Dear reviewer,

We thank you for the constructive comments on the earlier version of the manuscript. We have revised our manuscript following the comments and our response is as follows.

Reviewer 2:

First, I apologize to the editor and authors for the delay in writing this review.

The study by Lin et al. investigates an important aspect of stand-alone ocean and sea ice models, which is the relation between the simulated fields and the atmospheric forcings. In simple terms, the paper tries to assess whether a new and arguably better forcing (JRA-55-do) leads to better simulation compared to an older forcing (CORE2). I find this a very interesting question, with important implications also for fully coupled model setups, and I would like to see more studies on these technical but rather important aspects. Congratulations to the authors for pursuing such an interesting problem.

That said, I found the authors' methodological approach unsatisfactory in providing a convincing answer to the problem. The analyses presented here are formally correct, but several central aspects have been almost completely neglected, as illustrated in my comments below. Otherwise, the paper is very well written and structured, but I have some suggestions for improving figures and tables, which sometimes are not adequate.

In summary, I have several major concerns that in my opinion should be addressed before this manuscript is considered for publication. I hope the authors find these helpful for improving their study.

MAJOR COMMENTS:

My biggest concern is that this manuscript does not consider the tuning of the systems analyzed. Let me start by acknowledging that finding out specific details about namelist parameters and other technical information is not trivial when dealing with CMIP-type simulations. Nevertheless, this aspect is important for a study of this kind and cannot be neglected. Specifically, I am wondering whether the CORE2 and JRA-55-do simulations were run with the same model setup, or if the model was specifically tuned for a certain forcing. In my view, tuning is a fundamental step that, given the under-constrained nature and spatiotemporal variability of the sea ice model parameters, must be performed to accommodate a model configuration to a specific forcing. I would argue that the best setup for a study like this would be one where each simulation is optimized to obtain the best compatibility with a set of observations, given a specific atmospheric forcing (i.e., different parameters for CORE2 and JRA-55-do). If this is not the case and an identical model setup is adopted for all atmospheric forcings, we should at least be made aware of whether the model parameters

have been tuned under CORE2 or JRA-55-do, if tuned at all for an OMIP setup. For example, if a model configuration has been tuned under JRA-55 and this same configuration is then run under CORE2, it is not surprising that one would outperform the other. Based on the information currently in the paper, we cannot say anything regarding the previous considerations, which, in my view, is a substantial limitation that should be addressed to pursue the research question from the right angle.

Answer: Thanks for your comment. We recognize that tuning is a key aspect in climate models. The design of the CMIP6 OMIP simulations has been organized by the World Climate Research Programme (WCRP) Climate Variability and Predictability (CLIVAR) Working Group on Ocean Model Development Panel (OMDP), and ongoing research collaboration is done through the OMDP to develop OMIP2 (Griffies et al., 2016). Importantly, and to our understanding, the same configuration is used under two different atmospheric forcing datasets as mentioned in Tsujino et al. (2020). It is possible that some groups could have tuned for OMIP2 and then used the same setup for OMIP1, so part of the improvement with OMIP2 could be due to this experimental setup choice. From our analysis, the differences in the atmospheric forcing are transferred to the modeled surface fluxes and contribute to the improved ice concentration simulation (see our response to major comment 4, Fig. 5 below). Thus, the OMIP1 and OMIP2 datasets allow singling out how atmospheric forcing differences are translated into simulated sea ice differences.

Action: The information has been added in the revised introduction when introducing the OMIP models. We also added one sentence in the discussion part to address the possible effects from tuning.

I think the manuscript lacks an in-depth description of the model components used in the three systems considered, which could be helpful for a more detailed interpretation of the results. The only information available in Lin et. al. 2021 is that two systems employ different versions of CICE as sea ice model components (CMCC-CM2 and NorESM2). By digging in the MRI-ESM2 model description paper we discover the sea ice component of MRI. COM4.4 is also based on CICE. Given the modularity of CICE, the model version tells us nothing about the physical configuration used by each modeling center. A better description of the model configurations would allow linking differences in the model response to the reanalyses, to differences in the specific physics used. For example, I suspect that different radiative schemes lead to very different sea ice concentration conditions in summer.

Answer: We provided some information on the sea ice models in Table 1. For the MRI-ESM2-0 model, the sea ice component is MRI.COM4.4 and the thermodynamics are based on Mellor and Kantha (1989), which are different from the thermodynamics in CICE (Bitz and Lipscomb, 1999). We have added the EC-Earth3 and the MIROC6 models in the revised manuscript (see our response to the next comment, Figs. 1 to 4) and they are based on the LIM3 and COCO4.9 sea ice models, respectively. The radiative schemes in EC-Earth3 and NorESM2-LM are different, which can affect the summer sea ice concentration conditions in EC-Earth3

(Fig. 1) and NorESM2-LM (Fig. 6 below). We cannot have much more information on the radiative schemes of MIROC6 and MRI-ESM2-0.

Action: This sea ice model information has been added to the revised section 2. We have added a sentence in the manuscript recommending that, in such inter-comparison exercises, all the specificities/namelists of the sea ice models used should systematically be reported. These comments have been added in the discussion section.

Table 1. The details of five CMIP6-OMIP sea ice models evaluated in the study.

Model	Sea Ice Model	Sea Ice Component	References
CMCC-CM2-SR5	CICE4	-Energy-conserving thermodynamics; -Elastic-Viscous-Plastic (EVP) rheology; -Ice Thickness Distribution (ITD) with five thickness categories; 1 layer of snow and 4 layers of ice; -A Delta-Eddington multiple-scattering shortwave radiation treatment	Cherchi et al. (2019)
EC-Earth3	LIM3	-Energy-conserving halo-thermodynamics; -2-D EVP; -ITD with five thickness categories; 1 layer of snow and 2 layers of ice; -The impact of melt ponds is implicitly accounted for through imposed changes on the albedo activated when the surface temperature is 0°C.	Döscher et al. (2022)
MIROC6	COCO4.9	-Energy-conserving thermodynamics on only one layer for sea ice; -EVP; -ITD with five thickness categories;	Tatebe et al. (2019)
MRI-ESM2-0	MRI.COM4.4	-Energy-conserving thermodynamics based on Mellor and Kantha (1989); -EVP; -ITD with five thickness categories;	Yukimoto et al. (2019)
NorESM2-LM	CICE5.1.2	-Mushy-layer thermodynamics with prognostic sea ice salinity; -EVP; -ITD with five thickness categories; 3 layers of snow and 8 layers of ice; -A Delta-Eddington multiple-scattering shortwave radiation treatment, with melt ponds modeled on level, undeformed ice.	Seland et al. (2020)

When designing the study, the authors limit the number of models considered based on the availability of diagnostic variables: the dynamic and thermodynamic sea ice concentration tendencies. The main result of this choice is to limit the analysis to systems based on different flavors of one sea ice model (CICE). I think including more models might be interesting and more in line with the scope of CMIP6 and OMIP. The tendencies analysis, which is not central to this study, can be limited to the system with the appropriate diagnostic variables.

Answer: We have added EC-Earth3 and MIROC6 models for the sea ice concentration (Fig. 1) and ice drift evaluation (Figs. 2 to 4). They are based on the LIM3 and COCO 4.9 sea ice models, respectively. Including

these two models do not change our conclusions on the improved sea ice concentration and ice drift simulations and their connections to the atmospheric forcings. These two models did not provide the dynamic and thermodynamic sea ice concentration tendencies nor the surface fluxes.

Action: We have added Figs. 1 to 4 below to the revised manuscript as new Figs. A2, A8-A10 and revised the main text to include these two model results.

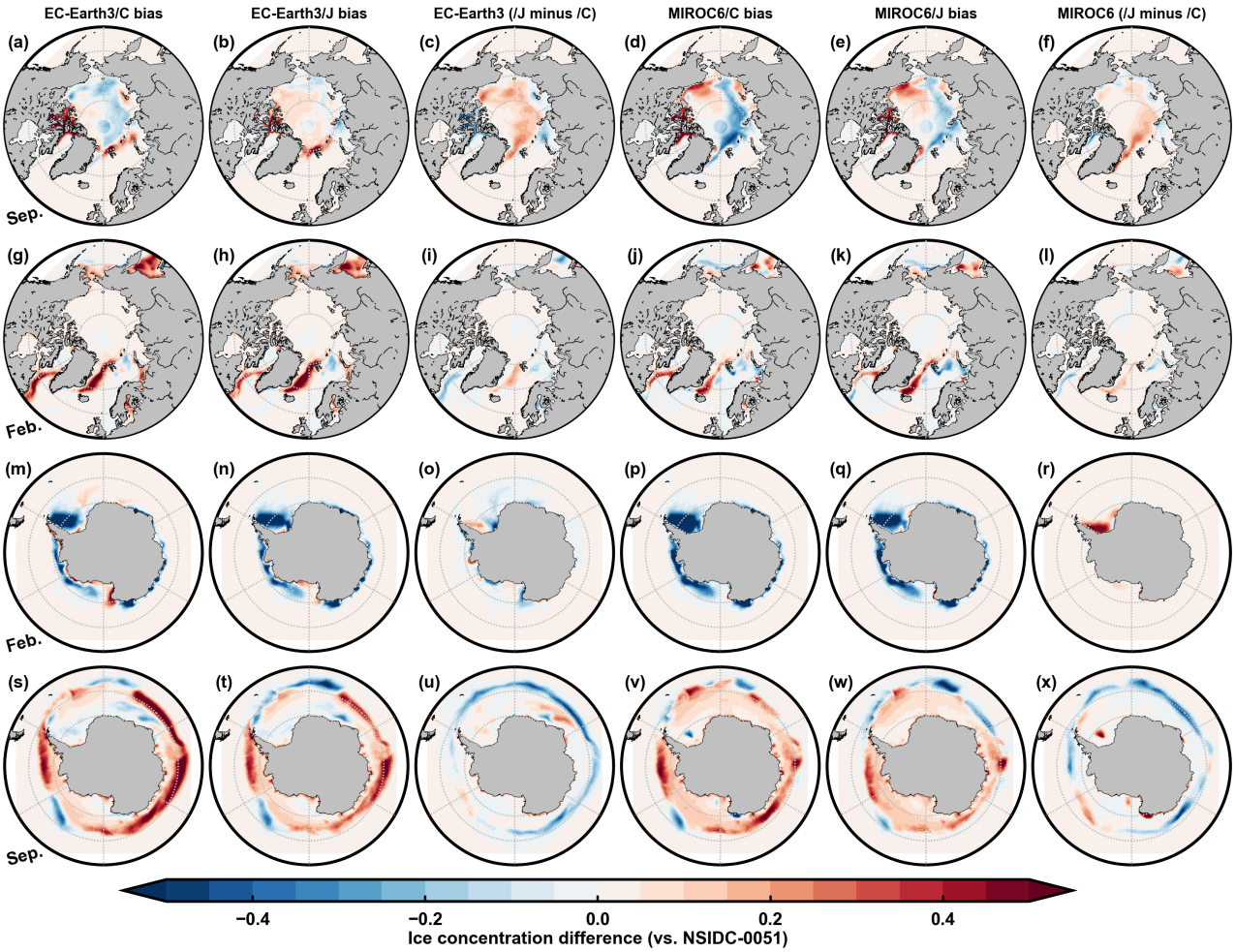

Figure 1. 1980-2007 September and February mean Arctic (a to l) and Antarctic (m to x) sea ice concentration differences between EC-Earth3/C and NSIDC-0051 (first column), EC-Earth3/J and NSIDC-0051 (second column), EC-Earth3/J and EC-Earth3/C (third column), MIROC6/C and NSIDC-0051 (fourth column), MIROC6/J and NSIDC-0051 (fifth column), and MIROC6/J and MIROC6/C (sixth column).

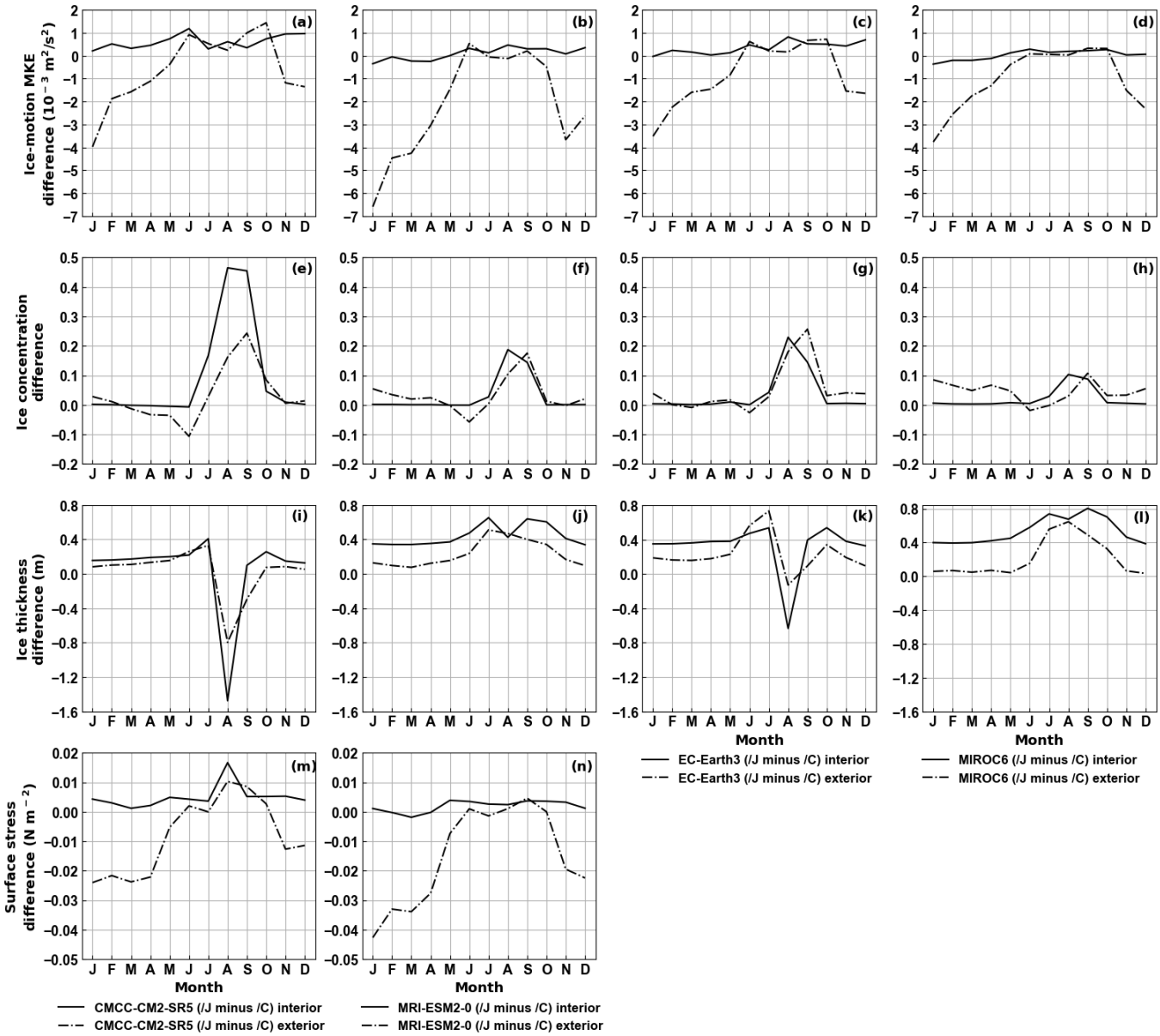

Figure 2. 2003-2007 monthly mean and spatially averaged Arctic ice kinetic energy (MKE) (a to d), ice concentration (e to h), ice thickness (i to l) and surface wind stress (m and n) differences between model/J and model/C. The first to fourth columns correspond to CMCC-CM2-SR5, MRI-ESM2-0, EC-Earth3, MIROC6 model results, respectively. The solid and dashed lines are spatial averages on the regions with ice concentration larger (interior) and smaller (exterior) than 80% in NSIDC-0051, respectively.

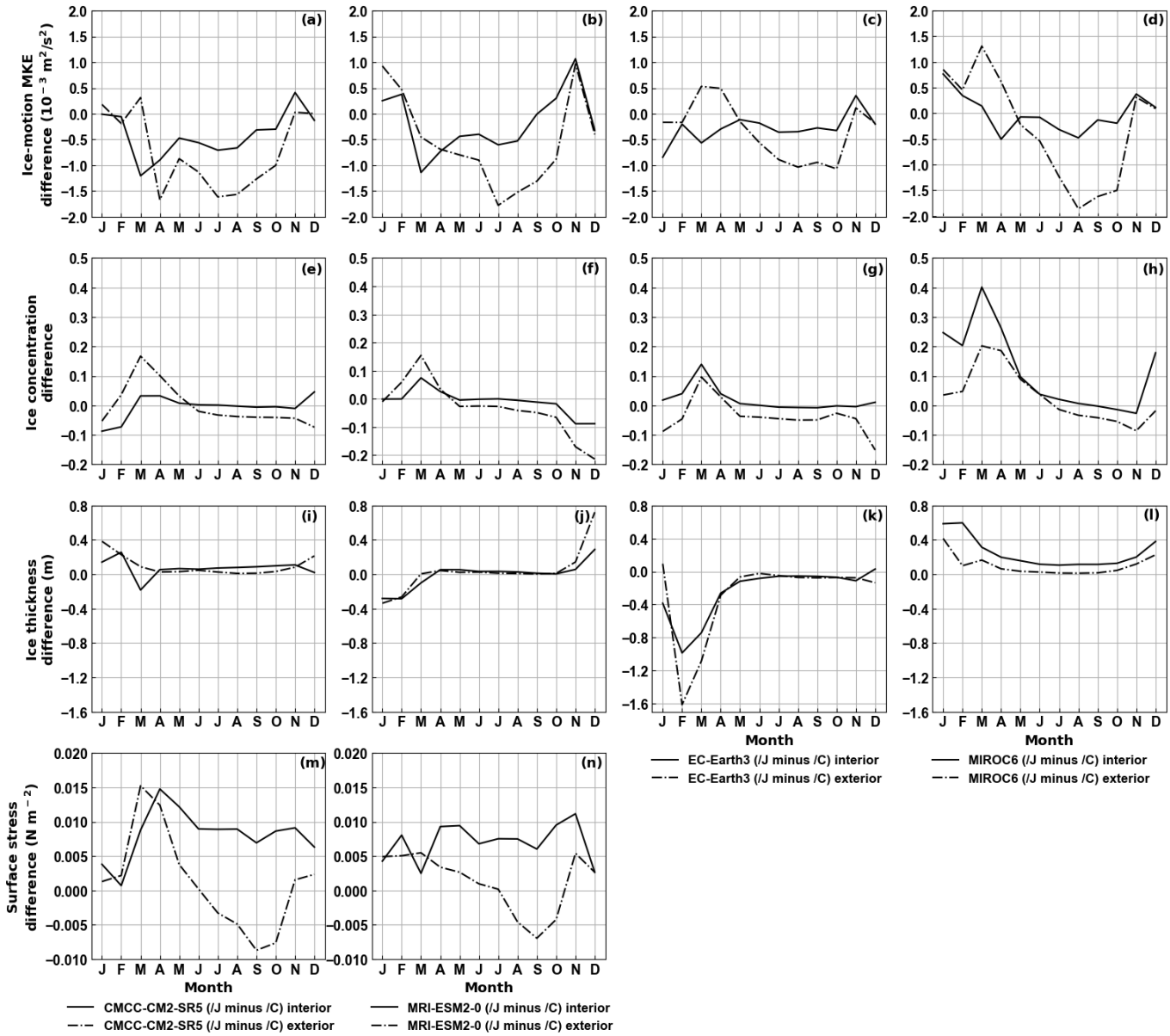

Figure 3. Same as Fig. 2 but for the Antarctic.

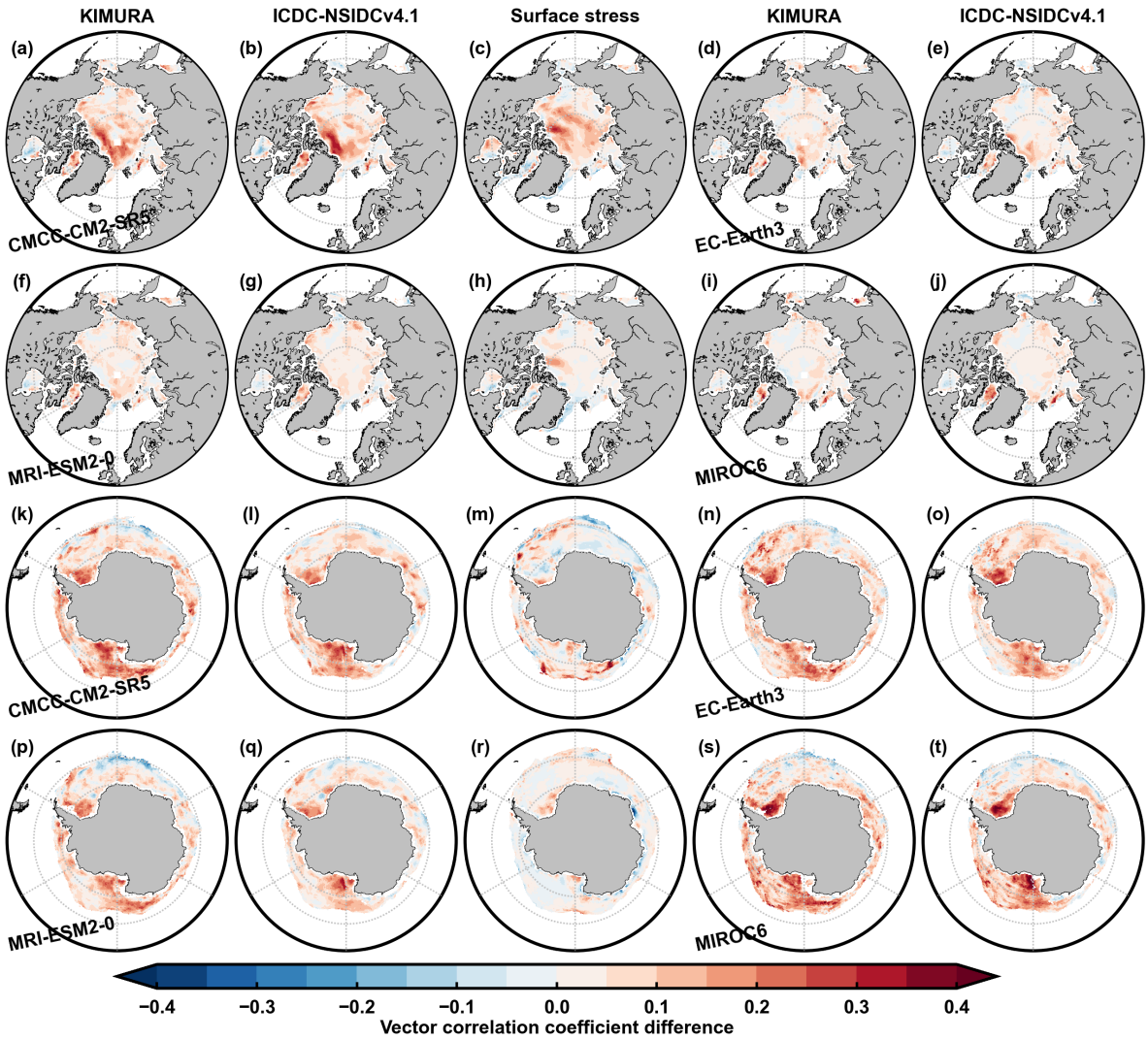

Figure 4. Differences of significant vector correlation coefficients during 2003–2007 at a level of 99% between model/J and model/C in the Arctic and Antarctic. The first, second, fourth and fifth columns are significant vector correlation coefficients between modeled ice drift and KIMURA/ICDC-NSIDCv4.1 data, and the third column are significant vector correlation coefficients between modeled (CMCC-CM2-SR5 and MRI-ESM2-0) ice drift and surface stress.

The paper lacks a direct comparison of the reanalysis fields. Also, the CORE2 and JRA55-do forcings have both been bias-corrected in the Arctic to avoid unrealistic model behaviors (e.g. too little sea ice in summer). The correction follows the work of Large and Yeager (2009). How does this bias correction impact your results? To what extent are the forcing converging?

Answer: Thanks for your suggestion. The surface air temperature, specific humidity, downward shortwave and longwave radiation fluxes during melting months, and wind speed during freezing months in COREII and JRA55-do are shown in Fig. 5 below. These months are shown because in general the ice concentration simulations are improved from OMIP1 to OMIP2 in summer due to surface heat flux changes and in winter due to the wind stress changes. More information can be found in Figs. 1 to 4, A1 to A4 of our previous manuscript. Compared to COREII, the downward shortwave radiation flux and specific humidity in JRA5-do in the central Arctic Ocean and the coastal region of the western Weddell Sea (Figs. 5g, h, q, r) are

smaller, the downward shortwave radiation flux in the Canadian Arctic Archipelago (CAA) and central Weddell Sea (CWS) regions and the air temperature in the CAA region are larger (Figs. 5h, r, f), and the surface wind speed on Antarctic sea ice in the inner part of the exterior region from 70° to 180°E is weaker (Figs. 5t). These differences in the atmospheric forcing are transferred to the modeled surface fluxes and contribute to the improved ice concentration simulation in those regions. Compared to OMIP1 simulations, the downward shortwave radiation flux and latent heat flux in OMIP2 in the central Arctic Ocean and the coastal region of the western Weddell Sea are smaller, the downward shortwave radiation flux in the CAA and CWS regions and the sensible heat flux in the CAA region are larger, and the surface wind stress on Antarctic sea ice in the inner part of the exterior region from 70° to 180°E is weaker (Figs. 4, 5, A4, A5, A6 in the previous manuscript).

The bias-corrected processes are done in both atmospheric forcings (Tsujino et al., 2018). We can find the JRA55-do forcing is much improved in the Arctic downward shortwave radiation fluxes and this improvement is transferred to the OMIP2 and contribute to an improved ice concentration simulation.

Action: We have added Fig. 5 below to the revised manuscript as a new Fig. 6 and added the explanation before on how the differences in the atmospheric forcings are transferred to the model results in the main text.

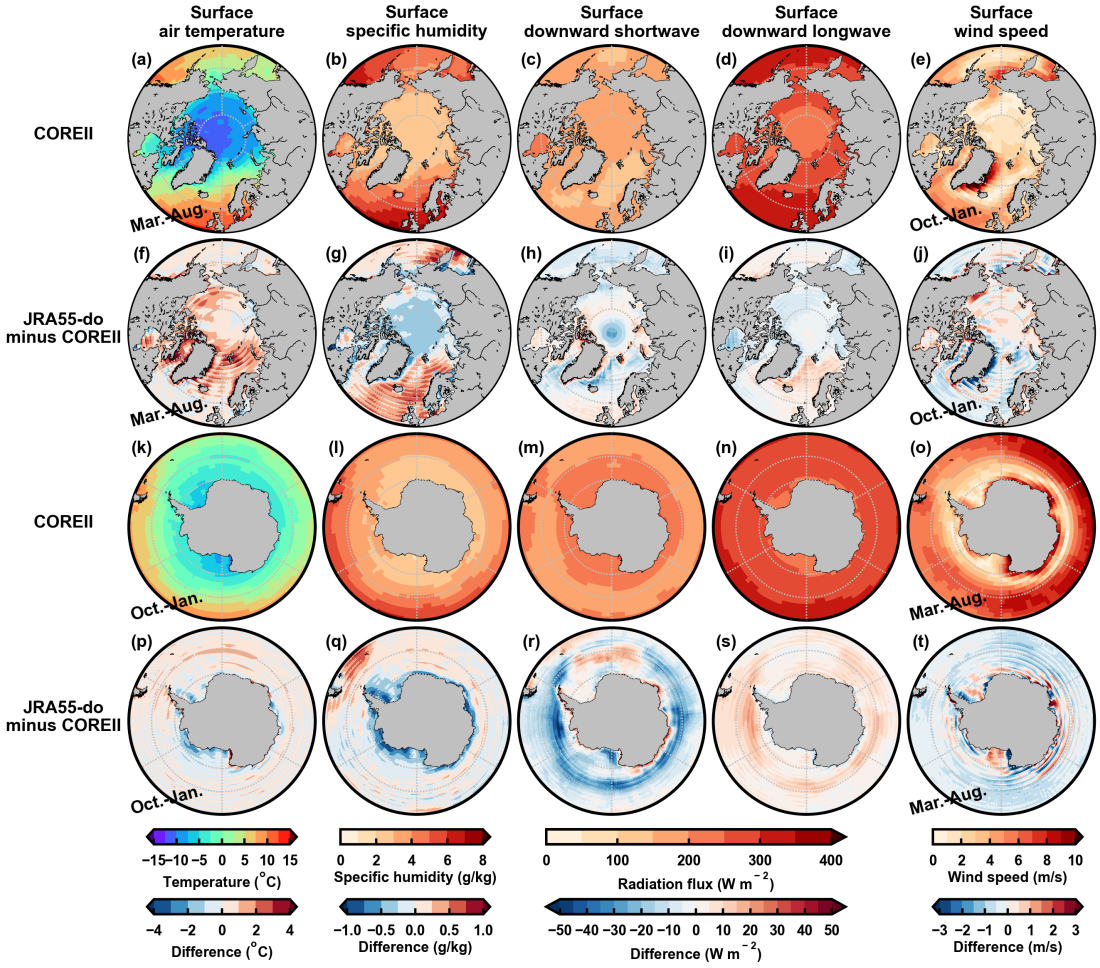

Figure 5. 1980-2007 March-August mean Arctic and October-January mean Antarctic surface air temperature (first column) and specific humidity (second column), downward shortwave (third column) and longwave radiation fluxes (fourth column), as well as October-January mean Arctic and March-August mean Antarctic surface wind speed (fifth column). The first and third rows correspond to COREII, and the second and fourth rows are differences between JRA55-do and COREII.

I believe more observations should be included in the analysis to allow a correct interpretation of the results. We know that in the Arctic different observational datasets are not always in agreement with each other and that identifying the best product is not obvious. I am surprised that this has not been done given that SITool includes multiple observational datasets for sea ice concentration, thickness, and drift. Illustrating the differences between reanalysis in relation to different observational products would certainly be an interesting addition to the study.

Answer: Thanks for pointing this out. Two observational datasets are included for the sea ice concentration (NSIDC-0051 and OSI-450), ice drift (KIMURA and ICDC-NSIDCv4.1) and thickness (Envisat and Icesat) evaluations (Figs. 6 to 9) in the revised manuscript. For the ice concentration and ice drift, observational uncertainties are small compared to the model biases. The conclusions on the improved ice concentration and ice drift simulations in OMIP2 do not change by comparing to different observational products (Figs. 6, 7 and 9). The ice thickness observations during 2003-2007 are restricted to a few months per year in both

Envisat and ICESat datasets (Figs. 7 and 8). The Envisat data includes ice thickness from November to April for the Arctic with coverage up to 81.5^{o}N and May to October for the Antarctic from 2003. The ICESat data includes 13 measurement campaigns for the Arctic and 11 for the Antarctic during 2003–2007, and these campaign periods are limited to the months of February–March, March–April, May–June, and October–November, with each campaign lasting roughly 33 d. The comparisons between individual models and the two observational references are thus restricted to these months when data are available. In general, the modeled sea ice thickness is close to the ICESat dataset, while modeled ice thickness is too thick in the Arctic and too thin in the Antarctic compared to Envisat data.

Action: These figures are added as new Figs. 1, 7, 8 and 9 in the revised manuscript. We have added these data information and the comparison results in the main text.

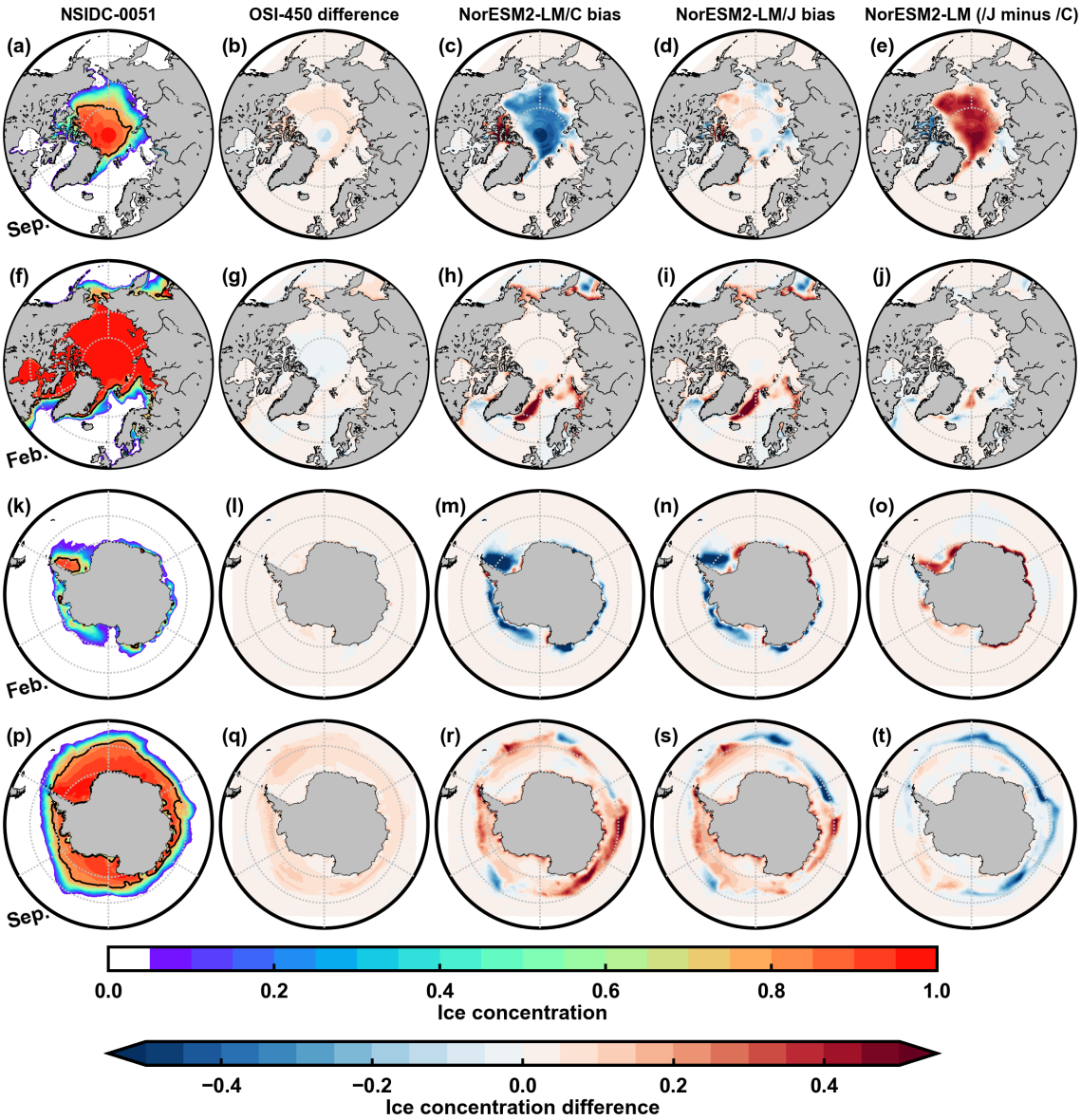

Figure 6. 1980-2007 September and February mean Arctic (a to j) and Antarctic (k to t) sea ice concentration from the NSIDC-0051 data (first column), differences between OSI-450 and NSIDC-0051 (second column), NorESM2-LM/C and NSIDC-0051 (third column), NorESM2-LM/J and NSIDC-0051 (fourth column), and NorESM2-LM/J and NorESM2-LM/C (fifth column). The black lines are contours of 80% concentration (a, f, k and p), which delineate the interior and exterior domains to compute spatial averages in Table 1.

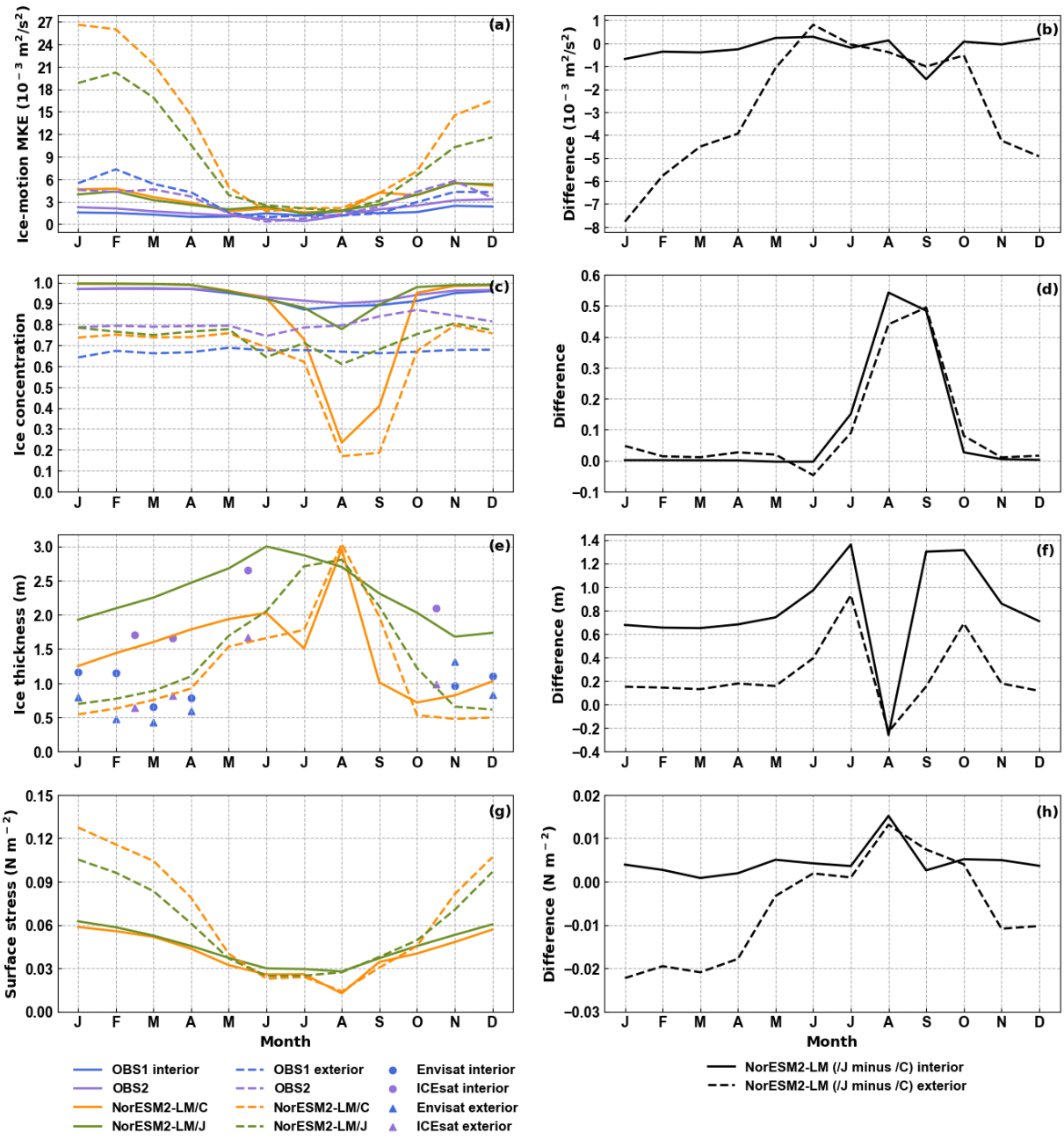

Figure 7. 2003-2007 monthly mean and spatially averaged Arctic ice kinetic energy (MKE) (a), ice concentration (c), ice thickness (e) and surface wind stress (g) from observations (blue and purple), NorESM2-LM/C (orange) and NorESM2-LM/J (green). Two observational datasets are included for ice MKE (KIMURA and ICDC-NSIDCv4.1), concentration (NSIDC-0051 and OSI-450) and thickness (Envisat and Icesat). The Envisat ice thickness data is provided from November to April and the coverage is limited up to 81.5°N. The measurement campaigns of Icesat ice thickness is for the months of February–March, March–April, May–June, and October–November, with each campaign lasting roughly 33 d. The solid and dashed lines are spatial averages on the regions with ice concentration larger (interior) and smaller (exterior) than 80% in NSIDC-0051, respectively. The differences between NorESM2-LM/J and NorESM2-LM/C ice MKE (b), ice concentration (d), ice thickness (f) and surface wind stress (h) in the interior (black solid) and exterior (dashed) regions are shown.

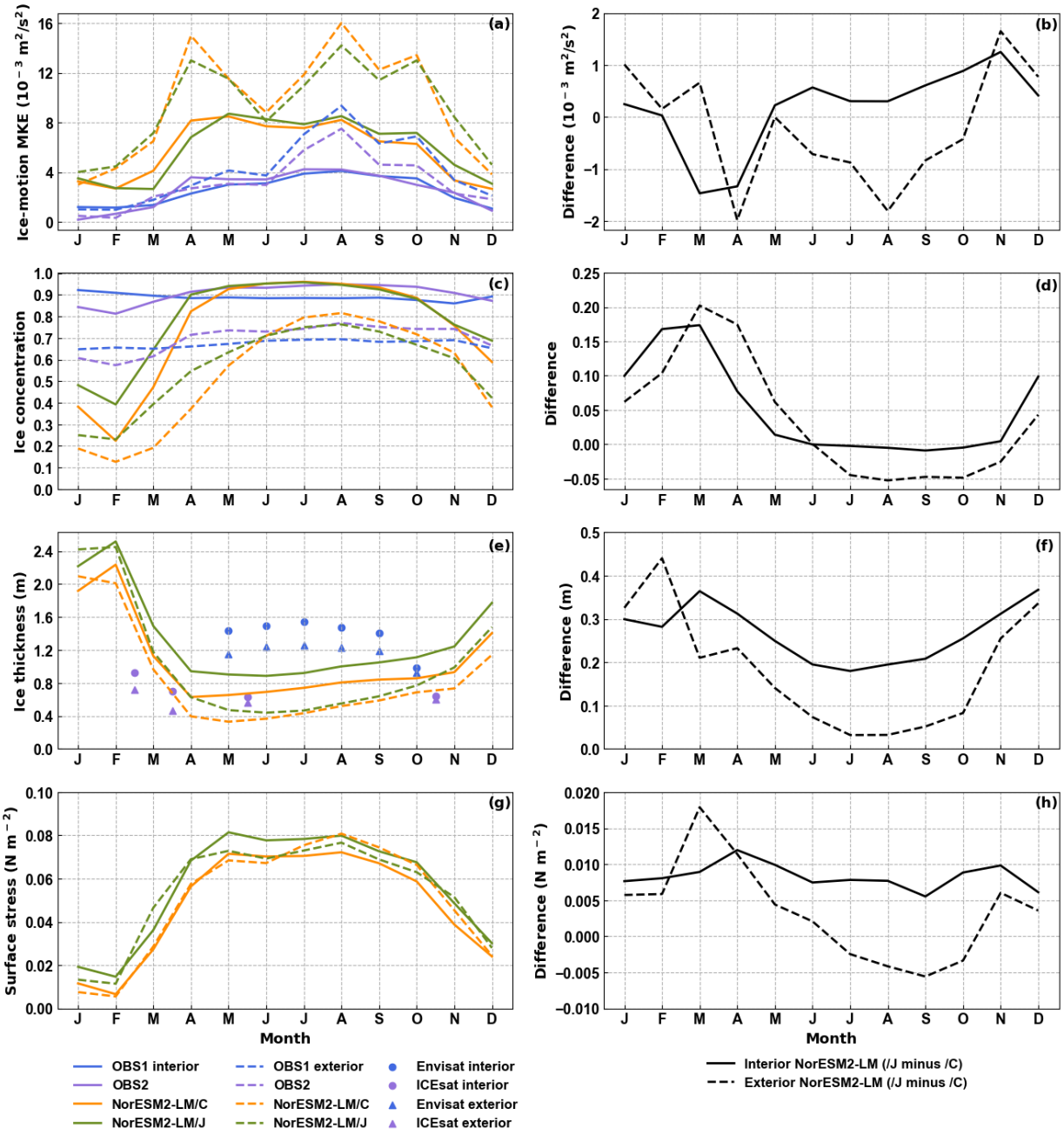

Figure 8. Same as Fig. 7 but for the Antarctic. The Envisat ice thickness data is provided from May to October.

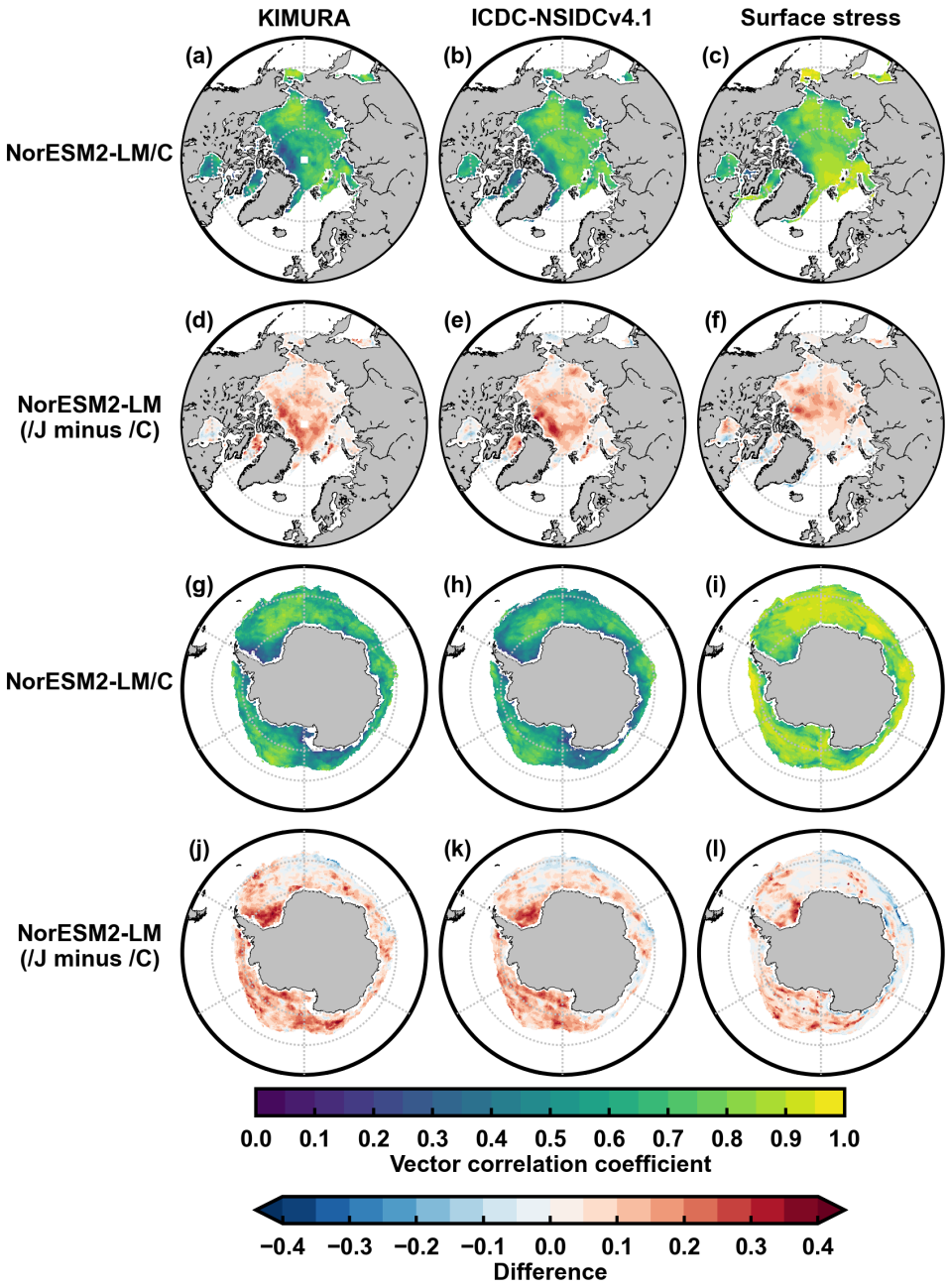

Figure 9. The significant vector correlation coefficients during 2003–2007 at a level of 99% between modeled ice drift (NorESM2-LM/C) and two observational data (KIMURA and ICDC-NSIDCv4.1), respectively, and between NorESM2-LM/C modeled ice drift and surface wind stress in the Arctic (a to c) and Antarctic (g to i). The second and fourth rows are the vector correlation coefficient differences by changing the modeled ice drift from NorESM2-LM/C to NorESM2-LM/J.

FIGURES AND TABLES:

Tables 1 and 2: I find this table very hard to read and overcrowded. It might be a personal preference, but I could read these data more easily when converted into plots and reorganized. In particular, the use of parenthesis and bold and italic font is confusing. I suggest fully rethinking this, and possibly working with colors/symbols instead of font styles.

Thanks for your comments. We have modified the tables and used different colors to mark them.

Figure 1 and beyond: Showing the results of just one model is, in my opinion, limiting. I understand the authors are concerned about having too many display items in the manuscript but storing relevant material in the appendix is not necessarily a solution. If the space is a concern, why not report only the maps with the difference between the two forcings in the main text, while moving the bias to the appendix? Also, the addition of the observed sea ice concentration in Fig 1 is not very insightful, or at least not a priority for the panel. The same is true in the following figures. Again, this is a suggestion and I realize it depends on personal preferences.

Thanks for your comment. We prefer to include the bias with respect to the observations in the figures. By providing the observational ice concentration and the bias with respect to the observations, we can identify where are the errors (positive and negative values). Then, the differences between OMIP2 and OMIP1 can inform us where are the improvements.

Figures 6 and 7: The color choice of the bar plots is very unhappy. Please consider differentiating the colors of the interior vs. exterior bars.

Thanks for pointing this out. We have changed the bar plots to lines and the interior and exterior regions are shown in solid and dashed lines, respectively, as shown in Figs. 7 and 8.

Yours sincerely,
Xia Lin, François Massonnet, Thierry Fichefet, Martin Vancoppenolle