1. Overview

This manuscript presents a large-scale intercomparison of 23 Arctic sea-ice thickness (SIT) products drawn from the SIN'XS database, spanning satellites, models, reanalyses, and multi-product datasets. The analysis includes evaluation against upward-looking sonar (ULS) reference data from the Beaufort Gyre Exploration Project (BGEP), a pairwise product intercomparison, spatial mapping, and a time-series trend analysis from 1995–2023. The breadth of the dataset assembled here is genuinely unprecedented, and the community will benefit from having such a resource. I also commend ESA for funding this effort and the authors for collating this database.

However, I recommend major revisions. My concerns fall into three broad categories: (1) methodological limitations that undermine the stated scientific goals, (2) incomplete and inconsistent treatment of both input products and reference data, and (3) a lack of substantive scientific interpretation throughout the results sections. These concerns are detailed below.

2. Major Comments

2.1 Seasonal Bias in the Validation Framework

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.

2.2 Incomplete and Inconsistent Use of Reference Data

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, Cryo-TEMPO, CryoVEx, OIB) and additional ULS sources as desirable extensions.

2.3 Uncertainties in the Reference Dataset Are Not Addressed

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.

2.4 Incomplete Product Selection and Inconsistent Justification

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.

2.5 Incomplete Coverage of Altimeter Missions in the Introduction

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 depth estimation, the rationale for excluding AltiKa deserves explicit treatment.

2.6 Ice Concentration Threshold

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.

2.7 Lack of Scientific Interpretation in Section 3.1

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.

2.8 Section 3.2 Pairwise Comparison Lacks Analytical Depth

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.

2.9 Domain Discontinuity Confounds the 1995–2023 Trend Analysis

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.

2.10 Inconsistent Treatment of Passive Microwave Products

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.

2.11 Multi-Product Category Heterogeneity in Trend Analysis

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.

3. 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.
- 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.
- 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.

- Line 115: Typo — "cconcentration" should be "concentration."
- Line 20: The copyright statement contains the placeholder "TEXT" and should be completed.
- Section 3.3, line 292: The extra period ("...-20°C (Bilello, 1961). .") should be 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.

4. Summary

This paper assembles an impressive and genuinely valuable ensemble of Arctic sea-ice thickness products, and the community will benefit from this resource. However, the scientific analysis does not yet do justice to the dataset. The validation framework is dominated by the seasonal cycle and therefore poorly suited to evaluating the trend-detection capabilities that the paper's primary findings depend on. The reference dataset is geographically restricted and its own uncertainties are not addressed. Key products are omitted without justification. The pairwise comparison and individual product evaluation lack the interpretive depth needed to provide actionable guidance to users or developers. And several methodological choices — the 15% SIC threshold, the 75 km averaging radius, the domain discontinuity in the long-term trend, and the conflation of heterogeneous product categories — require either revision or substantive sensitivity analysis.

I encourage the authors to address these concerns thoroughly. The dataset they have assembled deserves an analysis that fully exploits its potential, and with revision this paper could become a landmark community reference.