviewer for this suggestion. We have clarified that the intended extension concerns a broader range of sea-ice environments, including fast ice, brash ice, and both winter and summer ice conditions, rather than only first-year and multi-year ice.

The first sentence of the abstract reads: *Sea-ice thickness is a key component of the Arctic climate system, but yet a comprehensive assessment across observations and numerical models is still missing.*

There's a strong implication to the reader here that this is the missing comprehensive assessment. In the pitch to the editor for this becoming a highlight paper upon publication, the authors made this explicit: *We provide for the first time a reconciled estimate of Arctic sea-ice thickness which comprises the entire range of available products*.

Unless I've misunderstood, the CPOM sea ice thickness product described by Laxon et al. (2013) and Tilling et al. (2018) wasn't included? I think this limits any ability to say that the paper analyses 'the entire range of available products'. It's historically been one of the more popular sea ice thickness products and is publicly available and regularly updated. There are multiple products from some groups (UiT, LEGOS)- so why omit the CPOM product in what is implicitly (and to the editor and reviewers explicitly) described as a comprehensive assessment?

As explained above, the current version of the SIN'XS database does not include every existing sea-ice thickness product, as participation relied on voluntary data contributions from product providers. We have revised the manuscript to clarify the scope of the study and avoid implying that every available product is included, while noting that the database nevertheless provides one of the most comprehensive collections of Arctic sea-ice thickness products assembled to date.

There are thirty authors on this paper, which seems a lot for an article that investigates a dataset that was previously curated and submitted to a data journal as a separate publication. Normally I would regard this as an editorial concern rather than one for me as a reviewer, but I am curious that nine authors have defined their contribution to this manuscript solely as "Data Curation". What data curation was there to do, given the sea ice thickness data were already curated for the Scientific Data submission? If the SIN'XS data submitted to Scientific Data required the dedication of nine additional authorships purely to subsequent "curation", then this casts some doubt on the usability of the SIN'XS data product, and also these doubly-curated data should be published and archived in their own right alongside this paper. I'm on board with publication and detailed description of large or complex data in data journals, with subsequent analysis of the data in a more conventional journal such as TC. But those who self-describe as having solely curated the original data surely do not warrant authorship on a subsequent analysis paper to which they themselves declare they have not contributed analysis?

The Author Contributions statement has been revised to replace "data curation" with "providing data in the SIN'XS format", which more accurately describes the contributions made to this study.

I'm not an expert in this stuff, but I wonder whether L184 would be better phrased as 'regridded them to the EASE2 grid'? My loose understanding is that the EASE2 grids are specific sets of x/y coordinates where the coordinate space (in the NH) is a Lambert Azimuthal equal-area projection. So things are gridded to EASE2, but projected to Lambert Azimuthal? I appreciate this is confusing and sometimes even the NSIDC refers to the EASE "projections". If there's somebody on the authorship team (probably several people) that thinks 'Robbie is just wrong on this' then feel free to say that and ignore this comment.

We thank the reviewer for this suggestion. The terminology has been corrected, and the manuscript now refers to products being *regridded* onto the EASE2 grid.

I'm not sure what the p-values correspond to in Figure 1. Presumably this is some kind of hypothesis test for a linear trend? The way it's presented, the reader might assume that the p-value somehow corresponds to the R-squared value, but to my knowledge R-squared statistics do not inherently come with a p-value. I suppose one could construct a hypothesis test that given a hypothetical random distribution, and then calculate the p-value for an R2 statistic that's at least as extreme as the value produced. But that seems unnecessary here since there's obviously a relationship and the interesting science surrounds the strength of the relationship rather than its statistical significance as indicated by a p-value.

We have clarified that the reported *p*-values correspond to the significance of the fitted linear regression, testing whether the observed relationship differs significantly from zero. The text has been revised to avoid potential ambiguity regarding their interpretation.

'It also highlights the necessity of dual-frequency KuKa measurements (Willatt et al., 2025), which can be used to derive snow thickness directly'

In that paper we saw that almost no co-polarised power came from the ice surface, so insofar as the KuKa instrument we used simulates CRISTAL (and it may well not do because of issues of length-scale and beam-geometry), the results imply that CRISTAL would not have 'worked' on that particular snow. So if anything, that paper goes in the pile that casts doubt on the CRISTAL dual-frequency method (although such a classification is obviously superficial, and surface-to-satellite comparisons are hard as I've mentioned). There are many papers in the pile that supports the dual-frequency operating principle of CRISTAL (e.g. Landy et al., 2026), but I wonder if citing Kern et al. (2020) may actually be best here for a reader who's not familiar with upcoming missions.

We agree and have moderated the corresponding statement and updated the references to better reflect the current understanding of dual-frequency retrieval approaches.

'a daily parametrised snow density based on (Mallett et al., 2020)' I didn't intend for that equation to be used so widely, and there are unfortunately quite a few downsides to what I did. I recently published an easier to use 'updated' version of this equation that is similar in form but more methodologically robust (Mallett, 2025). Suggest that would be better.

We appreciate the recommendation regarding the updated parameterisation. However, in the context of the present study, snow density is not expected to be the dominant

source of uncertainty, and using the updated formulation would have only a limited impact on the results. We therefore retained the original parameterisation.

Other comments

"The observed pause in sea-ice extent decline in the last twenty years is statistically robust." This statement is quite vague. I feel like it's a statement about whether it is statistically significant by comparison to the overlay of internal variability on the forced response? If so, it should be more specific.

We have revised the text to clarify that we refer to the apparent pause in sea-ice thinning in relation to the pause in sea-ice extent decline reported by England et al. We also emphasise that this comparison is based on first-order statistics and that a more detailed attribution analysis is beyond the scope of the present study.

Should there be units of [m] on RMSD in Figure 6, and Bias in Figure 4, and the heatmap of Figure 3?

The missing units have been added to the corresponding figures.

L49: I'm not sure altimeters do cover the full range of SIT - they really struggle when it's below 20 cm or so, hence the mergers with SMOS etc

We agree that the original wording was too broad. The text has been revised to clarify the reduced sensitivity of radar altimetry over very thin sea ice and the complementary role of passive microwave retrievals.

Related to the above, why do none of the heatmaps contain BGEP measurements of less than 0.2m? My analysis of the BGEP mooring data (see figure above) definitely does produce mean thicknesses for some months in this 0- 0.2 m range

The analysis presented in the manuscript is based on monthly averages for the winter season (October to April) using a 0.1 m binning. Within this dataset, no monthly mean values below approximately 0.15 m were found. We have clarified this point in the manuscript.

Furthermore on Figure 1, I'm suspicious that the "core" of the data cloud on panel (a) is blown out, i.e. it's passed the max of the colorbar. This is perhaps happening on panel c

too, but to a lesser extent. Normally this phenomenon is indicated by a small arrow on top of the colorbar (extend keyword in matplotlib), but more usefully one could just normalise the colormaps so that colors represent % of datapoints- this would allow better representation of the range I think.

We adjusted the colorbar to reflect % of datapoints as per your suggestion.