RESPONSE TO REVIEWERS

We express our sincere gratitude to the reviewers for dedicating their time and thoughtful consideration, along with providing valuable feedback and suggestions for our manuscript.

In response to the comments received, we have incorporated text edits and also added our responses to the individual comments provided by the reviewers. Throughout this document, reviewer comments are indicated by plain blue text and our responses are presented in plain black text, addressing each comment individually. Additionally, the incorporated changes in the original manuscript are added through track-changes in the new submitted manuscript document.

The following major revisions have been made to the manuscript:

1. A **supplementary document** has been added, including two figures (Fig. S1 and S2) and one table (Table S1) to further support the manuscript and address reviewer comments.
2. A **new figure** (Fig. 9) has been incorporated into the manuscript to provide quantitative evidence for our conclusions.
3. The manuscript structure has been **revised for improved flow and clarity**. We expanded the explanations of methods (included a new section; Sect. 3.4), including nudged simulations, Monte Carlo methods, and proxy year selection approaches. Additionally, the discussion and conclusion sections were condensed to eliminate excess detail and reduce confusion.

Reviewer 1:

This study attempts to evaluate the accuracy of CMIP6 climate models in simulating Arctic sea-ice and snow thickness, comparing their results to year-long MOSAiC observations. It introduces the concept of "proxy years" based on sea-ice area and atmospheric conditions to align model data with the single-year MOSAiC observations (Snow Buoys/IMBs).

Thank you for your comments. We have responded to the comments individually.

I think the general concept of how we can use a single year of observations to evaluate climate models is worth exploring. But I struggled with the methodology and am not at all confident that the conclusions about model “biases” are valid. There are large scaling issues and significant natural climate variability at play that I don’t think have been appropriately accounted for in this comparison. The main objective is to “examine whether

discrepancies in the proxy-based ice and snow cycles arise from mismatching weather conditions or from insufficient process representations" but I am not convinced you have demonstrated that and I don't think we can use this analysis to really understand CMIP6 model biases.

We would kindly like to underline the primary objective of our study which is mentioned in **lines 146-148** "In this methodology-oriented study, we aim to address the above challenges by proposing a simple approach to perform meaningful comparisons of CMIP6 models with the field observations", also it is again clarified in **lines 621-625** "This study proposes a new proxy-year selection approach to perform meaningful comparisons of free-running CMIP6 models' sea ice and snow data with measurements relatively localized in time and space – here, using the unique MOSAIC time-period and observations. We propose this method in an exploratory study using a single ensemble member from the set of 10 selected CMIP6 models."

We nevertheless now introduce this point as early as in the abstract **(lines 50-53)** to improve clarity and differentiate our primary objective from that of the rationale of employing the nudged simulations **(lines 55-57)**, as well as in the main text (**lines 536-539**), as quoted below. Thank you for pointing this out.

Lines 536-539: "We utilize nudged simulations (see Sect. 3.4.1) to evaluate and compare our proxy-based methodology. By doing so, we assess whether discrepancies in the proxy-based sea ice and snow cycles stem from differences in weather conditions or from inadequate representation of key processes."

I think the nudged simulations offer the most enticing part of the current manuscript, so I would encourage the authors to focus more of the effort and write-up on that.

Thank you for acknowledging the use and importance of nudged simulations in our study. While nudging is certainly a preferable methodology, not every modeling center is able to perform these additional nudged simulations, and not all coupled climate models have nudging capabilities. This has also been kindly acknowledged by the second reviewer. However, our study does show that our proposed proxy methods, specifically the SIA-based method, offers a valid alternative to nudged simulation methods as model biases look similar when one uses nudging methodology or our proxy-selection methodology. As suggested, we now present the nudged simulation in more detail in the introduction:

Lines 156-172: "We verify the skill of the proxy-year selection criteria using two approaches. First, we employ a set of "nudged" simulations, where the atmospheric circulation is directly constrained to observations (e.g., Sánchez-Benítez et al., 2022; Athanase et al., 2024a, 2024b; Zhuo et al., preprint). These nudged simulations produce quite accurate analogues of the observed conditions during the MOSAiC campaign in the

climate models, as shown by Pithan et al. (2023). Comparing the annual cycles of sea ice and snow thickness obtained using the nudged simulations and those obtained using the proxy-year selection criteria thus reveals whether model-observations discrepancies arise from a mismatch in anomalous weather conditions, or from insufficient process representations. As not all GCMs have nudging capabilities, our proposed proxy method aims to offer a generalizable alternative to nudged simulations to enable such direct model-observation comparisons. At last, we compare our proxy-year selection methods with random selections generated using the Monte Carlo approach. Monte Carlo methods are frequently employed in climate research to statistically assess uncertainties related to a hypothesis (e.g., New & Hulme, 2000; Chen et al., 2022). In our case, it is employed to determine whether the proxy-year selection methods produce results that are significantly different from random selection."

General Comments:

L273: "averaging over a large number of observations contributes towards making it comparable."

This is quite a simplistic way of dealing with the considerable scaling issues. It's not just that the data represents different resolutions, it's also about how much a sea ice model should match with a relatively sparse and localized series of observations like the ones collected from MOSAiC in either the proxy analysis or the nudged simulation analysis.

Thank you for your comment.

We fully acknowledge that the sparse and localized nature of observations like those from the MOSAiC campaign presents challenges when comparing them to model simulations, particularly in terms of scaling. Ideally, a continuous and widespread observation sea-ice and snow thickness datasets across the entire Arctic over an extended period will be best for such model comparisons but this is currently neither available nor feasible due to logistical and resource constraints. While currently available sea ice and snow thickness observations are valuable, it is clear that the scaling issues between localized measurements and model grid cells exist and will persist until future satellite observations become available, offering more comparable data in terms of spatial coverage and resolution.

Following lines have been added to emphasize more on the above points and to address the comment.

Lines 259-265: “We underline that the MOSAiC campaign was designed with the aim to support climate model evaluation in the Arctic. An array of instruments was therefore deployed during the expedition over an area comparable to a typical model grid size, rather than from a single point, collecting observations over a more representative sample (Shupe et al., 2022). While not a perfect solution, our proposed methodology can offer potential for initial model evaluation, contributing to improved understanding and refinement of model performance.”

Reference:

Shupe, Matthew D., Markus Rex, Byron Blomquist, P. Ola G. Persson, Julia Schmale, Taneil Uttal, Dietrich Althausen et al. "Overview of the MOSAiC expedition: Atmosphere." Elem Sci Anth 10, no. 1 (2022): 00060.

L293: “For an accurate and comprehensive selection of proxy years with characteristics similar to the MOSAiC year, it is crucial to eliminate divergences from observations which arise from the free-running models’ different realizations of natural variability. Therefore, our method refines the selection process by excluding conditions (or years) vastly different from those observed during MOSAiC, ensuring the chosen years mirror the sea-ice and atmospheric conditions of the study period.”

I do not consider this approach to be the right one. I think it would be more helpful to really assess how the MOSAiC year compared with the full model ensemble spread, how typical was it etc, and to provide more insight into the probability of a model sea ice state agreeing with the sea ice observations considering the atmospheric variability across the model ensemble.

Thank you for your comment. We agree that the use of full model ensemble spread is a more popular and broadly employed method for model-observation comparison. However, as outlined in the manuscript, “This study proposes a new proxy-year selection approach to perform meaningful comparisons of free-running CMIP6 models’ sea ice and snow data with measurements relatively localized in time and space – here, using the unique MOSAIC time-period and observations. We propose this method in an exploratory study using a single ensemble member from the set of 10 selected CMIP6 models.” (**Lines 621-625)**

Our approach—comparing nudging with our proxy methodology—demonstrates that results are indeed comparable, specifically when using the SIA-based selection method. Incorporating the full ensemble spread, which includes years with vastly different sea ice conditions, would introduce significant variability that is not relevant to the specific conditions observed during the MOSAiC year. This would, in effect, dilute the value of the

comparison, as the spread in snow and sea ice conditions would not be representative of the MOSAiC year, undermining the purpose of our focused analysis.

Furthermore, the use of the Monte Carlo method in our study could be seen as reflecting a broader ensemble in a more controlled manner. By resampling and using a bootstrapping method across multiple models, we are able to validate the robustness of our proxy year selection methods and confirm its added value compared to random year selection (as detailed in Sect. 4.3) from the *historical* time-period. This, in combination with the physical year selection methods described in Sect. 3.1, serves to validate the reliability of our approach.

We also emphasize that this is an exploratory study aimed at developing a novel proxy year selection methodology, building upon the traditional model evaluation techniques. Given the experimental nature of our research, we have purposefully started with 10 models with one ensemble each (rationale for using a single ensemble can be found in Sec. 2.2). Future work could expand this approach to include a larger ensemble as well as greater number of models for further validation.

The nudged simulations I think are maybe the best idea here, but even with that you don't really compare before and after nudging, so it's hard to know how significant the nudging is compared to other factors. ECHAM6/FESOM contributed to CMIP6 but I don't think this is sufficient to be described as the non-nudged version? And you don't show that anyway.

The ECHAM6/FESOM model version used here is the exact same version as the one which contributed to CMIP6 and labelled "AWI-CM-1-1-MR". Indeed, this model (with or without nudging) is not evaluated in the present paper. However, the non-nudged version of the model for CMIP6 has been shown to be one of the best performing models in terms of both climatological mean states and variability (Merrifield et al. 2023, Lee et al. 2024). The nudged simulations of this model have also been used and evaluated in several studies (e.g. Sanchez-Benítez et al., 2022; Pithan et al., 2023; Athanase et al., 2024a; 2024b; Zhuo et al., preprint) covering several topics, demonstrating the excellent performance of these simulations (now mentioned in the introduction in lines 163-169). Indeed, in Sánchez-Benítez et al., (2022) a comprehensive discussion about the choice of nudging configuration is included. Sánchez-Benítez et al., (2022), indicates that: "Similar to previous studies (e.g. Takhsha et al., 2018; Wehrli et al., 2019), we have found that the global model climatology is not strongly affected by the nudging". Pithan et al., (2023) also demonstrated that the applicability of these simulations is confirmed for direct evaluation of coupled climate models with in situ observations, specifically during the MOSAiC expedition in the Arctic. The text has been modified in the introduction (**lines 156-172**) and the newly created section 3.4. (**lines 382-436**) to better introduce this point. We further added "Details on the nudging configuration and the good performance of

nudged simulations can be found in Sanchez-Benitez et al. (2022) and Pithan et al. (2023)". **(Lines 412-14)**

There are a lot of subjective statements in the methods about the benefit of the approach, I think it is best to stick to the methods description and let the reader decide on how suitable they are!

We removed several subjective statements in the Methods, and hope the description now reads more balanced.

The paper, especially the introduction needs help with the writing, lots of language issues and a lack of relevant citations and much in the way of literature review.

Thank you for your suggestion.

We have made changes to different sections of the paper in order to enhance its clarity and flow. Our introduction, in particular, is now well supported with relevant citations. The changes can be seen as track-changes in the latest version of the manuscript.

Specific comments:

L212: "The SIT values used throughout the study for all the climate models, are weighted by the "siconc". I think you need to check this as the ice thickness should be the ice thickness where you have ice and shouldn't need to be weighted by ice concentration, you are maybe mixing that up with the ice volume?

Thank you for pointing this out. This was clearly an oversight from our end, and we have made the changes addressing the comment. The corrected text now reads: "We focus on the *"sithick"* variable, representing simulated effective floe thickness." **(Line 218-219)**

L270: "Since GCMs provide a single mean value of SIT and snow thickness for each grid cell…" Actually some CMIP6 models and model runs like CESM do include the full ITD output!

Indeed, we have now corrected this into "Since ***most*** GCMs **(Line 245)**"

And again, just averaging everything doesn't ensure they are comparable quantities.

As addressed in our previous response:

We fully acknowledge that the sparse and localized nature of observations like those from the MOSAiC campaign presents challenges when comparing them to model simulations,

particularly in terms of scaling. Ideally, a continuous and widespread observation sea-ice and snow thickness datasets across the entire Arctic over an extended period will be best for such model comparisons but this is currently neither available nor feasible due to logistical and resource constraints. While currently available sea ice and snow thickness observations are valuable, it is clear that the scaling issues between localized measurements and model grid cells exist and will persist until future satellite observations become available, offering more comparable data in terms of spatial coverage and resolution.

Following lines have been added to emphasize more on the above points and to address the comment.

Lines 259-265: “We underline that the MOSAiC campaign was designed with the aim to support climate model evaluation in the Arctic. An array of instruments was therefore deployed during the expedition over an area comparable to a typical model grid size, rather than from a single point, collecting observations over a more representative sample (Shupe et al., 2022). While not a perfect solution, our proposed methodology can offer potential for initial model evaluation, contributing to improved understanding and refinement of model performance.”

Reference:

Shupe, Matthew D., Markus Rex, Byron Blomquist, P. Ola G. Persson, Julia Schmale, Taneil Uttal, Dietrich Althausen et al. "Overview of the MOSAiC expedition: Atmosphere." Elem Sci Anth 10, no. 1 (2022): 00060.

Reviewer 2:

This article presents a method for model-observation comparisons using a proxy year method, in the context of comparing observations of sea ice and snow thickness to in situ observations from the MOSAiC field campaign. Two criteria for the proxy year selection are considered: selecting years based on agreement in sea ice area, and in atmospheric conditions (in particular, the strength of the Arctic Oscillation). The snow depth and ice thickness in proxy years are compared to model simulations nudged to atmospheric conditions, and also to a bootstrap (Monte Carlo) distribution of randomly-selected years to examine the performance of proxy years compared to randomized years. Proxy years selected using the sea ice area criterion perform comparably to nudged model simulations for sea ice thickness and agree better with observations than randomly-selected values. However, regardless of nudging or proxy year selection, the models have difficulty representing the evolution of observed snow thickness.

I think the development of the proxy-year method is reasonably motivated; I agree with the authors that it would be helpful to have less resource-intensive approaches (relative to running nudged simulations) available. However, I think this article needs a number of major revisions and clarifications for the proxy-years method to be more replicable, and for more clarity on the applicability of this method, and in particular, on its limitations. I will detail my general and minor comments below.

Thank you for your positive feedback and suggestions! Please find our responses to your individual comments in the following segment.

General comments

Although the article tests two proxy selection criteria (SIA and AO) against two observed quantities (SIT and snow thickness), from what I see in the results (Fig 8), only the SIA proxy performs better than random selection, and only for SIT. The AO proxy, conversely, does not perform better than random selection, and snow thickness is not well-estimated overall. The authors do state this in the article results, but I think there are several instances in the discussion and conclusions where this is not stated clearly enough, in my view. I will identify some specific examples of this in the minor comments below, but overall, I think more care needs to be taken throughout the article to explicitly state which approaches are applicable for which situations, to clearly state where methods and models did not perform well, and to explicitly state which metrics are being used to evaluate performance. I think the article would benefit from some more emphasis on quantitative assessment of the performance of the method (e.g. quantify more explicitly the agreement with observations).

Thank you for your detailed comments regarding the issues raised in this general comment. We have carefully addressed the specific instances you pointed out, making the necessary clarifications and refining the explanations throughout the manuscript. These changes are reflected in both the track-change version of the manuscript and our response to the minor comments. In particular:

1. We have made the required changes to, ensuring that the points raised in the comments have been thoroughly addressed. We have also explicitly outlined which methods perform better/bad for specific variables, such as sea ice thickness and snow thickness. This ensures that the manuscript now clearly communicates where each method excels or falls short, allowing readers to better understand the results.
2. We appreciate your comment on adding more emphasis on the quantitative assessment of the accuracy of our proxy methods. Therefore, in response to the comment, we have now added a new figure (Figure 9) in the manuscript where we describe the absolute biases for sea ice thickness and snow thickness from the different methods employed as well as a supporting table (Table S1) in the supplementary section assessing the significance of the performance of various methods. Following are the new additions:

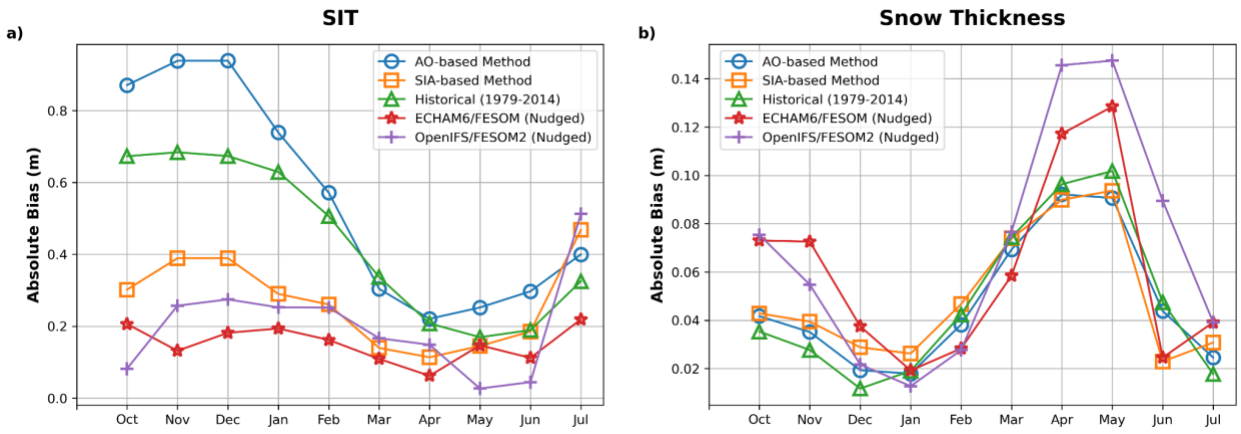


Figure 9: Monthly biases for SIT and Snow Thickness across different methods. Biases (absolute differences) are calculated as the absolute difference between annual cycles from observations and those obtained from simulations, using the two proxy-year selection methods, the two nudged models, and all historical simulations from the 10 selected CMIP6 models (1979-2014, the average of all single model biases is shown).

Table S1: Significance of the performance of various methods based on p-values. The p-values are derived from Mann-Whitney U-tests, which assess the statistical significance of the differences between proxy methods, nudged and *historical* cycles for each type of thickness (SIT and Snow). For instance, if the p-values are 0.01 or less, we

can say with 99% confidence level (significance) that the two methods are different. Else, we reject the hypothesis.

Note: Nudged simulations only have one year i.e. the MOSAiC year and is a mean based on three distinct ensemble members.

	SEA ICE THICKNESS	**SNOW THICKNESS**
SIA vs. Historical	0.002	0.791
AO vs. Historical	0.570	0.791
AO vs. SIA	0.002	0.909
AO vs ECHAM6/FESOM (Nudged)	0.001	0.623
SIA vs ECHAM6/FESOM (Nudged)	0.104	0.733
Historical vs ECHAM6/FESOM (Nudged)	0.0007	0.623
AO vs OpenIFS/FESOM2 (Nudged)	0.021	0.791
SIA vs OpenIFS/FESOM2 (Nudged)	0.850	0.791
Historical vs OpenIFS/FESOM2 (Nudged)	0.011	0.677

Following lines have also been added in the main text:

Lines 640-662: “Our results demonstrate that the SIA-based proxy year selection method outperforms both the AO-based proxy selection method and the *historical* simulations from free-running CMIP6 models. This is also evident from the lower biases in the SIA-based method (Fig. 9a), which shows its closer approximation to the observed data and comparable performance with the two atmospherically nudged simulations for SIT (Table S1). Notably, the SIT annual cycles obtained using the SIA-based method also differ significantly from those of the *historical* simulations (Table S1). However, for snow

thickness, no method offers a clear bias advantage over the others (Fig. 9b). This confirms that snow-related processes are insufficiently represented in both CMIP6 and nudged models, resulting in poor simulation accuracy across all methods when compared to observations. For the atmospherically nudged simulations, it is important to emphasize that the calculations in Fig. 9 and Table S1 are based on a single year (the MOSAiC year) and three distinct ensemble members, unlike the CMIP6 models which rely on only one ensemble member. Additionally, the *historical* simulations cover a period from 1979-2014. Therefore, beyond the unresolved snow processes, the biases observed in Fig. 9 may also be influenced by error compensations across different methods.

Overall, our findings highlight the potential of this experimental yet relatively simple approach in using the free-running CMIP6 models to achieve outcomes similar to the more precise, observationally constrained nudged simulations. Such methods are particularly valuable for institutions that lack the resources to produce their own nudged simulations, offering a reasonable alternative that maintains relatively better accuracy. Therefore, this study offers a particular methodology that may serve as one of the many comparison possibilities between coupled climate models and field observations."

I would appreciate more background motivation on the choice of the proxy year method as a method for this analysis. To my recollection, similar methods have been used in paleoclimate research; including some references there would be helpful, or some references to similar methods to the proxy year method, if possible.

Thank you for the comment. We have added the following lines and literature citations in our introduction section to address the above comment:

Lines 150-152: "Similar proxy-based selection methods are widely used in paleoclimatology studies, particularly for temperature-based proxy reconstructions (e.g., Miller et al., 2010; Burke et al., 2018)."

References:

Miller, Gifford H., et al. "Arctic amplification: can the past constrain the future?." Quaternary Science Reviews 29.15-16 (2010): 1779-1790.

Burke, Kevin D., et al. "Pliocene and Eocene provide best analogs for near-future climates." Proceedings of the National Academy of Sciences 115.52 (2018): 13288-13293.

Since this is an article proposing a method presumably for use in other research, I think it is crucial that the method be documented in sufficient detail for it to be reproduced. In its current state, the article is somewhat lacking in detail necessary for reproducibility and would benefit from some clarifying revisions. For example, for the SIA proxy, estimates

of contributions from first-year and multi-year ice are discussed, but it is unclear to me how the total seasonal difference in Fig 2b was calculated from these contributions. Is this the sum of the two differences in Fig 2a? Also, was the absolute difference used, or some other difference metric? Some additional details would be helpful here. Likewise, I would appreciate similar clarification for the AO proxy. In general, it is also not always clear to me when authors discuss result significance and inter-model differences, how these differences are being determined and compared.

Thank you for your comment. We have added more detailed explanation of both the proxy-selection methods:

1. **SIA-based method:**

Line 298-311: "The selection of sea-ice based proxy years is conducted using two key indicators: Firstly, we consider the seasonal difference in SIA (March minus September) which serves as an estimate of the first-year ice component. Secondly, we examine the SIA during the yearly minimum (September) to estimate the contribution of multi-year ice during those years (Fig. 2a). In Fig. 2b, the total seasonal difference represents the sum of (i) the absolute differences between the seasonal difference in models and the corresponding observational reference and (ii) absolute difference between the model and observational reference minima in September. The proxy years are selected by identifying those with the smallest combined differences (or total seasonal difference), ensuring that the selected years closely match the observed SIA values from the MOSAiC year. This method selects years with similar sea ice conditions to those observed during the MOSAiC year. By concentrating on these specific proxy years, we aimed to replicate the contributions of both the first-year and multi-year ice in shaping the sea-ice and snow distributions in a particular year."

2. **Similarly, we added details for AO-based selection criteria:**

Lines 333-346: "The AO index is calculated as the normalized leading mode of EOF analysis (loading pattern) of the geopotential height anomalies (at 1000hPa) polewards of 20°N latitude. Given this context, our evaluation considered the proximity of the seasonal AO index values in climate models to the corresponding reference values observed during the MOSAiC year. Specifically, we compare the simulated AO index values during the winter season (January–to–March) of a given year and the November values in the preceding year with the observational reference. For each CMIP6 model, we identify three years in the *historical* period which exhibited the smallest *total seasonal differences* (similar to the method applied for Fig. 2b) from the observed AO indices (capturing extreme AO trends during both winter and the preceding November) (Fig. 3). The selected proxy years are presumed to replicate the anomalous atmospheric conditions prevalent during the MOSAiC year. This methodology thus involves comparing

AO values in corresponding periods and selecting proxy years based on minimizing differences, thereby aiming to align with the observed AO dynamics."

This article emphasizes that the proxy-year method addresses three challenges for model-observation comparisons; spatial sampling differences, measurement locations shifting in time, and the fact that GCMs are not designed to simulate specific years. However, in my opinion, the approaches used to address the first two challenges are not novel. As such, I do not think there needs to be as much emphasis placed in the article on averaging a large number of observations or collocating measurements to model grid cells, since both of these are commonly-used approaches for model-observation comparisons. I do think the use of these approaches is worth mentioning, but I do not think they need as much emphasis as the use of the proxy-year method.

Thank you for your comment. The detailed explanation of the three challenges serves two purposes: *first*, to provide greater clarity regarding the process behind our proxy selection method and to strengthen the robustness of the methodology. *Second*, in response to the first reviewer's comment, we felt that including these details would be beneficial in addressing and clarifying their concerns as well.

However, as suggested, we have reduced the emphasis on the three challenges and have now clubbed them into two distinct categories: *Spatial and Temporal Differences*. Following changes are made in the manuscript:

Line 134-145: "Yet, there still remain certain challenges in using the MOSAiC-derived SIT and snow data for a comparison with coarse resolution GCM simulations:

1. **Spatial differences:** These arise from (i) discrepancies between point measurements of sea ice and snow in MOSAiC and the averaged values at grid cell resolution represented in GCMs and (ii) the drifting of MOSAiC ice flow in space during the MOSAiC year.
2. **Temporal differences:** The MOSAiC data represents a single year, capturing one realization of natural variability; while freely running climate models are not designed to simulate the characteristics of a specific observed year (e.g. atmosphere and ocean circulation that may lead to certain sea ice and snow patterns)."

Additionally, as this is an exploratory model evaluation study utilizing a novel proxy year approach, we agree that more emphasis should be placed on the challenge of comparing coupled climate models with in-situ observations that are limited to specific regions and timeframes. In light of this, we have added the following lines, which also respond to the first reviewer's comment:

Lines 259-265: “We underline that the MOSAiC campaign was designed with the aim to support climate model evaluation in the Arctic. An array of instruments was therefore deployed during the expedition over an area comparable to a typical model grid size, rather than from a single point, collecting observations over a more representative sample (Shupe et al., 2022). While not a perfect solution, our proposed methodology can offer potential for initial model evaluation, contributing to improved understanding and refinement of model performance.”

Reference:

Shupe, Matthew D., Markus Rex, Byron Blomquist, P. Ola G. Persson, Julia Schmale, Taneil Uttal, Dietrich Althausen et al. "Overview of the MOSAiC expedition: Atmosphere." Elem Sci Anth 10, no. 1 (2022): 00060.

Moreover, we follow similar proxy year selection methods as used in paleoclimate studies for simulating specific years based on certain proxies (in our case, SIA and AO):

Lines 150-152: “Similar proxy-based selection methods are widely used in paleoclimatology studies, particularly for temperature-based proxy reconstructions (e.g., Miller et al., 2010; Burke et al., 2018).”

References:

Miller, Gifford H., et al. "Arctic amplification: can the past constrain the future?." Quaternary Science Reviews 29.15-16 (2010): 1779-1790.

Burke, Kevin D., et al. "Pliocene and Eocene provide best analogs for near-future climates." Proceedings of the National Academy of Sciences 115.52 (2018): 13288-13293.

I understand that models represent a range of possibilities. However, I wonder about the representativeness of historical years being chosen as proxies for the MOSAiC observation time year, which took place after the historical period being observed. Given that there have been recent discussions of regime shifts in Arctic sea ice (e.g. Sumata et al., 2023), I wonder about the applicability of proxy years from earlier decades to more recent years, and what biases could be potentially introduced due to interdecadal differences. I would appreciate it if the authors would comment on this, and other possible confounding factors. Given the caveats associated with this method, I think the authors need to be more careful when making conclusions about overall model process representativeness based on proxy years.

Thank you for raising this important point. While we understand the concerns about regime shifts in Arctic sea ice, as discussed by Sumata et al. (2023) (which highlights that "the timing of the shift was preceded by a two-step reduction in residence time of sea ice in the Arctic Basin, initiated first in **2005 and followed by 2007**), we believe the use of proxy years from earlier decades remains valid in our study for following reasons:

1. **Model Simulations of Regime Shifts**: After reviewing the average SIT across the 10 models selected for our study, we find that none of the models accurately capture the regime shifts (i.e., the sharp decreases in SIT) noted in 2005 and 2007. Instead, there is a more gradual decline in SIT from 1979 to 2014 across all models. In fact, some models even show peaks in SIT following these regime shift years, which suggests that the models do not simulate these shifts in a consistent or significant way. This limitation is expected as Sumata et al. (2023) uses *in-situ* ULS Moorings for sea-ice draft which is more precise. Such interannual changes are difficult to replicate in coupled climate models.

2. **Possible Regional Differences**: The regime shifts described by Sumata et al. (2023) are based on regions within the Alaskan and Siberian sectors of the Arctic, which differ from the MOSAiC trajectory and the regions of interest in our study, which are predominantly in the Eurasian Arctic sector. Therefore, the interannual differences noted in that paper may not apply directly to the conditions along the MOSAiC track, where our focus lies.

3. **Proxy Year Selected Between 1979-2014**: It's also worth noting that only a few models in our analysis have proxy years extending beyond 2007. Specifically, one model (Figure 2b) in the SIA-based proxy selection and two models (Figure 3) in the AO-based selection include more than one year beyond 2007, the second year highlighted by Sumata et al. as marking a significant regime shift. This suggests that our study's conclusions are not heavily influenced by post-regime shift conditions in most models, if at all captured by them.

In light of these factors, we believe that the use of historical proxy years is still appropriate for our study.

Finally, I find the flow and organization of the article to be confusing at times, and some parts of the discussion appear to be outside of the intended scope of the article. Some points are also raised suddenly without being mentioned earlier in the article. For example, methodology validation is not mentioned until Section 3.4, when the Monte Carlo method is suddenly introduced. I would find a mention of this in the introduction to be helpful.

Thank you for the suggestion. We now include a paragraph in the introduction presenting the Monte Carlo and nudging methods, and explaining how they are used to evaluate the proxy-year selection approach.

Line 156-172: “We verify the skill of the proxy-year selection criteria using two approaches. First, we employ a set of “nudged” simulations, where the atmospheric circulation is directly constrained to observations (e.g., Sánchez-Benítez et al., 2022; Athanase et al., 2024a, 2024b; Zhuo et al., preprint). These nudged simulations produce quite accurate analogues of the observed conditions during the MOSAiC campaign in the climate models, as shown by Pithan et al. (2023). Comparing the annual cycles of sea ice and snow thickness obtained using the nudged simulations and those obtained using the proxy-year selection criteria thus reveals whether model-observations discrepancies arise from a mismatch in anomalous weather conditions, or from insufficient process representations. As not all GCMs have nudging capabilities, our proposed proxy method aims to offer a generalizable alternative to nudged simulations to enable such direct model-observation comparisons. At last, we compare our proxy-year selection methods with random selections generated using the Monte Carlo approach. Monte Carlo methods are frequently employed in climate research to statistically assess uncertainties related to a hypothesis (e.g., New & Hulme, 2000; Chen et al., 2022). In our case, it is employed to determine whether the proxy-year selection methods produce results that are significantly different from random selection.”

We have also reorganized the structure of Section 3 (Methods), adding a new sub-section, *‘3.4: Methods for Proxy-Year Evaluation’*. This section clearly delineates and provides a detailed discussion on both the evaluation methods: Nudging and the Monte Carlo method, both of which were introduced in Section 1.

The last paragraph of the introduction could potentially be expanded with slightly more detail about the contents of the sections. I will include some mentions of where detail could be added in the comments below. I would encourage the authors to consider if parts of the article could also be condensed or restructured for clarity of scope as well.

Thank you for your suggestions.

We have expanded the introduction section in response to the comments. The paragraphs leading up to the end of Sect.1 now provide a detailed overview of the background, motivation, and rationale behind our chosen methodology. We hope that the final paragraph now offers a more concise summary of the paper’s contents and better outlines its organization.

Lines 173-176: “This study is organized as follows: In Sect. 2, we describe the observation and model data sets. Sect. 3 details the proposed observation - model

comparison methods. Results of the comparison are given in Sect. 4 while Sect. 5 discusses the proposed methods and results and finishes with concluding remarks and ways forward (Sect. 6)"

Additionally, we have made changes to different sections of the paper in order to enhance its clarity and flow. The changes can be seen as track-changes in the latest version of the manuscript.

Minor comments:

161: You introduce the nudging here but not why it is used; I suggest briefly mentioning the purpose of the nudging here, because otherwise it seems unrelated to the proxy year method, and "another set of comparisons" is vague phrasing (comparisons to what?)

We agree with your comment, and now added a description of the motivation to use the nudged simulations (**Line 156-172**). Thank you.

208: I'm guessing the period 1979-2014 was chosen because the historical runs are limited to up to 2014, but maybe specify that in the text.

Added a line **(Line 217)** to specify the end year of *historical* simulations.

Why were the historical runs chosen over other years? Could you comment on the potential utility of other scenarios? My comment below on line 220 is related to this also.

The rationale for choosing *historical* experiments is that they are considered mostly closer to the observations and are widely used for such present-world model comparison studies when testing the model accuracy (as used in SIMIP Community, 2020; Xu et al., 2023; Shu et al., 2020 and Roach et al., 2020).

References:

Notz, D., & Community, S. I. M. I. P.: Arctic sea ice in CMIP6. Geophysical Research Letters, 47(10), e2019GL086749, 2020.

Roach, L.A., Dörr, J., Holmes, C.R., Massonnet, F., Blockley, E.W., Notz, D., Rackow, T., Raphael, M.N., O'Farrell, S.P., Bailey, D.A. and Bitz, C.M.: Antarctic sea ice area in CMIP6. Geophysical Research Letters, 47(9), p.e2019GL086729, 2020.

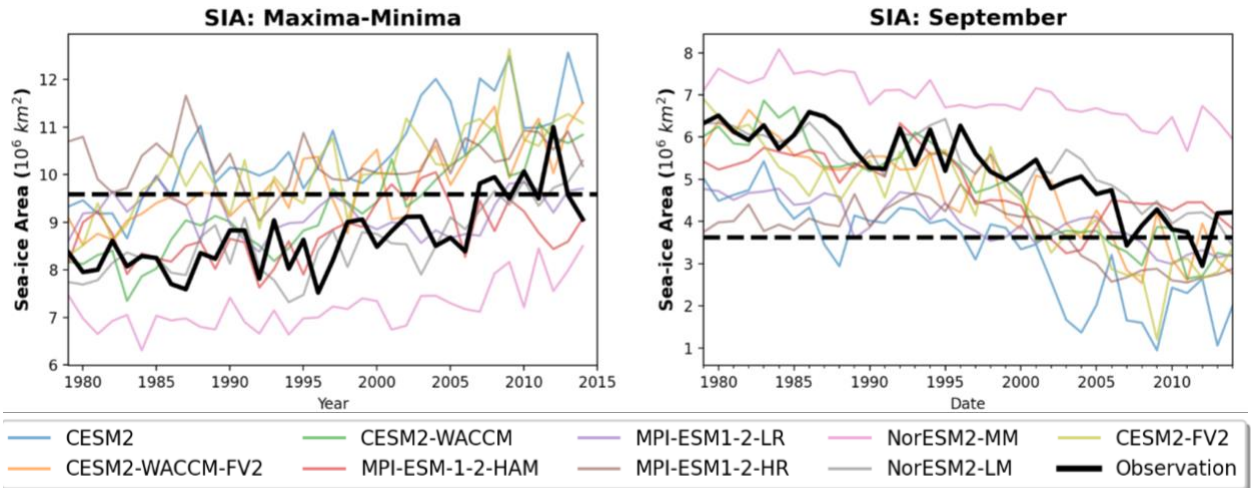
Xu, M., & Li, J. (2023). Assessment of sea ice thickness simulations in the CMIP6 models with CICE components. *Frontiers in Marine Science*, *10*, 1223772.

220-223: I have some questions as to how this decision was made, since it does seem like using the SSP585 scenario may be adding biased results. Which observed sea ice characteristics were not reached? Some clarification would be helpful.

Following lines have been added in Sec.2.2 for clarity:

“For NorESM2-MM, we have utilized the SSP585 scenario for selecting proxy years based on the SIA-criterion (Fig. 5), as its *historical* values simulated exaggerated sea ice characteristics compared to observations and other climate models (figure not included here, but see Seland et al., 2020). Therefore, using the warmer scenario for this model was specifically done to reduce SIA biases and align its behavior more closely with other models for comparison purposes.” (**Lines 227-232**)

Aforementioned figure, which is not included in the manuscript, can be found below. The figure shows NorESM2-MM (pink solid line) as an outlier when we carried out our analysis using *historical* simulations of this model. However, as seen in Fig. 2a, when using a warmer scenario, we find its better alignment to the observations.



We put forward two reasons to support the usage of the warmer scenario:

1. When using the SIA values for the *historical* experiments for NorESM2-MM, we found that the averages were exceptionally high compared to the other models. For this reason, and for a fairer comparison, we used the warmer SSP585 scenario which gave us comparable SIA values.

   According to **Seland et al., 2020 (also cited in the manuscript)**: *"NorESM2-MM shows too much sea ice in the central Arctic in September. In general, the model is colder in the Arctic than NorESM2-LM, and it has thicker sea ice in the Arctic Ocean. The Northern Hemisphere sea-ice volume in NorESM2-*

MM is 21 % (36 %) larger in March (September) compared with the NorESM2-LM (not shown)"

2. Due to the varying structures of coupled climate models, we believe that NorESM2-MM reaches similar sea-ice conditions as other models but later in the post-*historical* period. This assumption is based on the observation that the annual cycles of SIT and snow thickness in NorESM2-MM become more pronounced and exhibit similar patterns, amplitudes, and ranges to those in other models, particularly in the warmer scenario simulations. This convergence likely occurs because of the model's specific setup (as stated in Seland et al., 2020), which may lead to comparable sea-ice states in the later years.

249: Previous nudging studies are cited, but I think it would be helpful for clarity to explicitly state in the article that the nudging methods used here are motivated by previous studies.

We followed your suggestion and modified these lines accordingly in section 3.4.1 (**Lines 387-391** and **412-415**). The good performance of these simulations including comparison with MOSAiC data is now also directly mentioned in the introduction (**Lines 156-172**).

Figure 2: The colourbar in b) denotes the scale for the selected years, but I would also be curious to know the scale for the greyed-out values. Consider possibly including a colourbar for the greyscale values also, or highlighting the selected years in another way. In the caption, "highlighted" is also somewhat ambiguous, perhaps replace this phrasing with "in colour".

Figures are updated along with the captions, as suggested. The updated figures now only highlight the three selected proxy years.

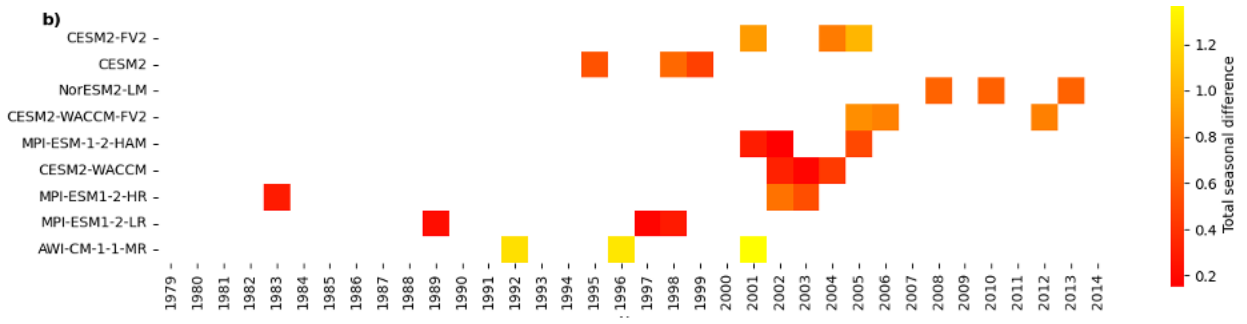


Figure 2: Selection of sea-ice based proxy years. (b) Heatmap showing the seasonal differences between each year in the *historical* CMIP6 models and the SIA values for the MOSAiC year. Colored are the three proxy years with lowest three differences, selected for each model between 1979-2014 based on their proximity to the annual SIA values.

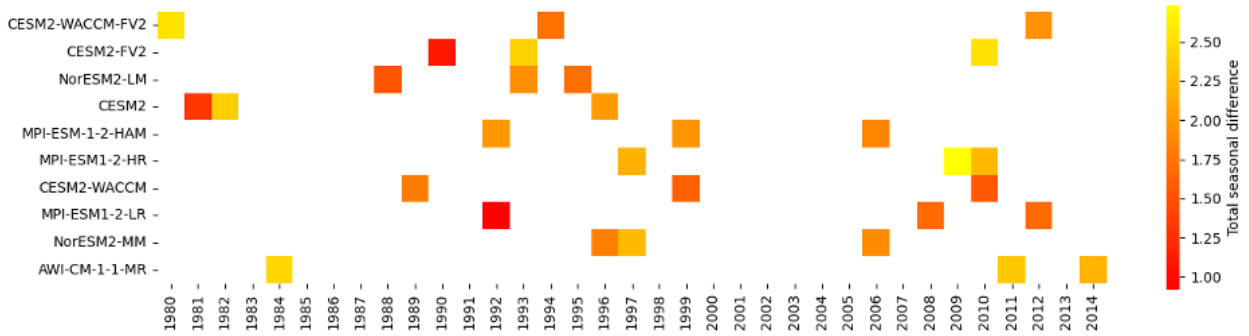


Figure 3: Atmospheric circulation-based proxy years. Heatmap showing the seasonal differences between each year in the *historical* CMIP6 models and the AO values for the MOSAiC year. Colored are the three proxy years with lowest three differences selected for each model between 1979-2014 based on their proximity to the seasonal AO values.

Figure 7: I realize you want to show daily variations here, but it would be nice to have monthly means of these plots available to be able to more directly compare with Fig 5 and 6.

Thank you for your interest and suggestions. The following figure has now been added in the supplementary section.

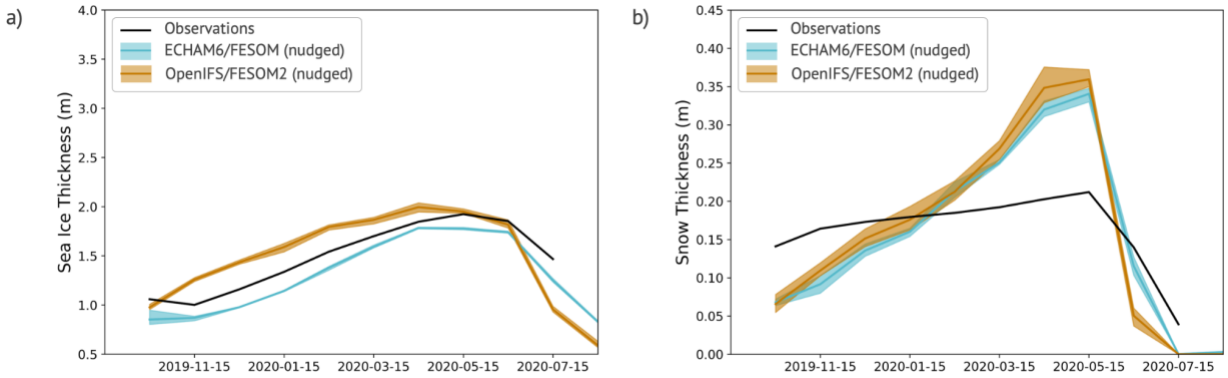


Figure S2: Annual cycles of sea-ice and snow parameters from the nudged simulations. (a) Monthly SIT and (b) Monthly snow thickness along the MOSAiC trajectory; the black line represents in-situ observations, and rest show nudged simulations in which the evolution of winds observed during the MOSAIC year is imposed. Climate models used are ECHAM6/FESOM (blue) and OpenIFS/FESOM2 (orange). Shaded areas are the ensemble range for each set of coupled climate simulations, thick lines are the ensemble mean.

545: Would it be possible to show the monthly accumulated precipitation/snowfall values in a supplemental figure? I am curious to see how closely they agree, given that precipitation is challenging to represent in models. This would also help illustrate how large the 0.005 m quantity is relative to the amount of precipitation.

Thank you for your keen interest in this comparison. As suggested, we have now added the daily and monthly snowfall figure (below) as a supplementary figure which was previously only mentioned in the text.

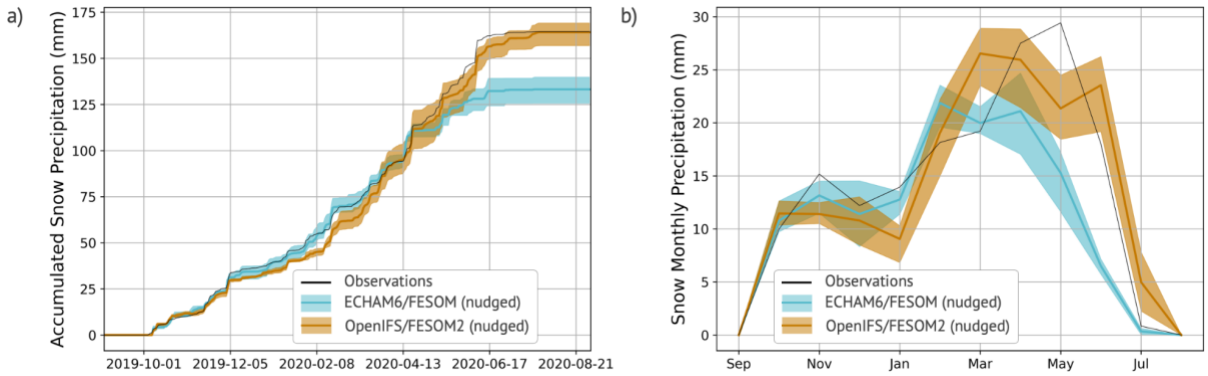


Figure S1: Snowfall data from the nudged simulation ensembles from ECHAM6/FESOM (orange) and OpenIFS/FESOM2 (blue) along the MOSAIC drift trajectory, (a) daily accumulated and (b) monthly averaged, compared to observations from ERA5 (black line).

Our manuscript mentions the above plot at two instances as follows:

Lines 564-570: “Yet, we note that precipitation and snowfall in nudged simulations follow closely to the observations, with monthly accumulated values comparable to ERA5 within a 0.005 m difference, to the exception of May when this difference reaches 0.015 m (Fig. S1). This confirms that biases in snow thickness accumulation are not primarily driven by the atmospheric flow, which is captured in the nudged runs, but rather by other processes of snow advection and melt insufficiently represented in CMIP6 and nudged models.”

Lines 666-673: “Seasonal snowfall variations in the nudged simulations match reasonably well those of ERA-5 (Fig. S1). We further note that snowfall in ERA-5 during the MOSAiC expedition slightly exceeded the in-situ observations (Wagner et al. 2022). This suggests that a comparable, yet marginally underestimated snowfall therefore is not the cause for the overestimated accumulation of snow on sea ice in the simulations. Instead, other processes such as snow advection and melt, which are insufficiently represented in CMIP6 and nudged models may likely contribute to this bias (e.g., Chen et al., 2021, Nicolaus et al., 2022, Pithan et al., 2023).”

This supplementary figure now supports the statements made in the text. In summary, the figure allows us to conclude the following:

1. The nudged simulations accurately capture snowfall, following generally well the ERA5 reanalysis data.
2. Therefore, divergences in snow thickness seasonality are not due to mismatches in atmospheric input, as the input is well reproduced.

3. Instead, other snow processes—such as snow advection, melt, and densification, which are not yet well-represented in climate models—likely contribute to the higher snow accumulations and the resulting mismatch in snow thickness cycles between the models and reference observations.

561: I think more detail is needed on how the Monte Carlo samples were generated and selected. E.g. were all the models pooled together for these 10,000 samples?

We added further details in Sec.3.4.2, as suggested.

Lines 427-436: "The Monte Carlo samples are generated independently for each of the 10 models. We apply the resampling technique known as bootstrap, which randomly selects three years during each of the 10,000 iterations for all the selected models. These samples are used to extract and calculate the seasonal cycle for each model's data. After generating results for all 10 models separately, they are combined to compute the Multi-Model Mean (MMM) by averaging the corresponding values across all models. This process generates a multitude of random possible combinations, encompassing various annual cycles for SIT and Snow Thickness. We thereby evaluate whether the proxy year selection methods perform better than randomly selected years."

Therefore, no models were pooled together during the Monte Carlo sampling process; each model was treated individually before combining for the MMM.

574-575: It may just be the phrasing confusing me here, but when you say the method "captures real-world scenarios, albeit rare ones", does that imply that more typical scenarios may be more difficult to capture?

Thank you for pointing this out. Lines are rewritten and now read as:

Line 596-598: "This alignment with observations can be seen as a positive aspect, indicating that our method captures real-world scenarios, including, in this case, even the rare ones (Rinke et al., 2021; Dethloff et al., 2022)."

605-607: "Both methods account for the observed spatio-temporal variables […] ensuring a closer approximation to the sea-ice and atmospheric conditions during the study period." I think this sentence may need to be rephrased, since in this article, you show that only the SIA criterion improves agreement with sea ice, with the AO criteria lying close to the bootstrap mean (as stated in lines 565-566)

The following lines have been rephrased to avoid confusion. Here, we are referring to the two mentioned challenges and how our methodology addressed them by making spatial (i.e. averaging over multiple autonomous buoy observations) and temporal (i.e. proxy-year selections based on either sea ice or atmospheric criteria) adjustments.

Line 628-630: “Both methods account for spatio-temporal variations and aim for a reasonable approximation to the observed sea-ice and atmospheric conditions during the study period (refer to Methods).”

610-612: “Our two proxy year selection methods demonstrate performance comparable to that of atmospherically nudged simulations”; I think you need to be careful here because the AO method does not appear to be as comparable to nudged simulations as the SIA method. By which metric are you defining “comparable” performance here?

Thank you for your comment. In response, we have now incorporated Figure 9 and Table S1 to offer a more quantitative assessment of the methods’ performance and their comparison with observations. Additionally, we have revised the manuscript to ensure clarity on the specific applicability of each method and the variables they address. The following changes have been made:

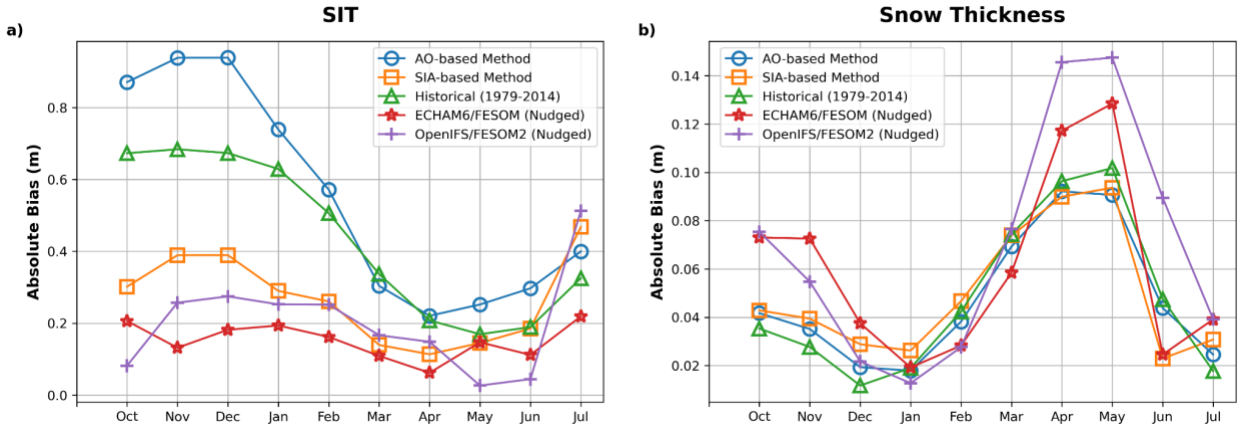


Figure 9: Monthly biases for SIT and Snow Thickness across different methods. Biases (absolute differences) are calculated as the absolute difference between annual cycles from observations and those obtained from simulations, using the two proxy-year selection methods, the two nudged models, and all historical simulations from the 10 selected CMIP6 models (1979-2014, the average of all single model biases is shown).

Table S1: Significance of the performance of various methods based on p-values. The p-values are derived from Mann-Whitney U-tests, which assess the statistical significance of the differences between proxy methods, nudged and *historical* cycles for each type of thickness (SIT and Snow). For instance, if the p-values are 0.01 or less, we can say with 99% confidence level (significance) that the two methods are different. Else, we reject the hypothesis.

Note: Nudged simulations only have one year i.e. the MOSAiC year and is a mean based on three distinct ensemble members.

	SEA ICE THICKNESS	SNOW THICKNESS
SIA vs. Historical	0.002	0.791
AO vs. Historical	0.570	0.791
AO vs. SIA	0.002	0.909
AO vs ECHAM6/FESOM (Nudged)	0.001	0.623
SIA vs ECHAM6/FESOM (Nudged)	0.104	0.733
Historical vs ECHAM6/FESOM (Nudged)	0.0007	0.623
AO vs OpenIFS/FESOM2 (Nudged)	0.021	0.791
SIA vs OpenIFS/FESOM2 (Nudged)	0.850	0.791
Historical vs OpenIFS/FESOM2 (Nudged)	0.011	0.677

Following lines have been added in the main text:

Lines 640-662: "Our results demonstrate that the SIA-based proxy year selection method outperforms both the AO-based proxy selection method and the *historical* simulations from free-running CMIP6 models. This is also evident from the lower biases in the SIA-based method (Fig. 9a), which shows its closer approximation to the observed data and comparable performance with the two atmospherically nudged simulations for SIT (Table S1). Notably, the SIT annual cycles obtained using the SIA-based method also differ significantly from those of the *historical* simulations (Table S1). However, for snow thickness, no method offers a clear bias advantage over the others (Fig. 9b). This confirms that snow-related processes are insufficiently represented in both CMIP6 and nudged models, resulting in poor simulation accuracy across all methods when compared to observations. For the atmospherically nudged simulations, it is important to emphasize that the calculations in Fig. 9 and Table S1 are based on a single year (the MOSAiC year) and three distinct ensemble members, unlike the CMIP6 models which rely on only one

ensemble member. Additionally, the *historical* simulations cover a period from 1979-2014. Therefore, beyond the unresolved snow processes, the biases observed in Fig. 9 may also be influenced by error compensations across different methods.

Overall, our findings highlight the potential of this experimental yet relatively simple approach in using the free-running CMIP6 models to achieve outcomes similar to the more precise, observationally constrained nudged simulations. Such methods are particularly valuable for institutions that lack the resources to produce their own nudged simulations, offering a reasonable alternative that maintains relatively better accuracy. Therefore, this study offers a particular methodology that may serve as one of the many comparison possibilities between coupled climate models and field observations."

624: Given that snowfall is brought up as a point of discussion here, I think it would be helpful to include it in a supplemental figure.

As mentioned in the above comment, the plot comparing the snowfall rates is now added as a supplemental figure (Fig. S1).

735: Annual maximum and variations of which specific quantities? Please clarify.

742: I think you should clarify here that you saw modest enhancements in aligning SIT with observed annual cycles, unless you're referring to other quantities as well, in which case you should specify for clarity.

Thank you for pointing out the confusion in our sentence phrasing. The entire paragraph in the conclusion section is now rewritten to provide more clarity and also addresses both the above comments:

Line 749-761: "Comparing the two proxy-year selection criteria we find that the SIA-based method yields the annual cycles of sea-ice thickness (SIT) closest to the observations. Annual cycles generated using this criterion exhibited relatively lower biases and narrower inter-model spreads when compared to AO-selected proxy years. Furthermore, our study evaluates the two proxy year selection methods by first comparing them with the nudged simulations and lastly by using the Monte Carlo method. These experiments show that SIA-based proxy year selections performed comparable to the nudged simulations as well as significantly better than randomly selected years in reproducing the observed SIT annual cycles. They also suggest that atmospheric conditions may not be the primary contributors to the model biases during the MOSAiC period. Our evaluation further reveals that neither the selection method nor the nudged simulations could accurately replicate the snow thickness annual cycle observed in-situ during MOSAIC, suggesting unresolved processes in nudged and CMIP6 simulations."

Minor edits by line

All the changes have been accepted and lines are modified accordingly.

159: “a comparison of MOSAiC dataset” should be “the MOSAiC dataset”

220: “ensemble.” -> “ensemble member.”

432: less -> fewer

435: fewer: consider instead “lower” or “lesser”

543: Fig 5b refers to SIT, did you mean Figures 6a and 6b for snow thickness?

748: hereinabove -> above

References:

Sumata, H., de Steur, L., Divine, D.V. et al. Regime shift in Arctic Ocean sea ice thickness. Nature 615, 443–449 (2023). https://doi.org/10.1038/s41586-022-05686-x