Responses to Reviewer #3 (Anonymous)

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.

Major comments

The assessment method used in Section 3.1 in scatter plots of monthly product values against BGEP reference data with correlation and RMSD statistics is dominated by the seasonal cycle of ice thickening rather than by interannual variability or trends. This has been studied in depth in Nab et al. (2024), and I strongly echo the concern raised there:

the seasonal signal in Arctic sea ice thickness is far larger than interannual or decadal variability, meaning that any product which faithfully reproduces the climatological seasonal cycle will score well in this framework, regardless of whether it captures year-to-year anomalies or long-term trends.

This is a critical internal inconsistency in the paper: the scientific goals are explicitly trend-oriented (Section 3.4, Figures 9–11), yet the validation framework tests a different and narrower dimension of product skill. I echo the concerns brought by reviewer Mallet to illustrate the severity of this problem in the use of a "broken" product constructed from zero radar freeboards plus the Warren snow climatology which achieves R^2 and RMSE scores comparable to the AWI, GSFC, and CPOM products under this framework. That result, demonstrated independently in the review of this manuscript, shows that the validation metrics as presented cannot distinguish between products that genuinely capture interannual thickness anomalies and those that do not.

I recommend the authors supplement the current analysis with anomaly correlation coefficients, calculated after removing the climatological mean seasonal cycle, consistent with the approach already used in Landy et al. (2022) and cited in the paper itself. This is not a prohibitive additional analysis, and it is far more appropriate for a paper whose headline findings are about trends. Additionally, the p-values reported in Figure 1 appear to test whether correlation differs from zero, a test that is trivially passed by seasonally structured data and that adds no meaningful interpretive information in this context. These should either be replaced with more informative statistics or their interpretation clarified.

Finally, I note that the widespread use of this particular validation rubric across the sea ice community, always at BGEP, always using the same seasonal statistics, raises the question of whether product developers have tuned toward these specific metrics over time. If so, the consistency between products observed in Figure 2 may partly reflect shared optimization targets rather than independently validated skill, masking true diversity between products.

We thank the reviewer for this very important comment. We agree that evaluating interannual variability independently of the dominant seasonal cycle is an interesting and valuable approach for assessing the ability of sea-ice thickness products to capture year-to-year variability. However, implementing such an analysis within the context of an ensemble of heterogeneous products requires additional methodological development, particularly regarding the definition and removal of product-specific seasonal cycles.

We confirm that this would be absolutely necessary to better evaluate the sea ice trends, but in fact the primary objective of this study was to assess the first-order consistency of sea-ice thickness products in terms of biases, consistency, and long-term behaviour, rather than to evaluate the actual sea ice interannual variability or trends. We acknowledge that this objective was not clear enough in the original manuscript and have corrected all the references to the ambiguous "sea ice trends" by more explicit references to the "trends shown by the sea ice products". We also consider the proposed anomaly-based evaluation to be highly relevant and intend to investigate its implementation in future developments of the SIN'XS assessment framework.

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.

The evaluation in Section 3.1 is limited to three ULS moorings in the Beaufort Sea. The authors acknowledge this limitation at line 256–258, but attribute the inability to extend the analysis to other regions solely to data availability within the SIN'XS database. Publicly available ULS data exist from, for example, the Fram Strait Arctic Outflow Observatory, and similar moorings in the Laptev Sea have been used in prior evaluation work by authors on this paper (see Landy et al., 2022). With 30 co-authors the decision to analyze only BGEP data warrants a more substantive justification than database scope.

This matters because the Beaufort Sea is a systematically biased validation region. As the paper itself notes (lines 302–303), it contains disproportionately thick multi-year ice advected from the Canadian Arctic Archipelago, and models are known to have their highest biases precisely in this region (Dupont et al., 2015, cited here). Conclusions drawn about relative product performance, particularly the finding that models outperform in the multi-year ice regime, and cannot be generalized from this single region. The paper's own concluding paragraph lists airborne campaigns (Icebird, CryoTEMPO, CryoVEx, OIB) and additional ULS sources as desirable extensions.

We fully agree that the current version of the SIN'XS database does not yet include all available reference datasets. The primary objective of the project was first to assemble a comprehensive collection of pan-Arctic and pan-Antarctic sea-ice thickness products. We also incorporated the most widely used reference datasets to enable large-scale product evaluations, although only the Beaufort Gyre moorings are presented in this study.

The extension of the database to include additional reference datasets, including ULS measurements from Fram Strait and the Laptev Sea, is already planned. We have added a statement in the manuscript to clarify this future development. We also note that interpreting Fram Strait observations is more challenging because of the coexistence of first-year ice, multi-year ice, and icebergs within the region.

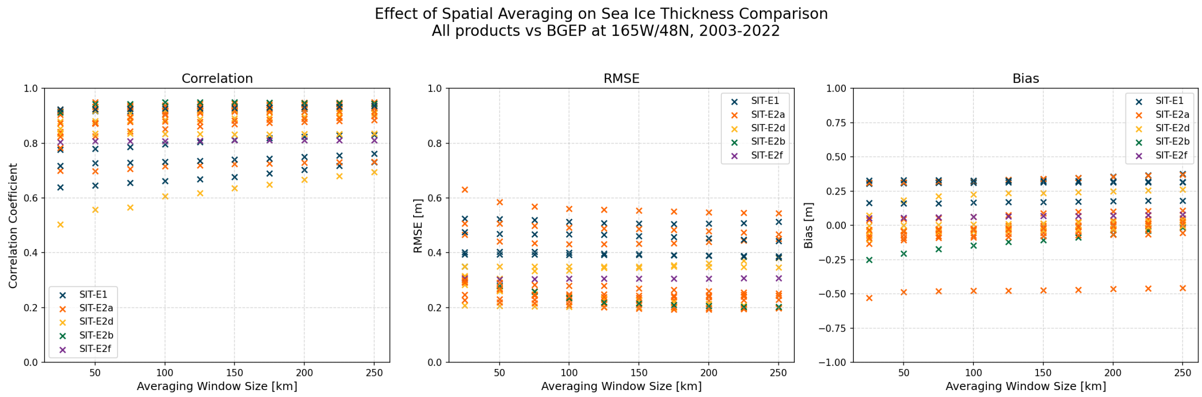
We also fully agree that the comparisons against BGEP moorings are not sufficient to draw conclusions regarding the actual sea ice trends. Thus, as explained in the previous answer we have replaced all the ambiguous "sea ice trends" by more explicit "trends shown by the sea ice products".

A significant omission from the paper is any treatment of uncertainties in the BGEP ULS reference data themselves. The ULS records are used throughout Sections 3.1–3.2 as ground truth, yet measurement uncertainties and spatial representativeness uncertainties are never discussed. This is a notable gap for a paper focused on uncertainty evaluation.

In particular, the choice of a 75 km averaging radius around each mooring (line 98) is presented without justification or sensitivity analysis. From experience, the choice of spatial averaging window makes a substantial difference to the resulting statistics: a smaller window reduces representativeness but increases sensitivity to local ice conditions, while a larger window averages over more products' coverage but risks including ice from different dynamic regimes. The authors should at minimum test several averaging radii (e.g., 25 km, 50 km, 75 km, 100 km) and report how bias, RMSD, and correlation change, to demonstrate the robustness of their results to this methodological choice.

Similarly, the conversion of ULS draft to sea-ice thickness requires assumptions about snow thickness, snow density, ice density, and sea-water density. The authors do perform a small perturbation analysis (±5 cm snow thickness, ±20 kg m^{-3} ice density; lines 93–95), which is appreciated, but the resulting uncertainty bounds in Figure 2 are described as small (0.04–0.06 m on average). The realism of these perturbation magnitudes should be justified more carefully, as snow thickness uncertainties on Arctic sea ice can substantially exceed ±5 cm.

We actually did perform a sensitivity study for the averaging radius as part of our analysis, but did not explicitly mention it yet to not overload the paper. The outcome is illustrated in this Figure:



which shows the dependency of RMSE, correlation and coefficient on the averaging radius for different product categories. The categories from the legend are SIT-E1 models/reanalyses, SIT-E2a → ku-band altimeters, SIT-E2b → laser altimeters, SIT-E2d → multi-product datasets, SIT-E2f → omitted from paper. From this we conclude that for most products there is little change for averaging radii above 75 km. This is an analysis for one BGEP mooring, but the results were similar for the others. A corresponding short sentence has been added to the paper (L163-L165 in the version without track changes).

The well-known and widely used CPOM CryoSat-2 product (Laxon et al., 2013; Tilling et al., 2018) is absent. This product is publicly available, regularly updated, and historically among the most widely cited in the literature. The paper includes three versions of the UiT CryoSat-2 product and four versions of the LEGOS product, but omits CPOM entirely. The claim of comprehensiveness is not sustainable under these circumstances, and the abstract and editorial pitch should be revised to reflect the actual scope of the database. I also note that the ICESat product selected for inclusion (GSFC-NSIDC_IS, Yi et al., 2011) is based on a study of Antarctic sea ice applied to the Arctic, whereas the heavily validated and widely referenced ICESat Arctic SIT product of Kwok and Cunningham (2008) — which was specifically developed and validated for the Arctic — is not included. This is a surprising omission given the field's reliance on that dataset, and it should be addressed or justified.

The SIN'XS database was intentionally populated using datasets voluntarily contributed by their respective data providers and the information was widely disseminated throughout the entire community. This approach was motivated by the desire to standardize the datasets in terms of format, projection type, and spatiotemporal sampling (monthly data in EASE2 projection with a resolution of 12.5 or 25 km in NetCDF format including sea ice thickness, snow thickness and if possible, uncertainties) to facilitate comparisons. We would not have been able to carry out this work for all available sea ice data (from satellite to model), and we also felt it was preferable for each provider to perform these transformations using the method they deemed most appropriate for their data. Consequently, many products, including well-established products such as the CPOM CryoSat-2 and the Arctic ICESat product of Kwok and Cunningham (2008), were not available for inclusion in the current version of the database. We have clarified this point in Section 2.1 and explicitly state that the database does not yet include all existing sea-ice thickness products.

Lines 35–40 enumerate the satellite altimeters used for sea-ice thickness estimation but omit several important instruments. AltiKa (SARAL/AltiKa), the Sentinel-3A and -3B satellites, and SWOT are not mentioned in this introductory survey despite being relevant to the field even if to mention in the study. Notably, the exclusion of a Ka-band altimeter (AltiKa) from the study itself is also unexplained. Given the discussion of dual frequency KuKa measurements (lines 275–278) and the role of Ka-band data in dual frequency snow thickness estimation, the rationale for excluding AltiKa deserves explicit treatment.

We thank the reviewer for pointing out this significant omission. The Introduction has been revised to include AltiKa, Sentinel-3A, Sentinel-3B, and SWOT among the satellite altimetry missions used for sea-ice thickness retrieval.

The study uses a 15% sea-ice concentration (SIC) threshold to define ice-covered grid cells (line 191). This threshold is standard in sea-ice extent products but is likely too low for SIT product comparisons. At 15% SIC, radar altimeter returns are heavily contaminated by open ocean backscatter, and the resulting thickness retrievals are unreliable. In the authors' own experience, and consistent with the broader altimetry literature, thresholds of 50% or higher are typically required to obtain robust thickness retrievals free of ocean wave contamination. The authors should analyze the sensitivity of their pairwise comparison and evaluation statistics to SIC threshold (e.g., comparing 15%, 30%, 50%) and justify their choice.

We agree that a 15% sea-ice concentration threshold would generally be too low if applied directly to sea-ice thickness retrievals. However, in this study the threshold was used only as a common mask to ensure that all products were compared over the same spatial domain. The individual sea-ice thickness products had already been processed by their respective producers using their own quality-control procedures and retrieval-specific masking criteria. Consequently, the additional 15% threshold has little practical impact on the analysed data while allowing the inclusion of marginal ice zone regions represented by models and passive microwave products. This clarification has been added to the manuscript.

Section 3.1 presents grouped statistics (bias, correlation, RMSD) for each product against BGEP data, but the discussion is largely descriptive. There is no evaluation of individual product strengths and weaknesses, no attempt to explain why certain products perform better or worse, and no guidance on how products should be used differently by downstream users. For example:

- Why do GSFC-NSIDC_IS and CCI_ENVISAT have the two worst RMSD values among satellite products? Is this attributable to their algorithm, their auxiliary assumptions, their sensor characteristics, or the time period they cover?
- Why do all model products show positive biases while satellite products show small negative biases? What physical or methodological mechanisms explain this?
- The finding that products sharing auxiliary assumptions with the BGEP reference conversion (same Warren snow climatology, same densities) tend to show lower biases is mentioned (lines 225–228) but not discussed as the potentially circular result it is. Products that share assumptions with the reference conversion should be expected to perform better simply due to methodological consistency, not due to superior measurement accuracy.

Without this interpretive layer, the evaluation in Section 3.1 does not allow users to make informed decisions about product selection, nor does it provide actionable feedback to product developers.

We agree that a more detailed interpretation of the behaviour of individual products would be valuable. However, such an analysis would require detailed knowledge of the retrieval assumptions adopted by each product, including snow thickness, snow density, ice density, and other processing choices, which are not consistently documented for all datasets included in the study.

The objective of this work was to provide a broad assessment of product consistency rather than a comprehensive evaluation of each individual retrieval algorithm. We have nevertheless expanded the discussion where appropriate and clarified the scope of the analysis in the revised manuscript.

The pairwise comparison in Section 3.2 reports bias, correlation, and RMSD between all product pairs but stops well short of interpretation. The key question — what do these differences tell us about the science? — is left largely unanswered. For instance:

- When two products based on the same CryoSat-2 input data (e.g., UiT_CS2_V2.1 vs. UiT_CS2_V2.2) show low RMSD, this is expected and uninformative. When they show surprisingly large differences, that is scientifically interesting and should be explored.
- Conversely, when products from entirely different sensor types (e.g., a laser altimeter and a passive microwave product) agree closely in certain regions or seasons, this convergence is potentially informative about the underlying truth. No such analysis is attempted.
- The comparison is restricted to March only. Many of the most scientifically and operationally relevant differences between products occur at the margins of the ice season — early freeze-up in October–November and late melt in May–June — when thin ice and mixed surface types create the greatest retrieval challenges. Restricting to a single month of peak winter ice limits the utility of this section substantially, and the authors should either expand to additional months or clearly justify the March-only choice.

We agree that grouping products according to their similarities provides a clearer overview of the main patterns emerging from the pairwise comparison. This organisation highlights common behaviours among products based on similar retrieval approaches.

A systematic interpretation of every pairwise comparison, however, is beyond the scope of the present study. Our objective is to characterise the overall consistency between product families rather than discuss the behaviour of every individual product combination. We also agree that a full analysis would require an evaluation among all the available months, but this would have overloaded the paper, and all the other sections (3.1, 3.3, 3.4) consider all the available months. Also, a late-melt analysis is not possible because most products are only provided until March.

The paper's most headline finding, a ~0.5–0.6 m decline in November SIT from 1995 to 2023, is based on data that change spatial domain over the analysis period. The LEGOS_30YRS product, which anchors the early period (pre-2010), only extends to

81.5°N, while CryoSat-2 products used post-2010 cover up to 88°N. Although the authors restrict the full-period analysis to below 81.5°N (Figure 10), this creates an asymmetric comparison: the early part of the trend reflects a southern-Arctic-only average, while the late period — even when similarly truncated — contains a richer mix of products.

The paper should include a quantitative sensitivity analysis demonstrating how the stated trend values change when: (a) the full analysis is restricted to below 81.5°N for all products throughout, and (b) when the same product (LEGOS_30YRS) is used as the sole backbone across the full period. Without this, it is difficult to determine whether the long-term trend is a robust geophysical signal or partly an artifact of the expanding product ensemble.

We acknowledge that the estimated trends depend on the products included in the analysis and that the confidence associated with the ensemble mean is influenced by the number and diversity of available products.

The purpose of this analysis was not to derive a new reference time series of Arctic sea-ice thickness, but rather to evaluate the consistency of the different products and assess their suitability for future ensemble approaches. While constructing a fully homogeneous long-term record remains an important objective, it is beyond the scope of the present study. We have therefore added a statement explicitly acknowledging this limitation and clarifying the interpretation of the reported trends.

Passive microwave products (BEC_SMOS, ESA_SMOS, UB_SMOS-SMAP) are correctly excluded from the pairwise comparison due to their 1 m thickness ceiling (line 206), and BGEP data are restricted to below 1 m for PMW evaluation (line 100). However, these products are still displayed alongside full-thickness products in Figure 2, where they predictably show the worst bias, correlation, and RMSD scores due to operating outside their design range. Showing them in the same panels without adequate visual separation or a clear accompanying explanation gives a misleading impression of their performance. The PMW evaluation should either be presented in a clearly separated sub-panel or discussed with sufficient context to prevent misinterpretation.

We agree that passive microwave products should be interpreted differently from products designed to retrieve the full range of sea-ice thickness. We have added a clarification: " The results obtained with SMOS confirmed that this methodology dedicated to thin ice (below 1m) in complementarity with altimetry is not adapted for the thicker ice encountered in Beaufort Gyre."

The "multi-product" category combines satellite-fusion products (ESA_CS2-SMOS, DMI_CS2-IS2, DMI_CS2-AMSR-AVHRR) with a model-assimilation hybrid (NERSC_TOPAZ4-CS2-SMOS) that extends back to 1994 and dominates the pre-2010 portion of the multi-product time series in Figure 10. Because this product uses model output corrected by satellite assimilation, its temporal behavior before 2011 is heavily model-influenced. Presenting the multi-product category as observationally constrained throughout the full record is misleading, and the trend attributed to this category should not be interpreted equivalently to the satellite altimeter trend.

We agree that the multi-product category combines products with different methodologies and levels of observational constraint. We have revised the manuscript to clarify that this category should not be interpreted as representing a homogeneous observational record and that the inclusion of hybrid assimilation products should be considered when interpreting the corresponding trends.

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.

Minor and technical comments

Writing and Grammar: The English prose throughout the manuscript requires careful editing. Several sentences are syntactically awkward or ambiguous. The authors should engage a native English speaker or professional editing service before resubmission.

The manuscript has been carefully proofread and revised to improve the clarity, grammar, and overall quality of the English throughout.

Lines 35–40: The list of altimeter missions is incomplete. AltiKa (SARAL), Sentinel-3A/B, and SWOT should be mentioned. The paper includes Sentinel-3 data (Table 1) but omits these missions from the introductory survey.

The Introduction has been updated to include AltiKa, Sentinel-3A, Sentinel-3B, and SWOT.

Line 49: The claim that "altimeters cover the full range of sea-ice thickness" is overstated. Altimeters struggle significantly with ice thinner than ~20 cm, which is why mergers with passive microwave (SMOS) products exist. This should be revised.

We agree that the original statement was too broad. The text has been revised to clarify the limitations of radar altimetry over very thin ice and to better explain the complementary role of passive microwave products.

Line 115: Typo — "cconcentration" should be "concentration."

The typographical error has been corrected.

Line 20: The copyright statement contains the placeholder "TEXT" and should be completed.

The placeholder has been removed from the revised manuscript.

Section 3.3, line 292: The extra period ("...-20°C (Bilello, 1961). .") should be corrected.

The typographical error has been corrected.

AltiKa/Ka-band: The paper discusses the importance of dual-frequency KuKa measurements (lines 275–278) but does not include any Ka-band product. The exclusion of AltiKa in particular should be explained given its relevance to the dual-frequency discussion.

No Ka-band freeboard product was available within the SIN'XS database for inclusion in the present analysis, which explains why AltiKa-derived products are not included. The database does contain a dual-frequency CryoSat-2/SARAL product, but the present study focuses exclusively on sea-ice thickness products. We have clarified this point in the revised manuscript.