We thank the reviewer for these useful comments which will certainly improve the manuscript. The reviewers' comments are reproduced in black, and our replies are provided in blue. When necessary, we have also included the modified section of the revised manuscript below our response.

Review RC2 (10 February 2021)

General Comments

This study compares PMIP3-simulated sea ice cover in the Southern Ocean, as well as that from two LOVECLIM experiments, against an updated catalogue of sea ice paleoproxy data. In that sense, it follows directly on Roche et al (2012) and Marzocchi and Jansen (2017). They focus on the summer season, as that season has had the fewest data constraints in previous studies. Additionally, they correlate the sea ice edge to SST and wind stress curl in the models.

Overall, I find the results of this study underwhelming. Although they provide a lot of results, I feel that not a lot of meaning is derived from them (i.e. Too much of a "figure tour" or "data tour").

We appreciate the comment and have now improved the manuscript to more clearly state our goals and results. Additionally, we have added PMIP4 simulations from six models to this intercomparison (plus an additional LOVECLIM simulation following the PMIP4 protocol), which will allow us to compare the LGM seasonal sea-ice cover differences between PMIP3 and PMIP4 models.

Assessing the paleo-proxy data is outside my area of expertise, but the model analysis does not feel like a substantial contribution beyond Marzocchi and Jansen (2017), who examined Southern Hemisphere sea ice controls in these same PMIP3 simulations.

We believe that our goals are different from the goal of Marzocchi and Jansen, as they analyse how sea ice impacts deep ocean circulation. Here, we want to i) assess the most likely seasonal sea-ice cover at the LGM by looking at both model simulations and proxies and ii) understand the processes that can impact the summer sea-ice distribution in the models.

In addition, our study provides regional comparisons between recently a updated compilation of proxy record estimates and simulated sea-ice extent and SST, which was not done in Marzocchi and Jansen (2017). Furthermore, as mentioned above, PMIP4 simulations are now also included in our study.

Indeed, I find it concerning that a serious flaw in one of those simulations reported in Marzocchi and Jansen (2017) (that of a bugrelated absence of wind-stress feedback on sea ice in MRI-CGCM3) is not mentioned here, even though wind-stress curl is one of the foci of the study.

We thank the reviewer for pointing this out. We have now included a sentence about the bug in the MRI-CGCM3 model in the results section, seen below:

Results section 3.4, lines 277-280:

"Our analysis suggests that the summer sea-ice edge in the MRI-CGCM3 is dynamically driven as its mean SO SST is close to the PMIP3 MMM, and its summer sea-ice edge is close to the maximum of the wind stress curl. However, this result should be taken with caution as in the MRI-CGCM3 simulation the coupling between sea ice and wind-stress at the ice/atmosphere interface was absent due to a model bug (Marzocchi & Jansen 2017)."

The two additional LOVECLIM simulations do not appear to be used in any significant way to get at causal mechanisms, even though these simulations differ substantially in both the design of the experiments and complexity of the model.

We acknowledge that there are experimental design differences between the LOVECLIM and PMIP3 model simulations presented here. We have thus removed the 2 LOVECLIM experiments from the PMIP3 group, and are presenting them separately. In line with other comments, we are now including PMIP4 LGM results, including one new simulation performed with LOVECLIM. We are thus now showing 3 different LOVECLIM simulations, in which the AMOC states differ. This allows us to assess inter and intra model differences, which is useful to understand the processes at play as well as to better constrain the LGM sea-ice extent. We also refer to this more in detail below in the Specific Comments section.

Finally, the authors do not provide much discussion of their results in the context of the previously-mentioned two papers, and given their claim that "the multi-model mean of austral summer and winter sea ice cover seem to provide good estimates of LGM conditions" appears to be at odds with the results presented in those papers for earlier generation (Roche et al, 2012) and the same model (Marzocchi and Jansen, 2017) data, I would have liked a bigger exploration of this difference.

Marzocchi and Jansen (2017) did not provide a multi-model mean of LGM sea-ice extent and did not show the spatial extent of the seasonal sea-ice in the models and proxy records. Roche et al. (2012) show the spatial extent of the seasonal sea-ice in the models and proxy records, but do not provide a multi-model mean and use results from PMIP2, whereas we are using the PMIP3 and PMIP4 LGM results. Finally, the paleo-data compilation that we are presenting includes all the recent Southern Ocean sea-ice records thus providing an updated view on proxy estimates of summer sea-ice extent compared to the two studies mentioned above. These differences will clearly be indicated in the Introduction of our revised manuscript.

Introduction, lines 57-63:

"PMIP2 LGM simulations suggested that simulated LGM Antarctic sea-ice cover did not reflect the zonal variability nor the seasonality seen in proxy reconstructions (Roche et al, 2012). PMIP3 LGM simulations have also been analyzed, however not regionally, with results highlighting large inter-model differences in annual-mean, minimum and maximum Antarctic sea-ice area, and suggesting most PMIP3 models underestimate austral winter sea-ice cover in comparison to proxy data (Sime et al., 2016, Marzocchi and Jansen 2017). Therefore, a regional analysis of seasonal LGM sea ice simulated by PMIP3 models is lacking. Furthermore, no seasonal sea-ice analysis of PMIP4 simulations under LGM boundary conditions has been performed yet."

I found it hard to interpret these differences on my own due to the small size and indistinguishable lines in the figures and vague descriptions of methodologies.

We are unsure exactly what figures and which descriptions of the methods this is referring to, however, all figures have been significantly improved.

Specific Comments

intro raises connections w/ SAM – why not connect results to SAM?

The SAM mode of variability tells us about the strength and position of the westerlies in a non-periodic way over a wide range of timescales (generally from inter-annual to multi-annual). We don't believe it would make sense to connect this to our results since we are looking at the mean state of sea-ice concentration over centuries to millennia when other drivers played on the SO wind system.

Line 55: while Marzocchi and Jansen didn't focus on seasonality of PMIP3 simulations, they did plot the seasonal comparisons btw PI and LGM and discuss the characteristics of the seasonality in the models

To acknowledge the seasonal sea ice characteristics discussed in Marzocchi and Jansen we have added following sentence to our introduction, shown below:

Introduction, lines 58-61:

"PMIP3 LGM simulations have also been analyzed, however not regionally, with results highlighting large inter-model differences in annual-mean, minimum and maximum Antarctic sea-ice area, and suggesting most PMIP3 models underestimate austral winter sea-ice cover in comparison to proxy data (Sime et al., 2016, Marzocchi and Jansen 2017)"

Within the introduction, we are now more explicit in placing our work into context with the prior related research. We provide these changes in the general comments section of this author response letter.

Given zonal asymmetries in sea ice edge, usefulness of hemispheric, zonal averages is unclear

We agree with the reviewer that there are some limitations when using zonal averages for analysis due to zonal asymmetries, but we cannot think of a better way to analyze the data.

Line 67-69: Multiple simulations from same model were averaged over – what if they involved different components? How reflective is the ensemble mean of the behaviour of any one simulation? (look into GISS-E2-R)

We thank the reviewer for pointing this out as we should have provided more information on these models in the manuscript. Three different models submitted two different LGM simulations (CCSM4, GISS-E2-R, MPI-ESM-P). The CCSM4 model runs differed in their initial states while the GISS-E2-R and MPI-ESM-P runs differed slightly in their physics. With both LGM runs yielding similar sea-ice extent for each of the three models and following Sime et al (2016) methodology, we decided to take an average of each model so that we would have one set of data per model.

We have now added the following sentences to section 2.1 of the Methods, lines 77-80:

"Three models submitted two different simulations to PMIP3 (CCSM4, GISS-E2-R, MPI-ESM-P). These simulations differed because of a difference in the initial state (CCSM4), or small changes in the physics of the model (GISS-E2-R and MPI-ESM-P). Following Sime et al. (2016), we chose to average the simulations for the models who submitted two LGM runs, yielding one output per model."

The LOVECLIM simulations that were chosen are not representative of the PMIP3 models for a number of reasons, and the differences in their attributes, the reasons for their selection and the implications for the results are not discussed much here. As a result, the comparison feels artificial and not very instructive. Looking further into the simulations in Menviel et al (2017), I can make a guess as to why these ones were chosen, on the basis of the performance of their carbon cycle models. However, the fact that these simulations were performed with prognostic CO2 concentrations (via a carbon cycle model) rather than prescribed emissions/concentrations, and were spun up in a transient fashion from 35ka BP rather than just equilibrated to fixed LGM conditions as most of the PMIP3 simulations would have been, and had anomalous hosing applied in either or both of the Northern and Southern Hemispheres has bearing on the interpretation of the results. However, only the hosing is mentioned, and the implications of these design choices on the sea ice distributions are not discussed.

We understand the concerns of the reviewer and now provide a better motivation to include the LOVECLIM simulations. Given their skills in representing the LGM oceanic tracer distributions, it is interesting to include LOVECLIM1 (now referred to as weakNA) and LOVECLIM2 (now referred to as weakNA_AB) in the inter-model comparison. These simulations are however not included in the PMIP3 pool of simulations. In addition, we are now presenting a LOVECLIM simulation, which follows the PMIP4 protocol. These three LGM simulations performed with LOVECLIM allow us to compare the

inter-model to intra-model range in sea-ice extent. These three simulations, which feature very different oceanic circulation states, highlight that the inter-model spread is much smaller than the intra-model spread, suggesting that an improved model-data fit within one model framework cannot provide (much) information on the oceanic circulation state.

I am concerned about the fact that at least one lower-resolution model (e.g. LOVECLIM, whose ocean is nominally 3degx3deg) was interpolated onto a higher-resolution grid (1degx1deg) for plotting the results. This artificially inflates the apparent resolution of the results and thus encourages attempts to interpret changes at smaller spatial scales than the model provides as real. Whether the resolution of any PMIP3 models was artificially inflated in this way is not clear.

We understand the concerns of the reviewer, however, we need all models to be on the same grid for some aspects of our analysis. To clarify the method of interpolation, we have adjusted the sentence in section 2.1 of our Methods, lines 101-102:

"To ease the comparison, we used bilinear interpolation to standardize each model to a 1˚x 1˚grid with the CDO software (Climate Data Operators, Schulzweida et al. 2014)."

Lines 82-97 I am not a specialist in the interpretation of sea ice proxy records, so I found this section confusing. My main source of confusion lay in interpreting the uncertainties related to the apparently weak signals (differences between 1-3% in diatom assemblages), given the RMSEP values were 10%. I'm assuming the two percentages were not referring to the same quantities. I'm not expecting this paper to provide an overview of this method, assuming the references already provide that, but a sentence or two to make these results interpretable to non-specialists would be appreciated.

Indeed, the reviewer rightly understood that these are different quantities referring to different metrics. First, Gersonde and Zielisnski (2000) defined two qualitative metrics to infer the past position of winter and summer sea-ice edges. Using sediment trap series and core-tops in the Atlantic sector of the Southern Ocean they showed that >3% of *Fragilariopsis curta+cyclindrus* and >3% of *F. obliquecostata* best track the mean position of winter and summer edges, respectively. Although these quantities appear low, it is worth noting that these species thrive only at very low SSTs (-1 to 1°C) and high sea-ice concentrations (> 70% of WSIC; Armand et al., 2005) and are generally not present (0%) in the open ocean. Second, Crosta et al. (1998) developed a diatom-based transfer function that uses a greater number of sea-ice-related species. It allows us to quantify sea-ice duration or sea-ice concentration with an error on the calibration step, which represents the capacity of the transfer function to reconstruct the distribution of the modern fields of the parameters (here winter and summer sea-ice concentrations). This error is around 1 month per year for sea-ice duration and ~10% for sea-ice concentration.

We tried to rephrase and simplify the whole paragraph to make it better understandable to non-specialists:

Section 2.2 of Methods, line 103-117:

"The numerical simulations are compared to a compilation of 149 proxy records covering the LGM (See Table S1 in the Supplement, Allen et al. 2011; Benz et al. 2016; Ferry et al. 2015; Gersonde et al. 2005; Ghadi et al. 2020; Nair et al. 2019; Xiao et al. 2016). Quantitative SST was reconstructed at 138 locations, proxies for winter sea-ice presence or concentration were available at 149 locations and proxies for summer sea-ice presence were available at 132 locations. SSTs were derived from diatom-based transfer functions (Crosta et al., 1998; Esper and Gersonde, 2014a) while winter and summer sea-ice extent were derived either from the relative abundance of sea-ice indicator diatoms, respectively the *Fragilariopsis curta* group and *F. obliquecostata* (Gersonde et al., 2005), or diatom-based transfer functions whenever possible (Crosta et al., 1998; Esper et al., 2014b). Relative abundances of the indicator diatoms above 3% are thought to indicate the presence of sea ice over the core site (mean sea-ice extent north of the core site) while relative abundances between 1 and 3% suggest the episodic presence of sea ice over the core site (mean sea-ice edge south of the core site but maximum sea-ice edge north of the core site). In this study, we characterize the relative abundance of >3% as evidence of sea ice and the relative abundance between 1 and 3% as evidence for possible sea ice. Quantitative values were considered to indicate the presence of winter sea ice when they were above the root mean square error of prediction (RMSEP) on the validation models, generally around 10% for winter sea ice (Crosta et al., 1998; Esper et al., 2014b). Quantitative values were always below the RMSEP of ~10% for summer sea ice in the validation model."

Why were two months selected to define sea ice maxima and minima ? Was this based on prior knowledge, or analyses of the seasonal cycle in the models?

We chose two months to define sea-ice maxima and minima because when looking at the sea ice output, some of the models had minima and maxima that persisted longer than one month. We have gone back and checked the sea-ice extent for both PMIP3 and PMIP4 models. The difference between the one-month minima and the two-month minima is relatively small and does not affect our conclusions (see tables below):

PMIP3 models	2 month average summer SIE (10^6 km^2)	1 month average summer SIE (10^6 km^2)
CNRM	0.06	0.0049
GISS-E2-R	2.39	1.92
IPSL-CM5A-LR	2.41	2.14
MIROC-ESM-P	3.53	2.86
MPI-ESM-P	4.48	4.17
MRI-CGCM3	12.54	11.98
FGOALS-G2	13.62	10.66
CCSM4	27.46	25.40

PMIP4 Models	2 month average summer SIE (10^6 km_2)	1 month average summer SIE (10^6 km_2)
MIROC-ES2L	0.4	0.14
IPSL-CM5A2	2.48	2.29
MPI-ESM1-2	3.64	3.16
AWI-ESM-1	14.66	14.32
LOVECLIM	15.54	14.31
CESM1.2	23.63	21.42
UoT-CCSM4	33.45	32.33

To reflect this, we have added in section 2.3 of the Methods, lines 122-124:

"We note that using a two month average leads to a larger summer and a smaller winter sea-ice extent, compared to what would be obtained from a one month average. However, we believe that a two month average is more appropriate for a comparison with proxy records."

Lines 104-106 I'm not entirely clear on the methodology here. Firstly, were zonal averages performed in each ocean basin region over the latitudes of 15% sea ice concentrations for each longitude division (= 1 grid cell) or over sea ice concentrations in each latitude band (from which the 15% was then calculated)?

We zonally averaged over each latitude band, then determined at which latitude the models reached a 15% sea-ice concentration. For the models that did not reach 15% sea-ice concentration at any latitude after the full zonal average, we then zonally averaged over each ocean basin and determined the latitude (within that ocean basin) in which the model reached 15% sea-ice concentration. We amended the manuscript as follows:

Section 2.3 of Methods, lines 126-130:

"The sea-ice edge is defined as the 15% sea-ice concentration isoline. We calculate this by zonally averaging across all longitudes for each latitude band, then determining at which latitude the model simulates a minimum of 15% sea-ice concentration. For model simulations that do not reach 15% of sea-ice concentration in some regions of the SO, we average only over the remaining regions with sufficient sea-ice cover. For model simulations that do not reach 15% sea-ice concentration in any region, we define the latitude of their sea-ice edge as the latitude of the Antarctic coast."

And then, when defining an hemispheric sea ice latitude, am I correct in understanding that a zonal average over all longitudes was not calculated and instead, an average (weighted or unweighted?) was performed over the individual ocean basin regions? If this is correct, why was this method chosen?

We calculated a zonal average over all longitudes. This was made more clear in the modified methods section highlighted above.

Lines 118-119 Why does the multi-model mean lead to an asymmetric region of variance north and south of the mean lines? Are you suggesting that the simulations are distributed in a non-Gaussian way? Based on visual inspection, I wonder how the mean was calculated. I'm assuming the multi-model mean was calculated by averaging over the latitude of sea ice margin for each cell's longitude range from each run, but it doesn't seem to match what I see in the figure. For example, between 150 and 180degE, two, maybe 3, runs extend past the latitude of the outermost blue points. That leaves 6ish runs south of those points, but the multi-model mean lies north of the points. If the two runs were far north of the points, I would understand this, but that is not the case. Rereading the text, I wonder if the authors are suggesting they performed a multi-model average of sea ice concentration in every grid cell and then calculated the 15% concentration margin from the ensemble-mean distribution.

This is correct. We calculated the mean sea-ice concentration among all the models in every grid cell, then calculated the 15% concentration isoline of that multi-model mean. As mentioned in your next comment, models with high sea-ice concentration up until their margin can pull this mean in one way or the other, and therefore the multi-model mean does not fall perfectly in the middle between the models.

This is now clearly indicated in section 2.3 of our Methods, lines 134-135:

"To calculate the multi-model mean (MMM), we average sea-ice concentration over each grid cell for all models (PMIP3, PMIP4, and LOVECLIM sensitivity runs separately). We then calculate the 15% sea-ice concentration isoline of each MMM."

If those more extensive runs had very high sea ice concentration up until their margins, it might explain the position of the mean, but it's not clear to me how the standard deviations of the 15% line could be derived from this calculation. Since assigning a cutoff to sea ice at the 15% concentration is a non-linear operation, I would expect to get a more Gaussian multi-model distribution for the latitudes of the 15% lines than performing the calculation on the ensemble means of the sea ice concentrations directly.

Similarly, we calculate the standard deviation in each individual grid cell. Each model and MMM has a sea ice concentration value for each individual grid cell, and therefore we can simply calculate the standard deviation for each individual grid cell.

This is also clarified in section 2.3 of our Methods, lines 136-140:

"To calculate the standard deviation, we similarly compute a standard deviation value for each individual grid cell, before adding and subtracting that standard deviation (σ) of sea-ice concentration from the MMM for each grid cell. The +/- 1σ then represents the 15% sea-ice concentration isoline calculated from the MMM +/- 1σ. Notably, this creates a non-symmetric standard deviation isoline as each grid cell has its own MMM (and σ) value, calculated independently from any surrounding grid cells."

Given the zonal asymmetries in the summer data and their relative absence in the models (based on inspection of the multi-model mean in Figure 1 and the performance of PMIP2 models in Roche et al, 2012), I'd like to see the analysis performed for Figure 2 calculated on a regional basis, rather than based on hemispheric averages. Such an analysis would be more likely to bring out discrepancies between the models and the data than the hemispheric average.

We understand the comment of the Reviewer, however given the limited proxy data and the extent of our analysis, which includes both zonal averages and spatial correlations (e.g. Figures 1 and 3, Table 3), we think that a regional assessment of SST vs sea-ice extent would not add much value.

Technical Comments

Two different definitions for LGM time period provided in the abstract and text

This has been corrected.

I find it very difficult to distinguish between the colours of the different lines in Figure 1, because they are so thin. If thickening was not chosen because of confusion in the plot, I think the line labels can easily be dropped to simplify the figure.

We have thickened the contour lines and have gotten rid of the line labels to make the contour lines more clear and distinguishable.

Also, without any latitudes and longitudes marked on the plot and only the southern tip of S. America included as a georeference, it takes a lot of work for someone who is not intimately familiar with Antarctic geography to translate the descriptions in the text (most of which seemed to be in longitudes) to their locations on the figure and compare the latitude values stated in the text with those in the figure.

We thank the reviewer for pointing this out. We have included the latitude and longitude gridlines and labels in Figure 1.

It would be helpful for the boundaries of the ocean basin regions to be marked in one figure. Their names are clear, but precisely where the boundaries between the basins are drawn and their northern and southern extents would be helpful in interpreting the results.

We thank the reviewer for pointing this out. We have included labels that clearly indicate each ocean basin in Figure 1.

Line 110: Sentence fragment “The proxy”

This has been corrected.

Line 214: Typo MRI-ESM-P should be MRI-CGCM3

This has been corrected.