August 31, 2026

The document is color coded in the following way:
Referee comments are in ***blue****.*
Our answers are in ***black****.*

Response to Referee #1 Salar Karam

We thank Referee #1 (Salar Karam) for the time and effort he spent reading the manuscript and providing constructive comments. We also appreciate the enormous number of potential suggestions for improvements. Please find our answers to the comments and descriptions of how we have addressed them below.

General comment:

There is great interest around the world to monitor the evolution of sea ice and mixed layer temperatures in the Arctic Ocean. Throughout the satellite observational record, we have seen a consistent decline in sea ice extent in the Arctic, with far-reaching implications for the global climate system. There is a range of models, reanalyses, and satellite products to evaluate these trends, however, critical evaluations of these products are often lacking. The submitted manuscript presents data from two regional models, two global climate models, and three reanalyses, and evaluates them against satellite measurements. The figures are generally clear, the metrics used to assess sea ice trends are appropriate, and the topic is within the scope of the journal.

However, the manuscript suffers from a lack of clearly defined research questions and objectives, which affects its focus and analytical depth. The abstract and introduction do not state specific aims, the justification for selecting datasets is incomplete, and the discussion summarises the results without connecting the findings to well-defined objectives. Mechanistic explanations for the observed biases are currently lacking, which makes it difficult to interpret the results and assess their significance. While this could be acceptable if the study's primary purpose is a very broad intercomparison, the manuscript could still provide concrete guidance—such as performance metrics or recommendations—that would help the community identify which models are most appropriate for particular scientific questions and applications in Arctic sea ice and mixed layer research.

This lack of clear direction makes it difficult to interpret the results, assess the significance of the findings, or understand how the results could inform the suitability of different models for specific applications in Arctic sea ice and mixed layer research. If the goal is to assess model realism, the manuscript could provide concrete guidance—such as performance metrics or recommendations—that would help the community identify which models are most appropriate for particular scientific questions.

Additionally, the manuscript reads somewhat as a first draft, with recurring scientific and formatting issues, including inconsistent acronym usage, misplaced citations, and inclusion of DOIs and URLs in the main text. While each of these issues is minor individually, their repetition throughout the manuscript negatively impacts readability and even the overall professionalism of the work.

Given these issues, I recommend rejection in its current form. However, with careful revision it has potential to make a meaningful contribution. The authors should explicitly define research questions,

justify dataset selection, restructure the discussion to address those questions with mechanistic interpretation, and address the recurring scientific writing and formatting issues. Addressing these points would strengthen the manuscript and make it suitable for resubmission.

The manuscript has been carefully revised and restructured based on the reviewer's comments, and we hope all formatting issues are now rectified. Research questions and objectives have been added to the introduction. The discussion is now part of a combined results and discussion section and the discussion and conclusions have been restructured to address the research questions. Additionally, the global climate models have been removed from the manuscript to better align with the special issue it was submitted to.

Major comments:

Abstract

The abstract provides a clear overview of the study and concisely summarises the main results, but it does not fully convey the purpose, nor the significance of the work. While the opening sentences frame the general context and the importance to understand model and reanalysis performance in the Arctic, there is no clear research question stated, only "better understand the strengths and limitations [...]", which is entirely too generic for a research question. Furthermore, although the results are summarised, there is no evaluation or discussion of these strengths and limitations, or mechanisms behind the observed trends. For example, the abstract reports that of the models, only HIRHAM-NAOSIM shows a significant sea ice decline, but it does not explain why these differences occur, or what they imply for model performance.

To improve clarity and focus, I recommend that the authors: (i) state specific research objectives, e.g., assessing the ability of models and reanalyses to reproduce sea ice variability and identifying underlying causes of potential biases. (ii) summarise the main results in the context of these objectives, i.e. not just stating the results, but what they reveal about the models' strengths and weaknesses. (iii) include a brief statement on the broader impact of their results. That is to say, how these findings inform understanding of e.g., sea ice variability, model development, or future projections.

The abstract has been completely revised following the guidelines provided by the reviewer for the manuscript.

Research questions have been added to the introduction, as stated below in the answer to the introduction comments.

Introduction

The introduction provides nice context on the Barents and Kara Seas, but several aspects could be improved for clarity and focus: (i) The detailed geographic description of the region (L. 30-42) is somewhat excessive. I would recommend condensing it to highlight only the features most relevant to modelling and sea ice variability. As it stands, it distracts from getting into the topic and motivation of the paper. (ii) The concept of "storylines" (Levine et al., 2024) is introduced but not really developed in the study. Unless this concept is explicitly connected to the results, I would recommend removing it. (iii) The study's aim is not stated clearly (L. 64-66); phrases such as "investigate the capability [...]" and "understanding the performance [...]" are too generic, and this is felt throughout the manuscript, which feels somewhat lacking a clear direction. In particular, it is not clear whether the focus is on long-term trends, interannual variability, seasonal variability, or mean state biases. As it stands, the manuscript attempts to address it all, but without a clearly defined focus, the analysis remains broad and does not explore any one aspect in depth. I recommend stating specific research questions (e.g., why do models/reanalyses fail to capture Barents-Kara sea ice variability/trends), and clearly defining the scope of the analysis. The authors should also clarify the expected contribution of the study (e.g., identifying underlying mechanisms driving biases, and informing improvements for specific model components that control sea ice and mixed layer properties in the Barents-Kara region).

Thank you for helping make our introduction better.

(i) The geographic description has been condensed

(ii) The concept of storylines and the global models have been removed to move our focus closer to the objective of the special issue.

(iii) We have made the research questions, aims, and objectives clearer in the introduction of the manuscript

Data and Methods

The methodology lacks a clear explanation for why these particular models and reanalyses were chosen. The only justification concerns the two CMIP6 models, where it is stated: "These models were chosen following the analysis described in (Levine et al., 2024), where the future simulation of CNRM represents a warm humid Arctic Ocean and cold dry continents, while NorESM represents a cold dry Arctic Ocean and warm humid continents, as described in (Mottram et. al (submitted))." However, it is not entirely clear why these models are appropriate for evaluating historical sea ice trends, since the stated motivation – that they represent extreme future Kara/Barents warming scenarios – relates to projections rather than the historical period analysed. The discussion reiterates the "warm/cold Arctic storyline" without linking it to past sea ice variability, which may be confusing to readers.

We have dropped the global CMIP6 models. The global models were originally added to have an additional category of models in the research to get more depth, using models evaluated by partners in the EU project PolarRES on both historical and future performance. The thought was that taking models from the two storylines would result in catching some of the spread in the whole CMIP6 ensemble.

We realise these CMIP6 models do not add any additional answers specifically for our research questions and the special issue, and we have dropped them to better focus on reanalyses and regional models. Removing the models with the coarsest resolution has also let us add more coastal grid cells to our area-averaged analyses.

Even more importantly, no justification is provided at all for the selection of the regional models or reanalyses. It would be helpful to understand whether these datasets were chosen based on prior evaluation, spatial or temporal resolution, regional performance, or indeed any criteria at all. Without a clear explanation for all datasets, I think it is difficult to assess the robustness of the comparisons or whether the results might simply reflect arbitrary choices rather than meaningful scientific patterns. I recommend that the authors provide criteria for the selection of all datasets, and either justify why the future-oriented storylines are relevant for assessing historical trends or remove that discussion from the manuscript.

Specific justification for the selection of models and reanalyses was actually missing in the previous version of the manuscript. We have added this justification at the start of Section 2.

When choosing the regional models and reanalyses, the aim was to choose models and reanalyses that use different model components to get a variety of models, while still keeping the amount of products relatively small. For the regional coupled models, the availability of model data within the PolarRES project also played an important role.

OSI-SAF is the observational reference for sea ice, and the global reanalysis ORAS5 is a global benchmark that is generally known to perform well (e.g. Zuo et al., 2019), and it also provides ocean boundary conditions for HIRHAM-NAOSIM, making it a good addition as a reference, as observational data from the ocean are sparse.

Meanwhile, the marine reanalyses are evaluated less than some others, but as they are used by the community, their evaluation and reporting are valuable. Topaz4b is a regional analysis assimilating SIC data, while RARE does not assimilate SIC data and instead constrains the sea ice with PIOMAS SIT. This difference in assimilation gives us variety between the two regional reanalyses chosen.

HIRHAM-NAOSIM and RASM are two regional climate models which were involved in the PolarRES project from which the present study has been funded. The HIRHAM-NAOSIM model runs were specifically made as part of the PolarRES project. The two model systems have different model components, giving us a variety of models. Both models have been evaluated on the Arctic Ocean before.

The coupled regional models take air-sea interactions into account, unlike the marine reanalyses. On the other hand, the marine reanalyses assimilate observations, so differences between the coupled regional models and the reanalyses might reflect the effects of data assimilation. Meanwhile, as the models are quite different in terms of their model components and configurations, their differences reflect the spread among models.

Discussion

The discussion is detailed and descriptive, but mainly just summarizes the results again. The authors identify relevant differences but do not sufficiently explore the causes of these differences or their implications, so the discussion ends up shallow and offers few novel insights. Another issue is a structural one, the authors do not sufficiently tie together the discussion of their sea ice analysis to their analysis of mixed layer properties. Rather it reads like two separate parts, with no connection between them.

We have restructured the manuscript so that there is now a results and discussion section and a conclusion section instead of a results section followed by a discussion and conclusion section. This is to more easily discuss the similarities and differences of the models. We have also added additional results to support a less descriptive approach. Firstly, we have added statistical tests throughout the results displayed in the tables to quantify whether differences between datasets are statistically significant or not. Secondly, we have added a new section to the results where we explore the correlation of sea ice and ML temperature with five different climate indices. This will help us explore drivers of the interannual variability, and how their relationships vary between the renanalyses and the models. The indices are North Atlantic Oscillation (NAO), Atlantic Multidecal Oscillation (AMO), Pacific Decadal Oscillation (PDO), the East Atlantic/Western Russia teleconnection (EA/WR) and the Northern Hemisphere Ocean Heat Content (NH OHC). The additional analysis also helps us tie together the sea ice and mixed layer temperature.

The new conclusion section now contains concrete suggestions for dataset choice and further work into improving the reanalyses and models.

On the discussion of reanalyses and sea ice representation (L. 415-427), the manuscript builds up to the conclusion that data assimilation improves sea ice representation, but that there are still deficiencies and that "they can not be used for everything". While technically correct, this is not a particularly novel insight and does not substantially advance understanding. Some opportunities for deeper analysis could perhaps include: (i) exploring why TOPAZ and ORAS – both of which assimilate satellite data – exhibit opposite biases in sea ice concentration (overestimation vs underestimation). This suggests that factors beyond assimilation, such as model physics or the assimilation scheme, play an important role, but this is not explored at all, and (ii) why RARE, despite not assimilating satellite data, is able to capture the negative trend, while largely failing to reproduce the mean state. More importantly the paragraph should explicitly connect the sea ice biases to the mixed layer properties discussed later. At present, sea ice variability is discussed in isolation, without relating it to the ocean state, and therefore lacks a mechanistic explanation for the identified biases. I think that establishing this connection is necessary to move beyond a descriptive comparison.

The addition described above helps us identify some underlying causes of inter-product differences. Using relationships between the SIC and MLT to the climate indexes help us find mechanistic explanations for the identified biases. Going into technical differences in model code is however outside the scope of this paper.

(i) We have explored this and added more statistical tests on differences, for example, for the mean values in tables; most of the differences in ORAS and TOPAZ are small and do not differ significantly from OSI–SAF. The TOPAZ 4b ice edge "tends to recede too fast in early summer and re-freeze too fast in early winter" according to its synthesis quality manual (Laurent Bertino and Jiping Xie, 2025), which is probably the reason for the seen differences.

(ii) RARE assimilates ocean data and is forced at the surface by the ERA5 reanalysis and steered toward PIOMAS by surface fluxes. Either of these might be the reason that a trend is seen.

The discussion of regional models and sea ice representation (L. 428-448), is largely descriptive and offers no additional insights from the results. It does not meet the analytical depth expected in a discussion. Statements such as "This might have to do with the differences between the basins [...]" are too general and vague, and are above all not supported by any analysis. The authors are expected to substantiate such claims or formulate concrete, process-based interpretations of their results. Additionally, the second paragraph (starting at L. 441) reads as general background and is not explicitly tied to the results of this study. As it currently stands, this paragraph is more appropriate in the introduction than the discussion. In the discussion, these mechanisms should be used to interpret the authors own findings. Furthermore, there is still no connection in these paragraphs relating sea ice biases to the ocean state or atmospheric variability or indeed any other metric, and therefore no analysis is provided for why e.g. RASM has the

smallest IIEE, or why HIRHAM-NAOSIM performs better for sea ice trends in the Barents Sea.

This paragraph has been rewritten and moved to the introduction.

Starting at L. 449, the discussion on the mixed layer temperature is mainly descriptive and restates results, rather than providing interpretations. High correlations between mixed layer temperature and sea ice area are noted, along with differences between datasets and regions, but these differences are not explored further. For example, the authors note that RARE shows higher correlations than the other reanalyses, but that discussion is immediately stopped with "more research would be needed [...]". Similarly, differences in the magnitude and spatial pattern of ML temperature trends between reanalyses are reported, but not investigated further. There is also a lack of mechanistic interpretation, for example, how ML temperature anomalies might influence sea ice melt or inhibit ice formation. The discussion could perhaps be strengthened by investigating the differences between the reanalysis products and analysing why RARE shows higher correlations, exploring the spatial patterns in ML temperature anomalies, as well as explicitly linking ML temperatures to observed sea ice variability more than reporting on high correlations.

The RARE issue has been analysed further in Sections 4.4 and 4.5 The anomalous ML temperatures coincide with regions of Atlantic water inflow.

The section on CMIP6 models and Barents-Kara sea ice (L. 461-477), adds some useful context, e.g. linking to Yamagami et al. (2022) and highlighting limitations due to model domain coverage and resolution. These insights are mostly descriptive though and I think again some mechanistic interpretation is missing, essentially, how do factors such as North Atlantic warming dynamically and quantitatively influence the observed mixed layer temperature anomalies and sea ice anomalies in this study? Another point to expand the discussion could be that HIRHAM-NAOSIM did not cover the Gulf Stream region, yet it showed better performances across most metrics compared to RASM. Could this mean that explicit representation of the Gulf Stream is not important? Or that other factors such as local ocean state variability or atmospheric variability are more important? Finally, references to the "storyline analysis" from Levine et al. (2024) are underdeveloped and do not really contribute to the interpretation of results. Storylines are mentioned briefly in the introduction, methods, and discussion, but are never analysed in relation to observed historical trends, sea ice biases, or ML anomalies. This concept, as presented, adds little analytical value and I think the text would be clearer if removed.

Thank you, these are important points. We have addressed the question related to the roles of oceanic and atmospheric factors by correlation analysis of climate indices and sea ice and mixed layer temperature anomalies. Please see details in our earlier response. It might well be that the different lateral boundary conditions in the Atlantic result in the better performance in HIRHAM-NAOSIM compared to RASM.

The final paragraph reiterates known challenges regarding Barents-Kara representation and model resolution but does not provide new insight or specific recommendations. Statements such as "further model development is required" are too generic and should be linked to concrete processes or mechanisms identified in this study. While noting that resolution alone does not solve the biases is useful, the manuscript does not analyze why higher resolution fails to address the discrepancies, e.g. are there specific ocean processes that are not represented well? Sea ice-ocean coupling? Vertical mixing? Statements such as "further model development is required" are too vague, especially after 24 pages of detailed analysis. The conclusions would benefit from more specific suggestions, for example: (i) which processes require improved representation (e.g., Atlantic water inflow, mixed layer heat exchange, ice-edge dynamics)? (ii) Are there model parameterizations, coupling approaches, or observational constraints that could be targeted for improvement? (iii) Could regional modeling experiments or sensitivity studies help disentangle the drivers of bias? Making these points explicit would strengthen the conclusions and provide some actionable steps for future work.

The rewritten conclusions give concrete suggestions on further research into how to develop the models.

Scientific writing

This manuscript contains a number of recurring style and formatting issues that affect readability. Acronyms are often introduced after their abbreviated forms, DOIs and direct URLs are included in the main text rather than in the reference list, and citations are frequently placed mid-sentence in ways

that disrupt flow. Table captions are consistently located below the tables instead of above. While each of these issues is minor individually, their repetition throughout the entire manuscript makes the text difficult to read and follow. The cumulative effect of these errors reduces the clarity and even the professionalism of the manuscript. I recommend a careful and thorough editorial pass to correct these issues and ensure consistent adherence to standard scientific writing conventions. Since these many of these errors are occurring throughout the manuscript, I will only list a few examples:

Acronyms are often introduced before their full names (e.g., L. 103: "HYCOM (Hybrid Coordinate Ocean Model) [...]"), which is inconsistent with conventional scientific writing. Standard practice is to present the full name first, followed by the acronym in parentheses (e.g., "Hybrid Coordinate Ocean Model (HYCOM) [...]"). This occurs in Sec. 2.1, 2.2, among others.

DOIs should not appear directly in the main text. This happens at e.g., L. 102. Similarly, direct URLs (e.g., L.119: "Arctic Great Rivers Observatory, https://arcticgreatrivers.org/") should not appear directly in the main text. All DOIs and URLs should be moved to the reference list or the data availability section. In-text citations should only include author(s) and year to maintain readability and a consistent citation style.

Citations often appear immediately after acronyms or mid-sentence, which interrupts the flow of the text (e.g., L. 91: "We use two CMIP6 (Eyring et al., 2016) Earth system models, [. . .]"). A clearer style is to place citations at the end of the sentence or clause describing the relevant data or concept (e.g., "We use two CMIP6 Earth system models (Eyring et al., 2016), [. . .]"). This ensures the text is more readable and not interrupted mid-sentence.

In some cases, text citations are incorrectly used instead of parenthetical citations (e.g., L. 84: "RASM Cassano et al., 2017 is a fully coupled atmosphere–land–ocean–sea ice model.") and should be rewritten as "RASM is a fully coupled atmosphere–land–ocean–sea ice model (Cassano et al., 2017)." Conversely, parenthetical citations are sometimes used where text citations are appropriate (e.g., L. 92: "These models were chosen following the analysis described in (Levine et al., 2024) [. . .]"), which should be rewritten as "These models were chosen following the analysis described in Levine et al., 2024 [. . .]".

The manuscript shows inconsistent use of parentheses in two contexts: (i) Acronyms with references: Examples such as L. 103: "HYCOM (Hybrid Coordinate Ocean Model) (Chassignet et al., 2003)" use separate parentheses for the acronym and the reference. For clarity, these should be combined in a single set of parentheses, separated by a semicolon: "Hybrid Coordinate Ocean Model version 2.2 (HYCOM; Chassignet et al., 2003)." (ii) Multiple references within text: At times, references are incorrectly formatted, e.g., L. 43: "(e.g., Li et al. (2022) and Dörr et al. (2024), among others)" Standard formatting combines such references into a single set of parentheses, separated by semicolons: "(e.g., Li et al., 2022; Dörr et al., 2024)"

We thank the reviewer for pointing out these issues in the text. The errors have now been fixed throughout the manuscript, especially in Sec. 2, where the main part of the issues were located.

Minor comments:

The title could be improved for clarity. The phrase "models versus reanalyses" suggests a competition, when in reality observations, reanalyses, and models, are all complementary tools. A more neutral and I think scientifically appropriate phrasing (e.g., "evaluation" or "comparison") would better reflect the study's aim. Beyond this minimal change, the title could be made more informative by indicating what is being evaluated (e.g., long-term trends, interannual variability, seasonal variability, mean state biases), and potentially why it matters (e.g., performance, processes).

We have improved the title to "Evaluation of sea ice variability in three ocean reanalyses and two regional models in the Barents and Kara Seas in 1991–2020".

Throughout the text, standardise "Barents-Kara Seas" vs "Barents and Kara Seas". Really either is fine, just be consistent, however, considering that the authors explicitly distinguish between the two

regions, I would write “Barents and Kara Seas”.

The Barents-Kara Seas were combined when referring to other studies where the seas are combined. This has now been changed to Barents and Kara Seas in all instances.

The manuscript uses season names inconsistently. At times “spring/MAM” and “autumn/ASO” are used, elsewhere just “spring”/“autumn,” elsewhere just “MAM”/”ASO”, and sometimes even “winter”/“summer.” Considering the already large number of acronyms in the manuscript, I recommend defining the seasons once in the Methods (e.g., spring = MAM, autumn = ASO) and then consistently using the terms “spring” and “autumn” throughout. This would improve readability, especially when referring to sea ice maxima and minima.

We have changed MAM and ASO to spring and autumn in the revised manuscript.

I recommend using Oxford commas throughout the manuscript where appropriate to improve clarity, for example in L. 18 and 44 to name a few.

We have considered this suggestion in the revised manuscript.

Wherever appropriate, I recommend adding a comma after parenthetical references when they occur at the end of a clause, to clearly signal the clause boundary. L. 24 would be changed to: “These observed changes characterise the region as a warming hotspot (Lind et al., 2018), and as [. . .]”.

This has been done.

E.g., L. 20, 44: I think the term “ocean basins” is not appropriate for the Barents and Kara Seas, which are two shallow shelf seas. Perhaps replace it with “shelf seas” or “marginal seas”.

We have now used either "marginal seas" or "regions" instead of the ocean basins

41: The phrase “disproportionately large influence” is somewhat unclear. Disproportionately large compared to what? I would consider removing the word "disproportionately" or explicitly state what this is being compared to.

The "disproportionately large influence" is a word-for-word quotation from Liu et al. (2025). We have clarified this in the revision.

50: European observations in the high Arctic go all the way back to Nansen’s expedition 130 years ago, so the phrase “Nowadays, observations [...]” is inaccurate. Additionally, the sentence is somewhat circular and redundant, as observations, reanalyses, and models, are the standard tools for studying the Arctic, and indeed the rest of the World Ocean. Consider rewriting to highlight that they are complementary tools to understand recent changes in the Arctic.

This section has been revised.

55: CMIP6 has already been introduced as an acronym earlier in the text, and therefore there is no need to reintroduce it.

The CMIP6 models has been completely removed in the revised version.

68, L. 75: Consider swapping the order of “coupled” and “regional” to emphasise that these are regional models that have coupled components.

Coupled regional models is the more precise term, since they consist of two (or more) regional climate subsystem models which were combined to a coupled climate model system. Regional coupled models would imply that they are regional versions of coupled (global) climate model systems.

153: σ typically refers to density anomaly, not density, which is denoted by ρ. Add the word “anomaly” after density.

Added.

153: Considering that we might expect model properties to be biased, are the results sensitive to the choice of density criterion? Do the results qualitatively change with other criteria?

We could expect differences in the MLD if other criteria are used. However, the impact of model biases (particularly those related to salinity) is minimised by using a fixed density-anomaly threshold rather than a variable threshold. While the exact MLD values may differ if we used another criterion, we do not expect the qualitative conclusions of the study to change. The density criterion adopted here ($\Delta\sigma = 0.03$ kg m^{-3}) is widely used in mixed-layer-depth studies and follows the methodology of Sallée et al. (2006, 2013).

155: It is unclear whether "The monthly mean temperature for each grid cell is calculated for this depth" is taken at the mixed layer base only, or as an integrated value from the surface down to the depth of the mixed layer base. The authors should clarify this, as it significantly affects interpretation of the mixed layer temperature. The correct way would be to integrate from the surface down to the base of the mixed layer.

This has been integrated from the surface down to the depth of the mixed layer that was calculated with the density difference criterion. This has been made clearer in the text.

158-165: The manuscript uses a significance threshold of $p < 0.1$ for trends, but a threshold of $p < 0.05$ for differences between correlations. Consider clarifying why different thresholds are used, and whether the choice of $p < 0.1$ for trends affects interpretation.

$p < 0.1$ was used for the trends to be more permissive, especially with the trend map hashing. We have made tests using a threshold of 0.05 as well, and the differences were small. The only place where $p < 0.05$ would change the results is the significance of the SIA trend in Table 2, where the p value of the HIRHAM-NAOSIM SIA trend is between 0.05 and 0.1 for the Barents Sea. This has been taken into account when interpreting the results.

For the correlations, the more common p-value of 0.05 was used.

We consistently use a p-value of 0.05 in the revised manuscript. For the maps using hashing for p-values, we use a different hashing (....) for p-values between 0.05 and 0.1.

191: The term "temporal pattern" is somewhat vague. I would specify whether this refers to inter-annual variability, seasonal timing, or long-term trends, as the paragraph also discusses mean values and decadal trends.

This has been clarified.

Fig. 1: I would add bathymetry to Figure 1. This would help illustrate key features, reducing the need for a lengthy description in the introduction.

The figure has been changed accordingly.

Fig. 2, 7: Figures 2 and 7 show the full seasonal cycle, but the manuscript mainly focuses on spring and autumn, and other seasons are not really discussed in detail. Consider whether these figures could be simplified and replaced with a simple time series showing only spring and autumn values, which might make the results clearer and more directly relevant to the study's focus. Alternatively, expanding the manuscript to consider the full seasonality.

We thank the reviewer for this suggestion. We have kept the figure, since the old and new correlation results in the study use the whole year, but we provide a supplementary figure, showing an annual anomaly time series, to better display year-to-year variations.

Additionally, the ticklabels for the years are not aligned with the start of the grid cells.

The tick marks are aligned to the middle of the month/year that they show. This was on purpose and we think that is clearer. This has not been changed.

Adding titles above the left and right panels showing "Barents Sea" and "Kara Sea" would improve readability and help orient the reader.

The figure has been changed accordingly.

The figures could be simplified by showing only one colorbar instead of two, which would reduce clutter.

The figure has been changed accordingly.

Finally, the figure could be even more simplified by removing the ticklabels showing years for the Kara Sea panels.

We feel that the thick labels improve the readability of the rightmost panels.

Fig. 3: Each subplot uses the same legend, so only one is necessary. I suggest keeping the legend in the top-left subplot and removing it from the other subplots to reduce redundancy. Additionally, the x- and y-axes should be the same for the Barents Sea and Kara Sea panels to make it easier to compare between regions.

The figure has been changed accordingly. It is now provided as supplementary figure.

Table 5. The correlations are calculated for summer (which you define as Apr–Sep) and winter (which you define as Oct–Mar), which include spring (MAM) and autumn (ASO) months that the rest of the manuscript focuses on. This shift in seasonal definition is somewhat confusing and not really explained in the manuscript. I recommend either aligning the correlation periods with the spring and autumn focus of the manuscript, or alternatively clarifying why a different seasonal grouping is used here.

The seasonal correlations have been removed as a part of refocusing the manuscript.

Response to Referee #2

We thank Referee #2 for the time and effort he/she spent reading the manuscript and providing constructive comments. Please find our answers to the comments and our plans for addressing them below.

Main points

Terms like clearly and somewhat are used too much and too loosely. They make the text more subjective and less quantitative.

We have decreased the use of these words in the revised version.

The OSI-SAF data is always the last value in the table or the last plot in the figure, but it is the first mentioned in the text. Rearrange the figures and tables so it comes first.

We have rearranged the tables and figures, and OSI SAF is now at the top. In consequence of the comments from referee #1, we have decided to remove the global models from the analysis.

Minor points

Line 10: Why defined like this "We focus on spring (March-April-May) and autumn (August-September-October) sea ice area"?

Sea ice in the Barents Sea usually reaches its maximum in April and its minimum in September `https://mosj.no/en/indikator/climate/ocean/sea-ice-extent-in-the-barents-sea-and-fram-strait/`. The seasons were defined around the minimum and maximum.

Line 127: ML not defined until line 151.

ML is now defined where it is first introduced.

Line 129: Is April the maximum sea ice? Some years March, some years April.

April is usually the maximum in the Barents Sea. We have added "usually" to the sentence

Line 145-149: Are these the definitions of overestimate and underestimate?

Yes, this is correct. We have made this clearer in the text.

Line 159: Why is $p < 0.1$ chosen? It doesn't inspire confidence.

This was an early choice during the analysis process, so that the hashing in the trend maps would not be too restrictive. Changing it to 0.05 would make the statistically significant areas slightly smaller. We have changed all p-value thresholds to 0.05 in the tables, and added "...." hashing to all figures showing p-values for values between 0.05 and 0.1.

Line 170: Why -3 in equation.

The term $N - 3$ arises from Fisher's asymptotic result that the Fisher-transformed correlation coefficient

$$z = \operatorname{atanh}(r)$$

is approximately normally distributed with variance

$$\operatorname{Var}(z) \approx \frac{1}{N-3}.$$

Line 194: Make is more clear what the comparison is here: "...clearly overestimating" compared to the observations (OSI-SAF).

We have tried to make this clearer when we rewrote the results section.

Line 207-208: Rephrase. Both show positive and negative phases. I think what you mean is NorESM has high frequency and lower amplitude interannual variability, but it is not clear. If this is what you mean, state the standard deviation.

The CMIP6 models have been completely removed in the revised version.

Line 231-270: This section is too vague and unquantified. Too much use of phrases like "much smaller", "somewhat larger", "rather close to", "relatively close to".

The Sea ice edge section has been completely rewritten.

Line 297: The point positive is really small, the significance here must be marginal. Is it significant at $p > 0.05$?

Yes, it is significant at $p > 0.05$. In the updated Figure 5 (Figure 4 in the revised manuscript), where "...." hashing has been added for p-values between 0.1 and 0.05. Using 0.05, the statistically significant area is slightly smaller.

Line 325-326: These two sentences are contradictory. How is the inner ice pack defined, 80%?

Yes, it is defined as $> 80\%$. We have made this clearer in the text.

Line 446-460: While this is true it is difficult, there is mounting evidence that interannual variability comes from the ocean. Given you discuss the link of mixed layer and sea ice it seems important to have a more in-depth discussion of work that has looked at the topic.

This paragraph has been rewritten and moved into the introduction following the suggestion of Reviewer #1.

References

Laurent Bertino and Jiping Xie: SYNTHESIS QUALITY OVERVIEW DOCUMENT (SQO) Associated to extended quality information document (QUID): CMEMS-ARC-QUID-002-003 QUID Version: 1.5 Associated to Product ID: ARCTIC_MULTIYEAR_PHY_ 002_003, 2025.

Liu, Y., Losch, M., Tremblay, L. B., and Janout, M.: Landfast ice in the Kara Sea stabilizes the Arctic halocline and may slow down Atlantification of the Eurasian Basin, Communications Earth and Environment, 6, https://doi.org/10.1038/s43247-025-02360-8, 2025.

Sallée, J.-B., Wienders, N., Speer, K., and Morrow, R.: Formation of Subantarctic Mode Water in the southeastern Indian Ocean, Ocean Dynamics, 56, 525–542, https://doi.org/10.1007/s10236-005-0054-x, 2006.

Sallée, J.-B., Shuckburgh, E., Bruneau, N., Meijers, A. J. S., Bracegirdle, T. J., and Wang, Z.: Assessment of Southern Ocean mixed-layer depths in CMIP5 models: Historical bias and forcing response, Journal of Geophysical Research: Oceans, 118, 1845–1862, https://doi.org/10.1002/jgrc.20157, 2013.

Zuo, H., Balmaseda, M. A., Tietsche, S., Mogensen, K., and Mayer, M.: The ECMWF operational ensemble reanalysis–analysis system for ocean and sea ice: a description of the system and assessment, Ocean Science, 15, 779–808, https://doi.org/10.5194/os-15-779-2019, 2019.