SITool (v1.0) - a new evaluation tool for large-scale sea ice simulations: application to CMIP6 OMIP

Xia Lin[1, 2], François Massonnet [1], Thierry Fichefet [1], Martin Vancoppenolle[3]

[1]Georges Lemaître Centre for Earth and Climate Research, Earth and Life Institute, Université catholique de Louvain, Louvain-la-Neuve, 1348, Belgium

[2]Southern Marine Science and Engineering Guangdong Laboratory (Zhuhai), Zhuhai, 519000, China

[3]Laboratoire d'Océanographie et du Climat, CNRS/IRD/MNHN, Sorbonne Université, 75252, Paris, France

Correspondence to: Xia Lin (xia.lin@uclouvain.be)

Abstract. The Sea Ice Evaluation Tool (SITool) described in this paper is a performance metrics and diagnostics tool developed to evaluate the skill of Arctic and Antarctic model reconstructions of sea ice concentration, extent, edge location, drift, thickness, and snow depth. It is a Python-based software and consists of well-documented functions used to derive various sea ice metrics and diagnostics. Here, the SITool version 1.0 (v1.0) is introduced and documented, and is then used to evaluate the performance of global sea ice reconstructions from nine models that provided sea ice output under the experimental protocols of the Coupled Model Intercomparison Project 6 (CMIP6) Ocean Model Intercomparison Project with two different atmospheric forcing datasets: the Coordinated Ocean-ice Reference Experiments version 2 (CORE-II) and the updated Japanese 55-year atmospheric reanalysis (JRA55-do). Two sets of observational references for the sea ice concentration, thickness, snow depth, and ice drift are systematically used to reflect the impact of observational uncertainty on model performance. Based on available model outputs and observational references, the ice concentration, extent, and edge location during 1980-2007, as well as the ice thickness, snow depth, and ice drift during 2003-2007 are evaluated. In general, model biases are larger than observational uncertainties and model performances are primarily consistent compared to different observational references. By changing the atmospheric forcing from CORE-II to JRA55-do reanalysis data, the overall performance (mean state, interannual variability and trend) of the simulated sea ice areal properties in both hemispheres, as well as the mean ice thickness simulation in the Antarctic, the mean snow depth and ice drift simulations in both hemispheres are improved. The simulated sea ice areal properties are also improved in the model with higher spatial resolution. For the cross-metric analysis, there is no link between the performance in one variable and the performance in another. The SITool is an open-access version-controlled software that can run on a wide range of CMIP6 compliant sea ice outputs. The current version of SITool (v1.0) is primarily developed to evaluate atmosphere-forced simulations and it could be eventually extended to fully coupled models.

1 Introduction

Most regional and global climate models now include an interactive sea ice model, reflecting the reality that sea ice plays a fundamental role in the polar environment, by influencing air-sea-ice exchanges, atmospheric and oceanic processes, and climate change. Large inter-model spread exists in the performance of sea ice simulations in the Coupled Model Intercomparison Project 5 (CMIP5) for both the Arctic and Antarctic (Massonnet et al., 2012; Stroeve et al., 2012, 2014; Turner et al., 2013; Zunz et al., 2013; Shu et al., 2015). Some improvements are identified in the CMIP6 models: (1) a more realistic estimate of sea ice loss for a given amount of CO_2 emissions and global warming in the Arctic (Notz et al., 2020), (2) reduced inter-model spreads in summer and winter ice area and improved ice concentration distribution in the Antarctic (Roach et al., 2020), (3) lower inter-model spreads in the mean state and trend of both the Arctic and Antarctic ice extents (Shu et al., 2020). However, sea ice projections and evaluations are still not systematic and to date, no tool allows precise tracking of sea ice model performance through time from one version to the next. The Earth System Model Evaluation Tool (ESMValTool) has been developed for routine evaluation of climate model simulations in CMIP including many components of the Earth system (Eyring et al., 2016, 2020). It is an efficient tool to obtain a broad view on the overall performance of a climate model, and it provides sea ice diagnostics on the ice concentration and extent, as well as relationships between sea ice variables. In addition to sea ice diagnostics, the Sea Ice Evaluation Tool (SITool) introduced in this paper provides systematic sea ice metrics for assessing large-scale sea ice simulations from various aspects.

The SITool has been designed to describe inter-model differences quantitatively and to help teams managing various versions of a sea ice model, detecting bugs in newly developed versions, or tracking the time-evolution of model performance. The SITool quantifies the performance of sea ice model simulations by providing systematic and meaningful sea ice metrics and diagnostics on each sea ice variable with thorough comparisons to a set of observational references. Arctic and Antarctic performance metrics and diagnostics on ice coverage, drift, thickness and snow depth are provided from seasonal to multi-decadal time scales whenever observational references are available. These sea ice metrics give a detailed view of sea ice state and highlight major deficiencies in the sea ice simulation. The SITool is written in the open-source language Python and distributed under the Nucleus for European Modelling of the Ocean (NEMO) standard tools. The SITool is provided with the reference code and documentation to make sure the final results are traceable and reproducible.

Here, the SITool version 1.0 (v1.0) is applied to evaluate the performances of Arctic and Antarctic historical sea ice simulations under the experimental protocols of the CMIP6 Ocean Model Intercomparison Project (OMIP, Griffies et al., 2016). OMIP provides global ocean-sea ice model simulations with a prescribed atmospheric forcing, which gives the opportunity to intercompare sea ice model performance under fully controlled conditions. In OMIP, two streams of experiments were

carried out: OMIP1, forced by the Coordinated Ocean-ice Reference Experiments version 2 interannual forcing (CORE-II, Large and Yeager, 2009), and OMIP2, forced by the updated Japanese 55-year atmospheric reanalysis (JRA55-do, Tsujino et al., 2018). The OMIP protocol ensures a close experimental setup among the different models. Models were run with both atmospheric forcings, when possible, to identify and attribute the influences of changed atmospheric forcings on sea ice characteristics. Tsujino et al. (2020) and Chassignet et al. (2020) evaluated the impact of atmospheric forcing and horizontal resolution on the global ocean–sea ice model simulations based on the experimental protocols of OMIP provided by model groups participated in this intercomparison project. Their studies focused on the evaluation of ocean components from sea surface height, temperature, salinity, mixed layer depth, kinetic energy to circulation changes. Some aspects of sea ice simulations are assessed in both hemispheres relative to an observational dataset. Tsujino et al. (2020) provide spatial maps of the 1980–2009 mean ice concentration and time series of ice extent in summer and winter, and Taylor diagrams of the interannual variations of ice extent under CORE-II and JRA55-do forcings. Chassignet et al. (2020) show spatial maps of the 1980–2018 mean ice concentration and ice thickness in summer and winter, and time series of annual mean ice extent and ice volume under different horizontal resolutions. In this paper, we focus on the sea ice in OMIP simulations available from the Earth System Grid Federation in a more systematic manner, including more sea ice variables (e.g., ice edge location, snow depth, and ice drift). The performance metrics and diagnostics (spatial maps and/or time series diagrams) for each ice variable are provided compared to two sets of observational references when data is available to appreciate the importance of observational uncertainty in the assessment.

This paper is organized as follows. The SITool (v1.0) with the details of sea ice metrics and diagnostics is described in section 2. The CMIP6 OMIP models and observational references are introduced in section 3. In section 4, the application of the SITool (v1.0) to CMIP6 OMIP and the results of the model performance are presented and discussed. Finally, conclusions and discussion are provided in section 5. Appendix A presents some additional sea ice diagnostics. The source code of the SITool (v1.0) used to assess the model skills is publicly available in the repository as shown in the section on "Code and data availability".

2 Overview of SITool (v1.0)

A schematic overview of the SITool (v1.0) workflow and its application in evaluating the CMIP6 OMIP model performance is shown in Fig. 1. The input sea ice data from model outputs and observations are detailed in section 3. The methods of the metrics calculation are discussed below in section 2.1 followed Massonnet et al. (2011) with some modifications. Namely, (1) more observational references are used to calculate the observational errors and the incorporation of observational errors is a prerequisite to do the

comparisons here, (2) ice edge location and snow depth metrics are included, (3) the method to calculate the vector correlation coefficient is updated. The SITool (v1.0) also produces additional sea ice diagnostics (spatial maps and time series diagrams) to help understand why metrics vary from one dataset to the next. Table 1 provides an overview of the diagnostic fields along with input variables,

125 output results and corresponding figures in this paper, python scripts in the repository, and comments. All the sea ice data from model outputs and observational references are regridded to the polar stereographic 25 km resolution grid using a kd-tree (k-dimensional, Bentley, 1975) nearest neighbour interpolation method provided by a python package (a component of the SITool workflow). The kd-tree is a binary search tree with a two-dimensional spatial index structure for use in this study. The

130 interpolation yields less than 5% error for each sea ice variable (not shown), which indicates that the results are not sensitive to the interpolation method used here. This interpolation allows point-by-point comparison and avoids the systematic bias of sea ice extent under different grids, due to differences in land-sea masks.

135 **Figure 1. Schematic overview of the SITool (v1.0) and its application to the CMIP6 OMIP model evaluation.**

Table 1. Overview of the diagnostic fields along with input variables, output results and corresponding figures in this paper, python scripts in the repository, and comments.

Diagnostic fields	Input variables	Output results (Figure(s))	Python scripts	Comments
Sea ice concentration	Ice concentration (%), grid cell area (m^2)	Metrics (Fig. 2), spatial maps (Figs. A1-A4)	siconc.py	Metrics: 1980-2007 mean state, interannual variability, and trend of ice concentration simulation; Spatial maps: 1980-2007 February and September mean ice concentration differences in both hemispheres
Sea ice extent	Ice concentration (%), grid cell area (m^2)	Metrics (Figs. 5a,b), time series diagrams (Figs. 3-4)	siext.py	Metrics: Similar to ice concentration evaluation; Time series diagrams: 1980–2007 mean seasonal cycle, monthly anomalies and trends of ice extent in both hemispheres
Sea ice edge	Ice concentration (%), grid cell area (m^2)	Metrics (Fig. 5c), time series diagrams (Fig. 6)	siedge.py	Metrics: 1980-2007 mean state ice edge location simulation; Time series diagrams: 1980-2007 mean seasonal cycle of IIEE in both hemispheres
Sea ice thickness	Ice thickness (m)	Metrics (Fig. 7a), spatial maps (Figs. A5-A6)	sithick.py	Metrics: 2003-2007 mean state ice thickness simulation; Spatial maps: 1980-2007 February (Arctic) and September (Antarctic) mean ice thickness differences
Snow depth	Snow depth (m)	Metrics (Fig. 7b), spatial maps (Figs. A7-A8)	sndepth.py	Metrics: 2003-2007 mean state snow depth simulation; Spatial maps: 1980-2007 February (Arctic) and September (Antarctic) mean snow depth differences
Sea ice drift (magnitude and direction)	Ice velocity in x-direction (m s^{-1}), Ice velocity in y-direction (m s^{-1})	Metrics (Fig. 10), spatial maps (Figs. 8-9, A9-A12)	sidrift.py	Metrics: 2003-2007 mean kinetic energy and vector correlations; Spatial maps: 2003-2007 significant ice-motion vector correlation coefficients; 2003-2007 February and September mean ice-motion mean kinetic energy differences in both hemispheres

2.1 Sea ice metrics and diagnostics

The general approach to derive metrics is by computing scaled absolute errors. We first compute the errors (in absolute value) between some simulated characteristics (e.g., sea ice extent) in individual models and the corresponding characteristic in observational references, respectively. Then, we scale these errors by a typical error to finally get the corresponding metric. The typical error is defined as the absolute difference of the relevant characteristic between two observational references when observations are available, and is therefore a proxy for observation uncertainty. Because our metrics are defined as scaled absolute errors, they are oriented positively meaning that lower values indicate better skill, and a value of 1 means that model error is comparable to observational uncertainty.

2.1.1 Sea ice concentration, extent, and edge location

The methods to calculate the metrics of ice concentration on the mean state, interannual variability, and trend in both hemispheres are introduced here. The consistent equations used to calculate the differences of the mean state (Meandiff), interannual variability (Stddiff) and trend (Trenddiff) between two datasets are shown below:

$$\text{Meandiff} = \frac{\sum_{n=1}^{n=12}\left[\frac{\sum_{i=1}^{i=N}|C_{0M}^{n,i} - C_{1M}^{n,i}| \times A_i}{\sum_{i=1}^{i=N}A_i}\right] \times D_n}{\sum_{n=1}^{n=12}D_n}, \tag{1}$$

$$\text{Stddiff} = \frac{\sum_{i=1}^{i=N}|\text{std}(C_0^i - C_{oM}^i) - \text{std}(C_1^i - C_{1M}^i)| \times A_i}{\sum_{i=1}^{i=N}A_i}, \tag{2}$$

$$\text{Trenddiff} = \frac{\sum_{i=1}^{i=N}|\text{trend}(C_0^i - C_{oM}^i) - \text{trend}(C_1^i - C_{1M}^i)| \times A_i}{\sum_{i=1}^{i=N}A_i}, \tag{3}$$

where n =1,...,12 and i =1,...,N denote the twelve months and the grid cells, respectively, C_{0M} and C_{1M} are monthly mean ice concentrations from two datasets used to do the comparison, A and D are grid cell area and the days in each month, respectively, C_0 and C_1 are monthly ice concentrations from two datasets, and 'std' is the abbreviation of standard deviation. For the mean state evaluation, we compute the monthly mean ice concentration over the study period (1980-2007 for the CMIP6 OMIP model evaluation), and calculate the absolute difference between individual models and observational references over 12 months at each grid cell as shown in equation (1). For the interannual variability and trend evaluation, we compute the standard deviation and linear regression on the monthly anomalies of ice concentration over the study period, and compute the absolute difference between each model output and the observational reference at each grid cell as shown in equations (2) and (3). Then we average these errors spatially weighted by grid cell areas. The typical errors are the differences between two observational references on the mean state, interannual variability, and trend by applying the same method shown before. The differences between individual models and observational references are computed and scaled by corresponding typical errors to get the metrics on ice concentration. The September (February) mean ice concentration differences between each model output and the observational reference, and between two observational references in both hemispheres are provided for diagnosis. These representative months of the summer and winter are selected because normally they respectively correspond to the minimum and maximum seasonal values of sea ice extent for both hemispheres in observations.

The ice extent is calculated as the total area of grid cells with the ice concentration above 15%. The same procedure is followed for ice extent metrics calculation as for ice concentration, except for the spatial averaging since ice extent is already an integrated quantity. The mean seasonal cycle, monthly anomalies, and trend of ice extent in both hemispheres from different models and two observational references are provided for diagnosis.

The integrated ice-edge error (IIEE) is the total area where the models and observational references disagree on the ice concentration being above or below 15% including both the ice extent error and a

200 misplacement error (Goessling et al., 2016). For the mean IIEE evaluation, we compute the monthly mean IIEE between each model output and the observational references over the study period. The typical error is the mean IIEE between two observational references themselves. The differences between each model output and the observational reference are computed and scaled by the typical error to get the metric on the ice edge location. The mean seasonal cycles of IIEE between each model output

205 and the observational reference, and between two observational references in both hemispheres are provided for diagnosis.

2.1.2 Sea ice thickness and snow depth

The same procedure is followed for ice thickness and snow depth metrics calculation as for ice concentration, except for the spatial averaging with equal weight. For the CMIP6 OMIP model

210 evaluation before 2007, the ice thickness and snow depth observations are limited to some months. Because the observational data are not complete to calculate differences between two observational references, the typical errors of ice thickness and snow depth are computed from the ice thickness and snow depth uncertainties of specific months from Envisat data. The mean winter (February for the Arctic and September for the Antarctic) ice thickness and snow depth from ESA's Environmental

215 Satellite (Envisat) radar altimeter data and the differences between model outputs and Envisat data are provided for diagnosis in this study. The mean ice thickness and snow depth differences of other months in both hemispheres can be provided for diagnosis in the future during other study periods when observational references are available. This is not included in this study due to the limited observations for the evaluation before 2007.

220 ### 2.1.3 Sea ice drift

The ice drift metrics include the evaluation of both the magnitude and direction of ice vectors by calculating the mean kinetic energy (MKE) and vector correlation of the ice vectors. The MKE is computed as:

$$\text{MKE} = \frac{1}{2}(u^2 + v^2), \hspace{3cm} (4)$$

225 where u and v are zonal and meridional components of ice drift, respectively. For the MKE evaluation, we compute the monthly mean MKE over the study period and calculate the absolute difference between individual models and observational references over the 12 months at each grid cell. Then we average these errors spatially with equal weight. The typical error is the difference between two observational references of the MKE by applying the same method discussed before. The differences

230 between each model output and the observational reference are computed and scaled by the typical error to get the metric on the ice drift magnitude.

The monthly mean ice vectors during the study period from individual models and observational references are correlated at each grid point by using a vector correlation measure, which is a generalization of the simple correlation coefficient between two scalar time series (Holland and Kwok, 2012). The vector correlation coefficient r^2 is computed by following the equations in Crosby et al. (1993), and the correlation coefficient is scaled (by a number 2) to keep it between 0 and 1 in our study. The nr^2 follows the chi-square distribution with four degrees of freedom, and the correlations are significant at a level of 99% when $nr^2>8$ with samples less than 64 based on the cumulative frequency distributions in Crosby et al. (1993). The significant correlation coefficients between individual models and observational references, and between two observational references are provided for diagnosis at each grid cell. Then we average these significant correlation coefficients spatially with equal weight. The typical correlation coefficient is a spatially averaged correlation coefficient between two observational references. As higher correlation coefficients indicate better skill, the typical correlation coefficients are scaled by the correlation coefficients between individual models and observational references to make it consistent with other metrics (lower values indicate better skill). The September (February) MKE differences and ice-motion vector correlation coefficients between each model output and the observational reference, and between two observational references in both hemispheres are provided for diagnosis.

3 Models and observational references

In this study, the SITool (v1.0) is used to evaluate the CMIP6 OMIP model skills in simulating the historical sea ice properties for both hemispheres. The CMIP6 OMIP models and a set of observational references providing ice concentration, thickness, snow depth, and ice drift are introduced in this section. Two sets of observational references for each sea ice variable are used for comparison.

The CMIP6 OMIP models used are shown in Table 2 with model details such as atmospheric forcing, ocean models, sea ice models, spatial resolution, and related references. A major improvement in JRA55-do atmospheric forcing relative to the CORE-II forcing is the increased temporal frequency from 6 to 3 hours and horizontal resolution from 1.875° to 0.5625°. The surface fields of JRA55-do forcing have been adjusted to match reference datasets based on high-quality satellite observations and several other atmospheric reanalysis products, as detailed in Tsujino et al. (2018). Nine models were run with either CORE-II or JRA55-do forcing; five of them were forced by both CORE-II and JRA55-do reanalysis; out of the four remaining models, one of them was forced by JRA55-do reanalysis only, and the other three were forced by CORE-II reanalysis only. The CMCC-CM2-HR4 (~0.25°) and CMCC-CM2-SR5 (~1°) models are different in spatial resolution, which provides an opportunity to identify the influence model resolution on sea ice simulation. The CORE-II forcing dataset has not been updated since 2009 and the two GFDL models only provide the model outputs until 2007. This is why the

Deleted: mean February and September

Deleted: the

Deleted: model outputs

Deleted: 1

evaluation period is chosen as 1980-2007 for ice concentration, extent, and edge location (the corresponding observations are available from 1980). The evaluation period is 2003-2007 for ice thickness, snow depth, and ice drift because some observational references are limited before 2003, and then the corresponding metrics are only on the mean state. The evaluation period can be extended in the future when different model and observational datasets are considered.

Table 2. The details of nine CMIP6-OMIP models evaluated in the study.

Model	Institution	Atmospheric forcing	Ocean Model	Sea Ice Model	Spatial Resolution	References
CMCC-CM2-HR4	CMCC	JRA55-do	NEMO3.6	CICE4	ORCA-0.25°	Cherchi et al. (2019)
CMCC-CM2-SR5	CMCC	CORE-II/JRA55-do	NEMO3.6	CICE4	ORCA-1°	
EC-Earth3	EC-Earth	CORE-II/JRA55-do	NEMO3.6	LIM3	ORCA-1°	EC-Earth consortium (2019)
GFDL-CM4	NOAA GFDL	CORE-II	OM4	SIS2	tripolar, ~0.25°	Held et al. (2019)
GFDL-OM4p5B	NOAA GFDL	CORE-II	OM4	SIS2	tripolar, ~0.5°	Zadeh et al. (2018)
IPSL-CM6A-LR	IPSL	CORE-II	NEMO-OPA	LIM3	eORCA-1°	Boucher et al. (2020)
MIROC6	JAMSTEC-AORI-NIES-RCCS	CORE-II/JRA55-do	COCO 4.9	COCO 4.9	tripolar, ~1°×(0.5-1)°	Tatebe et al. (2019)
MRI-ESM2-0	MRI	CORE-II/JRA55-do	MRI. COM4.4	MRI. COM4.4	tripolar, ~1°×(0.3-0.5)°	Yukimoto et al. (2019)
NorESM2-LM	NorESM	CORE-II/JRA55-do	BLOM	CICE 5.1.2	tripolar, ~1°×(0.25-1)°	Seland et al. (2020)

The observational reference products for sea ice concentration, thickness, snow depth, and ice drift used to compare with model simulations are summarized in Table 3. The first ice concentration product derives from the passive microwave data of the Scanning Multichannel Microwave Radiometer (SMMR), the Special Sensor Microwave Imager (SSM/I), and the Special Sensor Microwave Imager/Sounder (SSMIS), which are processed by using the NASA Team algorithm (NSIDC-0051, Cavalieri et al., 1996). The other product is based on the same raw data, but uses the EUMETSAT Ocean and Sea Ice Satellite Application Facility algorithm (OSI-450, Lavergne et al., 2019).

Our first ice thickness product is derived from the measurements of ESA's Envisat radar altimeter and provided by the Centre of Topography of Oceans and Hydrosphere (CTOH, Guerreiro et al., 2017). The other ice thickness product is from the measurements of the NASA's Ice, Cloud, and land Elevation Satellite (ICESat) Geoscience Laser Altimeter System (GLAS), and reprocessed separately for the Arctic (NSIDC-0393, Yi and Zwally, 2009) and Antarctic (Kurtz and Markus, 2012). The sea ice freeboard is less uncertain in observations than thickness, however, only five CMIP6 OMIP models at

present provide sea ice freeboard and the model's seawater densities, sea ice densities, and snow densities are not provided to calculate the freeboard. The Envisat data includes ice thickness and thickness uncertainties from November to April for the Arctic with coverage up to 81.5° N and May to October for the Antarctic from 2003. The ICESat data used here 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, October-November with each roughly 33 days. The comparisons between individual models and the two observational references are thus restricted to these months when data is available. The months chosen for the comparison are different from two ice thickness observational references, which can contribute to the differences in ice thickness performance metrics.

Table 3. Observational references used to compare with model simulations.

Variable (period)	Data name	References	Available online at:
Sea ice concentration (1980-2007)	NSIDC-0051	Cavalieri et al. (1996)	https://nsidc.org/data/nsidc-0051
	OSI-450	Lavergne et al. (2019)	http://osisaf.met.no/p/ice/
Sea ice thickness (2003-2007)	Envisat	Guerreiro et al. (2017)	http://ctoh.legos.obs-mip.fr/data/sea-ice-products/sea-ice-thickness
	ICESat	NH: Yi and Zwally (2009) SH: Kurtz and Markus (2012)	NH: https://nsidc.org/data/nsidc-0393 SH: https://earth.gsfc.nasa.gov/index.php/cryo/data/antarctic-sea-ice-thickness
Snow depth (2003-2007)	Envisat	Guerreiro et al. (2017)	http://ctoh.legos.obs-mip.fr/data/sea-ice-products/sea-ice-thickness
	SnowModel-LG	Liston et al. (2020) and Stroeve et al. (2020)	http://dx.doi.org/10.5067/27A0P5M6LZBI
Sea ice drift (2003-2007)	ICDC-NSIDCv4.1	Tschudi et al. (2019)	https://icdc.cen.uni-hamburg.de/en/seaicedrift-satobs-global.html
	KIMURA	KIMURA et al. (2013)	https://ads.nipr.ac.jp/vishop/

The Envisat thickness data also includes snow depth and associated uncertainty. The other snow depth product derives from a Lagrangian snow–evolution model (SnowModel–LG) forced by the European Centre for Medium–Range Weather Forecasts (ECMWF) 5th Generation (ERA5) atmospheric reanalysis, and NSIDC sea ice concentration and trajectory datasets (Liston et al., 2020; Stroeve et al., 2020). The SnowModel–LG data is only provided for the Arctic Ocean. The SnowModel–LG data used to do the comparison is in the same months as the Envisat data from 2003 to 2007.

The first ice drift product is processed by NSIDC and enhanced by the Integrated Climate Data Center (ICDC-NSIDCv4.1). This product derives from SMMR, SSM/I, SSMIS, and the Advanced Very High Resolution Radiometer (AVHRR) for the Antarctic. In addition to the above data, data of the Advanced Multichannel Scanning Radiometer-Earth Observing System (AMSR-E), observations of the

International Arctic Buoy Program (IABP), and ice drift derived from NCEP/NCAR surface winds are used for the Arctic Ocean. The second ice drift dataset is processed by Kimura et al. (2013) and derived from the AMSR-E data for both hemispheres from 2003.

355 The ice vectors are reprocessed before calculating the ice drift metrics. The ice vectors from observational references and models are rotated and interpolated to the polar stereographic grid. The monthly mean ice vectors of the observational references are computed when there are more than 10 days with valid daily drift data. The ICDC-NSIDCv4.1 ice drift data was shown to be biased low (i.e., too slow) relative to buoy data (Schwegmann et al. 2011; Barthélemy et al., 2018) and is therefore

360 corrected by multiplying the drift components with a correction factor of 1.357 (Haumann et al., 2016). The ice vectors from observational references and models are removed when ice concentrations are below 50%, or the data is closer than 75 km to the coast, or with a spurious value, to reduce the spatial and temporal noise by following Haumann et al. (2016).

4. SITool application and results

365 The SITool (v1.0) described in Section 2 is applied in this section to assess the performance of the sea ice simulations for both hemispheres carried out under the CMIP6 OMIP1 and OMIP2 protocols. Models forced by CORE-II atmospheric reanalysis data (OMIP1) or JRA55-do reanalysis data (OMIP2) are marked as <model name + /C, or /J>, respectively. The OMIP1 and OMIP2 model means shown below are from five models of CMCC-CM2-SR5, EC-Earth3, MIROC6, MRI-ESM2-0, and NorESM2-

370 LM providing both OMIP1 and OMIP2 model outputs. All the sea ice data from models and observational references are interpolated to the NSIDC-0051 polar stereographic 25 km resolution grid for comparison. The typical errors are the differences between two observational references for the ice concentration, extent, edge location, and ice drift, while typical errors of ice thickness and snow depth are calculated from the thickness and snow depth uncertainties of specific months from Envisat data.

375 **4.1 Sea ice concentration, extent, and edge location**

Figure 2 shows that model errors on ice concentration simulations are around two to five times the observational uncertainty and the ice concentration simulations are much closer to the NSIDC-0051 data (Fig. 2a) compared to the OSI-450 data (Fig. 2b). In general, the overall ice concentration simulations (mean state, interannual variability, and trend) in both hemispheres are improved under

380 OMIP2 protocol, forced by JRA55-do reanalysis. This is identified in Figs. 2a and 2b by comparing the 5 OMIP1 and OMIP2 model mean values (last two rows), and also by comparing five models' values separately under either OMIP protocol. The overall ice concentration simulations in both hemispheres are also improved in CMCC-CM2-HR4/J with higher spatial resolution of ocean-sea ice model compared to CMCC-CM2-SR5/J (first and third rows). The improvements on the overall ice

Deleted: 1
Deleted: 2
Deleted: . The
Deleted: because different observational months limit the calculation of differences between two observational references.¶
Deleted: /2
Deleted: increased
Deleted: /2

concentration simulations are not sensitive to the chosen observational reference and then robust. The improved ice concentration simulations are found compared to different observational references except for the interannual variability of the Antarctic ice concentration compared to the OSI-450 data as shown in the fifth column of Fig. 2b.

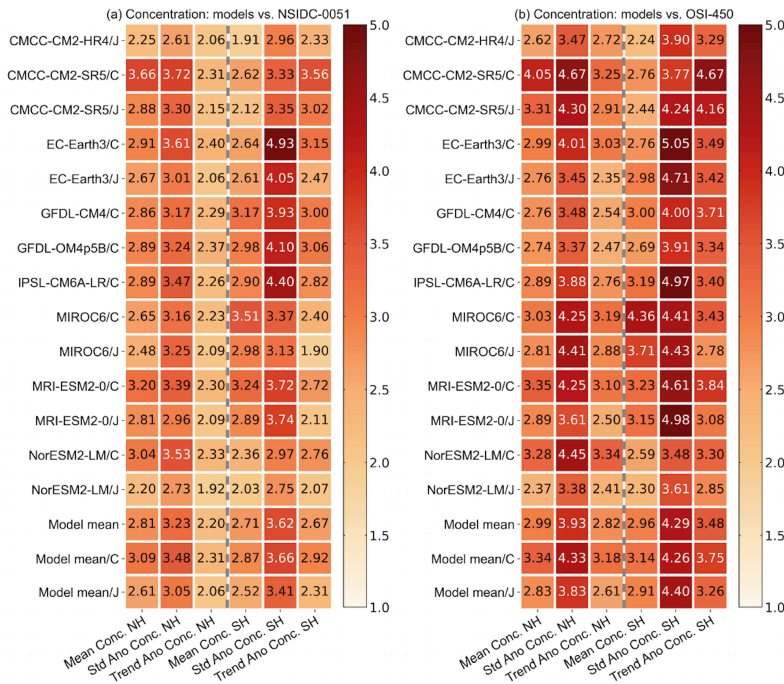



Figure 2. The ice concentration metrics of 14 model outputs under OMIP1 (/C) and OMIP2 (/J) protocols, 14-model mean (Model mean), 5-OMIP1-model mean (Model mean/C) and 5-OMIP2-model mean (Model mean/J) from CMCC-CM2-SR5, EC-Earth3, MIROC6, MRI-ESM2-0, and NorESM2-LM compared to (a) NSIDC-0051 and (b) OSI-450 data. The six columns correspond to model performance metrics on the mean state, standard deviation (Std Ano) and trend (Trend Ano) of monthly anomalies of the Arctic and Antarctic ice concentration during 1980-2007. Lower values indicate better skill.

The metrics on the interannual variability of ice concentration (second and fifth columns) are the highest among all metrics, which indicates relatively lower skill on the simulation of ice concentration variability in both hemispheres compared to the mean state and trend. The overall best performance on ice concentration simulations including the mean state, interannual variability, and trend is in NorESM2-LM forced by JRA55-do reanalysis for both hemispheres. To help understand the differences in the ice concentration metrics, the 1980-2007 September and February mean ice concentration differences between the OSI-450 and NSIDC-0051 data, and between model outputs and the NSIDC-0051 data are produced for both hemispheres in the appendix A: Figs. A1-A4.

Figures 3a and 3b reveal that the monthly ice extent differences between two observational references (observational uncertainty, black × vs. cyan +) are much smaller compared to the model bias (red lines vs. black/cyan marks) in both hemispheres. The negative ice extent biases under OMIP1 protocol in the summer of both hemispheres are reduced under OMIP2 protocol (Figs. 3a and 3b, red solid vs. dash-dotted) by changing the atmospheric forcing to JRA55-do reanalysis. The reduced mean ice extent biases in the summer under OMIP2 protocol are also identified in Tsujino et al. (2020) (see their Fig. 22 and Table D7). In the boreal winter, the 5-model mean ice extents under OMIP1 and OMIP2 protocols show no obvious difference (Fig. 3a, red solid vs. dash-dotted), and the ice extents among most models are close to the observational references (Fig. 3c) except for the MIROC6 (orange) and MRI-ESM2-0 (gray). In the austral winter, large spread exits for the ice extent simulation (Fig. 3d), and the positive ice extent bias under OMIP1 protocol (Fig. 3b, red solid) becomes a negative one under OMIP2 protocol (red dash-dotted). The absolute value of ice extent bias in the austral winter under OMIP2 protocol is not reduced compared to that under OMIP1 protocol (Fig. 3b, red dash-dotted vs. solid).

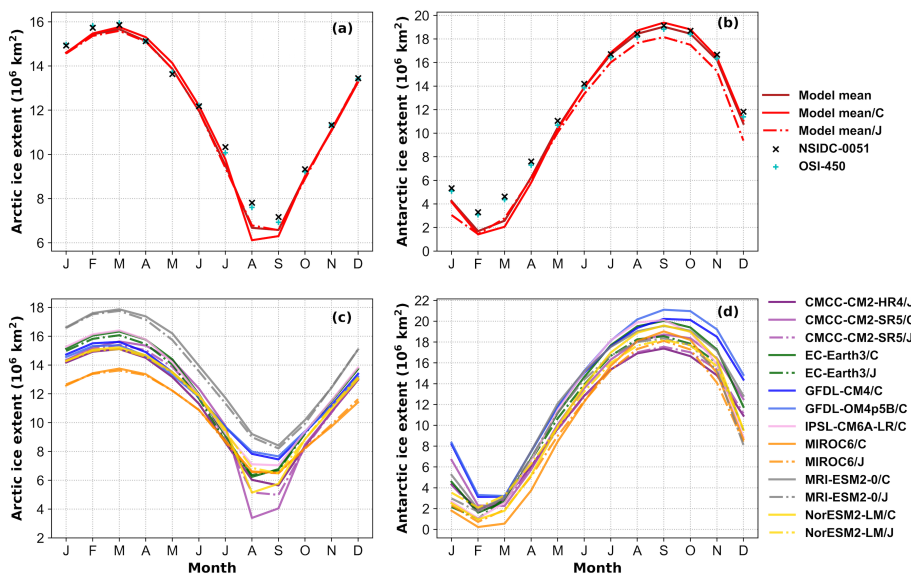



Figure. 3. The 1980–2007 mean seasonal cycle of ice extent (10^6 km^2) from 14-model mean (brick red solid), 5-model mean under OMIP1 and OMIP2 protocols (red solid and dash-dotted), NSIDC-0051 (black ×) and OSI-450 (cyan +) in the (a) Arctic and (b) Antarctic. The 5-model mean is from CMCC-CM2-SR5, EC-Earth3, MIROC6, MRI-ESM2-0, and NorESM2-LM. The mean seasonal cycle from 14 model outputs under OMIP1 (/C) and OMIP2 (/J) protocols are shown in the (c) and (d), and the model outputs under OMIP2 protocol from these five models are shown in dash-dotted lines.

The biases of 5-model mean ice extent monthly anomalies under OMIP1 protocol compared to the observational mean (green vs. black solid) are reduced under OMIP2 protocol (orange vs. black solid)

in both hemispheres as shown in Fig. 4. The standard deviations of the monthly anomalies of ice extent in both hemispheres are smaller under OMIP2 protocol than that under OMIP1 protocol. In the Arctic (Figs. 4a and 4b), the negative biases of ice extent monthly anomalies during 1980-1982 and after 1998, as well as positive bias during 1986-1990 are reduced in the OMIP2 model mean (orange vs. green solid). However, the declining trend of ice extent from the observational mean (black dashed) is close to the OMIP1 model mean (green dashed) but not the OMIP2 model mean (orange dashed). This can be caused by the error compensation of the negative ice extent biases to observational mean during 1980-1982 and after 1998 in the OMIP1 model mean. In the Antarctic (Figs. 4c and 4d), the reduced bias is obvious after 1988 in the OMIP2 model mean (orange vs. green solid). The increasing trend of the Antarctic ice extent in the observational mean (black dashed) is not shown in the OMIP1 and OMIP2 mean (green and orange dashed). The ice extent monthly anomalies in each model under OMIP1 and OMIP2 protocols are compared separately, and the improvements on the simulations of ice extent interannual variability are found in the OMIP2 model outputs of individual models (not shown). The improved interannual variability of ice extent in the OMIP2 simulations is also identified in Tsujino et al. (2020) (see their Figs. 22 and 23).

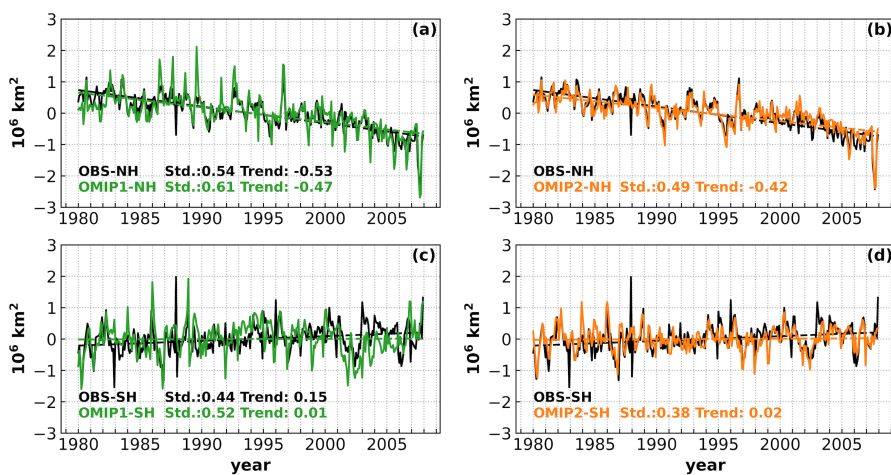
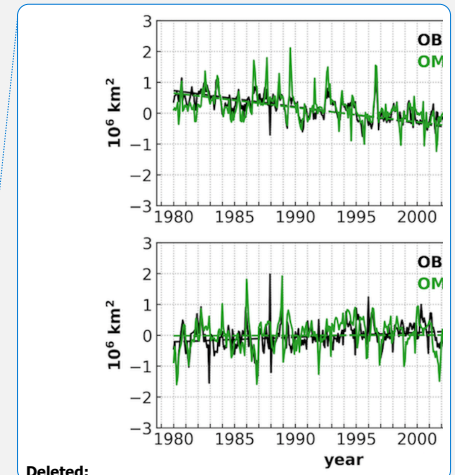



Fig. 4. The 1980–2007 monthly anomalies of ice extent (10^6 km^2) from the observational mean of NSIDC-0051 and OSI-450 (black solid), 5-model mean under OMIP1 or OMIP2 protocol (green vs. orange solid) in the Arctic (a, b) and Antarctic (c, d). The 5-model mean is from CMCC-CM2-SR5, EC-Earth3, MIROC6, MRI-ESM2-0, and NorESM2-LM. The dashed lines are the trends computed from linear regression over 1980-2007. The standard deviation (Std, 10^6 km^2) and trend (10^6 km^2/decade) of the monthly anomalies of ice extent are computed and displayed.

Figs. 5a and 5b shows that the model errors on ice extent simulation are much larger than the observational uncertainty in most cases, and the large values on the fifth columns are due to the very low typical error (0.0009×10^6 km^2) of the Antarctic interannual ice extent variability. In general, the ice extent simulations on the mean state and interannual variability for the Arctic, as well as the interannual

variability and trend for the Antarctic are improved under OMIP2 protocol, forced by JRA55-do reanalysis. This is identified in Figs. 5a and 5b by comparing the 5 OMIP1 and OMIP2 model mean values (last two rows), though there are several exceptions for the simulation of individual models under either OMIP protocol. The improved ice extent simulations are identified compared to different observational references.

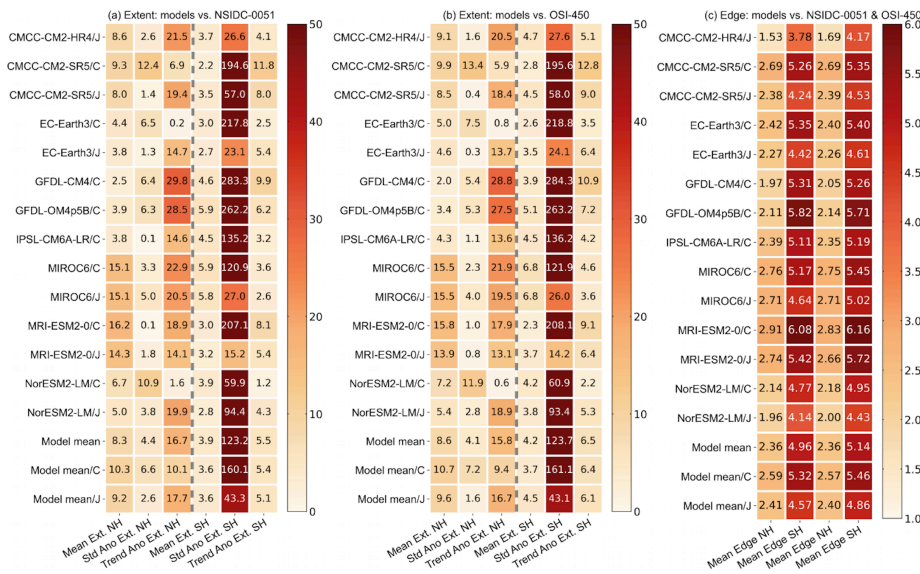



Figure 5. The ice extent metrics of 14 model outputs under OMIP1 (/C) and OMIP2 (/J) protocols, 14-model mean (Model mean), 5-OMIP1-model mean (Model mean/C) and 5-OMIP2-model mean (Model mean/J) from CMCC-CM2-SR5, EC-Earth3, MIROC6, MRI-ESM2-0, and NorESM2-LM compared to (a) NSIDC-0051 and (b) OSI-450 data. The six columns correspond to model performance metrics on the mean state, standard deviation (Std Ano) and trend (Trend Ano) of monthly anomalies of the Arctic and Antarctic ice extent during 1980-2007. (c) The mean state ice edge location metrics during 1980-2007 in both hemispheres compared to the NSIDC-0051 (first two columns) and OSI-450 data (last two columns). Lower values indicate better skill.

The simulation of Arctic ice extent trend under OMIP2 protocol is not better than that under OMIP1 protocol (the third columns in Figs. 5a and 5b), which is due to the error compensation of the monthly anomalies biases of the ice extent during different periods under OMIP1 protocol as explained in Figs. 4a and 4b. This error compensation can change the trend and make it close to the observational references even though the monthly anomalies are not well presented in the OMIP1 models. The unimproved Antarctic mean ice extent under OMIP2 protocol can also be found in Fig. 3b where the ice extent bias in the austral winter is not reduced under OMIP2 protocol. This is not consistent with what we found for the improvement in the ice concentration simulation under OMIP2 protocol, which is possibly because ice extent cancels out regional concentration differences. The overall best performance

on ice extent simulation including the mean state, interannual variability, and trend is in EC-Earth3/C for the Arctic and in MRI-ESM2-0/J for the Antarctic.

520 To gain insights in the spatial distribution of errors, we then apply the IIEE (Goessling et al., 2016) as introduced in section 2. In both hemispheres, the IIEEs between models and NSIDC-0051 are obviously much larger than that between two observational references as shown in Fig. 6. The largest model errors and model spreads are in the summer of both hemispheres. The IIEE under OMIP1 protocol is much reduced under OMIP2 protocol especially in the summer of both hemispheres (Figs. 6a and 6b, red solid
525 vs. dash-dotted) by changing the atmospheric forcing to JRA55-do reanalysis. In both hemispheres, the large IIEE in CMCC-CM2-SR5/J (light purple dashed) is reduced in CMCC-CM2-HR4/J (dark purple solid) with higher spatial resolution of ocean-sea ice model during all the seasons (Figs. 6c and 6d). To identify the ice edge location errors of various models, the contours of 15% concentration derived from the 1980-2007 September and February mean ice concentration are also shown for both hemispheres in
530 the appendix A: Figs. A1-A4.

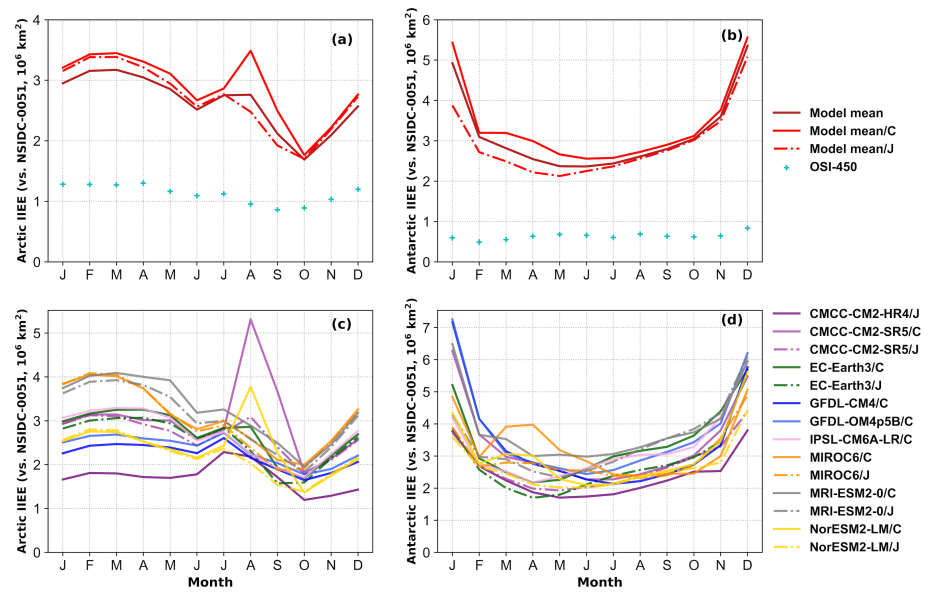



Figure. 6. The 1980–2007 mean seasonal cycle of the Integrated Ice-Edge Error (IIEE, vs. NSIDC-0051, 10⁶ km²) from 14-model mean (brick red solid), 5-model mean under OMIP1 and OMIP2 protocols (red solid and dash-dotted), and OSI-450 (cyan +) in the (a) Arctic and (b) Antarctic. The 5-model mean is from CMCC-CM2-SR5, EC-Earth3, MIROC6, MRI-ESM2-0, and NorESM2-LM. The mean seasonal cycle from 14 model outputs under OMIP1 (/C) and OMIP2 (/J) protocols are shown in the (c) and (d), and the model outputs under OMIP2 protocol from the five models are in dash-dotted lines.

(a) Models vs NSIDC-0051							
CMCC-CM2-HR4/2	8.6	2.6	21.5	3.7	26.6	4.1	CMCC-CM2-HR4/2 9.1
CMCC-CM2-SR5/1	9.3	12.4	6.9	2.2	194.6	11.8	CMCC-CM2-SR5/1 9.9
CMCC-CM2-SR5/2	8.0	1.4	19.4	3.5	57.0	8.0	CMCC-CM2-SR5/2 8.5
EC-Earth3/1	4.4	6.5	0.2	3.0	217.8	2.5	EC-Earth3/1 5.0
EC-Earth3/2	3.8	1.3	14.7	2.7	23.1	5.4	EC-Earth3/2 4.6
GFDL-CM4/1	2.5	6.4	29.8	4.6	283.3	9.9	GFDL-CM4/1 2.0
GFDL-OM4p5B/1	3.9	6.3	28.5	5.9	262.2	6.2	GFDL-OM4p5B/1 3.4
IPSL-CM6A-LR/1	3.8	0.1	14.6	4.5	135.2	3.2	IPSL-CM6A-LR/1 4.3
MIROC6/1	15.1	3.3	22.9	5.9	120.9	3.6	MIROC6/1 15.
MIROC6/2	15.1	5.0	20.5	5.8	27.0	2.6	MIROC6/2 15.
MRI-ESM2-0/1	16.2	0.1	18.9	3.0	207.1	8.1	MRI-ESM2-0/1 15.
MRI-ESM2-0/2	14.3	1.8	14.1	3.2	15.2	5.4	MRI-ESM2-0/2 13.
NorESM2-LM/1	6.7	10.9	1.6	3.9	59.9	1.2	NorESM2-LM/1 7.2
NorESM2-LM/2	5.0	3.8	19.9	2.8	94.4	4.3	NorESM2-LM/2 5.4
Model mean	8.3	4.4	16.7	3.9	123.2	5.5	Model mean 8.6
Model mean/1	10.3	6.6	10.1	3.6	160.1	5.4	Model mean/1 10.
Model mean/2	9.2	2.6	17.7	3.6	43.3	5.1	Model mean/2 9.6

Mean Ext. North Std Ano Ext. North Trend Ano Ext. North Mean Ext. South Std Ano Ext. South Trend Ano Ext. South   Mean Ext. North Std Ano Ext. Trend A:

The mean state ice edge location metrics in Fig. 5c shows that model errors on ice edge location simulations are around two to six times the observational uncertainty, and the ice edge location simulations in the Arctic are much better than that in the Antarctic. Zampieri et al. (2019) also show that the prediction skill of sea ice edge location is 30% lower in the Antarctic than in the Arctic from coupled subseasonal forecast systems. The lower prediction skill in the Antarctic can be related to more complicated ocean dynamic processes there, which decrease the persistence of ice areal changes (Ordoñez et al., 2018). The mean state ice edge location simulations in both hemispheres are improved under OMIP2 protocol, which is identified in Fig. 5c by comparing the 5-model mean values (last two rows), and also by comparing five models' values separately under either OMIP protocol. The mean state ice edge location simulations in both hemispheres are also improved in CMCC-CM2-HR4/J with higher ocean-sea ice model resolution compared to CMCC-CM2-SR5/J (first and third rows). The improved ice edge location simulations are identified compared to different observational references. The best performance on the mean state ice edge location simulations is in CMCC-CM2-HR4/J for both hemispheres.

4.2 Sea ice thickness and snow depth

Fig. 7 shows the mean state ice thickness and snow depth metrics, and the interannual variability and trend metrics are not included here because the observational record is too short to make such an assessment (Tilling et al., 2015). The model errors on the mean ice thickness and snow depth simulations are not obviously larger (even smaller in some models) than the observational uncertainty. The mean ice thickness simulation during 2003-2007 is improved in the Antarctic under OMIP2 protocol, forced by JRA55-do reanalysis. This is identified in Fig. 7a by comparing the 5-model mean values (last two rows), and also by comparing the five models' values separately under either OMIP protocol (an exception in NorESM2-LM compared to the Icesat data). The improved Antarctic mean ice thickness simulations are identified compared to different observational references. The best performance on the mean ice thickness simulation is in IPSL-CM6A-LR/C for the Arctic, while for the Antarctic the best performance is in CMCC-CM2-HR4/J compared to the Envisat data and in GFDL-OM4p5B/C compared to the ICESat data. The different model performance on the mean ice thickness simulations by comparing to two observational references are due to the different months chosen for the ice thickness comparison.

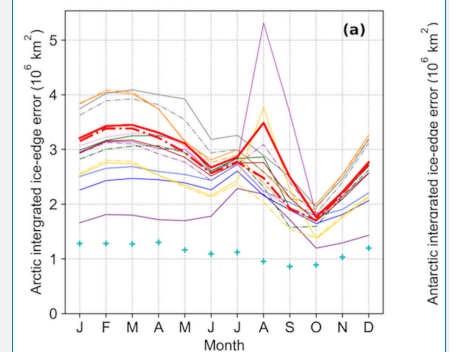



Figure. 6. The 1980–2007 mean seasonal cycle of the Integrated Ice-Edge Error (IIEE, 10⁶ km²) in the (a) Arctic and (b) Antarctic between 14 model outputs/OSI-450 and NSIDC-0051. The 14-model mean (brick red solid), 5-model mean under OMIP1 or OMIP2 protocol (red solid vs. dash-dotted), and OSI-450 IIEE (cyan +) are marked. The 5-model mean is from CMCC-CM2-SR5, EC-Earth3, MIROC6, MRI-ESM2-0, and NorESM2-LM, and the model outputs under OMIP2 protocol from these five models are shown in dash-dotted lines.

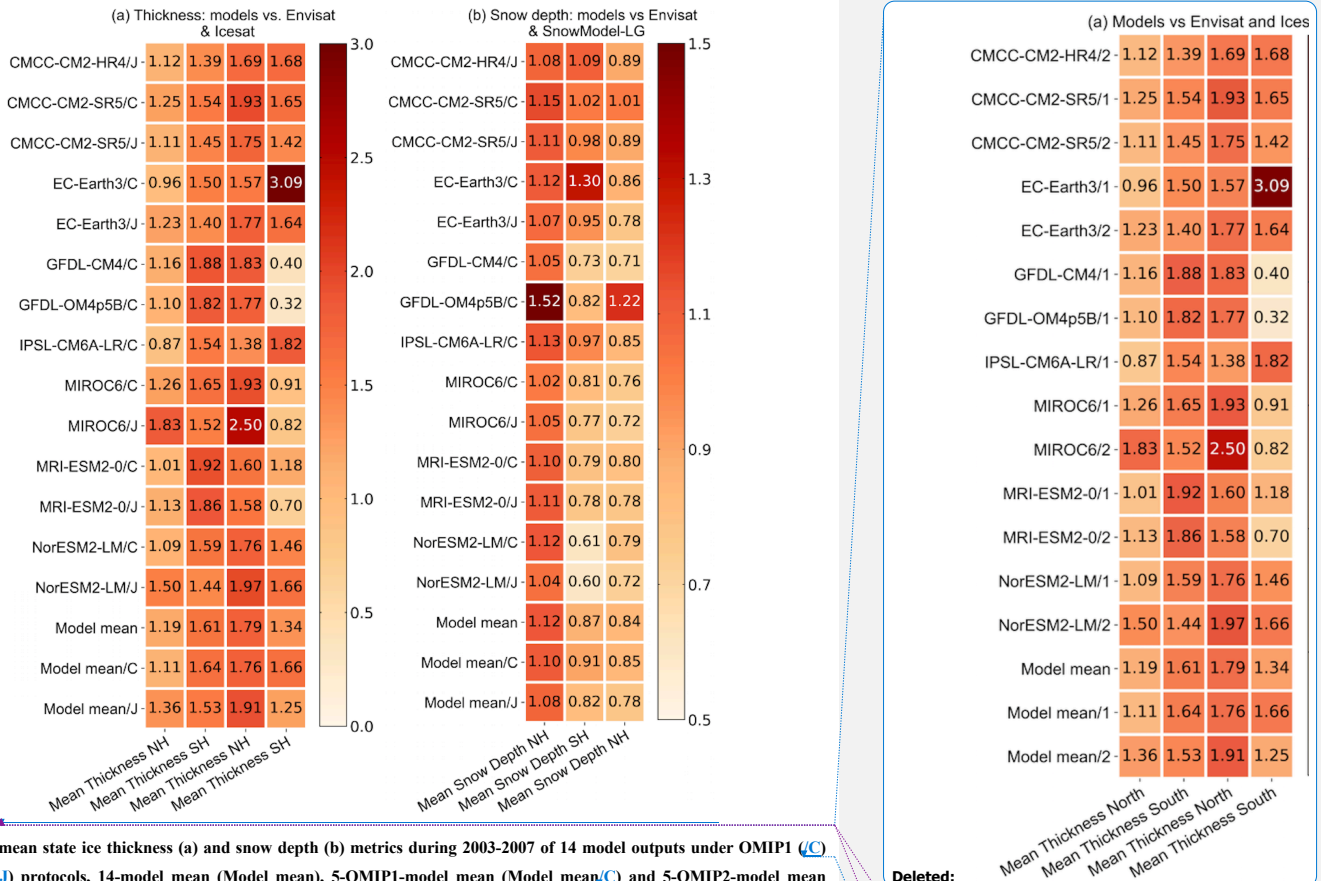



Figure 7. The mean state ice thickness (a) and snow depth (b) metrics during 2003-2007 of 14 model outputs under OMIP1 (/C) and OMIP2 (/J) protocols, 14-model mean (Model mean), 5-OMIP1-model mean (Model mean/C) and 5-OMIP2-model mean (Model mean/J) from CMCC-CM2-SR5, EC-Earth3, MIROC6, MRI-ESM2-0, and NorESM2-LM. The four columns in (a) correspond to ice thickness metrics in both hemispheres compared to the Envisat (first two) and ICESat data (last two), and the three columns in (b) correspond to snow depth metrics compared to the Envisat data in both hemispheres (first two) and the SnowModel-LG data in the Arctic (last one). Lower values indicate better skill.

The mean snow depth simulation during 2003-2007 in both hemispheres improved a bit under OMIP2 protocol, which can be found by comparing 5-model mean values under either OMIP protocol (last two rows) in Fig. 7b. The improvement on the mean snow depth simulation is relatively small compared to other ice metrics. The best performance on the mean snow depth simulation for the Arctic is in MIROC6/C compared to the Envisat data and in GFDL-CM4/C compared to the SnowModel–LG data, and for the Antarctic, the best performance is in NorESM2-LM/J (Fig. 7b). To help understand the differences in the ice thickness and snow depth metrics, the 2003-2007 winter-mean ice thickness and snow depth from Envisat data, and the differences between model outputs and Envisat data are produced for both hemispheres in the appendix A: Figs. A5-A8.

4.3 Sea ice drift

The magnitude and direction of simulated ice drifts are evaluated by calculating the MKE and the vector correlation from monthly mean ice vector data during 2003-2007. The vector correlation

630   coefficients are measured by using a generalization of the simple correlation coefficient between two scalar time series as introduced in section 2. The significant correlation coefficient at a level of 99% between ICDC-NSIDCv4.1 and KIMURA data, and between 14 model outputs and KIMURA data in the Arctic (Fig. 8) and Antarctic (Fig. 9) are displayed. The correlation coefficients are much lower between model outputs and the KIMURA data than that between two observational references. This is

635   obvious for the coastal regions of Greenland and Canadian archipelago in the Arctic and the Weddell Sea and the Ross Sea in the Antarctic, as well as the ice edge location of the Weddell Sea among some models. This implies that model errors on the ice-vector direction simulations are much larger than the observational uncertainty. The correlation coefficients are higher under OMIP2 protocol than that under OMIP1 protocol (third vs. second column in Figs. 8 and 9), which indicate the improvement on the ice-

640   vector direction simulation when forced by JRA55-do atmospheric forcing in both hemispheres.

Figure 10 shows that model errors on the mean ice drift simulations are larger than the observational uncertainty. In general, the ice drift simulations on the magnitude (Fig. 10a) and direction (Fig. 10b) in both hemispheres are improved under OMIP2 protocol, forced by JRA55-do reanalysis. This is identified from the 5 OMIP1 and OMIP2 model mean values (last two rows), and also by comparing

645   five models' values separately under either OMIP protocol (an exception in CMCC-CM2-SR5 of the Arctic ice-vector magnitude in Fig. 10a). The improved mean ice drift simulations under OMIP2 protocol are found compared to not only the ICDC-NSIDCv4.1 data but also the KIMURA data. The overall best performance on sea ice drift simulations including the magnitude and direction is in MIROC6/J for both hemispheres. To help understand the differences in the ice-motion magnitude

650   metrics, the 2003-2007 September and February mean ice-motion MKE differences between the ICDC-NSIDCv4.1 and KIMURA data, and between model outputs and the KIMURA data are produced for both hemispheres in the appendix A: Figs. A9-A12.

Deleted: of the ice vectors

Deleted: are identified compared to different observational references. The overall best performance on sea ice drift simulations

Deleted: /2

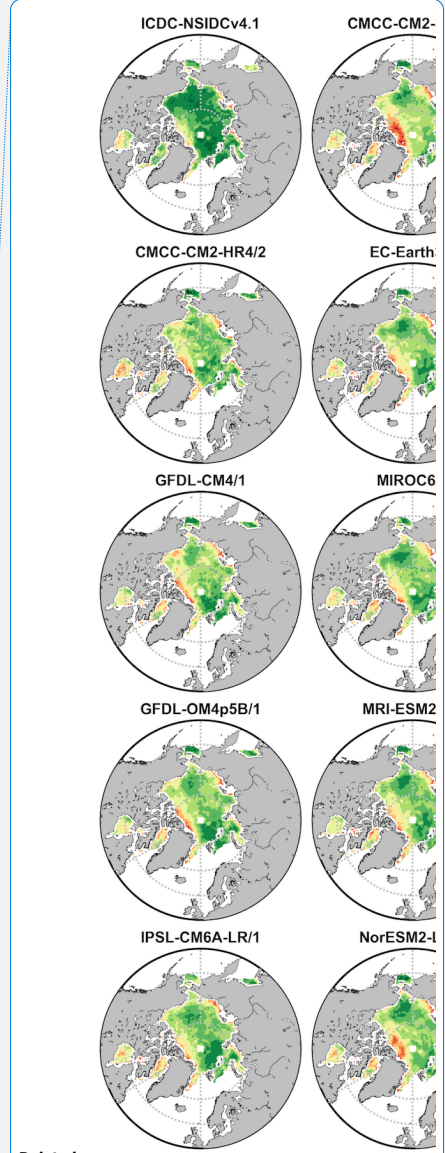
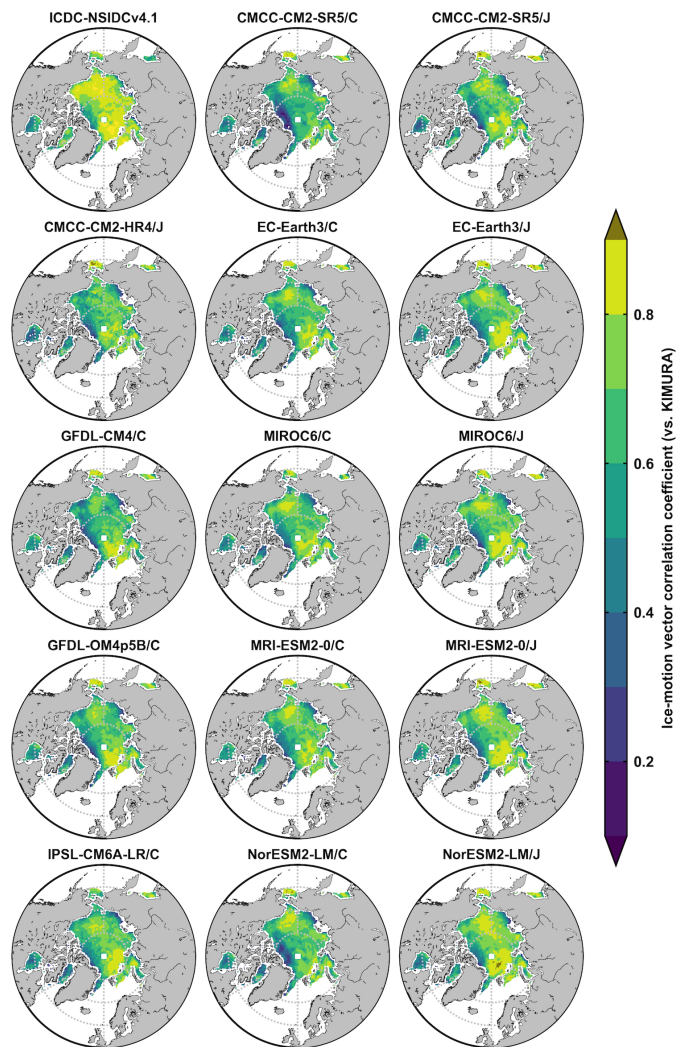



Figure 8. The significant Arctic ice-motion vector correlation coefficients from monthly mean data during 2003–2007 at a level of 99% between ICDC-NSIDCv4.1/model outputs and the KIMURA data ($m^2 s^{-2}$). The second and third columns are from 5 OMIP1 and OMIP2 model outputs of CMCC-CM2-SR5, EC-Earth3, MIROC6, MRI-ESM2-0, NorESM2-LM, respectively.

660

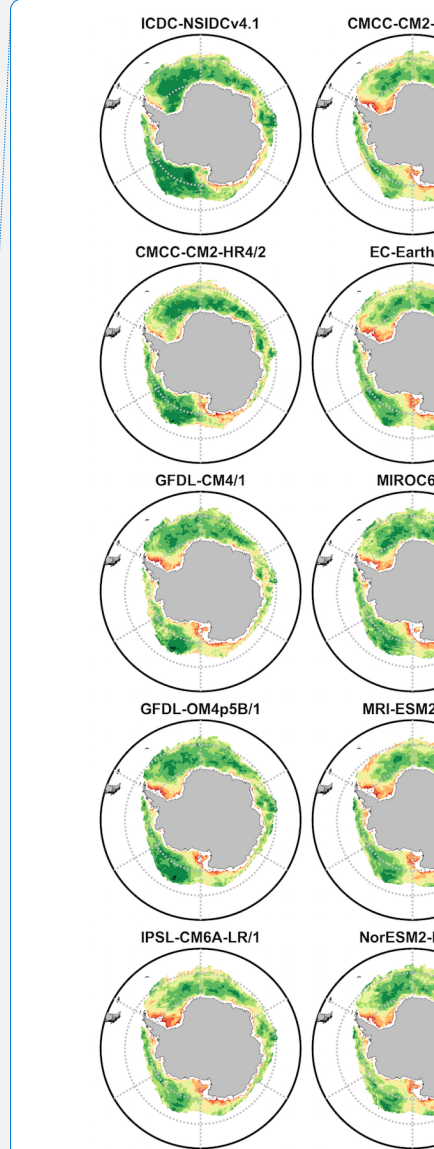
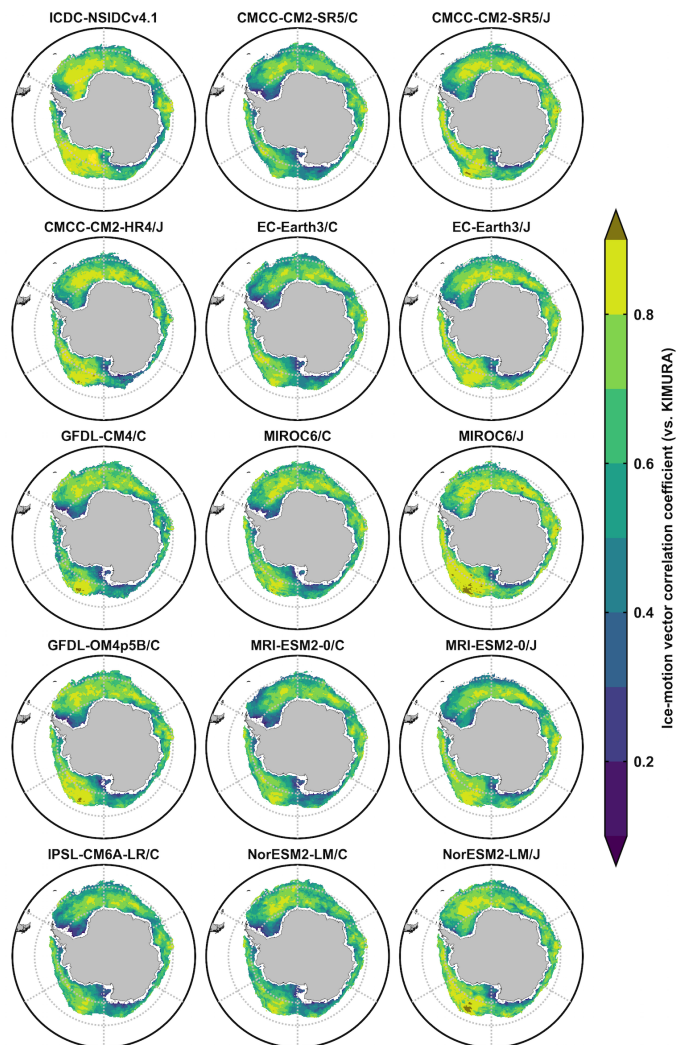
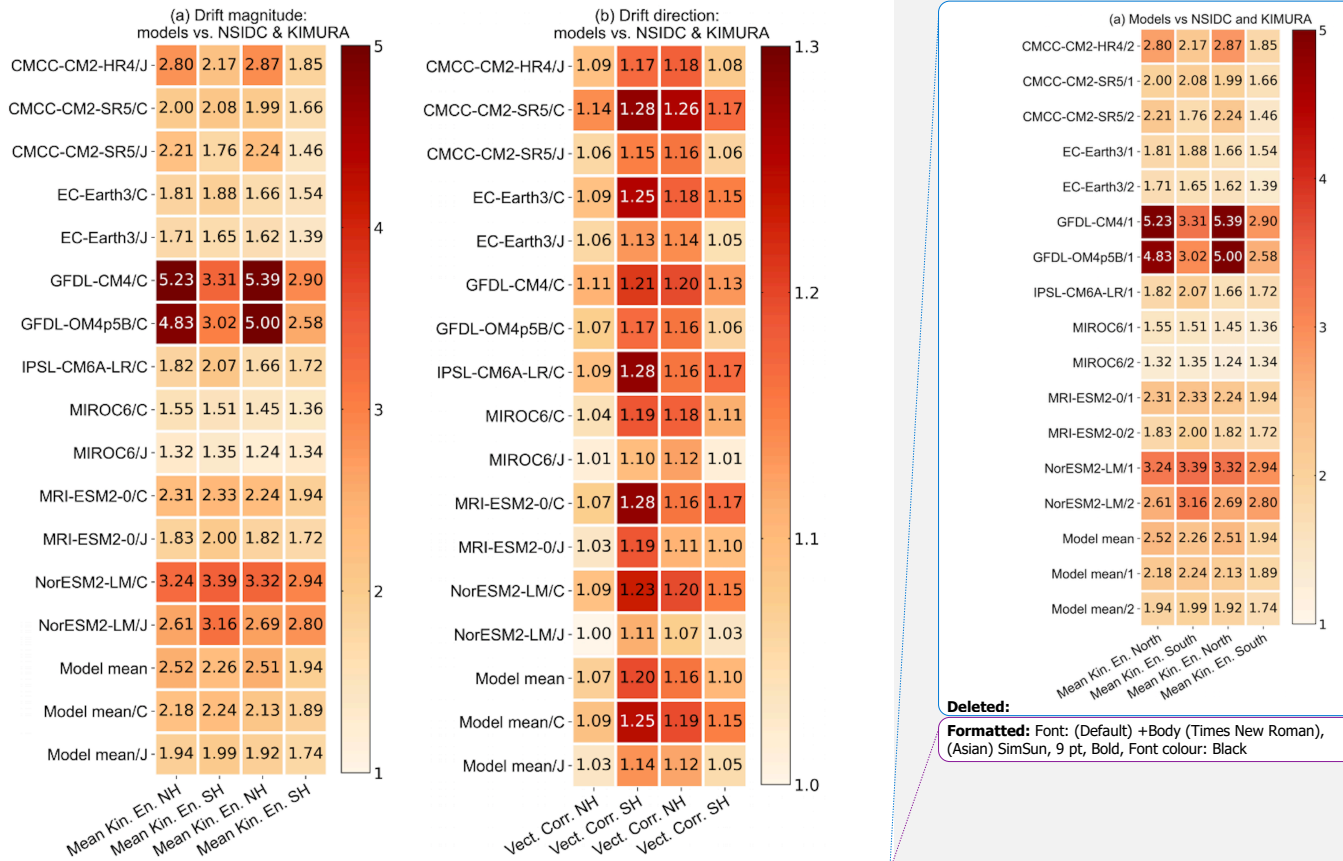



Figure 9. The significant Antarctic ice-motion vector correlation coefficients from monthly mean data during 2003–2007 at a level of 99% between ICDC-NSIDCv4.1/model outputs and the KIMURA data ($m^2\ s^{-2}$). The second and third columns are from 5 OMIP1 and OMIP2 model outputs of CMCC-CM2-SR5, EC-Earth3, MIROC6, MRI-ESM2-0, NorESM2-LM, respectively.

665

670 **Figure. 10. The ice drift metrics of 14 model outputs under OMIP1 (/C) and OMIP2 (/J) protocols, 14-model mean (Model mean), 5-OMIP1-model mean (Model mean/C) and 5-OMIP2-model mean (Model mean/J) from CMCC-CM2-SR5, EC-Earth3, MIROC6, MRI-ESM2-0, and NorESM2-LM. The four columns correspond to model performance metrics on the (a) mean kinetic energy (MKE) and (b) the vector correlations during 2003-2007 of the Arctic and Antarctic compared to the ICDC-NSIDCv4.1 (first two) and KIMURA data (last two). Lower values indicate better skill.**

675 **4.4 Cross-metric analysis**

From previous analyses, it seems that there is no best sea ice model simulation, but rather that each model has strengths and weaknesses. To further illustrate this aspect, the metrics of each sea ice variable are ranked in a cross-metric analysis, where the link between the model performance in one variable and the performance in another is clearly highlighted. By changing the atmospheric forcing from

680 CORE-II to JRA55-do reanalysis data, the sea ice model simulations are improved in general. In order to make the comparison simple, six models (CMCC-CM2-HR4, CMCC-CM2-SR5, EC-Earth3, MIROC6, MRI-ESM2-0 and NorESM2-LM) forced by the JRA55-do reanalysis were retained for this

analysis. The best and worst performing models for six ice variables are listed in Table 4. It is found that no single model performs best in all ice metrics as there is no link between performance in one variable and performance in another. For example, the NorESM2-LM/J is the best regarding the ice concentration and snow depth simulation, but the worst for the ice drift simulation in the Antarctic, and the MIROC6/J is the best regarding the ice drift simulation, but the worst for the ice extent and thickness simulation in the Arctic.

Table 4. The best (in bold) and worst performing models for the six evaluated sea ice variables, among the models which were run with JRA55-do reanalysis (CMCC-CM2-HR4, CMCC-CM2-SR5, EC-Earth3, MIROC6, MRI-ESM2-0 and NorESM2-LM). The numbers are derived from the performance metrics in Figs. 2, 5, 7, 10 (marked as () and []). The values given average the three metrics (mean state, interannual variability and trend) for ice concentration and ice extent from Figs. 2, 5a and b; the two metrics for the magnitude and vector correlations for ice drift from Fig. 10. The values for the ice edge location, thickness and snow depth are the metric of mean state from Figs. 5c and 7. In some cases, such as for the ice thickness evaluation, the best and worst performing models are different compared to different sets of observations.

	Arctic	Antarctic
Ice concentration (vs. NSIDC-0051) [vs. OSI-450]	**NorESM2-LM/J (2.3) [2.7]** CMCC-CM2-SR5/J (2.8) [3.5]	**NorESM2-LM/J (2.3) [2.9]** EC-Earth3/J (3.0) MRI-ESM2-0/J [3.7]
Ice extent (vs. NSIDC-0051) [vs. OSI-450]	**EC-Earth3/J (6.6) [6.2]** MIROC6/J (13.5) [13]	**MRI-ESM2-0/J (7.9) [8.1]** NorESM2-LM/J (33.8) [34.2]
Ice edge location (vs. NSIDC-0051) [vs. OSI-450]	**CMCC-CM2-HR4/J (1.5) [1.7]** MRI-ESM2-0/J (2.7) [2.7]	**CMCC-CM2-HR4/J (3.8) [4.2]** MRI-ESM2-0/J (5.4) [5.7]
Ice thickness (vs. Envisat) [vs. Icesat]	**CMCC-CM2-SR5/J (1.1)** **MRI-ESM2-0/J [1.6]** MIROC6/J (1.8) [2.5]	**CMCC-CM2-HR4/J (1.4)** **MRI-ESM2-0/J [0.7]** MRI-ESM2-0/J (1.9) CMCC-CM2-HR4/J [1.7]
Snow depth (vs. Envisat) [vs. SnowModel-LG]	**NorESM2-LM/J (1.0) [0.7]** CMCC-CM2-SR5/J (1.1) [0.9]	**NorESM2-LM/J (0.6)** CMCC-CM2-HR4/J (1.1)
Ice drift (vs. ICDC-NSIDCv4.1) [vs. KIMURA]	**MIROC6/J (1.2) [1.2]** CMCC-CM2-HR4/J (2.0)[2.0]	**MIROC6/J (1.2) [1.2]** NorESM2-LM/J (2.1) [1.9]

5 Conclusions and discussion

The SITool (v1.0), a performance metrics and diagnostics tool for CMIP6-compliant sea ice outputs, is introduced in this paper. The evaluation includes ice concentration, extent, edge location, thickness,

snow depth, and ice drift. The SITool (v1.0) provides rating scores for each sea ice variable in both hemispheres by comparing them to a set of observational references, using two observational references to account for the role of observational uncertainty in the evaluation process. In this paper, we evaluate the CMIP6 OMIP sea ice simulations with SITool (v1.0) to demonstrate the proof of concept and potentialities behind it. Specifically, we evaluate the performances of OMIP historical sea ice simulations (1980-2007 for sea ice areal properties, 2003-2007 for ice drift, thickness and snow depth).

Our main findings on CMIP6 OMIP simulations are summarized below. By changing the atmospheric forcing from CORE-II to JRA55-do reanalysis data, the improvement are identified in (1) the ice concentration simulations including the mean, interannual variability and trend in both hemispheres, (2) the ice extent simulations including the mean and interannual variability in the Arctic, as well as the interannual variability and trend in the Antarctic, (3) the mean ice edge location simulations in both hemispheres, (4) the mean ice thickness simulations in the Antarctic and the mean snow depth simulations in both hemispheres, (5) the ice drift simulations including the magnitude and direction in both hemispheres. By increasing the horizontal resolution of CMCC-CM2 ocean-sea ice model, the improvements are identified in the sea ice concentration (mean, interannual variability and trend) and the mean ice edge location simulations in both hemispheres.

In general, model errors are larger than observational uncertainty, and model performances on the ice concentration, extent, edge location, and ice drift simulations are consistent when comparing to different observational references. For the ice thickness and snow depth evaluation, the rating scores are not consistent compared to different observational references, which is due to the limited observations and to the fact that different months were chosen for comparison during 2003-2007. This finding shows that sea ice thickness and snow depth estimates are still at a more early stage of maturity compared to datasets of sea ice concentration or drift.

The improvements of mean ice concentration simulations in the summer for both hemispheres by changing the atmospheric forcing and increasing the horizontal resolution are also identified in Tsujino et al. (2020) and Chassignet et al. (2020). The reduced mean ice extent bias in boreal summer and much improved interannual variability of ice extent in OMIP2 simulations are also proved in Tsujino et al. (2020). For the mean ice thickness simulation, Chassignet et al. (2020) also shows that the improvement is not obvious by increasing the horizontal resolution of ocean-sea ice models. To understand the processes leading to the improvement of model simulations under different atmospheric forcings and ocean-sea ice models, we will discuss the sensitivity of sea ice simulation to CMIP6 OMIP model physics in an upcoming companion paper.

The metrics make a summary of the model performance on different aspects of the sea ice system to help detect the inter-model differences or track the time-evolution of model performance efficiently.

750 However, the usage of metrics comes at the risk of over-interpretation by summarizing all the complex behavior of models to one number. In fact, a good metric can be obtained for many wrong reasons, so that we do not recommend relying exclusively on these metrics to orient strategic choices regarding, e.g., sea ice model development. While it is running, SITool (v1.0) produces spatial maps (Figs. 8, 9 and Figs. A1-A12 in Appendix A) and time series diagrams (Figs. 3, 4, 6) that can be consulted by the
755 expert to understand the origin of one particular metric value.

While SITool (v1.0) is primarily designed to assess ocean-sea ice simulations forced by atmospheric reanalysis, it can also be used to evaluate coupled model simulations (e.g., CMIP6 historical runs). We draw the reader's attention to the fact that, in that case, several metrics may become less relevant and less easy to interpret. Indeed, a coupled model is not supposed to produce sea ice output that is in phase
760 with real observations due to the presence of irreducible climate internal variability. This is particularly true for the evaluation of sea ice thickness and snow depth, for which the limited time span (2003-2007) is likely not enough to draw robust conclusions regarding model performance.

Appendices A: Sea ice diagnostics

In this appendix, additional sea ice diagnostics are given to help understand why metrics vary from one
765 dataset to the next. The spatial distribution of the differences between model simulations and the observational reference is presented Figs. A1-A12 and the model simulations under OMIP1 and OMIP2 protocols are listed in the second and third columns, respectively. This includes the 1980-2007 September and February mean ice concentration differences (Figs. A1-A4), the 2003-2007 winter mean ice thickness (Figs. A5-A6) and snow depth differences (Figs. A7-A8) (February for the Arctic and
770 September for the Antarctic), and the 2003-2007 September and February mean ice-motion MKE differences (Figs. A9-A12).

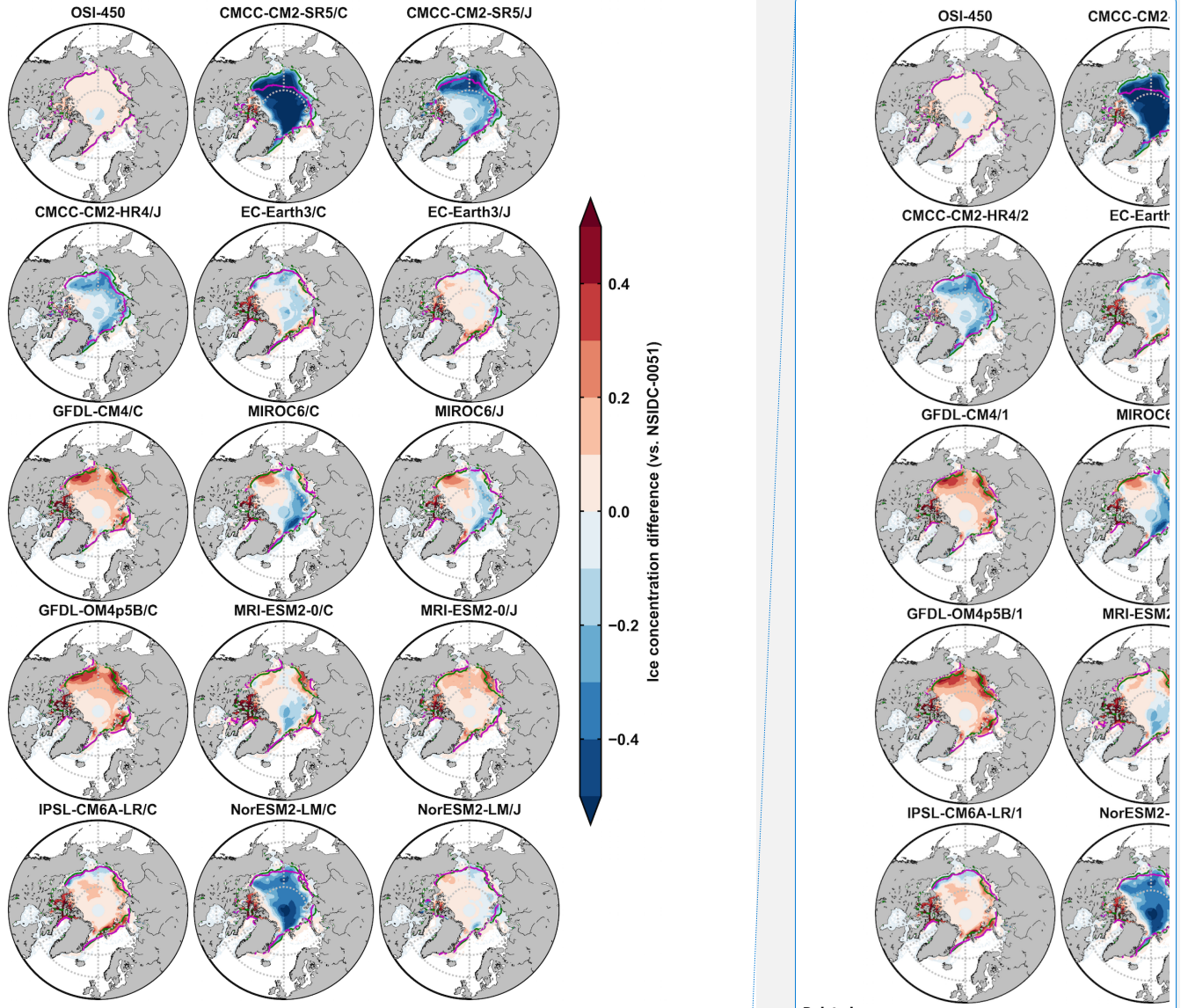
Deleted: ancillary
Deleted:
Deleted: and time series

775

Figure A1. The 1980-2007 September mean Arctic ice concentration differences between OSI-450/model outputs and the NSIDC-0051 data (colors), and contours of 15% concentration of the NSIDC-0051 data (green lines) and OSI-450/model outputs (magenta lines). The second and third columns are from 5 OMIP1 and OMIP2 model outputs of CMCC-CM2-SR5, EC-Earth3, MIROC6, MRI-ESM2-0, NorESM2-LM, respectively.

Figure A2. The 1980-2007 February mean Arctic ice concentration differences between OSI-450/model outputs and the NSIDC-0051 data (colors), and contours of 15% concentration of the NSIDC-0051 data (green lines) and OSI-450/model outputs (magenta lines). The second and third columns are from 5 OMIP1 and OMIP2 model outputs of CMCC-CM2-SR5, EC-Earth3, MIROC6, MRI-ESM2-0, NorESM2-LM, respectively.

785

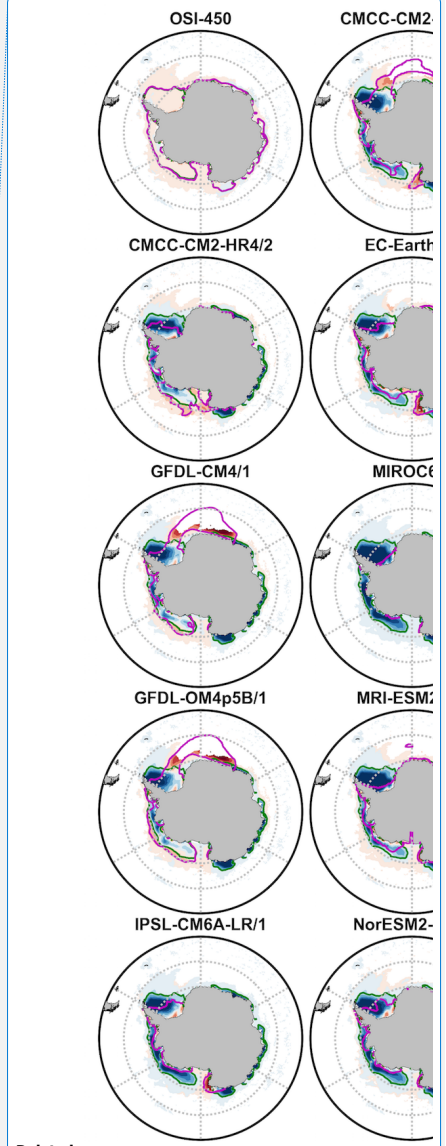
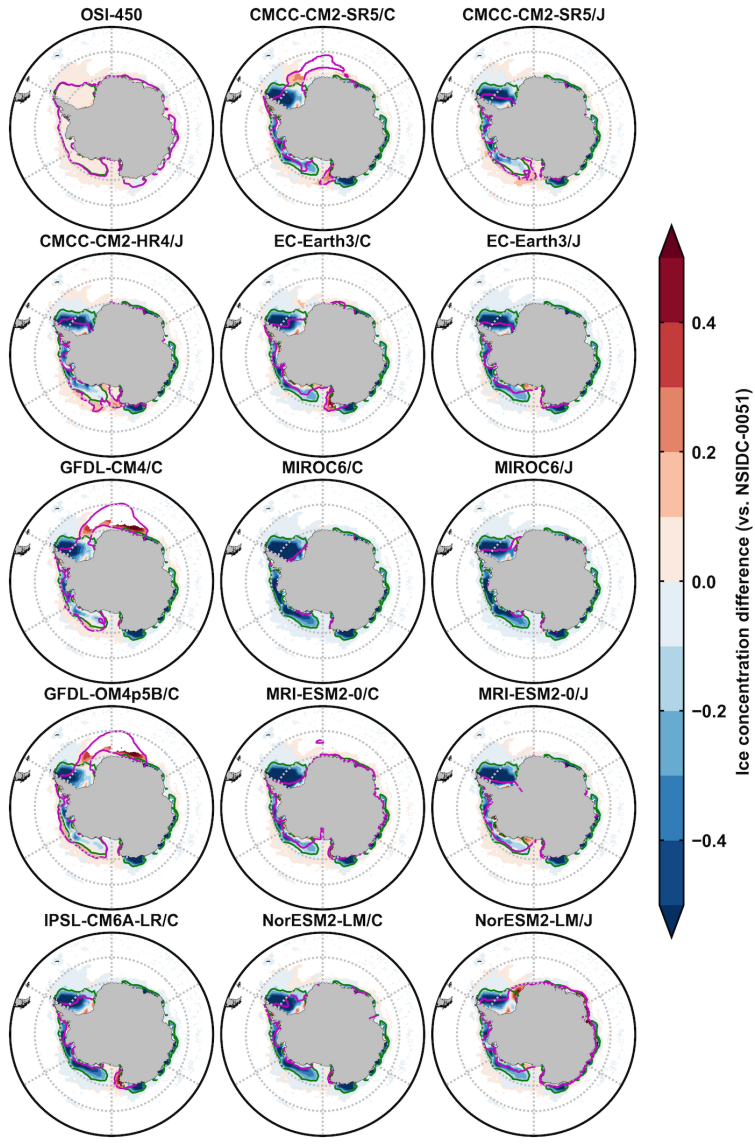



Figure A3. The 1980-2007 February mean Antarctic ice concentration differences between OSI-450/model outputs and the NSIDC-0051 data (colors), and contours of 15% concentration of the NSIDC-0051 data (green lines) and OSI-450/model outputs (magenta lines). The second and third columns are from 5 OMIP1 and OMIP2 model outputs of CMCC-CM2-SR5, EC-Earth3, MIROC6, MRI-ESM2-0, NorESM2-LM, respectively.

790

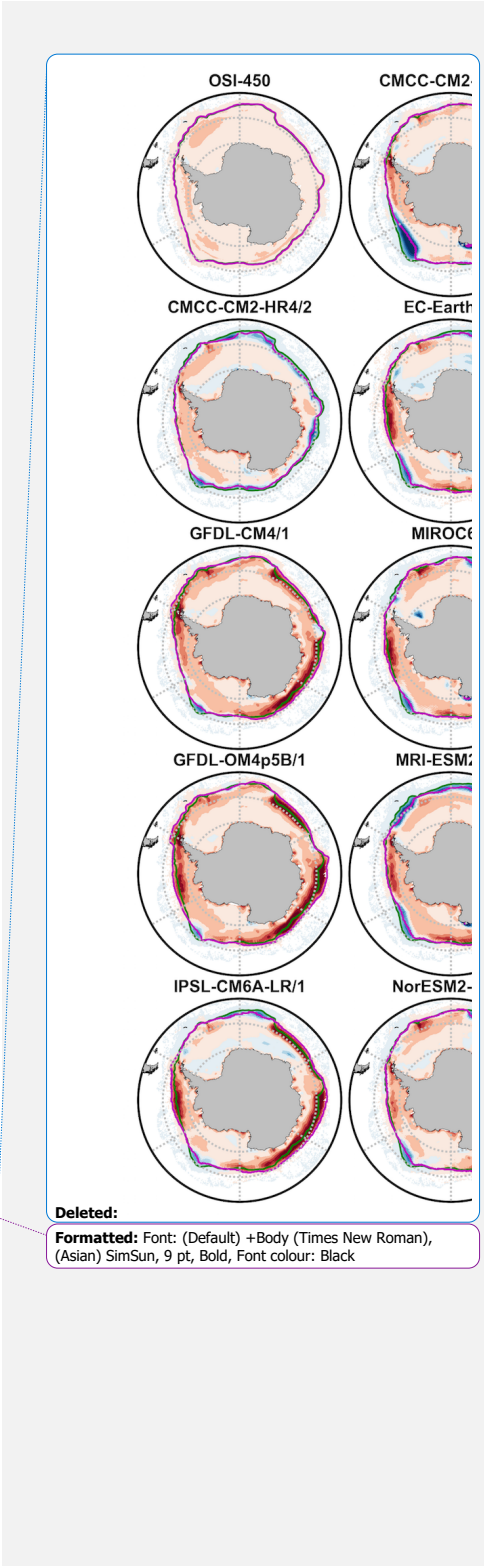
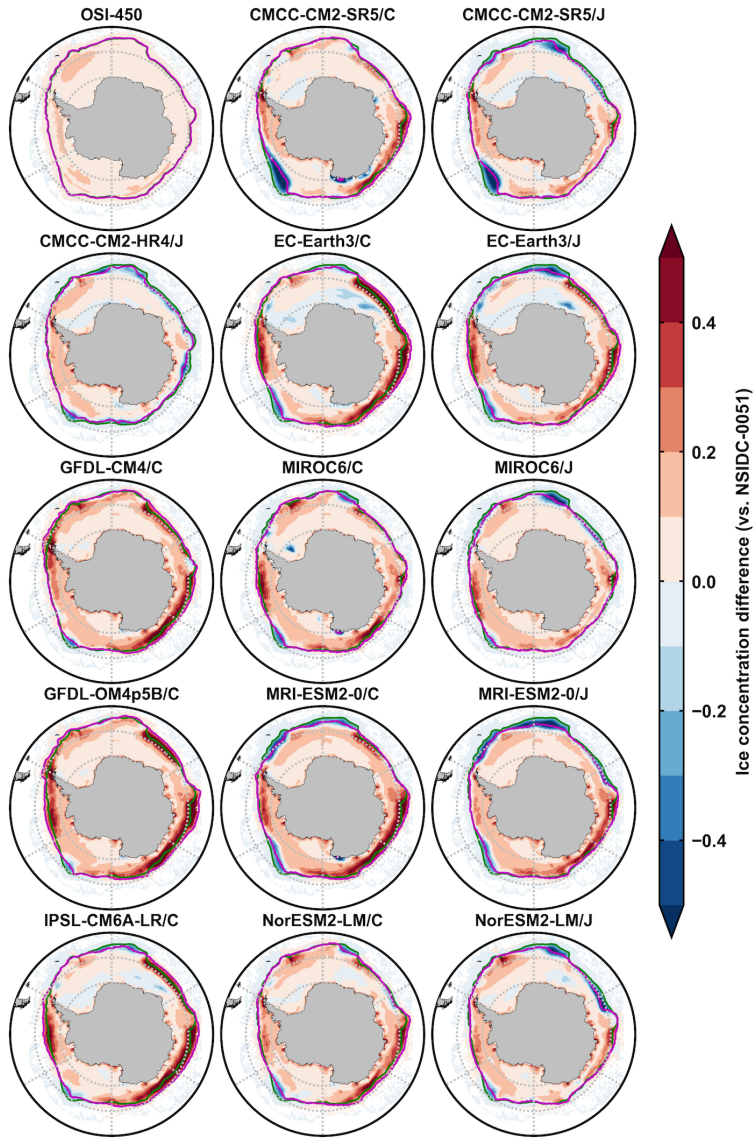



Figure A4. The 1980-2007 September mean Antarctic ice concentration differences between OSI-450/model outputs and the
795 **NSIDC-0051 data (colors), and contours of 15% concentration of the NSIDC-0051 data (green lines) and OSI-450/model outputs**
(magenta lines). The second and third columns are from 5 OMIP1 and OMIP2 model outputs of CMCC-CM2-SR5, EC-Earth3,
MIROC6, MRI-ESM2-0, NorESM2-LM, respectively.

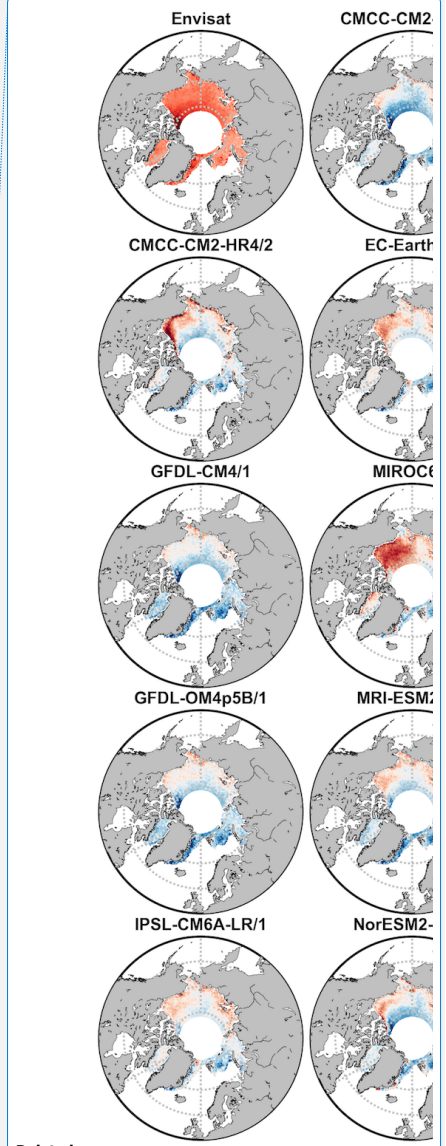
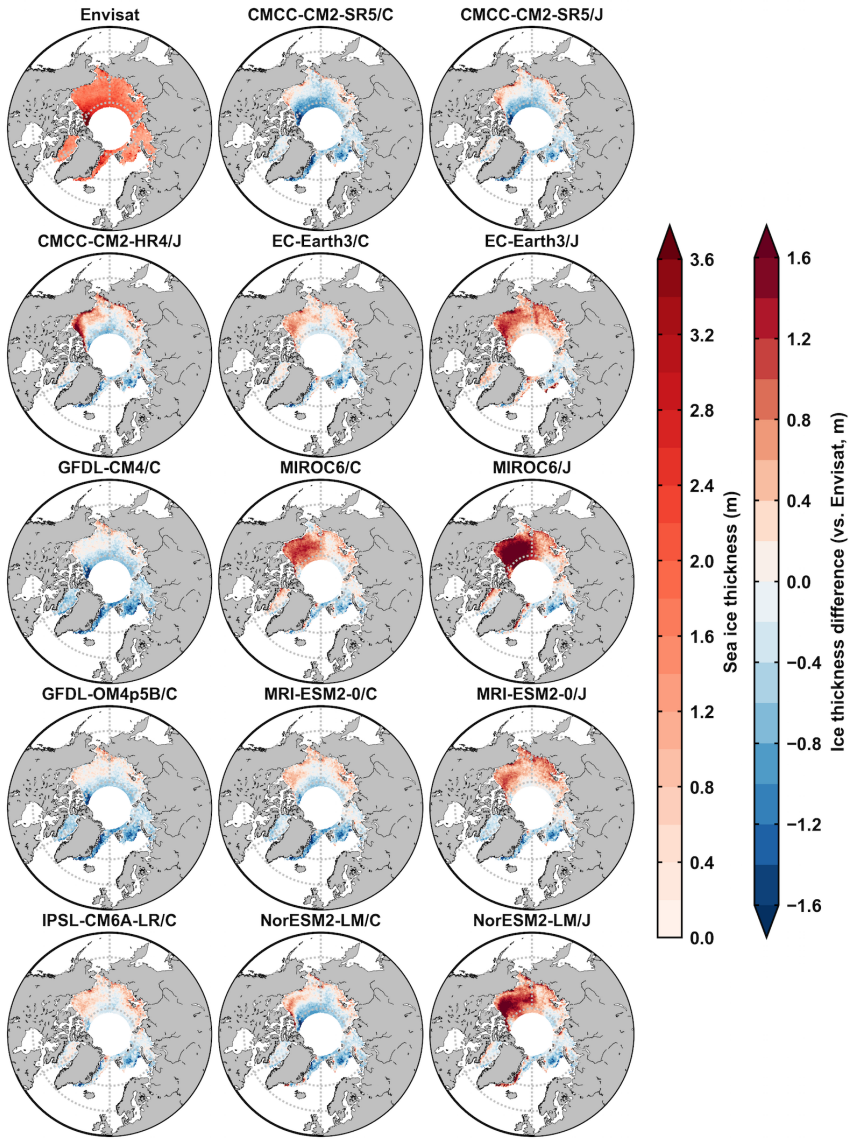
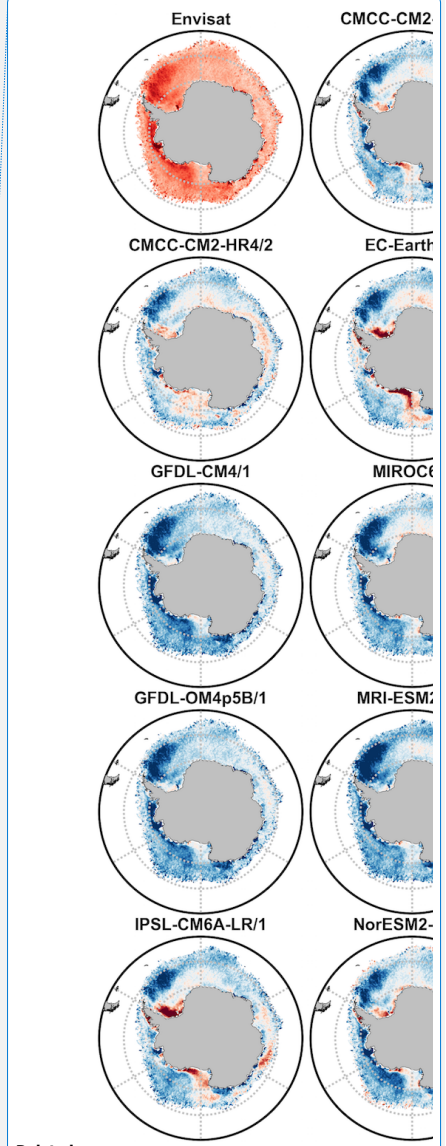
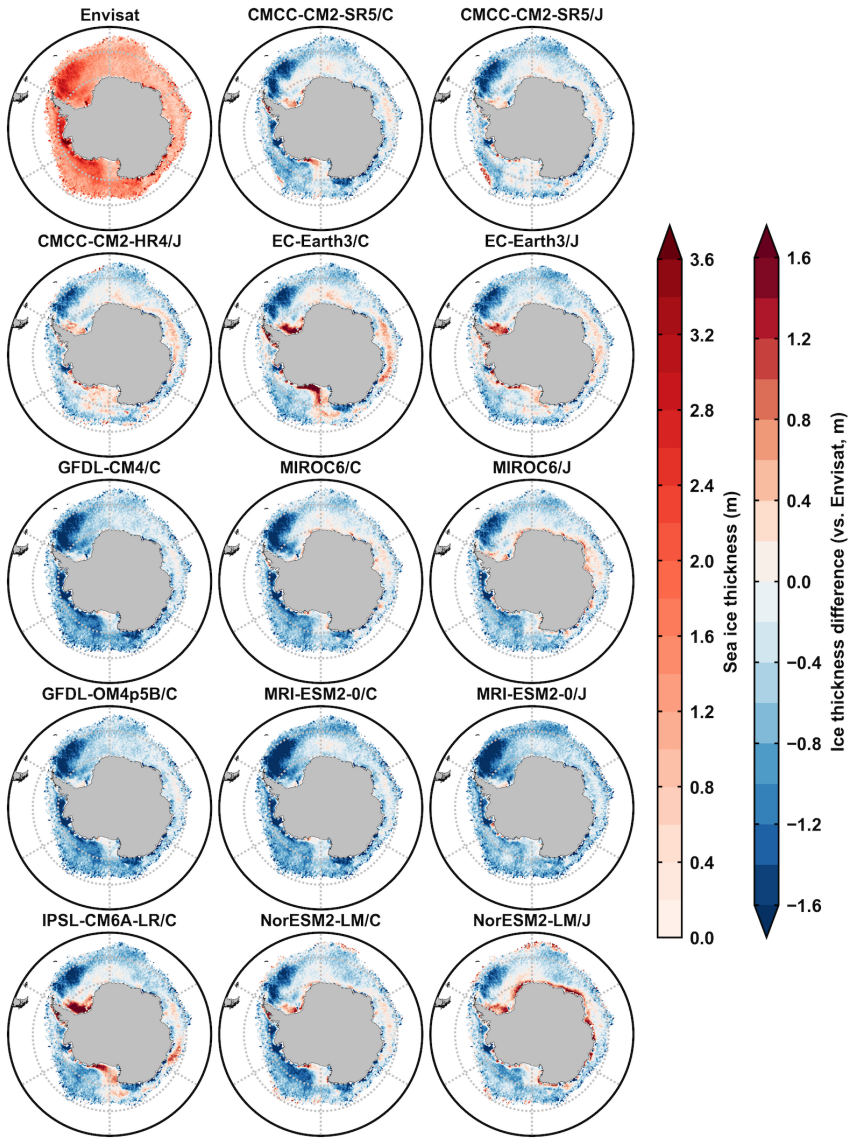
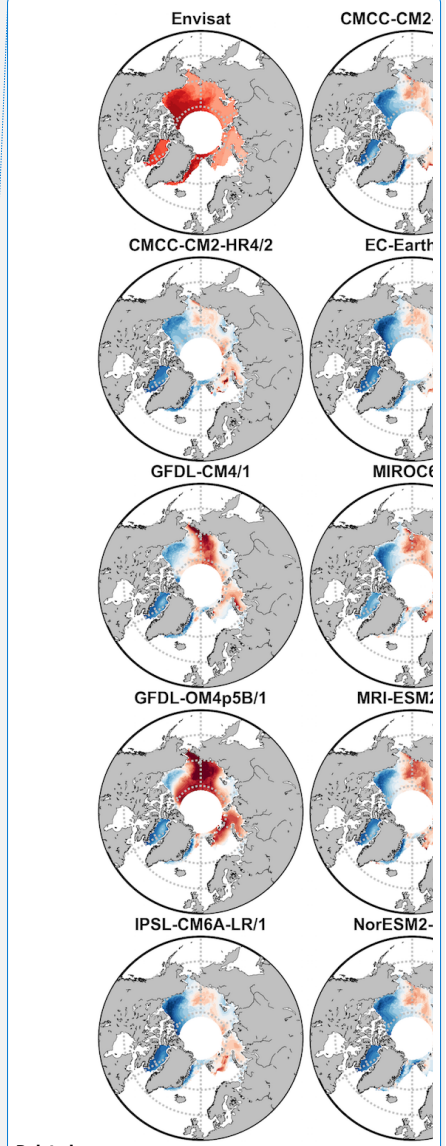
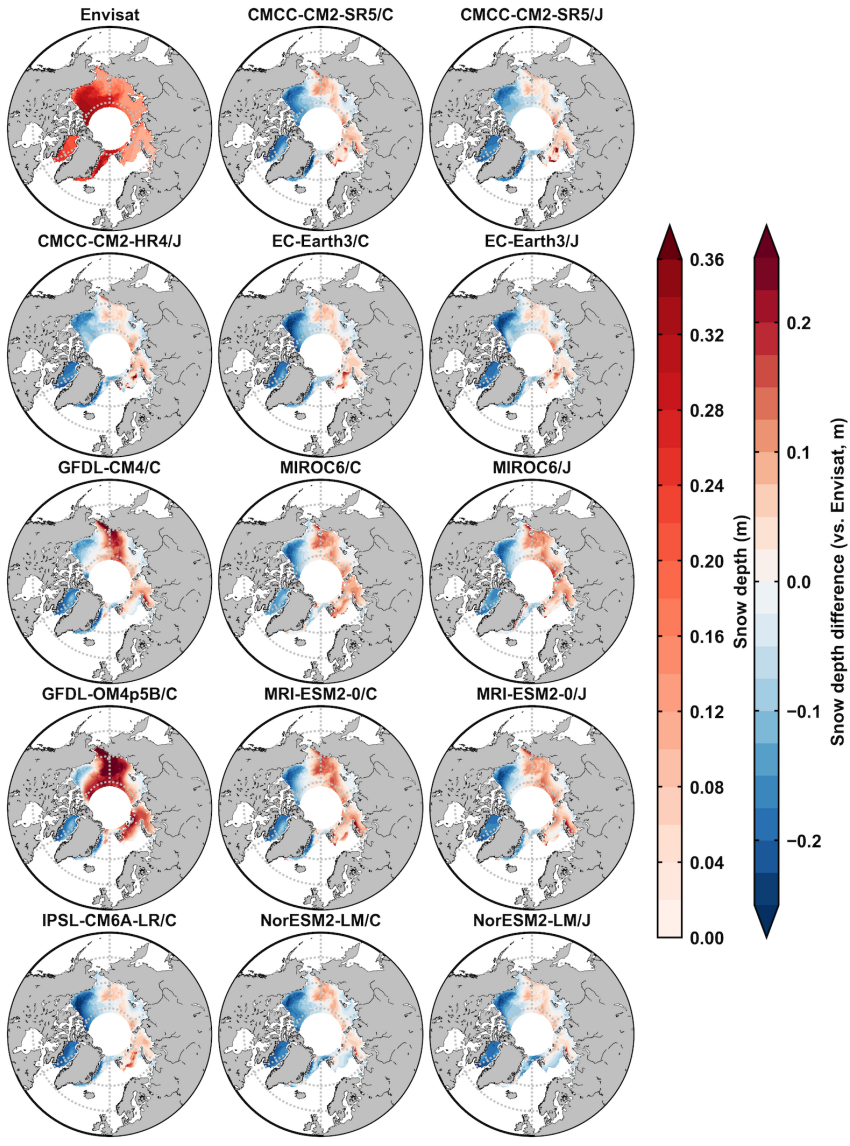
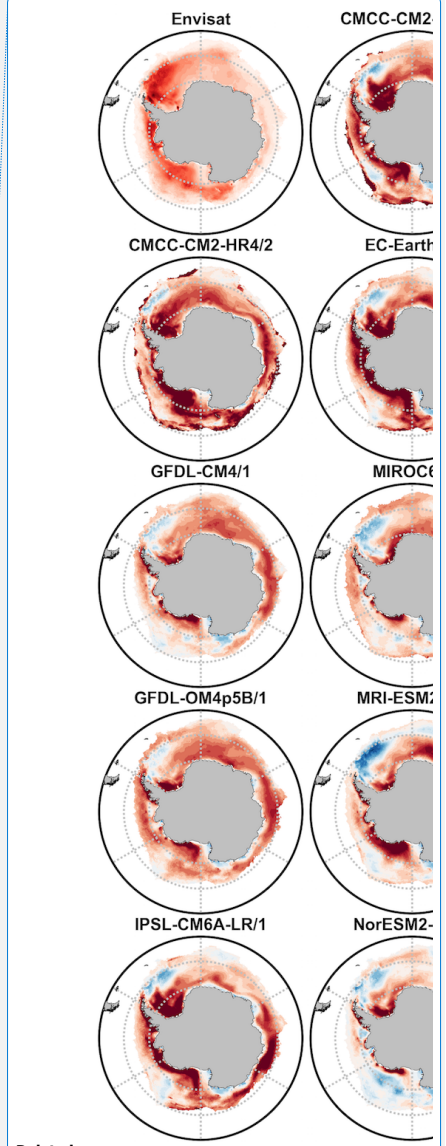
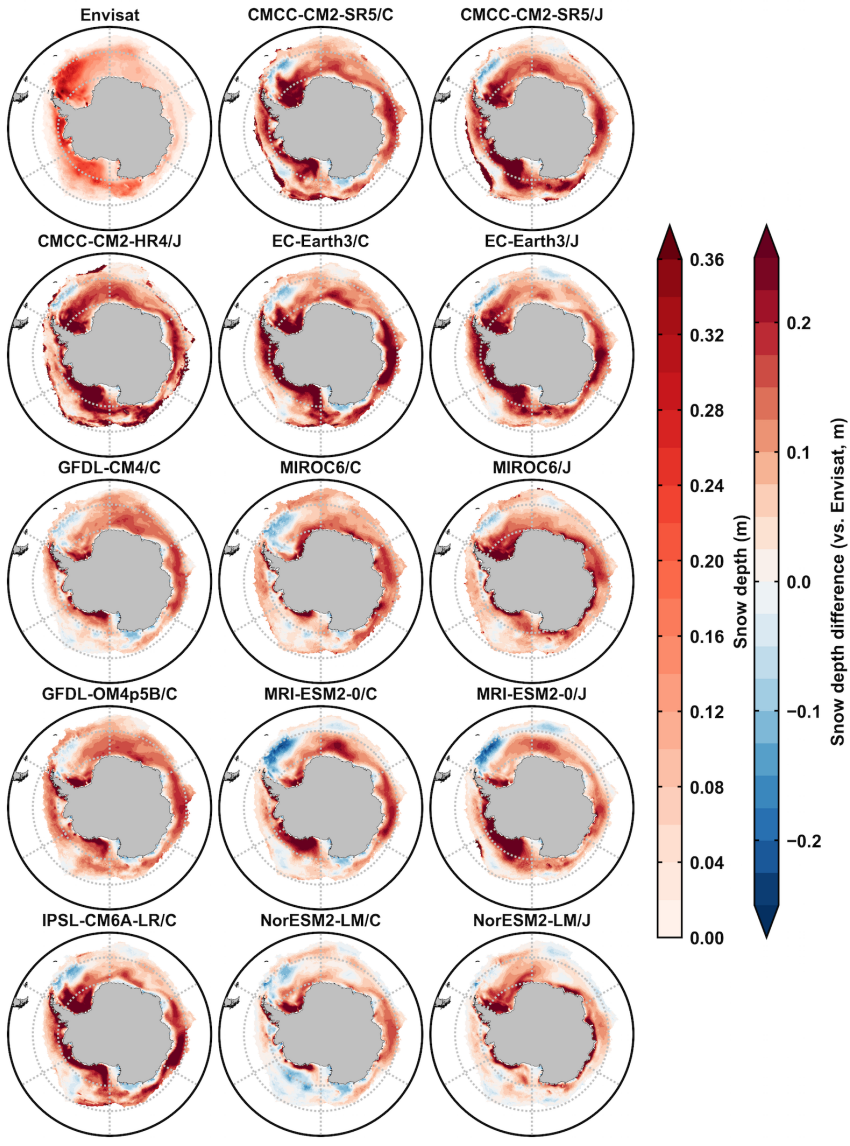
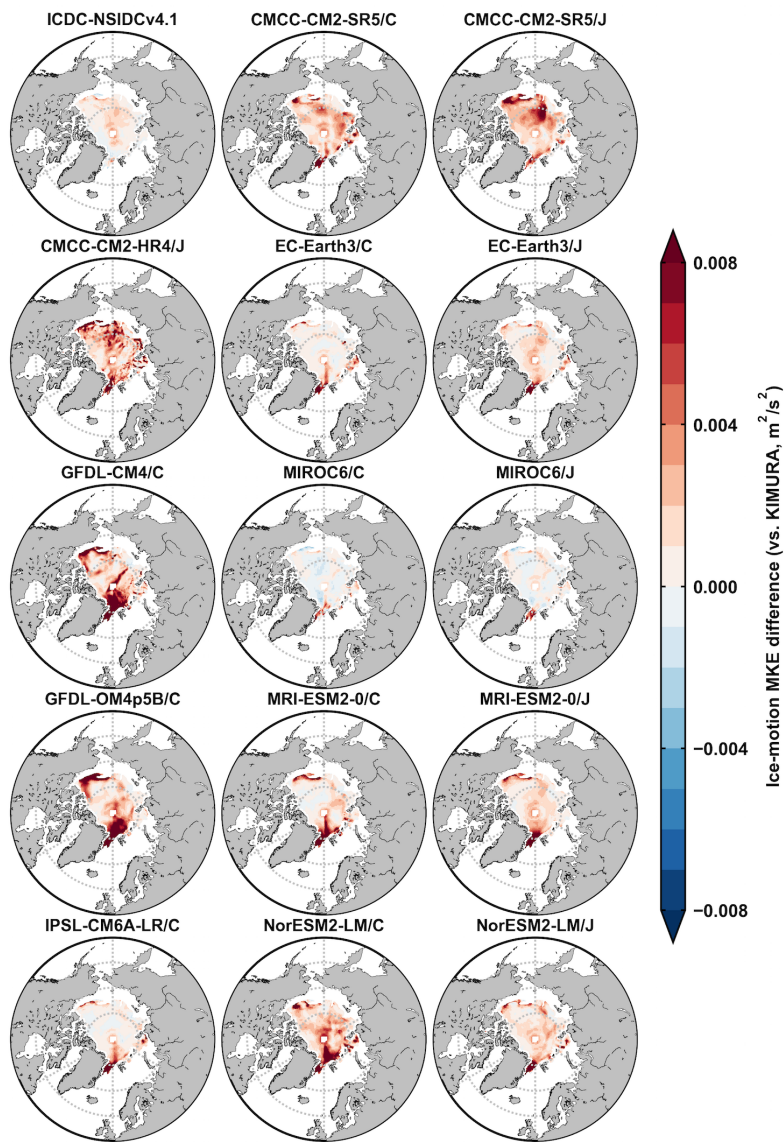
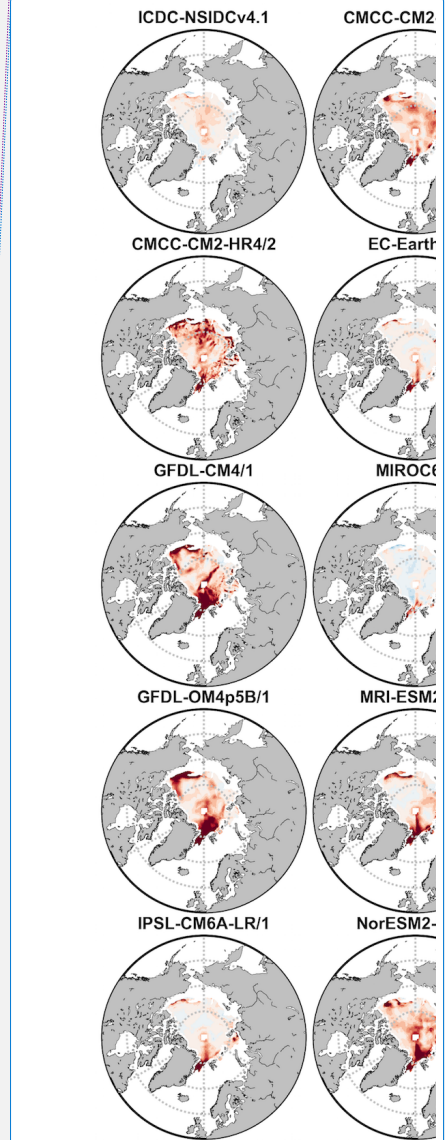
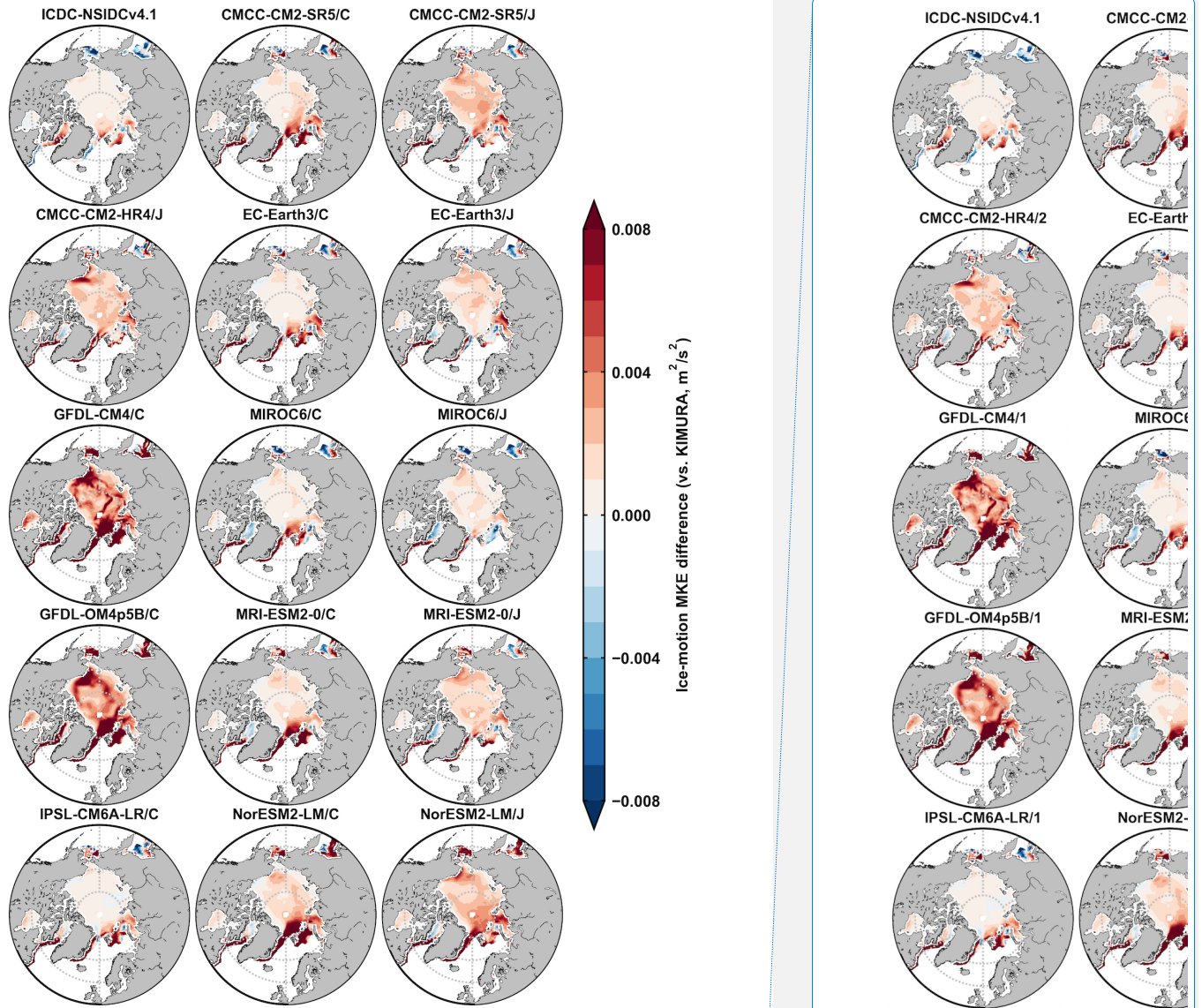
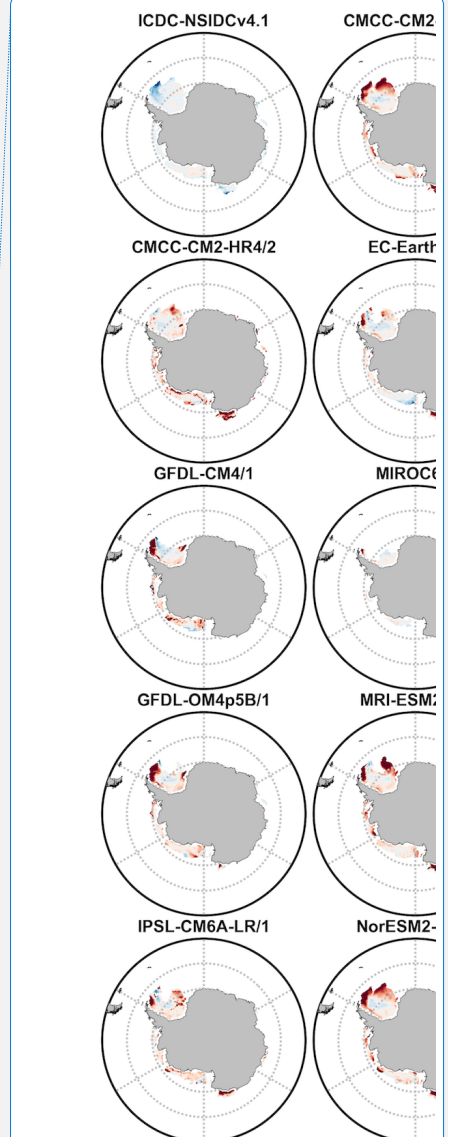
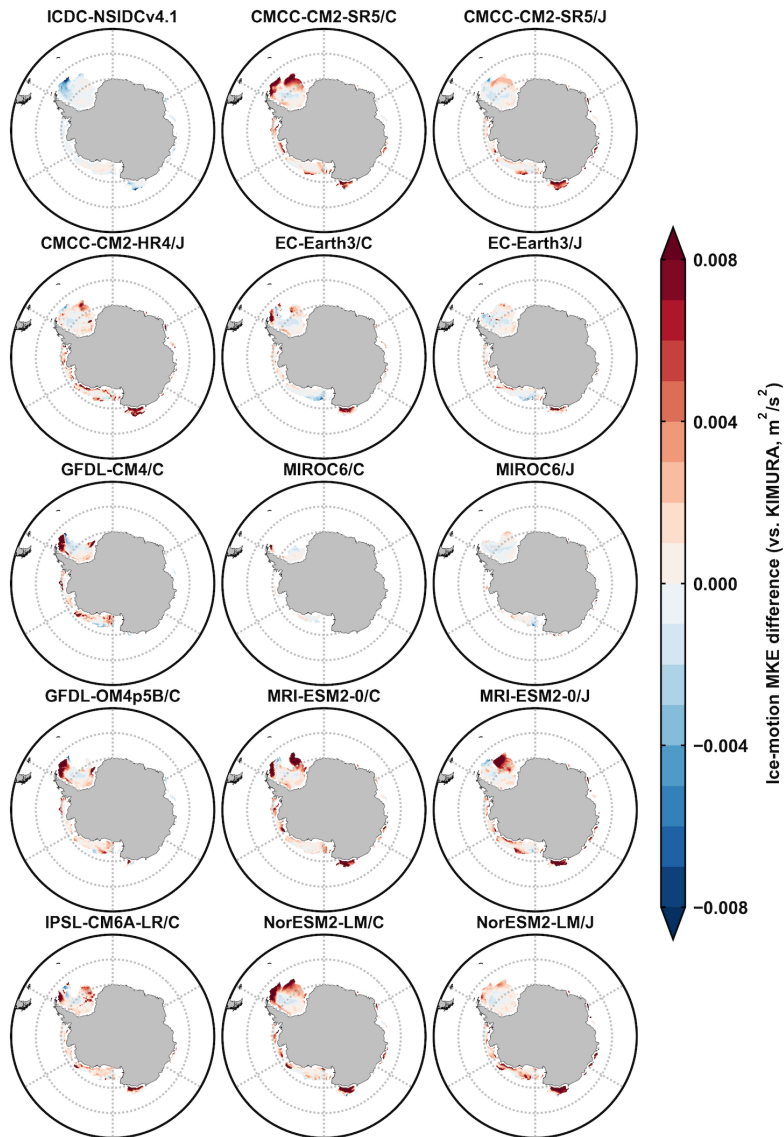
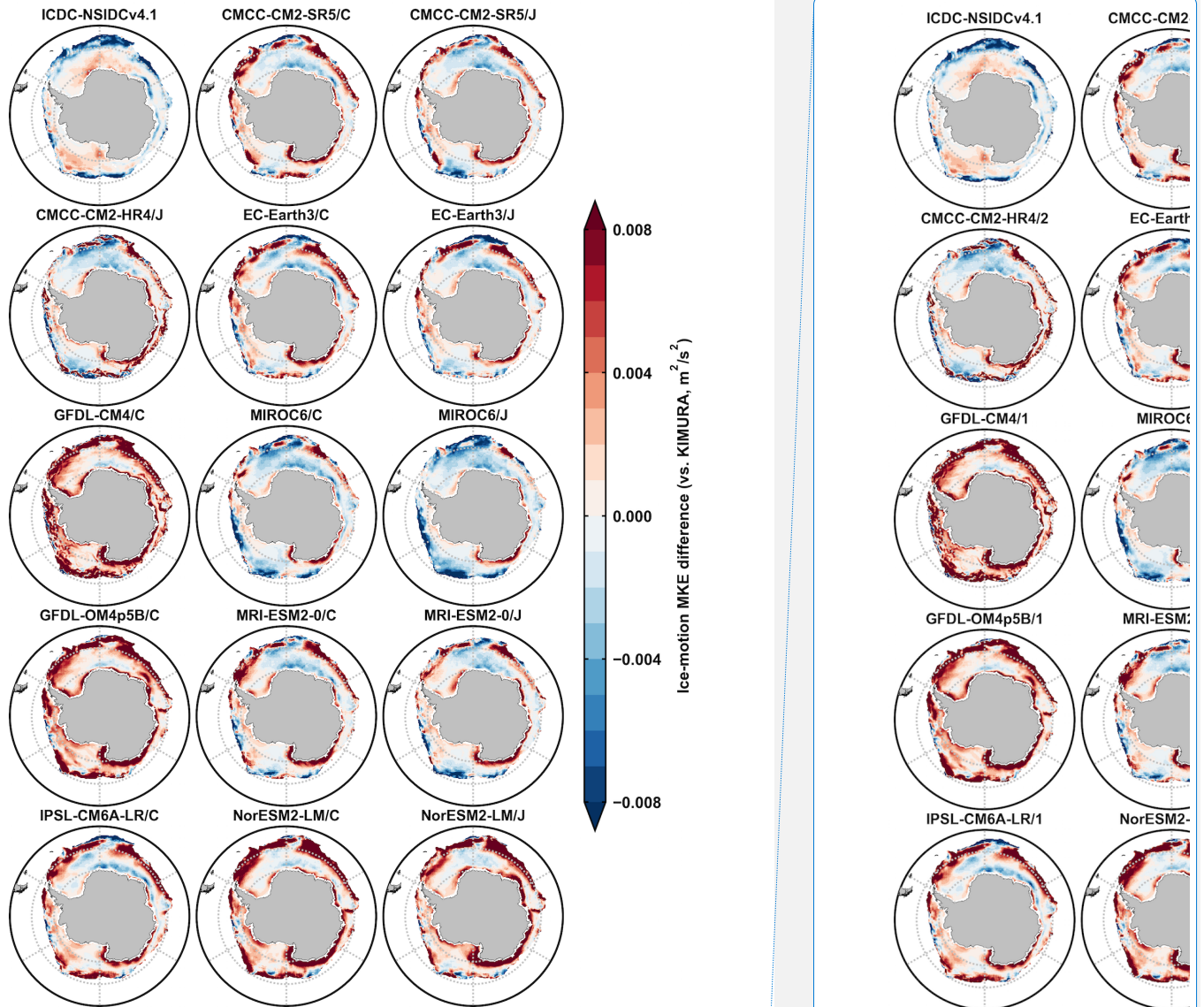



Figure A5. The 2003-2007 February mean Arctic ice thickness from Envisat data (first picture, m) and ice thickness differences between model outputs and Envisat data (m). The second and third columns are from 5 OMIP1 and OMIP2 model outputs of CMCC-CM2-SR5, EC-Earth3, MIROC6, MRI-ESM2-0, NorESM2-LM, respectively.

800

805 **Figure A6. The 2003-2007 September mean Antarctic ice thickness from Envisat data (first picture, m) and ice thickness differences between model outputs and Envisat data (m). The second and third columns are from 5 OMIP1 and OMIP2 model outputs of CMCC-CM2-SR5, EC-Earth3, MIROC6, MRI-ESM2-0, NorESM2-LM, respectively.**

810 **Figure A7. The 2003-2007 February mean Arctic snow depth from Envisat data (first picture, m) and snow depth differences between model outputs and Envisat data (m). The second and third columns are from 5 OMIP1 and OMIP2 model outputs of CMCC-CM2-SR5, EC-Earth3, MIROC6, MRI-ESM2-0, NorESM2-LM, respectively.**

815 **Figure A8. The 2003-2007 September mean Antarctic snow depth from Envisat data (first picture, m) and snow depth differences between model outputs and Envisat data (m). The second and third columns are from 5 OMIP1 and OMIP2 model outputs of CMCC-CM2-SR5, EC-Earth3, MIROC6, MRI-ESM2-0, NorESM2-LM, respectively.**

820 **Figure A9. The 2003-2007 September mean Arctic ice-motion MKE differences between ICDC-NSIDCv4.1/model outputs and the KIMURA data (m² s⁻²). The second and third columns are from 5 OMIP1 and OMIP2 model outputs of CMCC-CM2-SR5, EC-Earth3, MIROC6, MRI-ESM2-0, NorESM2-LM, respectively.**

825 **Figure A10. The 2003-2007 February mean Arctic ice-motion MKE differences between ICDC-NSIDCv4.1/model outputs and the KIMURA data (m² s⁻²). The second and third columns are from 5 OMIP1 and OMIP2 model outputs of CMCC-CM2-SR5, EC-Earth3, MIROC6, MRI-ESM2-0, NorESM2-LM, respectively.**

830 **Figure A11. The 2003-2007 February mean Antarctic MKE differences between ICDC-NSIDCv4.1/model outputs and the KIMURA data (m² s⁻²). The second and third columns are from 5 OMIP1 and OMIP2 model outputs of CMCC-CM2-SR5, EC-Earth3, MIROC6, MRI-ESM2-0, NorESM2-LM, respectively.**

835 **Figure A12. The 2003-2007 September mean Antarctic MKE differences between ICDC-NSIDCv4.1/model outputs and the KIMURA data (m² s⁻²). The second and third columns are from 5 OMIP1 and OMIP2 model outputs of CMCC-CM2-SR5, EC-Earth3, MIROC6, MRI-ESM2-0, NorESM2-LM, respectively.**

Code and data availability. The latest release of SITool (v1.0) is publicly available on Zenodo at https://doi.org/10.5281/zenodo.4621147 (Lin et al., 2021). The source code of the SITool (v1.0) is developed fully based on freely available Python packages and libraries and is released on the GitHub repository available at https://github.com/XiaLinUCL/Sea-Ice-Evaluation-Tool. CMIP6 OMIP data are

845 freely available from the Earth System Grid Federation. Observational references used in this paper are detailed in section 3 and listed in Table 3, and they are not distributed with SITool (v1.0) because SITool (v1.0) is restricted to the code as open-source software.

Author contributions. All authors contributed to the design and discussion of the study. XL performed the analysis and prepared the SITool codes with the help of FM and TF. XL led the writing of the paper

850 and all authors contributed to the editing of the manuscript.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgments. This work has been carried out as part of the Copernicus Marine Environment Monitoring Service (CMEMS) SI3 project. CMEMS is implemented by Mercator Ocean International in the framework of a delegation agreement with the European Union. Xia Lin also received support

855 from the National Natural Science Foundation of China (Grant Nos. 41906190, 41941007 and 41876220). François Massonnet is a F.R.S.-FNRS Research Fellow.

References

Alexander Haumann, F., Gruber, N., Münnich, M., Frenger, I. and Kern, S.: Sea-ice transport driving Southern Ocean salinity and its recent trends, Nature, 537(7618), 89–92, doi:10.1038/nature19101,

860 2016.

Barthélemy, A., Goosse, H., Fichefet, T. and Lecomte, O.: On the sensitivity of Antarctic sea ice model biases to atmospheric forcing uncertainties, Clim. Dyn., 51(4), 1585–1603, doi:10.1007/s00382-017-3972-7, 2018.

Bentley, J. L.: Multidimensional binary search trees used for associative searching, Commun. ACM,

865 18(9), 509–517, 1975.

Boucher, O., Servonnat, J., Albright, A. L., Aumont, O., Balkanski, Y., Bastrikov, V., Bekki, S., Bonnet, R., Bony, S., Bopp, L., Braconnot, P., Brockmann, P., Cadule, P., Caubel, A., Cheruy, F., Codron, F., Cozic, A., Cugnet, D., D'Andrea, F., Davini, P., de Lavergne, C., Denvil, S., Deshayes, J., Devilliers, M., Ducharne, A., Dufresne, J. L., Dupont, E., Éthé, C., Fairhead, L., Falletti, L., Flavoni, S., Foujols,

870 M. A., Gardoll, S., Gastineau, G., Ghattas, J., Grandpeix, J. Y., Guenet, B., Guez, L. E., Guilyardi, E., Guimberteau, M., Hauglustaine, D., Hourdin, F., Idelkadi, A., Joussaume, S., Kageyama, M., Khodri, M., Krinner, G., Lebas, N., Levavasseur, G., Lévy, C., Li, L., Lott, F., Lurton, T., Luyssaert, S., Madec, G., Madeleine, J. B., Maignan, F., Marchand, M., Marti, O., Mellul, L., Meurdesoif, Y., Mignot, J., Musat, I., Ottlé, C., Peylin, P., Planton, Y., Polcher, J., Rio, C., Rochetin, N., Rousset, C., Sepulchre, P.,

875 Sima, A., Swingedouw, D., Thiéblemont, R., Traore, A. K., Vancoppenolle, M., Vial, J., Vialard, J.,

Viovy, N. and Vuichard, N.: Presentation and Evaluation of the IPSL-CM6A-LR Climate Model, J. Adv. Model. Earth Syst., 12(7), 1–52, doi:10.1029/2019MS002010, 2020.

880 Cavalieri, D. J., Parkinson, C. L., Gloersen, P., and Zwally, H. J.: Sea Ice Concentrations from Nimbus-7 SMMR and DMSP SSM/I-SSMIS Passive Microwave Data, Version 1, [1980-2007], Boulder, Colorado USA, NASA National Snow and Ice Data Center Distributed Active Archive Center, doi: https://doi.org/10.5067/8GQ8LZQVL0VL, 1996, updated yearly.

Chassignet, E. P., Yeager, S. G., Fox-Kemper, B., Bozec, A., Castruccio, F., Danabasoglu, G., Horvat, C., Kim, W. M., Koldunov, N., Li, Y., Lin, P., Liu, H., Sein, D. V., Sidorenko, D., Wang, Q. and Xu, X.:
885 Impact of horizontal resolution on global ocean-sea ice model simulations based on the experimental protocols of the Ocean Model Intercomparison Project phase 2 (OMIP-2), Geosci. Model Dev., 13(9), 4595–4637, doi:10.5194/gmd-13-4595-2020, 2020.

Cherchi, A., Fogli, P. G., Lovato, T., Peano, D., Iovino, D., Gualdi, S., Masina, S., Scoccimarro, E., Materia, S., Bellucci, A. and Navarra, A.: Global Mean Climate and Main Patterns of Variability in the
890 CMCC-CM2 Coupled Model, J. Adv. Model. Earth Syst., 11(1), 185–209, doi:10.1029/2018MS001369, 2019.

Crosby, D. S., Breaker, L. C. and Gemmill, W. H.: A Proposed Definition for Vector Correlation in Geophysics: Theory and Application, J. Atmos. Ocean. Technol., 10(3), 355–367, 1993.

EC-Earth Consortium (EC-Earth): EC-Earth-Consortium EC-Earth3-Veg model output prepared for
895 CMIP6 ScenarioMIP, Version 20200601, Earth System Grid Federation, https://doi.org/10.22033/ESGF/CMIP6.727, 2019.

EUMETSAT Ocean and Sea Ice Satellite Application Facility: Global sea ice concentration climate data record 1979-2015 (v2.0), Norwegian and Danish Meteorological Institutes, doi: 10.15770/EUM_SAF_OSI_0008, 2017.

900 Eyring, V., Bock, L., Lauer, A., Righi, M., Schlund, M., Andela, B., Arnone, E., Bellprat, O., Carvalhais, N., Cionni, I., Cortesi, N., Crezee, B., L. Davin, E., Davini, P., Debeire, K., De Mora, L., Deser, C., Docquier, D., Earnshaw, P., Ehbrecht, C., K. Gier, B., Gonzalez-Reviriego, N., Goodman, P., Hagemann, S., Hardiman, S., Hassler, B., Hunter, A., Kadow, C., Kindermann, S., Koirala, S., Koldunov, N., Lejeune, Q., Lembo, V., Lovato, T., Lucarini, V., Müller, B., Pandde, A., Phillips, A.,
905 Predoi, V., Russell, J., Sellar, A., Serva, F., Stacke, T., Swaminathan, R., Vegas-Regidor, J., Von Hardenberg, J., Weigel, K. and Zimmermann, K.: Earth System Model Evaluation Tool (ESMValTool) v2.0 - An extended set of large-scale diagnostics for quasi-operational and comprehensive evaluation of Earth system models in CMIP, Geosci. Model Dev., 13(7), 3383–3438, doi:10.5194/gmd-13-3383-2020, 2020.

910 Eyring, V., Righi, M., Lauer, A., Evaldsson, M., Wenzel, S., Jones, C., Anav, A., Andrews, O., Cionni, I., Davin, E. L., Deser, C., Ehbrecht, C., Friedlingstein, P., Gleckler, P., Gottschaldt, K. D., Hagemann, S., Juckes, M., Kindermann, S., Krasting, J., Kunert, D., Levine, R., Loew, A., Mäkelä, J., Martin, G., Mason, E., Phillips, A. S., Read, S., Rio, C., Roehrig, R., Senftleben, D., Sterl, A., Van Ulft, L. H., Walton, J., Wang, S. and Williams, K. D.: ESMValTool (v1.0)-a community diagnostic and performance
915 metrics tool for routine evaluation of Earth system models in CMIP, Geosci. Model Dev., 9(5), 1747–1802, doi:10.5194/gmd-9-1747-2016, 2016.

Goessling, H. F., Tietsche, S., Day, J. J., Hawkins, E. and Jung, T.: Predictability of the Arctic sea ice edge, Geophys. Res. Lett., 43(4), 1642–1650, doi:10.1002/2015GL067232, 2016.

Griffies, S. M., Danabasoglu, G., Durack, P. J., Adcroft, A. J., Balaji, V., Böning, C. W., Chassignet, E. P., Curchitser, E., Deshayes, J., Drange, H., Fox-Kemper, B., Gleckler, P. J., Gregory, J. M., Haak, H., Hallberg, R. W., Heimbach, P., Hewitt, H. T., Holland, D. M., Ilyina, T., Jungclaus, J. H., Komuro, Y., Krasting, J. P., Large, W. G., Marsland, S. J., Masina, S., McDougall, T. J., George Nurser, A. J., Orr, J. C., Pirani, A., Qiao, F., Stouffer, R. J., Taylor, K. E., Treguier, A. M., Tsujino, H., Uotila, P., Valdivieso, M., Wang, Q., Winton, M. and Yeager, S. G.: OMIP contribution to CMIP6: Experimental and diagnostic protocol for the physical component of the Ocean Model Intercomparison Project, Geosci. Model Dev., 9(9), 3231–3296, doi:10.5194/gmd-9-3231-2016, 2016.

Guerreiro, K., Fleury, S., Zakharova, E., Kouraev, A., Rémy, F. and Maisongrande, P.: Comparison of CryoSat-2 and ENVISAT radar freeboard over Arctic sea ice: Toward an improved Envisat freeboard retrieval, Cryosphere, 11(5), 2059–2073, doi:10.5194/tc-11-2059-2017, 2017.

Held, I. M., Guo, H., Adcroft, A., Dunne, J. P., Horowitz, L. W., Krasting, J., Shevliakova, E., Winton, M., Zhao, M., Bushuk, M., Wittenberg, A. T., Wyman, B., Xiang, B., Zhang, R., Anderson, W., Balaji, V., Donner, L., Dunne, K., Durachta, J., Gauthier, P. P. G., Ginoux, P., Golaz, J. C., Griffies, S. M., Hallberg, R., Harris, L., Harrison, M., Hurlin, W., John, J., Lin, P., Lin, S. J., Malyshev, S., Menzel, R., Milly, P. C. D., Ming, Y., Naik, V., Paynter, D., Paulot, F., Rammaswamy, V., Reichl, B., Robinson, T., Rosati, A., Seman, C., Silvers, L. G., Underwood, S. and Zadeh, N.: Structure and Performance of GFDL's CM4.0 Climate Model, J. Adv. Model. Earth Syst., 11(11), 3691–3727, doi:10.1029/2019MS001829, 2019.

Holland, P. R. and Kwok, R.: Wind-driven trends in Antarctic sea-ice drift, Nat. Geosci., 5(12), 872–875, doi:10.1038/ngeo1627, 2012.

Kimura, N., Nishimura, A., Tanaka, Y. and Yamaguchi, H.: Influence of winter sea-ice motion on summer ice cover in the Arctic, Polar Res., 32(1), doi:10.3402/polar.v32i0.20193, 2013.

Kurtz, N. T. and Markus, T.: Satellite observations of Antarctic sea ice thickness and volume, J. Geophys. Res. Ocean., 117(8), doi:10.1029/2012JC008141, 2012.

Large, W. G. and Yeager, S. G.: The global climatology of an interannually varying air - Sea flux data set, Clim. Dyn., 33(2–3), 341–364, doi:10.1007/s00382-008-0441-3, 2009.

Lavergne, T., Macdonald Sørensen, A., Kern, S., Tonboe, R., Notz, D., Aaboe, S., Bell, L., Dybkjær, G., Eastwood, S., Gabarro, C., Heygster, G., Anne Killie, M., Brandt Kreiner, M., Lavelle, J., Saldo, R., Sandven, S. and Pedersen, L. T.: Version 2 of the EUMETSAT OSI SAF and ESA CCI sea-ice concentration climate data records, Cryosphere, 13(1), 49–78, doi:10.5194/tc-13-49-2019, 2019.

Liston, G. E., Itkin, P., Stroeve, J., Tschudi, M., Stewart, J. S., Pedersen, S. H., Reinking, A. K. and Elder, K.: A Lagrangian Snow-Evolution System for Sea-Ice Applications (SnowModel-LG): Part I—Model Description, J. Geophys. Res. Ocean., 125(10), doi:10.1029/2019JC015913, 2020.

Massonnet, F., Fichefet, T., Goosse, H., Vancoppenolle, M., Mathiot, P. and König Beatty, C.: On the influence of model physics on simulations of Arctic and Antarctic sea ice, Cryosphere, 5(3), 687–699, doi:10.5194/tc-5-687-2011, 2011.

Massonnet, F., Fichefet, T., Goosse, H., Bitz, C. M., Philippon-Berthier, G., Holland, M. M. and Barriat, P. Y.: Constraining projections of summer Arctic sea ice, Cryosphere, 6(6), 1383–1394, doi:10.5194/tc-6-1383-2012, 2012.

Monthly mean sea-ice motion vectors derived from Tschudi, M., Meier, W. N., Stewart, J. S., Fowler, C., and Maslanik, J.: Polar Pathfinder Daily 25 km EASE-Grid Sea Ice Motion Vectors, Version 4.1,

Boulder, Colorado USA, NASA National Snow and Ice Data Center Distributed Active Archive Center, Doi: https://doi.org/10.5067/INAWUWO7QH7B, [last access date: July 30, 2020] were provided in netCDF format (file version fv0.01) by the Integrated Climate Data Center (ICDC, icdc.cen.uni-hamburg.de) University of Hamburg, Hamburg, Germany, 2019.

965 Notz, D. and Community, S.: Arctic Sea Ice in CMIP6, Geophys. Res. Lett., 47(10), 1–11, doi:10.1029/2019GL086749, 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, Geophys. Res. Lett., 47(9), 1–10, doi:10.1029/2019GL086729, 2020.

970 Schwegmann, S., Haas, C., Fowler, C. and Gerdes, R.: A comparison of satellite-derived sea-ice motion with drifting-buoy data in the Weddell Sea, Antarctica, Ann. Glaciol., 52(57 PART 1), 103–110, doi:10.3189/172756411795931813, 2011.

Seland, Ø., Bentsen, M., Olivié, D., Toniazzo, T., Gjermundsen, A., Graff, L. S., Debernard, J. B., Gupta, A. K., He, Y. C., Kirkeväg, A., Schwinger, J., Tjiputra, J., Schanke Aas, K., Bethke, I., Fan, Y., 975 Griesfeller, J., Grini, A., Guo, C., Ilicak, M., Karset, I. H. H., Landgren, O., Liakka, J., Moseid, K. O., Nummelin, A., Spensberger, C., Tang, H., Zhang, Z., Heinze, C., Iversen, T. and Schulz, M.: Overview of the Norwegian Earth System Model (NorESM2) and key climate response of CMIP6 DECK, historical, and scenario simulations., 2020.

Shu, Q., Song, Z. and Qiao, F.: Assessment of sea ice simulations in the CMIP5 models, Cryosphere, 980 9(1), 399–409, doi:10.5194/tc-9-399-2015, 2015.

Shu, Q., Wang, Q., Song, Z., Qiao, F., Zhao, J., Chu, M. and Li, X.: Assessment of Sea Ice Extent in CMIP6 With Comparison to Observations and CMIP5, Geophys. Res. Lett., 47(9), 1–9, doi:10.1029/2020GL087965, 2020.

Stroeve, J., Barrett, A., Serreze, M. and Schweiger, A.: Using records from submarine, aircraft and 985 satellites to evaluate climate model simulations of Arctic sea ice thickness, Cryosphere, 8(5), 1839–1854, doi:10.5194/tc-8-1839-2014, 2014.

Stroeve, J., Kattsov, V., Barrett, A., Serreze, M., Pavlova, T., Holland, M. and Meier, W. N.: Trends in Arctic sea ice extent from CMIP5, CMIP3 and observations, Geophys. Res. Lett., 39(16), 1–7, doi:10.1029/2012GL052676, 2012.

990 Stroeve, J., Liston, G. E., Buzzard, S., Zhou, L., Mallett, R., Barrett, A., Tschudi, M., Tsamados, M., Itkin, P. and Stewart, J. S.: A Lagrangian Snow Evolution System for Sea Ice Applications (SnowModel-LG): Part II—Analyses, J. Geophys. Res. Ocean., 125(10), doi:10.1029/2019JC015900, 2020.

Tatebe, H., Ogura, T., Nitta, T., Komuro, Y., Ogochi, K., Takemura, T., Sudo, K., Sekiguchi, M., Abe, 995 M., Saito, F., Chikira, M., Watanabe, S., Mori, M., Hirota, N., Kawatani, Y., Mochizuki, T., Yoshimura, K., Takata, K., O'ishi, R., Yamazaki, D., Suzuki, T., Kurogi, M., Kataoka, T., Watanabe, M. and Kimoto, M.: Description and basic evaluation of simulated mean state, internal variability, and climate sensitivity in MIROC6, Geosci. Model Dev. Discuss., 1–92, doi:10.5194/gmd-2018-155, 2018.

Tilling, R. L., Ridout, A., Shepherd, A. and Wingham, D. J.: Increased Arctic sea ice volume after 000 anomalously low melting in 2013, Nat. Geosci., 8(8), 643–646, doi:10.1038/ngeo2489, 2015.

Tsujino, H., Urakawa, S., Nakano, H., Small, R. J., Kim, W. M., Yeager, S. G., Danabasoglu, G., Suzuki, T., Bamber, J. L., Bentsen, M., Böning, C. W., Bozec, A., Chassignet, E. P., Curchitser, E., Boeira Dias, F., Durack, P. J., Griffies, S. M., Harada, Y., Ilicak, M., Josey, S. A., Kobayashi, C., Kobayashi, S., Komuro, Y., Large, W. G., Le Sommer, J., Marsland, S. J., Masina, S., Scheinert, M., Tomita, H., Valdivieso, M. and Yamazaki, D.: JRA-55 based surface dataset for driving ocean–sea-ice models (JRA55-do), Ocean Model., 130, 79–139, doi:10.1016/j.ocemod.2018.07.002, 2018.

Tsujino, H., Urakawa, S., Griffies, S. M., Danabasoglu, G., Adcroft, A. J., Amaral, A. E., Arsouze, T., Bentsen, M., Bernardello, R., Böning, C. W., Bozec, A., Chassignet, E. P., Danilov, S., Dussin, R., Exarchou, E., Giuseppe Fogli, P., Fox-Kemper, B., Guo, C., Ilicak, M., Iovino, D., Kim, W. M., Koldunov, N., Lapin, V., Li, Y., Lin, P., Lindsay, K., Liu, H., Long, M. C., Komuro, Y., Marsland, S. J., Masina, S., Nummelin, A., Klaus Rieck, J., Ruprich-Robert, Y., Scheinert, M., Sicardi, V., Sidorenko, D., Suzuki, T., Tatebe, H., Wang, Q., Yeager, S. G. and Yu, Z.: Evaluation of global ocean-sea-ice model simulations based on the experimental protocols of the Ocean Model Intercomparison Project phase 2 (OMIP-2), Geosci. Model Dev., 13(8), 3643–3708, doi:10.5194/gmd-13-3643-2020, 2020.

Turner, J., Bracegirdle, T. J., Phillips, T., Marshall, G. J. and Scott Hosking, J.: An initial assessment of antarctic sea ice extent in the CMIP5 models, J. Clim., 26(5), 1473–1484, doi:10.1175/JCLI-D-12-00068.1, 2013.

Yi, D. and Zwally H. J.: Arctic Sea Ice Freeboard and Thickness, Version 1, [2003-2007], Boulder, Colorado USA, NASA National Snow and Ice Data Center Distributed Active Archive Center, doi: https://doi.org/10.5067/SXJVJ3A2XIZT, 2009, updated 2014-04-15.

Yukimoto, S., Kawai, H., Koshiro, T., Oshima, N., Yoshida, K., Urakawa, S., Tsujino, H., Deushi, M., Tanaka, T., Hosaka, M., Yabu, S., Yoshimura, H., Shindo, E., Mizuta, R., Obata, A., Adachi, Y. and Ishii, M.: The meteorological research institute Earth system model version 2.0, MRI-ESM2.0: Description and basic evaluation of the physical component, J. Meteorol. Soc. Japan, 97(5), 931–965, doi:10.2151/jmsj.2019-051, 2019.

Zadeh, N. T., Krasting, J. P., Blanton, C., Dunne, J. P., John, J. G., McHugh, C., Radhakrishnan, A., Rand, K., Vahlenkamp, H., Wilson, C., Winton, M.: NOAA-GFDL GFDL-OM4p5B model output prepared for CMIP6 OMIP omip1, Version 20200601, Earth System Grid Federation, https://doi.org/10.22033/ESGF/CMIP6.8622, 2018.

Zunz, V., Goosse, H. and Massonnet, F.: How does internal variability influence the ability of CMIP5 models to reproduce the recent trend in Southern Ocean sea ice extent?, Cryosphere, 7(2), 451–468, doi:10.5194/tc-7-451-2013, 2013.