Reply to the comments on manuscript egusphere-2024-1896

Edel et al. “Reconstruction of Arctic sea ice thickness (1992–2010) based on a hybrid machine learning and data assimilation approach”

Dear Editor and Reviewers,

Thank you for your thoughtful comments and for the time and effort you have dedicated to reviewing our manuscript. We appreciate the valuable feedback, which has significantly helped us improve the quality and clarity of our work.

In response to your suggestions, we have carefully revised the manuscript to address the points raised. Below, we provide a summary of the major changes:

1. The comparison between SIT TOPAZ4-ML and BGEP has been made more objective, with significant adjustments throughout the section to improve clarity and precision.
2. Concerns regarding satellite uncertainties have been explicitly addressed in a newly added paragraph, providing a clearer understanding of potential limitations.
3. Textual inaccuracies and figure-related issues (such as figure order and referencing) have been corrected.
4. PIOMAS data has been incorporated into Figure 7, as well as Figure R4 (in this review document), to enhance the completeness of the analysis.
5. The overall writing style has been revised to ensure clarity and facilitate easier reading.

Here we present a point-by-point response to the reviewers’ comments and suggestions, with the original comments from reviewers in black font, and our replies in blue font.

We believe these revisions have strengthened the manuscript and are confident that it now meets the expectations of the journal. We appreciate your continued consideration of our submission and look forward to your feedback on the revised version.

Sincerely,

Léo Edel and co-authors

Reviewer 1

General comments:

This paper proposes to use a neural method applied to analysis increments to reconstruct the Arctic sea ice thickness field over non-observed periods. The combination of the use of analysis increments and the EOF decomposition used in the neural network is original. Over a test period, e.g. independent of the neural network training period, this method considerably reduces the bias present in the experiment with no assimilation. The magnitude of the long-term trends obtained with this hybrid method are in line with estimates from state-of-the-art models. This method with EOFS shows results comparable to those obtained with a similar method performed with full signal but at a much lower computing cost and resources.

Even if the results obtained with this NN method are no more convincing than those obtained with the rudimentary bias correction used in the paper, it still performs better over periods of high variability and offers many more options.

These results are convincing with satellite data, but the results remain inconclusive or even contradictory with in situ (BGEP) data. The presentation of these results needs to be more balanced in different parts of the paper. Further work is needed to understand this inconsistency.

There are many inaccuracies in the text and errors in the figures that make it difficult to read. The English should be carefully revised. The writing style of the paper is uneven, the final sections, discussion and conclusion are better written than the previous ones.

My knowledge of neural methods is too general, and the details given on the neural network method used in this paper need to be reviewed by a more qualified person.

Given the implementation of the NN does not compromise the results, this manuscript should be suitable for publication after a few minor revisions.

Thank you once again for your insightful suggestions and for the time you have dedicated to reviewing our manuscript. As outlined in the introduction, we have carefully revised the comparison with the mooring data (BGEP) and resolved the inconsistencies throughout the manuscript. Additionally, inaccuracies in both the text and figures have been addressed, and the writing style has been improved for clarity and readability.

We sincerely hope that the revisions meet your expectations, and we are grateful for your valuable contributions to improving the quality of this work.

Specific comments:

Abstract

L11-12 : this statement is only true for the mooring NPEO and not for the others mooring; mitigation on the results with in situ comparison must appear in the abstract

>> You are right. In the abstract, the following sentence has been added: “In contrast, when applied in the Beaufort Gyre, our method approaches the performance of a basic correction algorithm.”

As well as in the text, see section 4.3.1. paragraph 1, 2, 4. (Lines ~288-310)

1 – Introduction

l.21: rewording : “While in situ observations offer unparalleled accuracy”

>> Reworded as "ground truth".

Figure 1: either ULS and ICESat appear in the illustration, or it must be removed from the legend

>> The caption has been updated accordingly.

2 – Datasets

l.97 : introduce the reference of ERA5 here at the first mention?

>> The reference has been added to the first mention of ERA5, and removed from the 2nd mention.

l.103-l.106: missing https:// in front of all “doi” references.

>> "https://doi.org/" is necessary in front of all "doi" references. They have been added, but do not contribute to a smooth reading of this paragraph.

l.115-l.117: its not clear why this preprocessing is necessary, the preprocessing step is meant to do what?

>> The next sentence explains what this preprocessing step is meant to do: "This decision was made to maintain consistency between the two TOPAZ4 runs and ensure uniformity in sea ice extent across both datasets."

To complete the explanation, I can add that having the same sea ice extent between the 2 TOPAZ4 runs allows the algorithm to focus solely on the SIT correction and not on the sea ice extent correction.

The sentence has been rephrased: "This step ensures consistent sea ice extent across the two TOPAZ4 runs, allowing the ML algorithm to concentrate solely on adjusting the SIT."

l.140-141: Figure S3 contains locations of “Transdrift” buoys that are never used in the paper.

>> The buoys “Transdrift” have been removed from Figure S3.

l.150: interpolation or extrapolation?

>> As the data provider (https://nsidc.org/data/nsidc-0393/versions/1) and the user guide use the term interpolation, we also use "interpolation" in the paper. In addition, data surrounding the polar hole on all sides confirms the previous statement.

l.151: provide reference for Envisat dataset

>> The reference "(Hendricks et al. 2018)" has been added.

159: put citation in bracket.

>> Done

Methods for sea ice thickness adjustment

L171: irregular? To what extent? is this method can be used with sparse observations such as in situ data?

>> Yes indeed, the *in situ* ocean profiles are assimilated as shown in Xie et al. (2017) provided that these data are representative, but the *in situ* sea ice data have not been assimilated because their coverage is too poor. While exploring the capabilities of this approach, we assimilated temperature profiles from ITP buoys, which displayed artifacts of unusually high SIT, and thus were not included in the final study.

Since this point is already made in the submitted text, we do not suggest changes.

l.172: '… decomposed independently using EOF with several components ranging from four to height" ? not clear, EOF is decomposed on four or height modes?

>> Because each input feature is decomposed separately, the number of modes changes depending on the feature.

The sea ice thickness and sea ice age are decomposed in height modes, while the rest of the inputs are decomposed in four modes.

The sentence has been modified as follows: "... decomposed independently using either eight EOFs (sea ice thickness and age) or four EOFs (all other variables).".

l.173: what is the criterion of this threshold? Linked to the algorithm, computing resources...?

>> As stated in the text, the criterion is based on the importance of each input feature, which is determined by the machine learning during initial runs.

The training is conducted multiple times, each time excluding one variable. If excluding a variable does not significantly impact the prediction (for one given PC), it is then considered of low importance.

After 14 runs with different exclusions, the ML algorithm is run again without the variables that demonstrated low importance.

l.174: More comment on this figure S5 is needed as the 24 modes decomposition seems to show degraded results compared to the 8 one, e.g. for the bias and RMSE for instance.

>> While bias is indeed slightly worse for 24 modes compared to 8, the correlation shows a slight improvement, and the RMSE is also marginally reduced for 24 modes.

Overall, enhancement in one statistical indicator appears to result in a degradation in another, suggesting comparable outcomes regardless of the number of modes used in this study. Theoretically, we expect a higher number of modes to enhance the capacity of bias correction. Since using all EOFs converge to zero truncation error, such degradations are certainly accidental and not significant.

This last paragraph has been added to the caption of Figure S5.

l.185: input variables have much greater impact on the prediction, how this impact is assessed? Are some variables more important than others?

>> This impact has been assessed through multiple tries and has not been thoroughly reported.

Obviously, the structure of the ML algorithm changes the results, yet the hyperparameters (i.e. number of units of LSTM, rate of dropout) do not modify the prediction as much as the input features.

l.196: rewording “sanity check” is not adapted here

>> This has been changed to: "comprehensive assessment".

Results

Figure 4 is cited before Figure 3 in the text, reorder the figures; the ‘a)’ is not specified in the figure, idem for figures 3, 5, 9…

In all figures (4, 6, 8, 9) showing timeseries, the year labels are out of line with the annual cycle of the graph, review the positioning of these labels

>> The previously Figure 3 and 4 have been switched.

The 'a)' have been removed as they were not relevant.

On some timeseries, the labels are oriented relative to the end of the label rather than the center.

l.204: rewording "withdrawn”

>> This has been replaced with "omitted".

l.209: rewording “observations at hand”

>> This has been changed to "available observations".

l.211: What “SIT errors accumulate in the absence of SIT data for assimilation” means? Biases are larger during summer because of much thinner ice in TOPAZ-FR?

>> Firstly, Data assimilation reduces the SIT errors weekly, assuming that observations are considered the truth. When there is no data assimilation, the model diverges from reality and the SIT errors increase.

Secondly, the SIT in TOPAZ4-FR consistently appears thinner across all seasons, despite significant inter-annual variability. Yet large biases can be observed even when the average SIT of TOPAZ4-RA and TOPAZ4-FR are similar (i.e. summers 2012, 2014, 2016).

l.212-213: SIT bias is lower during 2020-2022 because of less volume in TOPAZ-RA, how can you explain this sudden change from 2019 to 2020 in the volume of TOPAZ-RA?

>> This is a fair and interesting question that would require further investigation and a more thorough examination of the TOPAZ4 model.

The September 2019 sea ice area minimum was a posteriori not a historical low like the year 2020, yet it was lower than the three previous years, so a sudden change of volume is not unrealistic. As the free run does not indicate such a decrease, we believe that the assimilation of CS2SMOS data – which also shows lower volume after October 2019 – imposes the sudden change.

The comparison with PIOMAS, following recommendations of review #2, indicates that PIOMAS does not exhibit the sudden change in SIT from 2019 to 2022 as in TOPAZ4-RA. The difference can be related to the difference in data assimilation between the two systems, but also differences in model setup (atmospheric and ocean inputs). Since the SIT uncertainties are high, we cannot claim with confidence that the sharp decrease did or did not happen.

l.216: not in the Canadian Archipelagoes

>> Indeed! The sentence has been modified:

"close to the north of Greenland and the Fram Strait, while it depicts too thick sea ice in the Beaufort Gyre and Canadian Archipelago. "

l.217: rewording “The amplitude of this error varies slightly according to seasons, yet remains observable at all times”

>> This has been reworded as: "The magnitude of this error fluctuates slightly with the seasons but remains a systematic feature."

l.226: choosing a contiguous period makes a possible dependencies with the training period by the temporal autocorrelation, clarify or remove this sentence.

>> Since the SIT on any given day is closely related to that of neighboring days, the temporal autocorrelation of SIT data is high over short time scales (approximately ±2 months). Using a test period with randomly selected days could increase the likelihood of having similar conditions between the training and testing phases. By choosing two contiguous periods for training and testing, we ensure that any similarity between the periods occurs only once, thereby reducing the risk of having similar days in both phases.

The sentence has been modified as "Due to the high temporal autocorrelation of SIT data over short time scales (±2 months), we chose two contiguous periods for the test and training datasets, rather than using the method of random shuffling, to minimize dependencies between them.".

Figure5: RMSE against what? TOPAZ-RA? Over the entire period 2011-2013? How the EOF error is estimated? Why use another colorbar (black background) than Fig 3? tighten the limits of the colourbar, the field is difficult to discern at low values

>> All RMSEs are defined as differences with respect to TOPAZ-RA.

The legend has been adjusted accordingly: "RMSE of SIT bias (m) over the test period (2011-2013) of left) ML-adjusted error, middle) EOF error, right) baseline error against the bias between TOPAZ4-FR and TOPAZ4-RA.".

Figure 3 shows a diverging colormap whereas Figure 5 uses a sequential colormap, this is why the value at 0 is different. The colorbar has been adjusted (0, 1.5m) to observe low values more easily.

l.240: “outperforming” is overstated…

>> Replaced with "more accurate than".

Merge Figure 6 and Figure 5 Top?

>> Since Figure 5 focuses only on the test period (2011-2013) and Figure 6 illustrates the temporal evolution from 2011 to 2022, we prefer to keep those two figures separated.

l.246-248 : Figure 6 is introduced (l.247-248) after her first citation l.246.

>> The first sentence conveys the main point of the paragraph, while the second sentence introduces Figure 6. Although I understand your comment, I believe it is more logical for the reader to keep the structure as it is.

l.252: the baseline is obviously at a disadvantage in these years, where TOPAZ-RA shows a much lower volume than the average for previous years

>> Indeed, it is one of the limitations of the baseline, expressed in the section 5 Discussion.

l.253: sentence to be rephrased"… at times higher than the main peak."

>> This has been rephrased: "occasionally thicker than the winter maximum."

l.255: "the thin ice melts first and the surviving thick ice causes the average to increase where the ice is still present » why don't use the volume then?

>> For consistency between the spatial and temporal analyses, it is preferable to continue showing sea ice thickness. In addition, this is the primary quantity we aim to evaluate, without conflating it with sea ice concentration.

For completeness, we have included a plot showing the sea ice volume in the appendix.

l.274: "For this task, the most…" useless statement as in situ data have been already described.

>> This sentence has been moved to section 2.5: Validation data: Mooring data.

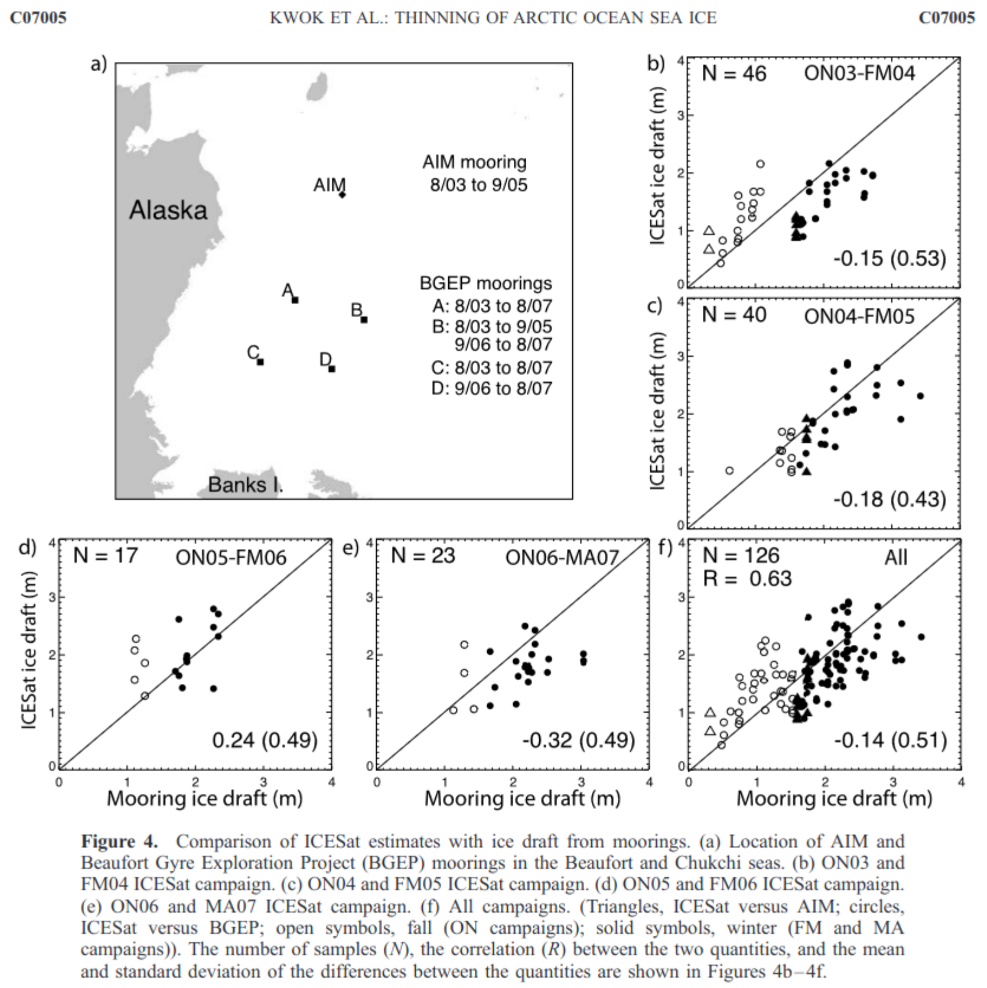
l.277: Figure 8 cited before Figure 7, reorder the figures.

>> Figures have been reordered.

l.277: Figure 7 seems to show a clear improvement between TOPAZ-ML and TOPAZ-FR. How does ICESat compare with BGEP?

>> Analyses between ICESat-1 and BGEP have been done in Kwok et al. 2009, Thinning and volume loss of the Arctic Ocean sea ice cover 2003–2008. Here is Figure 4:

Comparison between ICESat-1 and BGEP buoys are marked with open circles (summer: October-November) and solid circles (winter: February-March or March-April). Across all years, sea ice draft differences of up to 1m can be observed (Fig. Rx 4f). On average, ICESat-1 underestimates sea ice draft compared to buoys in winter (solid circles) and overestimates sea ice draft in summer (open circles).



C07005 KWOK ET AL.: THINNING OF ARCTIC OCEAN SEA ICE C07005

Figure 4. Comparison of ICESat estimates with ice draft from moorings. (a) Location of AIM and Beaufort Gyre Exploration Project (BGEP) moorings in the Beaufort and Chukchi seas. (b) ON03 and FM04 ICESat campaign. (c) ON04 and FM05 ICESat campaign. (d) ON05 and FM06 ICESat campaign. (e) ON06 and MA07 ICESat campaign. (f) All campaigns. (Triangles, ICESat versus AIM; circles, ICESat versus BGEP; open symbols, fall (ON campaigns); solid symbols, winter (FM and MA campaigns)). The number of samples (*N*), the correlation (*R*) between the two quantities, and the mean and standard deviation of the differences between the quantities are shown in Figures 4b–4f.

Figure R1: From Kwok et al. 2009.

The comparison to BGEP is obviously a difficulty for both models and remote sensing data. Qualitatively both model and remote sensing overpredict medium values of SIT, possibly for different reasons.

l.276 and l.289: this statement is wrong, Table 3 shows baseline is generally statistically better than ML for all BGEP buoys and, for BGEP A, the unique improvement compared to baseline is the reduction of the bias during summer. Further the free run overall shows better statistics than ML experiment. The ML shows systematic better statistics for the NPEO measurements.

>> Indeed, thank you for your comment. This has been modified in the text, see section 4.3.1. paragraph 1, 2, 4. (Lines ~300-335).

Moreover, a paragraph in section 4.3. has been added to underline the difficulties of such a validation exercise and to highlight the differences between observation datasets.

l.293: what is the "full spatial coverage"? free run is systematically better than adjustments for BGEP C biases.

>> "Full spatial coverage" meant when considering the whole Arctic area and not one specific buoy location.

What we mean is "All the scores are poorer when compared to buoy locations compared to studies conducted on the test period over the whole Arctic (section 4.2.), but the adjustments are never worse than the free run".

The statement was true for a previous TOPAZ4-ML simulation but does not hold for the results presented in this paper. This sentence has been removed from the paper (see previous comment).

Figure 7 : the reduction of SIT along the Siberian coast with TOPAZ4-ML is not clear, this qualitative description is difficult to assess. A map of differences, using ICESat as a reference, is more relevant for discussion and gives the possibility to a quantitative estimate of the differences. Why not mask the ICESat field north of 86°N, unless these values have some significance?

>> The reduction of SIT is indeed unclear and has been removed from the text.

We are not equipped for a proper quantitative comparison at the exact time and location of satellite data but wanted instead a qualitative comparison with the full data coverage to show interpretable features like the ridging near land masses and islands, the background colour of first year level ice. For that, we preferred to keep the full extent of the data despite ENVISAT's large polar hole.
Since ICESat includes data north of 86° N, and there is little to no data available for comparison at these latitudes, it would be deplorable to exclude them, despite the reduced quality due to interpolation.

Because PIOMAS has been added to this plot, the following sentence has been added to the paragraph: "Additionally, while PIOMAS SIT appears to be lower than ICESat-1 and Envisat along the coasts of Siberian and Alaska, it is generally consistent with satellite observations in the Central Arctic along 80°N."

L.303: Compared to ICESat, Envisat shows thinner SIT in the Barents Sea in Figure 7, in contrast to the Figure S6. What is the interest of using Envisat data if these estimates are not considered valid and are not subsequently used?

>> Both ICESat and Envisat have high uncertainty, yet are considered valid and, more importantly, are the only remote sensing observations available for this time period. Their inclusion in this work is crucial, particularly to highlight the challenges in reaching a consensus on historical sea ice thickness.

Since the true SIT is unknown, this section is intended as an inter-comparison rather than a definitive validation exercise. Line 303 only identifies the SIT in the Barents Sea as unrealistic, without discrediting Envisat data for the rest of the Arctic or other months.

In addition, Envisat's altimeter is possibly closer to the type of measurement obtained from CryOSAT-2's altimeter measurements, which, as part of CS2SMOS, has been used for training the ML. Assuming that the differences between radar and lidar altimeters prevail, we could have expected features of TOPAZ-ML to be closer to the Envisat measurements than to ICESat. This was, however, not obvious.

l.305: references of these reports?

>> Tilling et al. 2019 has been cited in section 2: Validation data: remote sensing. This reference has been added to line 305. An additional reference to Paul et al. 2018 has been added in both paragraphs.

l.313: provide the estimate of PIOMAS

The estimation of PIOMAS has been computed for the period 1992-2022 (before it was extracted from Schweiger et al. 2014 for the period 1979-2018) and reveals a trend of –3583 km^3/decade. The sentence has been modified as follows:

"The year-round trend is –3 153 km^3/decade according to our reconstruction, while the PIOMAS model reconstruction (Schweiger et al., 2014) estimates a slightly steeper trend of –3 583 km^3 /decade."

As well as the sentences in the section 5 lines 355-357:

"PIOMAS shows trends of -2.7 and 3.2 +/- 1 x10^3 km^3/decade for April and September, respectively, from 1979 to 2018 (Johannessen et al., 2020) [Fig. 5.24.]. In comparison, over the period 1992-2022, PIOMAS indicates –3.0 and -3.8 +/-1 x10^3 km^3/decade while TOPAZ4-ML shows trends of –3 120 and –2 960 km^3/decade for April and September, respectively. Although the two datasets align well for April, a notable discrepancy emerges in September, with PIOMAS indicating a more pronounced downward trend."

Figure9: Top : daily SIT from TOPAZ-ML?

>> Correct. This has been added to the caption.

l.325, Page 16: is this change from bimodal to unimodal distribution present in other experiments than TOPAZ-ML?

>> To clarify this point, the two following figures have been added to this document.

This shift from bimodal to unimodal distribution is also present in other experiments, however not to the same extent as TOPAZ4-ML. As stated in the manuscript, the SIT distribution is significantly different in TOPAZ4-BL and TOPAZ4-FR than in TOAPZ4-ML. We do not have sufficient observational support to believe that one is more realistic that the other.

In the FR bimodal distribution is also visible, yet there is one thickness that clearly prevails (usually between 1 and 2m), except between 2000 and 2004, when we can distinguish one mode around 1m and the second one around 2/2.5m. In the BL, we observed more distinctly the two modes between 1995 to 2007.

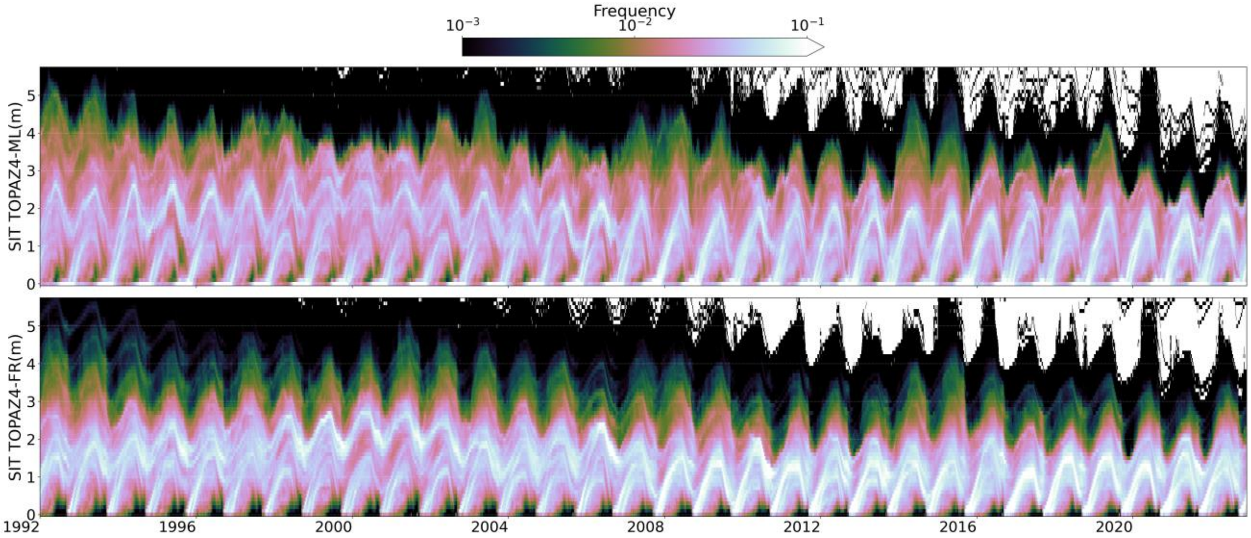


Figure R2: Distribution of daily SIT from TOPAZ4-ML (top) and TOPAZ4-FR (bottom).

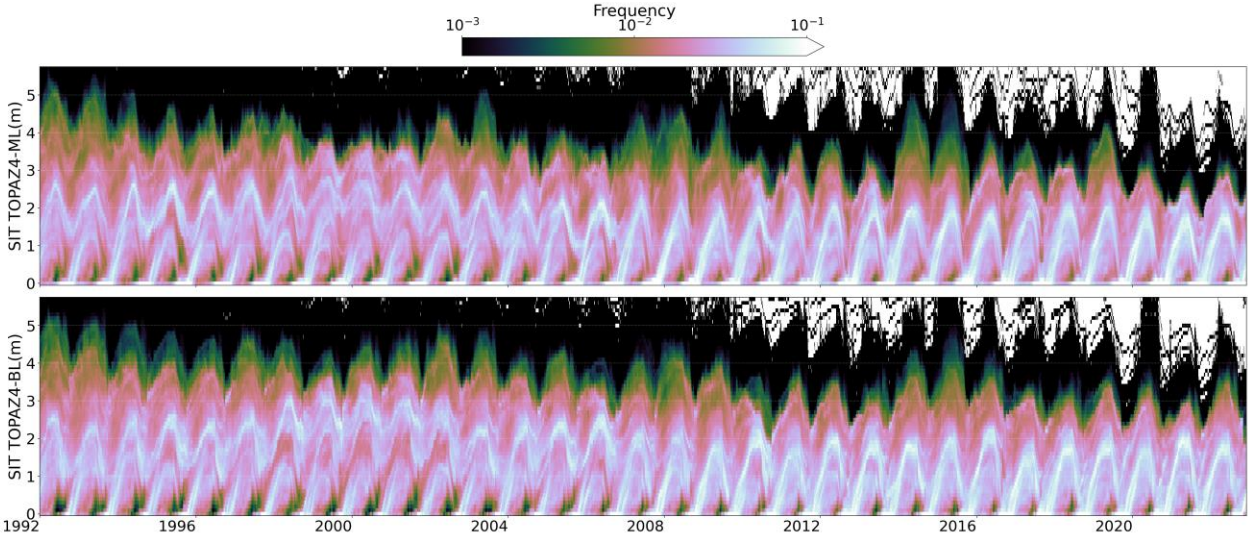


Figure R3: Distribution of daily SIT from TOPAZ4-ML (top) and TOPAZ4-BL (bottom).

5- Discussion

l.364: legend of the figure S2 is not complete : The cumulative explained variance is noted in the upper right corner of each EOF. The 95th percentile of the absolute values is noted on the upper right corner of each PC subplot

>> Thank you for this correction. This has been added to the caption of figure S2.

l.373: correction: a large part of the biases

>> This has been modified accordingly.

l.398: why this second peak is only observed twice in TOPAZ-RA ?

>> As explained in this paragraph, the second peak is more pronounced when thin sea ice is erroneously placed in the central Arctic or when thicker sea ice is located outside of the central Arctic. Consequently, reducing these types of errors would also diminish the

magnitude of the second peak. We believe that because TOPAZ4-RA is more accurate, then this second-peak feature is less prominent.

6– Conclusions

l.418: a low year-round bias of -10.0cm …

>> Modified.

l.421: ML doesn't show an overall improvement compared to the free run. See the comment in section 4.

>> This sentence has been modified "Applying our algorithm before 2011, the evaluation with independent mooring data indicates improvement in the central Arctic compared to TOPAZ4 free run, while the Beaufort Gyre results show a slight decline in performance."

Appendix C: Overestimation of the adjustment with ML algorithm compared to the bias correction between 2 and 5 m.

>> The comparison was mentioned in the previous sentence to lighten the phrasing. If you believe it is essential for the reader's understanding, it can be included. For now, for the sake of clarity, we will maintain the current phrasing.

Figure C1: last sentence : … with less than …

>> This has been corrected.

References:

Kwok, Ron, et al. "Thinning and volume loss of the Arctic Ocean sea ice cover: 2003–2008." *Journal of Geophysical Research: Oceans* 114.C7 (2009).

Reviewer 2

General comments:

The manuscript "Reconstruction of Arctic sea ice thickness (1992-2010) based on a hybrid machine learning and data assimilation approach" by Edel et al. presents a new study in which a Machine Learning (ML) model is used to predict the principal components of sea ice thickness Data Assimilation (DA) increments, where increments were generated from the assimilation of CS2SMOS sea ice thickness into the TOPAZ4 ocean-sea ice model between 2011-2022. This ML model is then used to bias-correct sea ice thickness in TOPAZ4 simulations over the period 1992-2010. The authors also provide an assessment of historical sea ice thickness trends from this bias-corrected model.

The study is novel and builds nicely on past work in hybrid DA+ML for learning numerical model errors. The study also adds a new component of learning the principal components of the increments, rather than the increments themselves. This has the nice feature of working in low dimensional space and allows fast training of ML models. I found some parts of the manuscript a little difficult to follow, particularly with Figure references bouncing back and forth. For example, Figure 4 gets referenced before Figure 3, which meant I was re-reading sections because I thought I had missed references to an earlier figure. This happens a few times throughout the manuscript. I suggest reordering the figures so that they appear in the same order as they are referenced in the text. I also have some general concerns about the validation of historical sea ice thickness against different altimeters, given the inherent inconsistencies between these products. Furthermore, given that PIOMAS has long been the 'industry standard' for evaluating sea ice thickness/volume over the historical period, I think more comparisons with PIOMAS would greatly strengthen this work. I appreciate the author's references to PIOMAS in terms of historical trends etc in the discussion, but it would be great to see actual PIOMAS data included in e.g., Figures 4-7. Otherwise, I think the manuscript is in good order and will be ready to publish after some minor revisions. Please see my questions/comments below. My thanks to the authors for their interesting work and I look forward to reading the revised version.

We greatly appreciate your thoughtful feedback and the time you have taken to review our manuscript.

As mentioned in the introduction, we have corrected inaccuracies in the text and figures to enhance clarity and flow. Additionally, we have addressed the inconsistencies between the validation products in a newly added paragraph to provide clearer guidance for the reader. Your suggestion to develop the comparison with PIOMAS was particularly valuable and appreciated, as it offers a well-known reference point that enhances the context for readers.

We appreciate your thoughtful feedback and hope the revisions meet your expectations.

Minor comments:

L6: "with and without CS2SMOS assimilation" is a little misleading here. It's effectively the same simulation, but you just have no summer data, so there's no assimilation happening over this summer period. Unless I've missed something, currently this reads as if you run two independent simulations and learn errors in both of these model runs.

>> We run two independent simulations, one assimilating CS2SMOS (and other products detailed in section 2.2. TOPAZ) and another without any assimilation. We obtain errors from the differences between these two runs.

This has been rephrased: “In this study, we train a machine learning (ML) algorithm to learn the systematic SIT errors between two simulations of the model TOPAZ4 over 2011-2022, one with CS2SMOS assimilation and another without any assimilation, to predict the SIT error and extrapolate the SIT prior to 2011.”

L35: suggest changing "parameter" to "quantity"

>> Changed

L46: suggest changing "remain important" to "remain significant"

>> Changed

L50: not sure what is meant by "spatial and temporal distributions" here. I would suggest adding some specific differences, such as retracking algorithms for satellite altimeters (e.g., Landy et al., 2020 vs Tilling et al., 2018).

>> For clarity purposes, it has been changed to “spatial and temporal coverage” and your point has been added: “Similarly, large deviations are observed when comparing satellite products (Sallila et al. 2019) or diverse in-situ datasets, mostly due to differences in spatial and temporal coverage (Lindsay et al. 2015, Labe et al. 2018), and in data processing methods, such as retracking algorithms for satellite altimeters (e.g. Tilling et al. 2018, Landy et al. 2020).”.

L54: "53 cm to 38 cm and down to 20 cm in March" is a little hard to read. Does this mean March SIT bias goes from 53 cm to 20 cm in March? Where does the value of 38 come in?

>> This means the mean March SIT bias goes from 53 cm to 20 cm.

Average over the entire year, the mean SIT bias goes from 53 cm to 38cm.

This sentence has been rephrased: "Assimilating CS2SMOS data in the coupled ocean-sea-ice model TOPAZ corrects a low SIT bias of roughly 16 cm, reducing average RMS errors from 53 to 38 cm and further to 20 cm in March."

L56-57: without reading the Brajard paper, it's not clear what the "iterative" nature of their method refers to. I would just say something like "Brajard et al (2020) introduced a method to combine DA and ML to build a hybrid numerical model. The present study is applying this approach to 'rewind' a climate record."

>> The sentences have been changed based on your suggestions. Indeed, the iterative nature of the method could bring undesired confusion.

L62: note that we released a follow-up paper from Gregory et al., 2023, showing that we can implement the machine-learned sea ice concentration increments into numerical simulations to both significantly reduce the model bias, and also develop a new data augmentation framework for online refinement of ML models (see Gregory et al., 2024).

>> Noted! Thank you for your work and the nice read.

L64: could you clarify what you mean here by "distorting" physical variables?

>> "distorting" refers to altering or misrepresenting the original relationships between the physical variables when they are reduced in dimensions.

We mean that the complex relationships between the variables remain accurate and true to their original form, even after use of EOF.

"distorting" has been changed to "altering" to convey the idea that relationships remain unchanged more explicitly.

L67: suggest rephrasing to "extrapolate the SIT errors prior to 2011"

>> It has been rephrased: "extend the SIT estimates to periods before 2011."

L83: suggest rephrasing to "The combination of CS2 and SMOS better handles their individual deficiencies in accurately resolving thin (<1m) and thick (>1m) sea ice floes, respectively"

>> The sentence has been modified according to your suggestion.

L84: suggest rephrasing "estimate of the total spectrum of sea ice" to "a more accurate representation of the true sea ice thickness distribution"

>> The phrase has been changed as follows:

"This advanced merged product provides the first accurate representation of the true sea ice thickness distribution, with such temporal continuity and spatial coverage.".

L85: suggest clarifying for first-time readers that altimeters can indeed measure surface elevation in summer, it's just that (besides Landy et al 2022) our current processing routines are largely insufficient to distinguish sea ice leads from surface melt ponds

>> The sentence has been clarified: "Due to challenges in differentiating between sea ice leads and surface melt ponds during the melting season, the observation period is limited to October through April, starting in 2010."

L89: could you clarify what is the "Arctic Ocean operational forecast"? Is this an operational forecast system run out of Nansen, using TOPAZ4?

>> TOPAZ is used by MET Norway for operational forecast and distributed by the Copernicus Marine Services.

L91: suggest clarying "coupled with a single-thickness-category sea ice model"

>> This has been added to the sentence.

L93: can you provide more information on how you generate the 100 ensemble members? Is this through e.g., perturbed sea ice physics? or something else?

>> The 100 ensemble members were initially from a long free run, biweekly picking one from snapshots in the summer months. Secondly, during the ensemble spin-up, the initial members had been freely driven by the perturbed atmosphere forcing from ECMWF for more than six months. Then, after several months of data assimilation, they were officially implemented to derive the reanalysis.

It is not to directly perturb sea ice physics, but it is a good suggestion, and we will test this perturbation strategy in the next reanalysis version. So far, most of the atmosphere-forcing fields have been perturbed by adding 2-D random noises with a 250 km spatial correlation scale and a 3-day time scale.

L95: i'm not sure "archive" is the right word here. Do you mean simply "to generate real-time forecasts..."

>> Thank you for this comment. "Archive" does seem wrong in this sentence. "Generate" is used.

L102: suggest removing "weekly" since you mention that assimilation is performed weekly on L107

>> "weekly" has been removed line 102.

L150: why fill the ICESat-1 polar hole? Can you provide more details on how you do the interpolation?

>> The interpolation is done by the data provider: Yi et al. 2009. To prevent this misunderstanding, the sentence has been updated with the reference:

"… is filled through interpolation (Yi and Zwally, 2009)."

L156-159: can you say more about the robustness of using these various altimeters to validate your historical simulation? Given that you are learning DA increments from the assimilation of CS2SMOS, you are hoping that your historical simulations will match what CS2SMOS would have been if it had been active over 1992-2010. One of the main reasons we haven't yet seen a continuous thickness record from the combination of various altimeters spanning the 1990s—present is because of their inherent differences in instrumentation, footprint resolution, sampling biases etc. Is it therefore fair to compare your modeled SIT to these data?

>> The reviewer's concerns are justified and clearly articulated. Indeed, these altimeter datasets present inherent differences in instrumentation which result in high SIT uncertainties.

While comparing modeled SIT from ICESat-1 and Envisat with our product is not ideal, they are the only observational datasets with coverage across the Arctic during these past periods.

In this study, our goal is not to determine which product is superior, but to verify that our product exhibits similar SIT patterns and spatial distribution than past observations.

To address this concern regarding fairness, we have added a paragraph in section 4.3.:

" In the following section, we use several validation datasets (described in section 2) as a series of indicators to assess the reliability of our sea ice thickness estimations. Unfortunately, the absence of a universal ground truth for sea ice thickness makes validation challenging. By presenting diverse sources of SIT, we aim to provide a comprehensive view of the legitimacy of our correction. Given the strengths and limitations of each product, we recommend the reader be mindful of differences in sea ice thickness related to various observation types, including measurement methods, processing techniques, and associated uncertainties, as well as any potential inconsistencies between products."

We can also cite the following paper, which aim to provide a SIT from various satellites from 1995 to 2021:

Bocquet, Marion, et al. "Arctic sea ice radar freeboard retrieval from the European Remote-Sensing Satellite (ERS-2) using altimetry: Toward sea ice thickness observation from 1995 to 2021." *The Cryosphere* 17.7 (2023): 3013-3039.

L163: can you clarify what you mean by "apply a strong adjustment" here

>> We mean "applying an adjustment which reduces the bias by a significant proportion". "Strong" has been replaced by "substantial". "Apply a substantial correction" could be clearer, but I prefer using the word "adjustment", because "correction" could imply that we know what "correct" is, which is not the case of SIT.

L173: is the choice "arbitrary"? Your choice was driven by the analysis conducted in Figure S5 no?

>> The choice itself was not arbitrary, but the threshold used to select the most important variables is arbitrary.

Line 173 concerns the selection of the variables, while Figure S5 concerns the number of modes for the EOF decomposition of the target variable (the SIT bias between TOPAZ4-DA and TOPAZ4-FR).

Table 2: Rather than having a 'x' in each box, could this be replaced with the percentage of variance explained by each PC for each variable? Just to maximize the information gained from the table (if it looks too messy then maybe it could be a supplementary figure).

>> This is a good suggestion. Yet, the specific importance of each variable has not been computed, as the SHAP package (commonly used) is not compatible with the LSTM layers of TensorFlow. Although such an analysis would provide valuable insights, it is currently not feasible.

L187: by "the uncertainty associated with the nonlinear estimation", do you mean the uncertainty associated with a given choice of ML architecture and inputs? More generally, in this paragraph, can you not characterize the uncertainty given that you have 100 ensemble members? Conceivably you could train 100 different networks which all start from identical random weights and get a good estimate of prediction uncertainty coming from the internal variability across your inputs

>> Yes, we mean the uncertainty associated with a given choice of ML architecture and inputs. More precisely, we refer to the uncertainty associated with the sensitivity of the input features. We have 100 members of perturbed inputs to assess the sensitivity of the ML.

Yes, conceivably we could train 100 different networks and estimate the uncertainty arising from the training process of the ML algorithm. This work has not been done due to time constraints.

However, the uncertainty associated with the TOPAZ4 FreeRun model and the uncertainty associated with the changes in sea ice conditions cannot be computed with the current resources. Under these circumstances, the total uncertainty cannot be determined.

L200: for "the corresponding month", do you compute bias over a running window, e.g., 30 days? Or is the bias on each day of the month identical, and then it suddenly changes on the first of the next month?

>> The bias is identical for the whole month, then suddenly changes on the first of the next month. While applying a 30-day running window could improve smoothness and enhance the baseline, we believe the current approach is satisfactory for the scope of this study.

Section 4.1: suggest swapping paragraphs one and two, since paragraph one talks about Figure 4 and paragraph two talks about Figure 3.

>> The figures have been swapped and we chose to keep the order of the paragraphs.

L246-247: the statements "strong accordance" and "noticeable differences" seem contradictory here

>> "noticeable" has been changed with "discernible".

L252: "the degree of agreement varies from year-to-year" - are you referring to the training or test period here? as the agreement looks very consistent year-to-year over the training period

>> We are referring to the test period. The sentence has been modified: “As anticipated, the ML algorithm closely aligns with TOPAZ4-RA during the training period, although the degree of agreement varies from year to year during the test period, supporting the assumption that the latter is largely independent of the training period.”

Figure 4-7: as I mention previously, I think it would be very insightful to show maps and time series of PIOMAS, to understand what spatio-temporal differences people would expect to get if they adopted your product for SIT evaluation over PIOMAS in the future.

>> This is a remarkably good suggestion. PIOMAS has been added in Figure 7 and Figure S3 (S6 in the first submission).

Comments on PIOMAS agreement with TOPAZ4-ML have been added in the description of Fig. 7. Line 346 “Additionally, while PIOMAS SIT appears to be lower than ICESat-1 and Envisat along the coasts of Siberian and Alaska, it is generally consistent with satellite observations in the Central Arctic along 80 N.”.

Additionally, Figure 6 has been reproduced with PIOMAS (Fig. R4, shown in this document) and a time series from 1992 to 2022 has been produced (added to the Supplements Fig. S9) displaying both PIOMAS and TOPAZ4-ML.

The following sentence commenting Supplements Fig. S9 has been added line 356:

“Compared to TOPAZ4-ML, PIOMAS indicates an earlier onset of the melting period while exhibiting similar average of SIT throughout the time series, except for the period after 2020 (Supplements Fig. S9).”

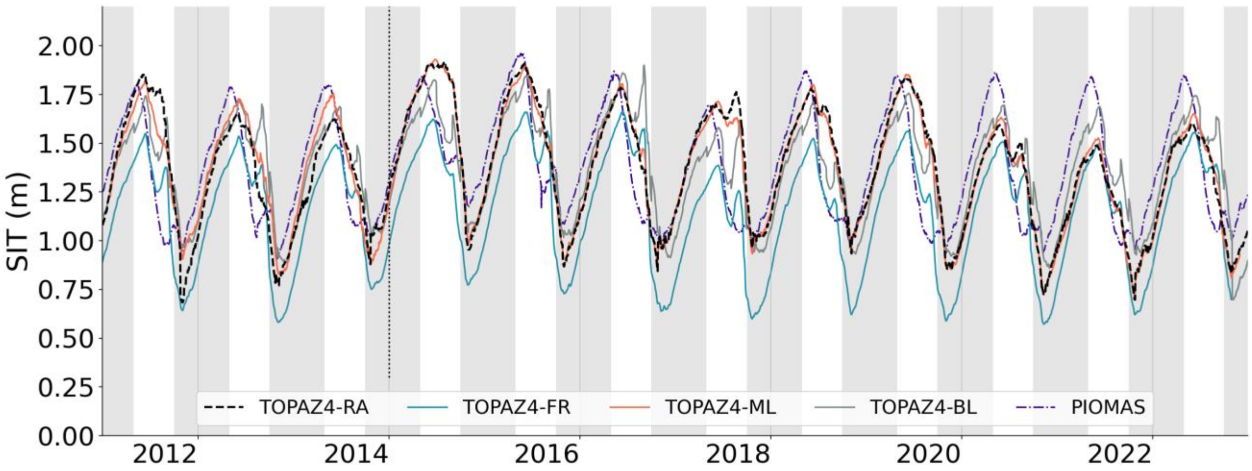


Figure R4: Same as Figure 6, with the addition of PIOMAS.

Figure 6. I think this figure would be nice split into 6a and 6b, where 6b shows RMSE.

>> Thank you for your thoughtful suggestion. While we recognize its potential value, we have decided not to incorporate it into the current version of the manuscript.

L254-259: if I understand correctly, this feature in the time series comes from the fact that, on each day, you compute average SIT only over grid cells where sea ice present? As you would expect this feature to go away if you include thickness=0 grid points? Can you briefly give your motivation for averaging in this way?

>> Indeed, the average sea ice thickness including grid cell = 0 do not show this feature. Yet they are overall smoother, and the mean quantity is not representative of the quantity of sea ice. Because of this, we were motivated to average with only sea ice thickness > 0.

L276-277: the statement that the ML approach is in better agreement with in situ data than the baseline or free-run seems contradictory with Table 3, which suggests that the baseline is better at each BGEP mooring and across all seasons. Meanwhile the ML is best at NPEO. It would be interesting to add CS2SMOS to Table 3 (understandably computed over a different time window), to understand what is the lowest bias we could reasonably expect to achieve if the ML model perfectly predicted the SIT increments.

>> This statement is indeed contradictory with Table 3.

The whole section has been modified: see section 4.3.1. paragraph 1, 2, 4. (Lines ~300-335). This also led to adding one sentence in the abstract "In contrast, when applied in the Beaufort Gyre, our method fails to exceed a trivial correction.".

And to change one sentence in the conclusion: "Applying our algorithm before 2011, the evaluation with independent mooring data indicates improvement in the central Arctic compared to TOPAZ4 free run, while the Beaufort Gyre results show a slight decline in performance."

Moreover, a paragraph in section 4.3. has been added to underline the difficulties of such a validation exercise and to highlight the differences between observation datasets.

Figure 7: it's a little hard to know the take-home message of this figure. Is the hope that TOPAZ4-ML looks like ICESat-1, or Envisat? or neither?

>> The hope is that TOPAZ4-ML exhibits similar patterns to ICESat-1 or Envisat, which were not perceived in TOPAZ4-FR or TOPAZ4-BL.

The first sentence of the paragraph has been changed to express the take-home message more clearly: "A qualitative comparison between remote sensing data and TOPAZ4-ML (Fig. 7) exhibits a close agreement of SIT and spatial distribution patterns, indicating that our reconstruction effectively behaves as a coherent correction when applied in the past."

L383: do all of the TOPAZ4 runs use 100 member ensembles? including the ML implementation? If so, were all results based on ensemble means? Presumably with 100 members we could investigate the internal variability aspect in more detail to understand regime shifts.

>> Yes, all the TOPAZ4 runs use a 100-member ensemble during the reanalysis processing. However, due to the space limit, the 100-member ensembles before assimilation are stored twice a year for restarting to extend the reanalysis run. Besides them, every week, the ensemble mean and a few diagnostic files, such as the parameters reflecting optimization efficiency, the program log, and the observation thinning and optimizing. So, unfortunately, under the current files' condition, we cannot dig out the internal variability to explain the

regime shift due to the sparse time samples. But it will be a good reason to help us apply more disk space to keep all the ensemble states as much as possible.

References:

Landy et al. 2020. Sea ice roughness overlooked as a key source of uncertainty in CryoSat-2 ice freeboard retrievals. JGR Oceans

Tilling et al. 2018. Estimating Arctic sea ice thickness and volume using CryoSat-2 radar altimeter data. Advances in Space Research

Gregory et al. 2024. Machine learning for online sea ice bias correction within global ice-ocean simulations. GRL

Editor review

Comments:

1) Clarify method further with limitations, assumptions etc. For example I am not clear if your approach is affected by overfitting (i.e. by using 8 PCA for something like 10+ variables). I am also unclear about the stationarity assumption when we know 2010s are so different in behaviour than 1990s.

2) I am not totally clear about the validation against BGEP as to your accounting for representation issues. Please at least discuss this as it is a very common mistake to ignore this and draw incorrect statements. My recommendation is to not just assume a constant conversion between draft and thickness when comparing to BGEP but use a more appropriate snow and ice densities as a function of months. See approach by Nab et al. (2024) in paper cited. Please cite Nab et al (2024, accepted in GRL) who is doing an in depth analysis of satellite vs BGEP comparison:

Nab, Carmen Julia, Robbie Mallett, Connor Nelson, Julienne Christine Stroeve, and Michel Tsamados. "Optimising interannual sea ice thickness variability retrieved from CryoSat-2." Authorea Preprints (2024).

3) Please compare your new model reanalysis against other recent satellite derived products that have extended their findings back into the early 1990s such as Bocquet et al (2023) and Soriot et al (2024):

Bocquet, Marion, Sara Fleury, Fanny Piras, Eero Rinne, Heidi Sallila, Florent Garnier, and Frédérique Rémy. "Arctic sea ice radar freeboard retrieval from the European Remote-Sensing Satellite (ERS-2) using altimetry: Toward sea ice thickness observation from 1995 to 2021." The Cryosphere 17, no. 7 (2023): 3013-3039.

Soriot, Clement, Martin Vancoppenolle, Catherine Prigent, Carlos Jimenez, and Frédéric Frappart. "Winter arctic sea ice volume decline: uncertainties reduced using passive microwave-based sea ice thickness." Scientific Reports 14, no. 1 (2024): 21000.

Here we present a point-by-point response to the editor's comments and suggestions, with the original comments from reviewers in black font, and our replies in blue font.

1) Clarify method further with limitations, assumptions etc.

>> Those points are addressed in the section 5 (discussion).

A high number of input features for one output (for each model of the eight model) does not automatically mean “overfitting” as the training is interrupted before overfitting happens.

To prevent over-fitting, our method includes dropout layers in the architecture of our ML model. We also did numerous trainings with different parameters (among others: the number of input features and the number of epochs). Overall, reducing the number of input features and epochs led the ML algorithm to be unable to correctly predict the PC over the training period.

The Principal Components predictions can be observed in Fig. A1 and the comparison to excluded data does not show sign of overfitting.

Stationarity: The behaviour of sea ice is indeed evolving dramatically in absolute terms, but not relative to our input data, which includes variables such as the ice age, also a marker of the regime shift. So, the assumption of stationarity-relative-to-all-input data becomes more reasonable. Indeed, the comparison to independent data ultimately confirms that this stationarity is justified over the period 1992-2022.

The stationarity assumption is a necessary assumption of this method. To the best of our knowledge, it is impossible to verify, we still provide insight into the validity of this assumption with two different ways.

1) the EOF decomposition does not change significantly when we remove the earlier years of the testing period (2011-2013). While it cannot reveal much for the period prior to 2007, it shows that our assumption is valid for the special year of 2012.

2) We show that our method can adjust the thickness of thick sea ice (>6 meters) (Fig. C1) and thus should have a good performance even in earlier period dominated by multi-year ice (before 2007).

2) I am not totally clear about the validation against BGEP as to your accounting for representation issues. Please at least discuss this as it is a very common mistake to ignore this and draw incorrect statements.

>> We thank the editor to bring to our attention the possibility to convert the BGEP ULS sea ice draft to thickness with the use of appropriate snow and ice densities as a function of months.

Although this approach has been used by Nab et al. (2024), no constants are available for monthly conversion. As it is out of the scope of this study to investigate snow and ice densities (which include downloading 2 datasets from Aaboe et al. (2021) and Liston et al. (2020)), we chose to keep the BGEP ULS conversion as it was in the previous version of the manuscript. Additionally, the validation against BGEP ULS SIT is not the main element of this study, and we do not expect the conversion factor to have a insignificant impact on the conclusions of our work.

One sentence has been added about this possibility, and we cite the study by Nab et al. (2024) in section 2.5 “Validation data: Mooring data”:

“A more precise conversion from sea ice draft to thickness is possible by using appropriate snow and ice densities (Nab et al. 2024).”

3) Please compare your new model reanalysis against other recent satellite derived products that have extended their findings back into the early 1990s such as Bocquet et al (2023) and Soriot et al (2024)

>> We thank you for this suggestion as it will greatly improve the manuscript and the relevance of our dataset in the sea ice community.

We added a subsection entitled “Comparison with other datasets” (4.3.2.) with a figure showing the monthly mean of sea ice thickness in March and October and associated trends for the 4 following datasets: TOPAZ4-ML, PIOMAS, Bocquet et al (2023) and Soriot et al. (2024).

At the introduction of the section, we state clearly that we do not aim to determine which dataset is the most accurate, but only to present differences appropriately.

Different trends are discussed as function of the season and the type of datasets (model-based or observation-based).