Simulation of oceanographic and meteorological parameters in arctic and subarctic seas by CMIP5 climate models: model validation and selection

Natalia Gnatiuk[1], Iuliia Radchenko[1], Richard Davy[2], Evgeny Morozov[1], Leonid Bobylev[1]

[1]Nansen International Environmental and Remote Sensing Centre, St. Petersburg, 199034, Russia
[2]Nansen Environmental and Remote Sensing Center, Bergen, N-5006, Norway

Correspondence to: Natalia Gnatiuk (natalia.gnatiuk@niersc.spb.ru)

Abstract.

Currently, there are a large number of climate models that give projections for various oceanic and meteorological parameters in the Arctic. However, their estimates often differ in absolute values in specific sea areas in comparison with the historical reanalysis data. The main goal of this research was to find out the methodology of selection of the optimal model ensemble that most accurately reproduces the state of abiotic parameters inherent in six target arctic and sub-arctic seas, viz. the Barents, Bering, Greenland, Labrador, North and Norwegian seas.

Here, we present the validation of 34 CMIP5 atmosphere-ocean General Circulation Models (GCM) over the historical period 1979-2005. Furthermore, we propose a procedure of model ranking and selection, which is based on the model's skill to represent several important oceanographic and meteorological parameters in the arctic and subarctic seas: the sea surface (i) water temperature and (ii) salinity (averaged over the top 30 m); (iii) wind speed at a height of 10 m above the surface; (iv) ocean surface current speed; and (v) surface downwelling shortwave radiation. The validation of the GCMs against reanalysis data includes analysis of the interannual variability, seasonal cycle, spatial biases and temporal trends of the simulated parameters. In total, 30 combinations of high-skilful models were selected for 5 variables over 6 study regions. The results show that there is no mutually optimal combination of models, nor is there a one top-model, that has a skill in reproducing either the regional climatic-relevant features of the whole Arctic region or all combinations of the considered parameters in target seas. Thereby, according to our methodology for each 'variable – target sea' combination, a unique best model subset was selected with the number of included models varying from 7 to 11.

The paper presents a comparison of the selected best-model sub-sets and the ensemble of all available models with the respective reanalysis data. The selected best-model sub-sets show a better performance vs. full-model ensemble in more than 70% cases that confirms the advisability of using the proposed model ranking method.

1 Introduction

Today climate models are state-of-the-art tools for the prediction of the future status of the climate system components on decadal and centennial time scales (Otero et al., 2018; Taylor et al., 2012). In particular, the modern coupled atmosphere-ocean General Circulation Models (GCMs) include the main climate system components such as the atmosphere, ocean, land and sea-ice, and therefore, represent more realistically the processes of their interactions. Thus, the fifth phase of the Coupled Model Intercomparison Project (CMIP5) gives the opportunity to use data of more than 30 GCMs (Taylor et al., 2012). The GCMs provide a large number of the meteorological and oceanographic parameters allowing to perform a comprehensive assessment of possible climate change impacts on marine ecosystems in the future. However, most of the studies addressing the CMIP models intercomparison show that the GCMs outputs usually vary significantly (Almazroui et al., 2017; Fu et al., 2013; Gleckler et al., 2008). Therefore, it is important to find a reliable approach for both model quality intercomparison and selection of optimal models for each specific scientific task and region.

The main goal of the paper is to find a reliable approach for CMIP5 model selection, in particular, those climate models that simulate most efficiently the state of abiotic parameters relevant to living conditions of phytoplankton communities inherent in a number of seas at subpolar and polar latitudes (viz. the Barents, Bering, Greenland, Labrador, North and Norwegian seas). Such a specific task is selected as a case study to have the results that would be applied for projections of abiotic factors affecting the dynamics of phytoplankton communities.

It is well established that the method of ensemble averaging can be used to reduce systematic model biases in the individual climate models (Flato et al., 2013; Gleckler et al., 2008; Knutti et al., 2010; Pierce et al., 2009; Reichler and Kim, 2008; Stocker et al., 2010). Furthermore, in case it is not possible to calibrate a model for a selected region, one of the main recommendation from climate model developers is to take into consideration more than one climate model (Flato et al., 2013; Gleckler et al., 2008; Pierce et al., 2009; Stocker et al., 2010). There are two main approaches to employing climate model ensembles: (i) use of the full-ensemble average data (Flato et al., 2013; Gleckler et al., 2008; Reichler and Kim, 2008; Stocker et al., 2010); and (ii) selection of an ensemble of the best models from the entire set of available climate models based on a comparison with observational data for a historical period (Herger et al., 2018; Stocker et al., 2010; Taylor, 2001). These two approaches are equally used depending on a specific scientific task: (i) full-ensemble averaging for future trends analysis, and (ii) selection of the best models ensembles for regional climate features analysis. However, when there are many climate models available (e.g., in our study the number of models available varied from 25 to 30 depending on the climate variable), then the averaging method will result in very strong smoothing of data, and poor reproduction of the interannual variability. So that only the long-term trend of a given variable will be well captured. We assume that a climate model that successfully represents the present-day conditions will also succeed in the future projections. Therefore, we chose the second approach, e.g., a selection of climate models that properly simulate the current regional features, including the spatial distribution, of the meteorological and oceanographic parameters under study (sea surface temperature and salinity, surface wind speed at 10 m, ocean surface current

speed, and surface downwelling shortwave radiation). At that, it was important to define the appropriate methodology for selection of the best model ensembles.

There are many approaches for the selection of an optimal set of climate models. One approach suggests choosing the models with focus only on some key climatological parameters, such as air temperature, precipitation and sea level pressure (Almazroui et al., 2017; Duan and Phillips, 2010; Pierce et al., 2009; Sarr and Sarr, 2017). This approach assumes that if the models skillfully reproduce these key parameters, they also must be good at reproducing the regional climate in general. Another approach, which is used in this study, is to select a unique combination of models for each study variable (Agosta et al., 2015; Anav et al., 2013; Fu et al., 2013; Gleckler et al., 2008). In order to select such a unique combination of models, it is necessary, firstly, to perform a validation of climate models through comparing GCMs outputs with the respective observations over a historical period, and then to identify the appropriate climate models based on statistical measures, i.e. to sort or rank the tested models. However, there are no generally accepted solutions for this task. For example, Agosta et al. (2015) ranked the CMIP5 models using only one statistical metric, viz, a climate prediction index, which is the ratio of the root mean square error to the standard deviation of observation data. Gleckler et al. (2008) evaluated the CMIP5 models and ranked them through analyzing the climatology of the annual cycle, inter-annual variability, and relative errors. They found that the performance of the analysed models varies for different parameters. Das et al. (2018) assessed 34 CMIP5 models using the following three criteria: the mean seasonal cycle, temporal trends, and spatial correlation. On this basis the models were selected using a cumulative ranking approach. Fu et al. (2013) and Ruan et al. (2019) applied a score-based method using multiple criteria for the assessment of CMIP5 model performance: mean value, standard deviation, normalized root mean square error, linear correlation coefficient, Mann-Kendall test statistic Z, Sen's slope, and significance score. Further, Ruan et al. (2019) selected the top 25% ranked CMIP5 models for composing a multi-model ensemble for air temperature projections over the Lower Mekong Basin. Fu et al. (2013) and Ruan et al. (2019) ranked the employed models using a weight criterion from 0.5 to 1.0. Ruan et al. (2019) reported that the introduction of multiple criteria results in less uncertainties in the models' performance in comparison with the respective observation data. However, Fu et al. (2013) and Ruan et al. (2019) did not consider the feature of spatial distribution of variables.

We decided to compile and improve the previously applied approaches that is to employ the multiple criteria ranking method following Fu et al. (2013) and Ruan et al. (2019) studies but (i) taking into consideration the Agosta et al. (2015) climate prediction index, (ii) analysing the features of spatial distribution of target variables (spatial biases and trends), (iii) ranking the models with the percentile method (25^{th},50^{th}, 75^{th}) that is widely used in statistical analysis, and, finally, (iv) selecting the top 25% ranked CMIP5 models following Ruan et al. (2019).

As the target arctic and subarctic seas differ in physical and geographical conditions, we performed the validation and selection model procedure for each sea individually. Moreover, we analyzed the specific marine areas with the stable localizations of intense growth of phytoplankton species both in spring (e.g. diatoms) and in summer-autumn (e.g. coccolithophores Kondrik et al., 2017; Smyth et al., 2004). Thus, the target regions permitted to identify the CMIP5 models that represented most closely

the cumulative state of the physical environmental factors (abiotic parameters) characterizing the conditions, under which the aforementioned blooms occurred. Such a specific task eventuated in the results that can be useful for further improvements of marine ecological models encompassing the phytoplankton community as well as for modelling the dynamics of physical parameters relevant to surface water environment at high-latitude seas under conditions of changing climate.

2 Materials and method

2.1 Data

Thirty-four CMIP5 GCMs outputs for the historical period 1979-2005 were used in this study. The data are freely available on the ESGF portal (https://esgf-node.llnl.gov). The list of climate models used is presented in Table 1. We analyzed five oceanographic and meteorological variables, namely the sea surface temperature (SST) and salinity averaged over 0-30 m (SSS), surface wind speed at a height of 10 m (WS), ocean surface current speed (OCS), and shortwave downwelling solar radiation (SDSR). These abiotic parameters are known to affect the phytoplankton life cycle in sub-polar and polar latitudes (Iglesias-Rodríguez et al., 2002; Raitsos et al., 2006; Winter et al., 2013). The availability of the CMIP5 GCMs analysed in this study are listed in Table1: in total, we used 25 models for OCS, 28 for SSS, SST, SDSR, and 30 for WS. For validation of the climate models outputs we used atmospheric and oceanic reanalyses: (i) Era-Interim from the European Centre for Medium-Range Weather Forecasts (https://apps.ecmwf.int) (Dee et al., 2011) for the surface wind speed at 10 m, sea surface temperature, and shortwave downwelling solar radiation for the period from 1979 to 2005; and (ii) GLORYS2V4 from the European Copernicus Marine Environment Monitoring Service (http://marine.copernicus.eu) for the sea surface salinity and ocean surface current speed for the period 1993-2005. The period for verification of the employed climate models was chosen based on the length of the reanalysis data and the limitations inherent in the "historical" runs of the GCMs, which usually terminate in 2005. The selected reanalyses are widely used in the literature and have been shown to be consistent with independent observational data (Agosta et al., 2015; Dee et al., 2011; Garric et al., 2017; Geil et al., 2013).

2.2 Study regions

The target regions are six arctic and subarctic seas: the Barents, Bering, Greenland, Labrador, North and Norwegian seas. Only specific areas were selected in each target sea relying on the results obtained by Kazakov et al. (2018) for the coccolithophore *Emiliania huxleyi* blooms based on the Ocean Colour Climate Change Initiative dataset version 3.0 (https://esa-oceancolour-cci.org/) for the period from 1998 to 2016. The selection of the listed seas and the specific areas within them was prompted by several reasons: firstly, in the context of global climate change, the subarctic and arctic seas are characterized by one of the most pronounced changes in environmental parameters due to the so called Arctic amplification, and, secondly, in the target water areas, summer-autumn phytoplankton blooms (e.g. *Emiliania huxleyi*) have a steady localization, while in other parts of the investigated seas the localization of phytoplankton blooms is variable from year to year. For identifying the specific study

areas, on the raster image with all blooming events during 1998-2016 we masked polygons that confine the territories seas where blooms occurred more than one 8-day period (Fig. 1). Besides, the periods for model validation were selected based on a sea-specific blooming periods, which include all summer months and, in some cases, beyond them: June-September for the Barents and Labrador seas, June-August for the Greenland Sea, May-July for the North Sea, May-August for the Norwegian Sea, and January-December for the Bering Sea (Kazakov et al., 2018). Thus, it is noteworthy that the results of the performed comparison of models can be used not only in terms of marine ecology-related issues but also for the purposes of forecasting of the region-specific climate interactions during the vegetation season, taking into account that the selection of the climate models was carried out individually for each sea/sea zone.

2.3. **Model evaluation metrics**

The CMIP5 climate models were validated against the reanalysis data in order to assess how well they reproduce the regional features of distribution the selected parameters/variables. The validation methodology for the GCMs outputs included the analysis of the climatological-mean seasonal cycle, interannual variability and trends, and analysis of spatial distributions of climatological-mean biases and trends for selected parameters averaged over the blooming period in each sea.

a) *The seasonal cycle* was analyzed using the multi-year averaged monthly variables for all months of year (i.e., a sample size of 12). Basic statistical measures were calculated, such as the root-mean-square deviation (RMSD), the correlation coefficient (r), and the standard deviation (SD) (Fu et al., 2013; Gleckler et al., 2008; Kumar et al., 2015; Ruan et al., 2019). In addition, following Agosta et al. (2015) we calculated the climate prediction index (CPI) for the seasonal cycle, which is a ratio of the model root mean square error to the standard deviation of observation data. This model evaluation statistics weighs the simulated data against the observations and often used to validate model data (Agosta et al., 2015; Golmohammadi et al., 2014; Moriasi et al., 2007; Murphy et al., 2004; Stocker, 2004).

b) *The interannual variability* of the parameters was analyzed based on monthly variables solely for blooming periods (the sample size varied according to sub-region and parameter combination, e.g., a sample size for SST in the Barents Sea was 108 – monthly variables from June to September during 1979-2005). The same statistical measures for analysis of the seasonal cycle were used, viz. RMSD, r, SD, and CPI.

c) *The spatial distribution of biases and trends* between the model outputs and the reanalysis data were calculated for temporal-averaged data in each grid point of the target marine zone.

2.4. Percentile score-based model ranking method

For ranking models and selection of the best model sub-set, we proposed and employed the percentile score-based model ranking method, which is a compilation of the previously applied model ranking and the selection approaches with some modifications (see also Introduction). Following Fu et al. (2013) and Ruan et al. (2019), we used the multiple criteria for model selection (RMSD, r, SD). Following Agosta et al. (2015) we analysed the climate prediction index (CPI), and considered the differences in spatial distributions of biases and trends between the model outputs and the respective reanalysis data. Further, we ranked the models based on the percentile method (25th, 50th, 75th) for each obtained statistical metrics based on the

amplitude of its values. Finally, we selected the top 25% ranked CMIP5 models following Ruan et al. (2019) for each considered oceanographic and meteorological parameter, and target region. Thus, for example, for a sample of 28 models, the top 25% is a sub-set of 7 models that showed the best total-score. However, if two or more models show the same score they all are included in the selected best model sub-set. Thus, the number of included models in selected best model subsets varying from 7 to 11.

Figure 2 illustrates an example of the percentile score-based ranking approach applied to RMSD of the sea surface temperature in the Barents Sea. We divided the obtained statistical measures into 4 groups based on the amplitude of the values and assigned a score to each model according to its group: (i) models considered as very good (less than 25th percentile of the statistical metrics distribution) were given a score of 3; (ii) good models (between 50th and 25th percentile) were given a score of 2; (iii) satisfactory models (between 75th and 50th percentile) were given a score of 1; and (iv) unsatisfactory models (more than 75th percentile) were given a score of 0. In the case of the correlation coefficient, it is vice versa, very good models with correlations scores above 0.75 were ranked with a score of 3, and so forth.

For ranking models based on the obtained differences in the spatial distributions of biases and trends between model outputs and reanalysis, we analysed the absolute values of the median and the amplitude of the spatial variation in model biases. For example, Figure 3 displays the box plots of spatial variability in SST biases relevant to the target area in the Barents Sea for the vegetation season (June-September) and the study period 1979-2005. The median bias varies from -6.6 (model #20) to 1.5 K (model #24) among the models, whereas the amplitude bias has a wide spread of values from 7.3 (model #21) to 16.5 K (model #3). Thus it can be concluded from Fig. 3 that the analysis of spatial distribution of biases is very important, e.g., if we compare model #2 (ACCESS1-3) with model #3 (CanESM2), we can see that the medians of these two models have a small difference (0.28 K), while, the amplitude of spatial values for model #3 is much higher than that for model #2. Application of the percentile score-based method to modes #2 (ACCESS1-3) and #3 (CanESM2) resulted in inclusion of the former in the best-model sub-set, whereas the latter was placed beyond it (Fig. 4).

Table 2 presents all calculated statistics that were used to rank GCMs for SST in the Barents Sea as well as the final total score for each model. The spread of the total assigned scores is from 9 to 35. Based on this range we selected the top 25% of GCMs. Thus, the best model ensemble for SST for the Barents Sea is the 8-model set: ACCESS1-0; ACCESS1-3; GFDL-CM3; HadGEM2-ES; MIROC-ESM; MIROC-ESM-CHEM; MPI-ESM-LR; MPI-ESM-MR. Additionally, we identified the top-model for SST in this region – MIROC-ESM. The same procedure was performed for other target seas/zones and variables.

3 Results and discussion

The selected best CMIP5 model sub-sets for five oceanographic and meteorological variables, viz. the sea surface temperature (SST) and salinity averaged over 0-30 m (SSS), surface wind speed at a height of 10 m (WS), ocean surface current speed (OCS), and shortwave downwelling solar radiation (SDSR) in the Barents, Bering, Greenland, Labrador, North and Norwegian

seas are presented in Fig. 4. Each number of the heat map shows the final skill score for one model-variable intersection. For each individual column, its own colour gradation was applied based on percentile ranking approach; therefore, the same numbers can have different colours on the heat map. For example, for OCS in the Barents Sea, the spread of the final model scores is from 7 to 26, whereas for SSS it is from 8 to 34. Therefore, even model #3 CanESM2 has the total score 26 for SSS (which is higher than that (25) for OCS), this model was not included in the SSS best model sub-set and has red color, whereas for OSC it is included in the best model sub-set and has green color. The final skill scores of the models, which were selected as the best model sub-sets are highlighted in bold blue, and their total number is indicated at the bottom of each column.

Analysing the heat map, one can conclude, that there is no an optimal model ensemble, or a one top-model, which could properly simulate all parameters over target seas/regions. However, some climate models show good results for many cases, e.g., ACCESS1-3; ACCESS1-0; GFDL-CM3; GISS-E2-R; GISS-E2-R-CC; HadGEM2-AO; HadGEM2-CC; HadGEM2-ES; INMCM4. The models that have higher biases across the majority of the target regions are CMCC-CM; FGOALS-g2; IPSL-CM5A-LR; IPSL-CM5A-MR; IPSL-CM5B-LR; MIROC5; MRI-ESM1.

Such heterogeneity of climate models ability to equally reproduce the regional climate features residing in different seas can be explained by various reasons. Climate models are often tuned to adequately reproduce global processes and globally averaged values. An insufficient number of long time series of observations is available for model calibration, especially for marine tracts. GCMs errors increase to the poles because of the convergence of meridians at the poles. In addition, the target arctic and sub-arctic seas are essentially different in terms of their physical and geographical conditions, which could also cause the ability of the GCMs to reproduce well the conditions in some seas and fail in others.

In order to analyse how well the selected best-model sub-sets represent five studied parameters, we analysed the spatial distribution of biases between the selected model ensemble and the respective reanalysis data in six target seas, viz, the Barents, Bering, Labrador, Greenland, Norwegian and North seas (Figure 5a-e). Thus, fewer biases in SSS are determined in the case of the Labrador, Greenland and Norwegian seas (±0.5 psu); high positive biases observed in the Bering Sea next to the coastline: up to 1.5-4 psu, this overestimation is possibly due to insufficiently accurate parameterization of the river runoff in the sub-arctic region (Figure 5a). SSS is underestimated in waters next to the coastline in the Barents and North seas (1.5-2.5 psu), which is probably due to some overestimation of river runoff or underestimation of salty atlantic water. The selected CMIP5 models simulate SDSR (Figure 5b) well almost in all target seas: the biases in SDSR in the Barents Sea vary from 5 to 14 W m^{-2} (≈4-10 %), in the Bering Sea – from 2 to10 W m^{-2} (≈2-9 %), in the Greenland Sea – from 0 to 12 W m^{-2} (≈0-7 %), in the North Sea – from 1 to17 W m^{-2} (≈0-7 %), in the Norwegian Sea – from 4 to 9 W m^{-2} (≈2-5 %), only in the Labrador Sea the CMIP5 models overestimate SDSR and the biases much higher – from 20 to 29 W m^{-2} (≈11-15 %). The selected GCMs simulate WS well in all studied seas: the biases in WS are not more than 1 m s^{-1}, only in some places of the Bering and North Seas' coastal regions, the biases in WS simulations are up to about 1.5 m s^{-1} (Figure 5c). Concerning SST, we also obtained quite good results for the selected models. Low biases were observed mainly over the entire territory of the North and Norwegian seas constituting ±0.5° C (Figure 5d). Near the English Channel models overestimate the temperature by ≈2° C in

the North Sea probably due to the influence of warm water from the English Channel, and models slightly underestimate the temperature by ≈1° C near the coastline in the Norwegian Sea. In the Labrador Sea, the CMIP5 models simulate SST with lower biases in the northern and north-western parts of the sea – ±0.5° C (Figure 5d). However, in the southern and south-western parts of the sea, the models underestimate SST by ≈1-2° C, which is possibly due to the influence of the cold Labrador Current. In the Greenland Sea, the models underestimate SST by ≈1-1.5° C in the west probably also due to the influence of the cold Greenland Current and overestimate SST by ≈2° C in the south apparently due to overestimation of contribution of the warm Atlantic water (the North-Atlantic Current). In the Barents Sea, the models overestimate north-western part of the sea probably due to the influence of the warm atlantic water, and in the southern part of the study area, the models underestimate SST by ≈1-2° C probably due to some underestimation of the influence of coming warm atlantic waters. Finally, the CMIP5 models simulate the surface ocean current speed with rather large biases, especially in the Bering Sea and closer to the Bering Strait (-0.19…0.14 m s^{-1}), where the models mainly overestimate OCS (Figure 5e). Smaller biases in the modeling of the OCS by CMIP5 models found for the Barents and Greenland seas – from -0.06 to 0.03 m s^{-1}. The biases in the other studied seas vary from -0.17 to 0.06 m s^{-1}.

To examine our percentile score-based model ranking method we analysed the spatial distribution of biases and trends for the full-model ensemble, selected best-model sub-set and top-model vs. reanalysis data for each target sea and parameter combination. Figure 6 illustrates the case for SST in the Barents Sea, and in the Supplements we present maps for all variables and target regions. As seen in Fig. 6a, the full 28-model ensemble underestimates the SST in the target region while the top-model, MIROC-ESM, overestimates it. The selected 8-model ensemble shows smaller biases (± 1 K) in SST for the most part of the sea. Illustrated in Fig. 5b, the spatial distribution of SST trends (the difference between model data and reanalysis data) indicates that the full 28-model ensemble overestimates the trends for the whole sea (model-reanalysis errors are 0.03-0.07 K yr^{-1}), the top-model MIROC-ESM partly underestimates the SST trend, but for the larger area it reveals reanalysis small trends (± 0.01 K yr^{-1}) that are similar to Era-Interim. As for the selected 8-model ensemble, the spatial variability of errors in trends in SST varies from -0.01 to 0.06 K yr^{-1}, although for the major part of the study region the errors are in the range -0.01 to 0.02 K yr^{-1}. Analysis of comparison of all selected model sub-sets (see Supplements) shows that, in general, the selected best-model ensemble assures somewhat better performance (with regard to the biases between model and reanalysis data) than either the full-model ensemble or the single top-model do. Comparing the full-model ensemble, selected sub-set models or/and top-model performance in terms of biases and trends, the selected best-model ensembles are more skilful in parameter simulations, respectively in 74% (biases) and 83% (trends) cases. The performance of the selected models proved to be equal to the full-model ensemble and top-model efficiency, respectively in 13% (biases) and 10% (trends) cases, and they are less skilful in the simulations in 13% (biases) and 7% (trends) cases.

4 Conclusions

A percentile score-based model ranking method has been presented for selection of optimal model ensembles from a total of 34 CMIP5 models, for five different climate-relevant variables (SST, WS, SSS, OCS, SDSR) in six arctic and subarctic seas, viz. the Barents, Bering, Labrador, Greenland, North, and Norwegian seas. The best model ensembles for each parameter and each target sea were selected (in total 30 combinations of most-skilful models) based on different statistical measures: the root mean square error, correlation coefficient, standard deviation, RMSD-observations standard deviation ratio, spatial biases and trends. Our results show that there is no any optimal model ensemble or a one top-model, which could best simulate all parameters across all target seas. Despite the fact that the Arctic is often considered as one single region in many studies, our results show that CMIP5 climate models do not have consistent performance across such a large area. However, the selected best model ensembles show quite good results with lesser biases in smaller study regions, i.e., some specific arctic seas.

To assure best implementation of the model selection results, it is essential to select climate models that properly simulate the spatial distribution of the chosen variables. Therefore, we suppose that the spatial distribution of biases and trends in the considered parameters are as well important as other statistical metrics within the framework of the model selection procedure performed. Based on our results, we can also conclude that it is essential not only to analyse spatially averaged values, but also the spatial distribution of their amplitudes.

The results of examination of the percentile score-based model ranking method proposed in this paper generally reveal a better performance of the selected best model ensemble vs. the full-model ensemble or a single best model for different variables and target regions.

We can conclude that the range of different factors is important for model selection, including the spatial pattern of model biases, and that the proposed methodology is a way of enhancing the model selection procedures sophistication that promises a better chance to identify more skillful models for the features we are interested in. Thus, the proposed method can be used for analyses to be done for other seas/regions with the purpose to evaluate the performance climate models in terms of various atmospheric and oceanic parameters at different scales.

Author contribution

NG, RD, LB: methodology development. NG, IR: development of the paper concept. IR, NG, EM: data processing and figures producing. All authors contributed to the writing and discussion of the manuscript.

Competing interests

The authors declare that they have no conflict of interest.

Acknowledgements

We express our gratitude for the financial support of this study provided by the Russian Science Foundation (RSF) under the project 17-17-01117.

N. Gnatiuk and I. Radchenko thank Dmitry Pozdnyakov and Dmitry Kondrik for the invitation to participate in the project as well as for very useful discussions of the results obtained.

We acknowledge the members of the Coupled Model Intercomparison Project phase 5, the European Centre for Medium-Range Weather Forecasts, the European Copernicus Marine Environment Monitoring Service, and we extend our gratitude to the modelling groups specified in Table 1).

References

Agosta, C., Fettweis, X. and Datta, R.: Evaluation of the CMIP5 models in the aim of regional modelling of the Antarctic surface mass balance, Cryosph., 9, 2311–2321, 2015.

Almazroui, M., Nazrul Islam, M., Saeed, S., Alkhalaf, A. K. and Dambul, R.: Assessment of Uncertainties in Projected Temperature and Precipitation over the Arabian Peninsula Using Three Categories of Cmip5 Multimodel Ensembles, Earth Syst. Environ., 1(2), 23, doi:10.1007/s41748-017-0027-5, 2017.

Anav, A., Friedlingstein, P., Kidston, M., Bopp, L., Ciais, P., Cox, P., Jones, C., Jung, M., Myneni, R. and Zhu, Z.: Evaluating the Land and Ocean Components of the Global Carbon Cycle in the CMIP5 Earth System Models, J. Clim., 26(18), 6801–6843, doi:10.1175/JCLI-D-12-00417.1, 2013.

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M. A., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A. C. M., van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A. J., Haimberger, L., Healy, S. B., Hersbach, H., Hólm, E. V., Isaksen, L., Kållberg, P., Köhler, M., Matricardi, M., McNally, A. P., Monge-Sanz, B. M., Morcrette, J.-J., Park, B.-K., Peubey, C., de Rosnay, P., Tavolato, C., Thépaut, J.-N. and Vitart, F.: The ERA-Interim reanalysis: configuration and performance of the data assimilation system, Q. J. R. Meteorol. Soc., 137(656), 553–597, doi:10.1002/qj.828, 2011.

Duan, Q. and Phillips, T. J.: Bayesian estimation of local signal and noise in multimodel simulations of climate change, J. Geophys. Res., 115(D18), D18123, doi:10.1029/2009JD013654, 2010.

Flato, G., Marotzke, J., Abiodun, B., Braconnot, P., Chou, S. C., Collins, W., Cox, P., Driouech, F., Emori, S., Eyring, V., Forest, C., Gleckler, P., Guilyardi, E., Jakob, C., Kattsov, V., Reason, C. and Rummukainen, M.: Evaluation of Climate Models, in Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assess- ment Report of the Intergovernmental Panel on Climate Change, pp. 741–866, NY., 2013.

Fu, G., Liu, Z., Charles, S. P., Xu, Z. and Yao, Z.: A score-based method for assessing the performance of GCMs: A case study of southeastern Australia, J. Geophys. Res. Atmos., 118(10), 4154–4167, doi:10.1002/jgrd.50269, 2013.

Garric, G., Parent, L., Greiner, E., Drévillon, M., Hamon, M., Lellouche, J.-M., Régnier, C., Desportes, C., Le Galloudec, O., Bricaud, C., Drillet, Y., Hernandez, F. and Le Traon, P.-Y.: Performance and quality assessment of the global ocean eddy-permitting physical reanalysis GLORYS2V4., 19th EGU Gen. Assem. EGU2017, Proc. from Conf. held 23-28 April. 2017 Vienna, Austria., p.18776, 19, 18776, 2017.

Geil, K. L., Serra, Y. L., Zeng, X., Geil, K. L., Serra, Y. L. and Zeng, X.: Assessment of CMIP5 Model Simulations of the North American Monsoon System, J. Clim., 26(22), 8787–8801, doi:10.1175/JCLI-D-13-00044.1, 2013.

Gleckler, P. J., Taylor, K. E. and Doutriaux, C.: Performance metrics for climate models, J. Geophys. Res. Atmos., 113(D6), D06104, doi:10.1029/2007JD008972, 2008.

Golmohammadi, G., Prasher, S., Madani, A. and Rudra, R.: Evaluating Three Hydrological Distributed Watershed Models: MIKE-SHE, APEX, SWAT, Hydrology, 1(1), 20–39, doi:10.3390/hydrology1010020, 2014.

Herger, N., Abramowitz, G., Knutti, R., Angélil, O., Lehmann, K. and Sanderson, B. M.: Selecting a climate model subset to optimise key ensemble properties, Earth Syst. Dyn., 9(1), 135–151, doi:10.3929/ETHZ-B-000246202, 2018.

Iglesias-Rodríguez, M. D., Brown, C. W., Doney, S. C., Kleypas, J., Kolber, D., Kolber, Z., Hayes, P. K. and Falkowski, P. G.: Representing key phytoplankton functional groups in ocean carbon cycle models: Coccolithophorids, Global Biogeochem. Cycles, 16(4), 47-1-47–20, doi:10.1029/2001GB001454, 2002.

Kazakov, E., Kondrik, D. and Pozdnyakov, D.: Spatial data assimilation with a service-based GIS infrastructure for mapping and analysis of E. Huxleyi blooms in arctic seas, in Sixth International Conference on Remote Sensing and Geoinformation of the Environment., 2018.

Knutti, R., Furrer, R., Tebaldi, C., Cermak, J., Meehl, G. A., Knutti, R., Furrer, R., Tebaldi, C., Cermak, J. and Meehl, G. A.: Challenges in Combining Projections from Multiple Climate Models, J. Clim., 23(10), 2739–2758, doi:10.1175/2009JCLI3361.1, 2010.

Kondrik, D., Pozdnyakov, D. and Pettersson, L.: Particulate inorganic carbon production within *E. huxleyi* blooms in subpolar and polar seas: a satellite time series study (1998–2013), Int. J. Remote Sens., 38(22), 6179–6205, doi:10.1080/01431161.2017.1350304, 2017.

Kumar, D., Mishra, V. and Ganguly, A. R.: Evaluating wind extremes in CMIP5 climate models, Clim. Dyn., 45(1–2), 441–453, doi:10.1007/s00382-014-2306-2, 2015.

Moriasi, D. N., Arnold, J. G., Liew, M. W. Van, Bingner, R. L., Harmel, R. D. and Veith, T. L.: Model evaluation guidelines for systematic quantification of accuracy in watershed simulations, Am. Soc. Agric. Biol. Eng., 50(3), 885–900, 2007.

Murphy, J. M., Sexton, D. M. H., Barnett, D. N., Jones, G. S., Webb, M. J., Collins, M. and Stainforth, D. A.: Quantification of modelling uncertainties in a large ensemble of climate change simulations, Nature, 430(7001), 768–772, doi:10.1038/nature02771, 2004.

Otero, N., Sillmann, J. and Butler, T.: Assessment of an extended version of the Jenkinson–Collison classification on CMIP5 models over Europe, Clim. Dyn., 50(5–6), 1559–1579, doi:10.1007/s00382-017-3705-y, 2018.

Pierce, D. W., Barnett, T. P., Santer, B. D. and Gleckler, P. J.: Selecting global climate models for regional climate change studies, Proc. Natl. Acad. Sci., 106(21), 8441–8446, 2009.

Raitsos, D. E., Lavender, S. J., Pradhan, Y., Tyrrell, T., Reid, P. C. and Edwards, M.: Coccolithophore bloom size variation in response to the regional environment of the subarctic North Atlantic, Limnol. Oceanogr., 51(5), 2122–2130, doi:10.4319/lo.2006.51.5.2122, 2006.

Reichler, T. and Kim, J.: How Well Do Coupled Models Simulate Today's Climate?, Bull. Am. Meteorol. Soc., 89(3), 303–312, doi:10.1175/BAMS-89-3-303, 2008.

Ruan, Y., Liu, Z., Wang, R. and Yao, Z.: Assessing the Performance of CMIP5 GCMs for Projection of Future Temperature Change over the Lower Mekong Basin, Atmosphere (Basel)., 10(2), 93, doi:10.3390/atmos10020093, 2019.

Sarr, A. and Sarr, A.: Multi-Scale Characteristics of Precipitation and Temperature over West Africa Using SMHI-RCA Driven by GCMs under RCP8.5, Am. J. Clim. Chang., 06(03), 455–486, doi:10.4236/ajcc.2017.63024, 2017.

Smyth, T. J., Tyrrell, T. and Tarrant, B.: Time series of coccolithophore activity in the Barents Sea, from twenty years of satellite imagery, Geophys. Res. Lett., 31(11), n/a-n/a, doi:10.1029/2004GL019735, 2004.

Stocker, T. F.: Models change their tune, Nature, 430(7001), 737–738, doi:10.1038/430737a, 2004.

Stocker, T. F., Qin, D., Plattner, G., Tignor, M. and Midgley, P., Eds.: Meeting Report of the Intergovernmental Panel on Climate Change Expert Meeting on Assessing and Combining Multi Model Climate Projections, IPCC Working Group I Technical Support Unit, University of Bern, Bern, Switzerland., 2010.

Taylor, K. E.: Summarizing multiple aspects of model performance in a single diagram, J. Geophys. Res. Atmos., 106(D7), 7183–7192, doi:10.1029/2000JD900719, 2001.

Taylor, K. E., Stouffer, R. J., Meehl, G. A., Taylor, K. E., Stouffer, R. J. and Meehl, G. A.: An Overview of CMIP5 and the Experiment Design, Bull. Am. Meteorol. Soc., 93(4), 485–498, doi:10.1175/BAMS-D-11-00094.1, 2012.

Winter, A., Henderiks, J. and -, B.: Poleward expansion of the coccolithophore Emiliania huxleyi, J. Plankt., 36(2), 316–325, 2013.

Table 1. CMIP5 models used for simulation of selected parameters: SST – sea surface temperature in K, WS – near-surface wind speed in m s^{-1}, SDSR – surface downwelling shortwave solar radiation in W m^{-2}, SSS – sea surface salinity (averaged over 30 m) in psu, OCS – surface ocean current speed in m s^{-1} (models available for respective variable are marked as "+")

Model	**ID**	**Modelling Center** (acronym, full name, and country)	**Resolution** (^{o}lon x ^{o}lat)	**SST**	**WS**	**SDSR**	**SSS**	**OCS**
ACCESS1.0	1	CSIRO-BOM, Commonwealth Scientific and Industrial Research Organisation, Australia and Bureau of Meteorology, Australia	1.25 x 1.875	+	+	+	+	+
ACCESS1.3	2			+	+	+	+	+
CanESM2	3	CCCma, Canadian Centre for Climate Modelling and Analysis, Canada	2.7906 x 2.8125	+	+		+	+
CMCC-CM	4	CMCC, Centro euro-Mediterraneo sui Cambiamenti Climatici, Italy	0.7484 x 0.75	+	+	+	+	+
CMCC-CMS	5		3.7111 x 3.75	+	+	+	+	+
CNRM-CM5	6	CNRM-CERFACS, Centre National de Recherches Meteorologiques, France and Centre Europeen de Recherche et Formation Avancees en Calcul Scientifique, France	1.4008 x 1.40625	+	+	+	+	+
CSIRO-Mk3.6.0	7	CSIRO-QCCCE, Commonwealth Scientific and Industrial Research Organization, Australia and Queensland Climate Change Centre of Excellence, Australia	1.8653 x 1.875		+	+	+	+
EC-EARTH	8	EC-EARTH, EC-EARTH consortium, Europe	1.1215 x 1.125	+				
GFDL-CM3	9	NOAA GFDL, National Oceanic and Atmospheric Administration, Geophysical Fluid Dynamics Laboratory, USA	2 x 2.5	+	+	+	+	+
GFDL-ESM2G	10			+	+	+	+	+
GFDL-ESM2M	11			+	+	+	+	+
GISS-E2-H	12	NASA GISS, National Aeronautics and Space Administration, Goddard Institute for Space Studies, USA	2 x 2.5	+	+	+	+	+
GISS-E2-H-CC	13			+	+	+	+	+
GISS-E2-R	14			+	+	+	+	+
GISS-E2-R-CC	15			+	+	+	+	+
HadCM3	16	MOHC INPE, Met Office Hadley Centre, UK and Instituto Nacional de Pesquisas Espaciais, Brasil	2.5 x 3.75		+			
HadGEM2-AO	17		1.25 x 1.875	+	+	+	+	+
HadGEM2-CC	18			+	+	+	+	+
HadGEM2-ES	19			+	+	+	+	+
IPSL-CM5A-LR	20	IPSL, Institut Pierre-Simon Laplace, France	1.8947 x 3.75	+	+	+	+	+

IPSL-CM5A-MR	21			+	+	+	+	+
IPSL-CM5B-LR	22			+	+	+	+	+
MIROC5	23	MIROC, Atmosphere and Ocean Research Institute, the University of Tokyo, National Institute for Environmental Studies, and Japan Agency for Marine-Earth Science and Technology, Japan	1.4008 x 1.40625	+	+	+	+	
MIROC4h	24		0.5616 x 0.5625		+			
MIROC-ESM	25	MIROC, Japan Agency for Marine-Earth Science and Technology, Atmosphere and Ocean Research Institute, the University of Tokyo, and National Institute for Environmental Studies, Japan	2.7906 x 2.8125	+	+	+	+	
MIROC-ESM-CHEM	26			+	+	+	+	
MPI-ESM-LR	27	MPI-M, Max Planck Institute for Meteorology, Germany	1.8653 x 1.875	+	+	+	+	+
MPI-ESM-MR	28			+	+	+	+	+
MRI-CGCM3	29	MRI, Meteorological Research Institute, Japan	1.12148 x 1.125	+	+	+	+	+
MRI-ESM1	30				+			
NorESM1-M	31	NCC, Norwegian Climate Centre, Norway	1.8947 x 2.5	+		+	+	
NorESM1-ME	32			+		+	+	+
INM-CM4	33	INM, Russian Academy of Sciences Marchuk Institute of Numerical Mathematics, Russia	1.5 x 2		+	+		
FGOALS-g2	34	LASG-CESS, Institute of Atmospheric Physics, Chinese Academy of Sciences; and Tsinghua University, China	2.7906 x 2.8125					+
Total number of available CMIP5 models				**28**	**30**	**28**	**28**	**25**

Table 2. Results of the CMIP5 model performance for SST in the Barents Sea.

(Numbers in brackets indicate the models' scores. RMSD is the root-mean-square deviation, K; r is the correlation coefficient between models and reanalysis; RSR is the RMSD-observations standard deviation ratio; $|SD_{dif}|$ is the modulus of the standard deviation difference (model minus reanalysis), K; $|Tr_m|$ is the modulus of spatial trend median difference (the model minus reanalysis), K yr^{-1}; $|Tr_a|$ is the modulus of spatial trend amplitude difference (the model minus reanalysis), K yr^{-1}; $|Br_m|$ is the modulus of spatial bias median difference (the model minus reanalysis), K; $|Br_a|$ is the modulus of spatial biases amplitude difference (the model minus reanalysis), K).

Model acronym	ID	Seasonal cycle (averaged over the territory)				Interannual variability (averaged over the territory)				Spatial trends (Tr) and biases (Br)				Total score
		RMSD	r	RSR	$\|SD_{dif}\|$	RMSD	r	RSR	$\|SD_{dif}\|$	$\|Tr_m\|$	$\|Tr_a\|$	$\|Br_m\|$	$\|Br_a\|$	
ACCESS1-0	1	0,26*(3)*	0,99*(2)*	0,13*(3)*	0,08*(3)*	1,17*(3)*	0,68*(3)*	0,81*(3)*	0,02*(3)*	0,06*(2)*	0,01*(3)*	0,07(3)	14,7(2)	**33**
ACCESS1-3	2	0,37*(3)*	0,99*(3)*	0,19*(3)*	0,03*(3)*	1,02*(3)*	0,75*(3)*	0,71*(3)*	0,19*(3)*	0,01*(3)*	0,01*(3)*	0,57*(3)*	16,1*(1)*	**34**
CanESM2	3	1,76*(2)*	0,98*(2)*	0,88*(2)*	0,28*(0)*	2,21*(2)*	0,64*(3)*	1,54*(2)*	1,12*(3)*	0,10*(1)*	0,04*(3)*	0,85*(3)*	17,2*(1)*	24
CMCC-CM	4	5,15*(0)*	0,96*(1)*	2,58*(0)*	1,73*(1)*	7,06*(0)*	0,28*(3)*	4,90*(0)*	0,63*(0)*	0,06*(2)*	0,18*(0)*	6,64*(0)*	13,1*(2)*	9
CMCC-CMS	5	4,40*(0)*	0,97*(2)*	2,20*(0)*	1,34*(1)*	5,94*(0)*	0,56*(3)*	4,12*(0)*	0,59*(0)*	0,01*(3)*	0,02*(3)*	5,58*(0)*	14,1*(2)*	14
CNRM-CM5	6	0,64*(3)*	0,99*(2)*	0,32*(3)*	0,55*(1)*	1,59*(3)*	0,73*(3)*	1,10*(3)*	0,81*(2)*	0,08*(2)*	0,00*(3)*	0,49*(3)*	16,4*(1)*	29
EC-EARTH	7	0,41*(3)*	0,99*(2)*	0,21*(3)*	0,13*(2)*	1,43*(3)*	0,64*(3)*	0,99*(3)*	0,38*(3)*	0,13*(1)*	0,12*(1)*	0,14*(3)*	18,1*(0)*	27
GFDL-CM3	8	1,34*(3)*	0,99*(3)*	0,67*(3)*	0,20*(3)*	1,71*(3)*	0,80*(3)*	1,19*(3)*	0,22*(3)*	0,00*(3)*	0,09*(1)*	1,39*(3)*	11,1*(3)*	**34**
GFDL-ESM2G	9	3,23*(1)*	0,98*(2)*	1,62*(1)*	0,27*(2)*	3,72*(1)*	0,69*(3)*	2,58*(1)*	0,29*(3)*	0,04*(3)*	0,04*(3)*	3,46*(1)*	13,9*(2)*	23
GFDL-ESM2M	10	2,60*(2)*	0,99*(2)*	1,30*(2)*	0,61*(3)*	3,42*(2)*	0,68*(3)*	2,37*(2)*	0,25*(2)*	0,01*(3)*	0,08*(2)*	3,10*(2)*	15,7*(1)*	26
GISS-E2-H	11	3,39*(1)*	0,97*(3)*	1,70*(1)*	0,41*(3)*	4,09*(1)*	0,83*(3)*	2,84*(1)*	0,18*(3)*	0,05*(2)*	0,04*(3)*	3,86*(1)*	11,4*(3)*	25
GISS-E2-H-CC	12	3,68*(1)*	0,96*(2)*	1,84*(1)*	0,56*(3)*	4,62*(1)*	0,72*(3)*	3,20*(1)*	0,12*(2)*	0,03*(3)*	0,02*(3)*	4,36*(1)*	10,8*(3)*	24
GISS-E2-R	13	3,34*(1)*	0,96*(2)*	1,67*(1)*	0,04*(1)*	3,83*(1)*	0,72*(3)*	2,66*(1)*	0,84*(3)*	0,05*(2)*	0,07*(2)*	3,34*(1)*	15,1*(2)*	20
GISS-E2-R-CC	14	3,38*(1)*	0,96*(2)*	1,69*(1)*	0,07*(1)*	3,78*(1)*	0,75*(3)*	2,62*(1)*	0,83*(3)*	0,03*(3)*	0,05*(2)*	3,29*(2)*	13,6*(2)*	22
HadGEM2-AO	15	1,28*(3)*	0,99*(2)*	0,64*(3)*	0,01*(3)*	1,51*(3)*	0,73*(3)*	1,05*(3)*	0,13*(3)*	0,02*(3)*	0,05*(2)*	1,33*(3)*	19,8*(0)*	31
HadGEM2-CC	16	1,70*(2)*	0,99*(2)*	0,85*(2)*	0,16*(2)*	2,34*(2)*	0,62*(3)*	1,62*(2)*	0,35*(3)*	0,05*(2)*	0,05*(2)*	1,66*(3)*	19,1*(0)*	25
HadGEM2-ES	17	0,30*(3)*	0,99*(3)*	0,15*(3)*	0,08*(3)*	0,98*(3)*	0,77*(3)*	0,68*(3)*	0,00*(3)*	0,05*(2)*	0,04*(3)*	0,09*(3)*	17,5*(1)*	**33**
IPSL-CM5A-LR	18	3,66*(1)*	0,98*(2)*	1,83*(1)*	0,31*(3)*	4,59*(1)*	0,70*(3)*	3,19*(1)*	0,18*(3)*	0,01*(3)*	0,03*(3)*	4,32*(1)*	18,4*(0)*	22
IPSL-CM5A-MR	19	2,22*(2)*	0,99*(2)*	1,11*(2)*	0,67*(1)*	2,57*(2)*	0,73*(3)*	1,78*(2)*	0,80*(2)*	0,06*(2)*	0,05*(2)*	1,91*(2)*	16,0*(1)*	23
IPSL-CM5B-LR	20	5,03*(0)*	0,96*(1)*	2,52*(0)*	1,71*(1)*	6,90*(0)*	0,36*(3)*	4,79*(0)*	0,69*(0)*	0,00*(3)*	0,03*(3)*	6,51*(0)*	17,6*(0)*	11
MIROC-ESM	21	1,40*(3)*	0,99*(3)*	0,70*(3)*	0,04*(3)*	1,63*(3)*	0,82*(3)*	1,13*(3)*	0,06*(3)*	0,01*(3)*	0,08*(2)*	1,51*(3)*	11,8*(3)*	**35**
MIROC-ESM-CHEM	22	0,97*(3)*	0,99*(3)*	0,49*(3)*	0,05*(3)*	1,34*(3)*	0,82*(3)*	0,93*(3)*	0,13*(3)*	0,07*(2)*	0,05*(3)*	1,08*(3)*	15,1*(2)*	**34**
MIROC5	23	2,42*(0)*	0,98*(2)*	1,21*(0)*	0,51*(1)*	5,69*(2)*	0,51*(3)*	3,95*(2)*	0,64*(2)*	0,18*(0)*	0,08*(2)*	5,14*(0)*	19,8*(0)*	14
MPI-ESM-LR	24	1,27*(3)*	0,99*(3)*	0,63*(3)*	0,04*(3)*	1,54*(3)*	0,81*(3)*	1,07*(3)*	0,21*(3)*	0,02*(3)*	0,04*(3)*	1,33*(3)*	16,3*(1)*	**34**
MPI-ESM-MR	25	0,91*(3)*	0,99*(2)*	0,45*(3)*	0,05*(3)*	1,47*(3)*	0,71*(3)*	1,02*(3)*	0,11*(3)*	0,05*(2)*	0,04*(3)*	0,96*(3)*	17,2*(1)*	**32**
MRI-CGCM3	26	2,88*(2)*	0,99*(3)*	1,44*(2)*	0,08*(2)*	2,54*(1)*	0,82*(3)*	1,77*(1)*	0,34*(3)*	0,00*(3)*	0,07*(2)*	2,30*(2)*	11,9*(3)*	27
NorESM1-M	27	1,53*(2)*	0,99*(2)*	0,77*(2)*	0,76*(2)*	2,56*(2)*	0,64*(3)*	1,78*(2)*	0,31*(2)*	0,05*(2)*	0,07*(2)*	2,33*(2)*	13,7*(2)*	25
NorESM1-ME	28	1,72*(2)*	0,99*(2)*	0,86*(2)*	0,78*(2)*	2,79*(2)*	0,57*(3)*	1,94*(2)*	0,39*(2)*	0,02*(3)*	0,02*(3)*	2,58*(2)*	15,0*(2)*	27

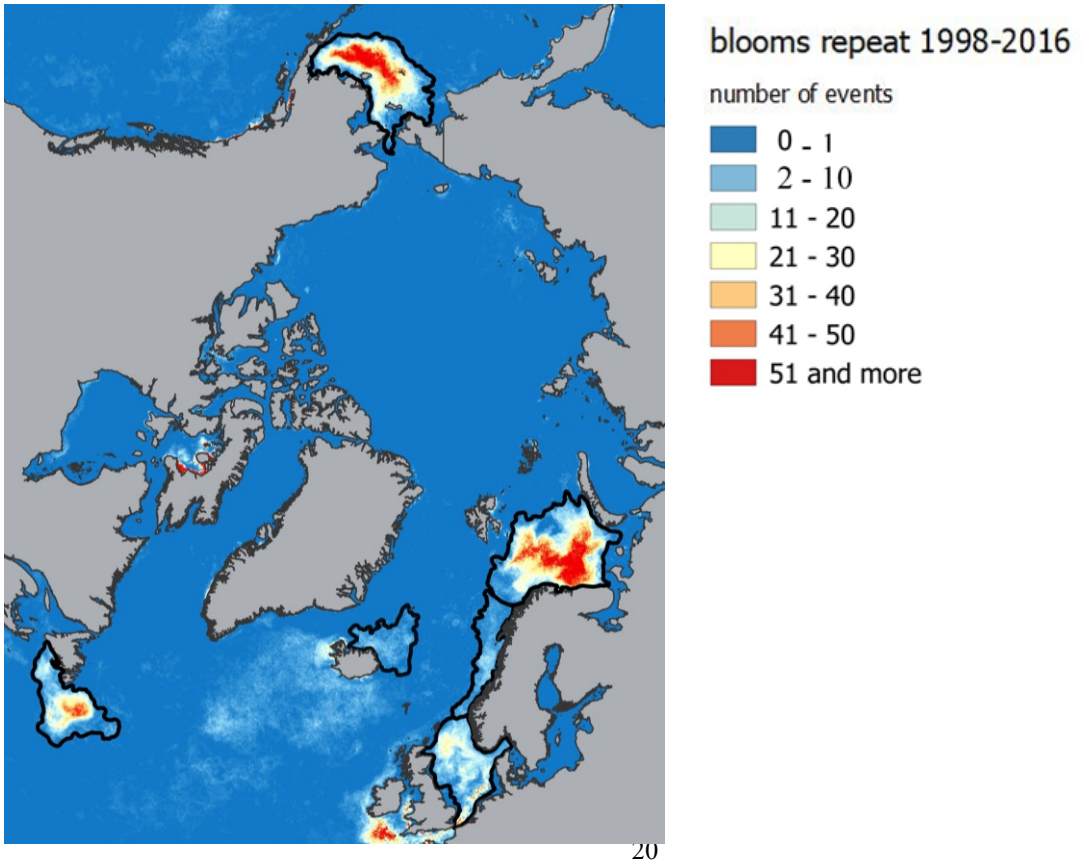

Figure 1: Spatial distribution of *Emiliania huxleyi* blooms occurrence based on the Ocean Colour Climate Change Initiative dataset version 3.0 (Kazakov et al., 2018) **for the Barents, Bering, Labrador, Greenland, North, and Norwegian seas. Black lines confine the territories where blooms occurred more than one 8-day period and show target sea areas.**

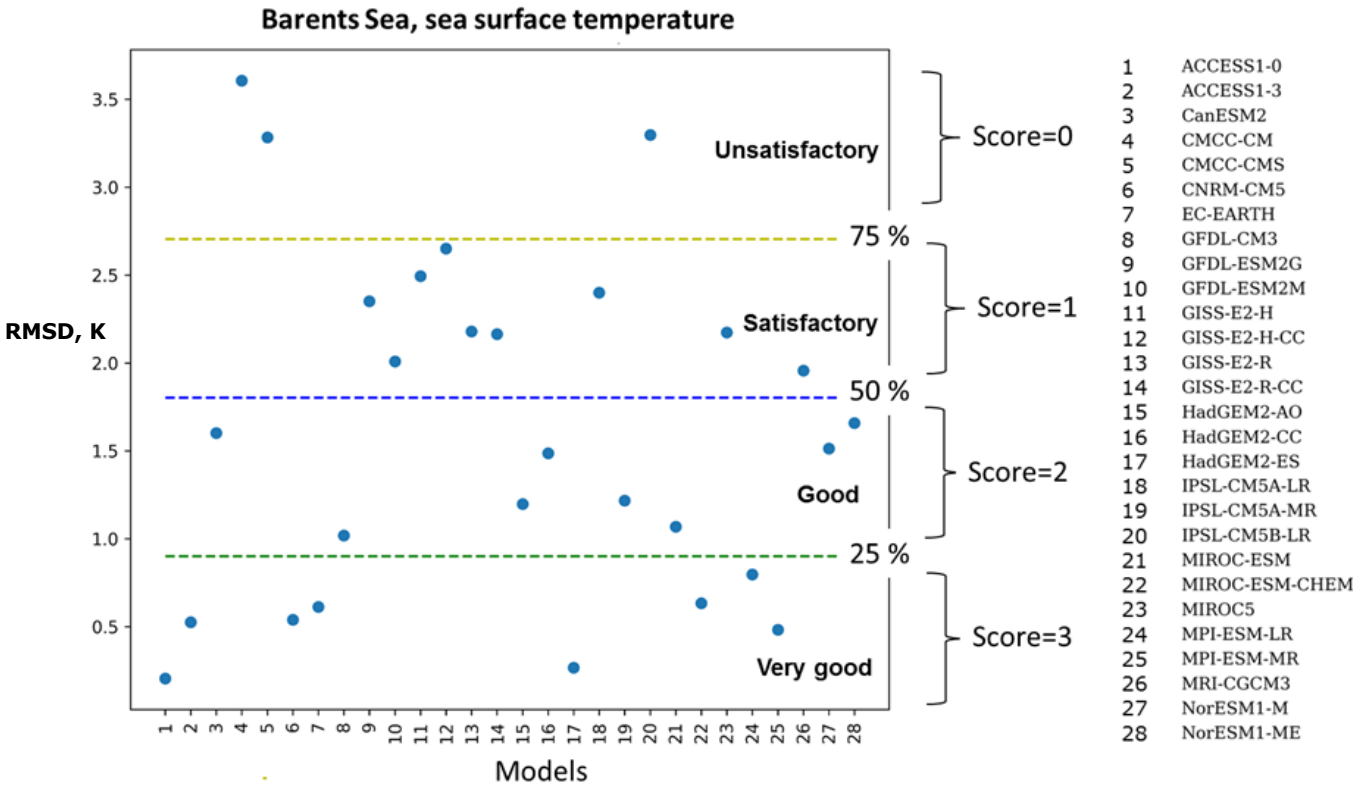

Figure 2: A schematic representation of the percentile score-based model ranking method (Division of RMSD values distribution of 28 models into four groups that are limited by 25th, 50th and 75th percentiles and the relative assignment of scores from 3 to 0 to each group accordingly - very good, good, satisfactory and unsatisfactory).

5

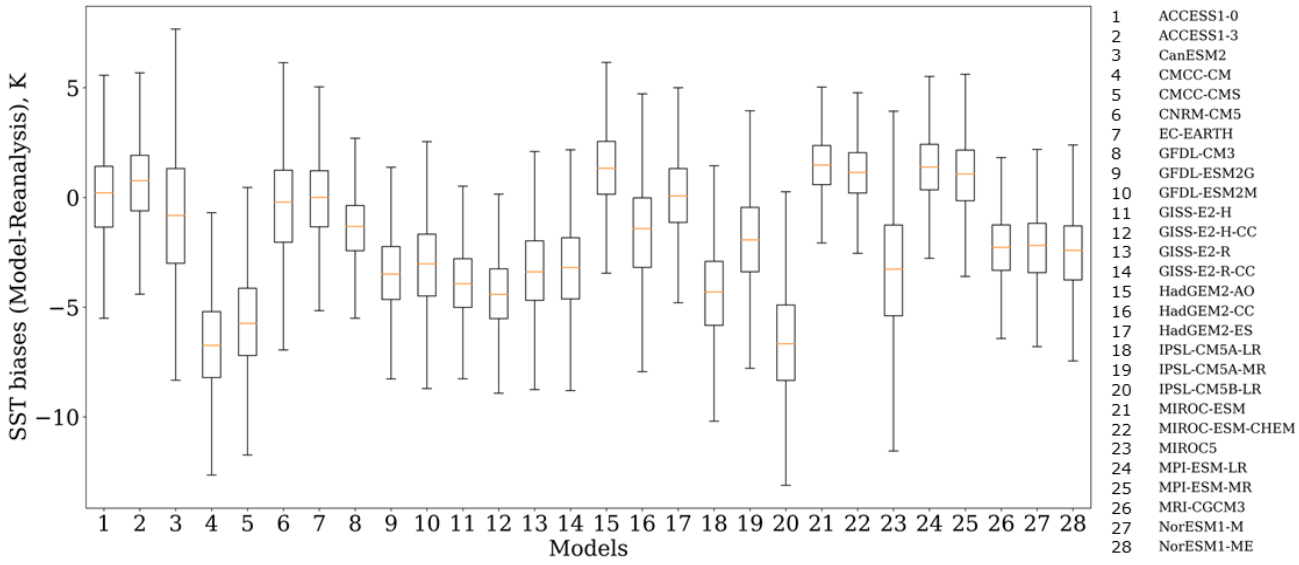

Figure 3: Box plots of the spatial variability of SST biases, which are calculated as the difference between the model and reanalysis data in the Barents Sea over the vegetation season and the time period 1979-2005. Each box spreads from the lower quartile Q1 to the upper quartile Q3 of biases, the orange lines represent the medians. The lower "whiskers" are represented as Q1-1.5 Standard deviation and the upper "whiskers" are represented as Q3+1.5 Standard deviation.

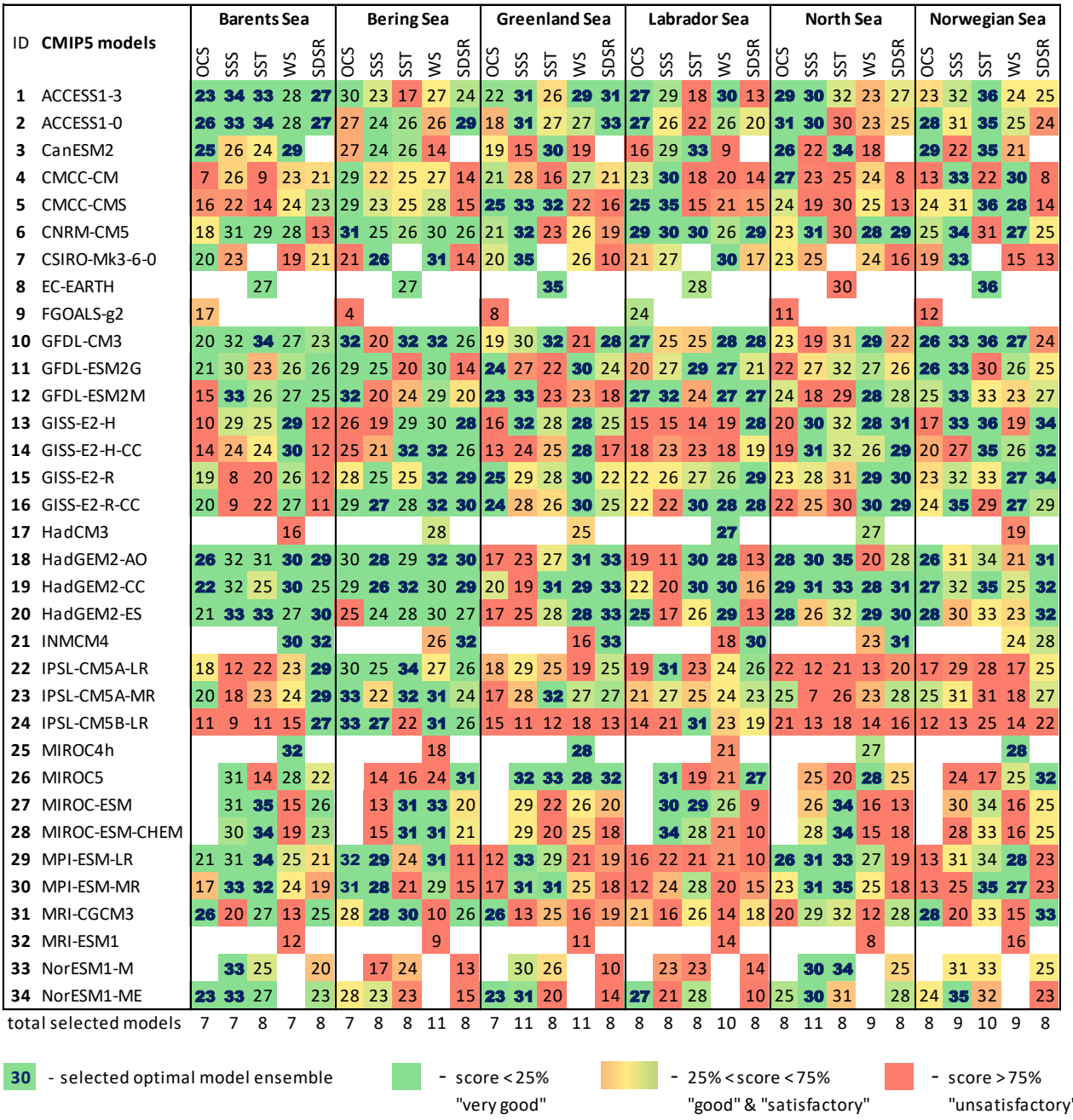

Figure 4: Heat map with the final model scores obtained using the percentile score-based model ranking method for the five variables (sea surface temperature (SST, K) and salinity averaged over 0-30 m (SSS, psu), surface wind speed at 10 m (WS, m s^{-1}), ocean surface current speed (OCS, m s^{-1}), and shortwave downwelling solar radiation (SDSR, W m^{-2}) for the Barents, Bering, Greenland, Labrador, North, and Norwegian seas based on different statistical metrics (Figure 2, Table 2). The white areas indicate that the model was not considered due to partial or complete unavailability of hindcasts, and future projections (RCP4.5, RCP8.5) data.

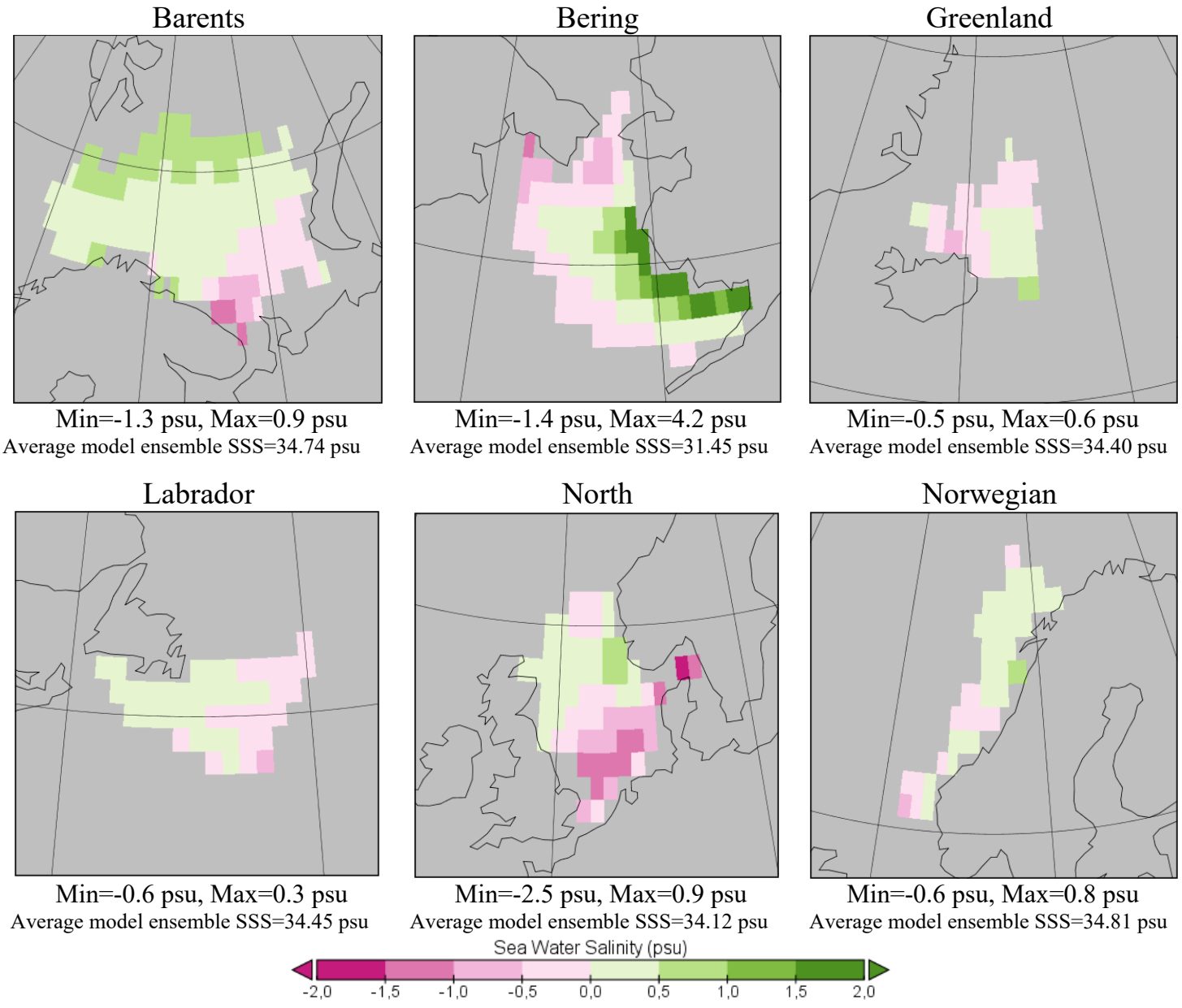

Figure 5a. Spatial distribution of biases in sea surface salinity models and reanalysis in six target seas averaged over the vegetation season and the time period 1993-2005.

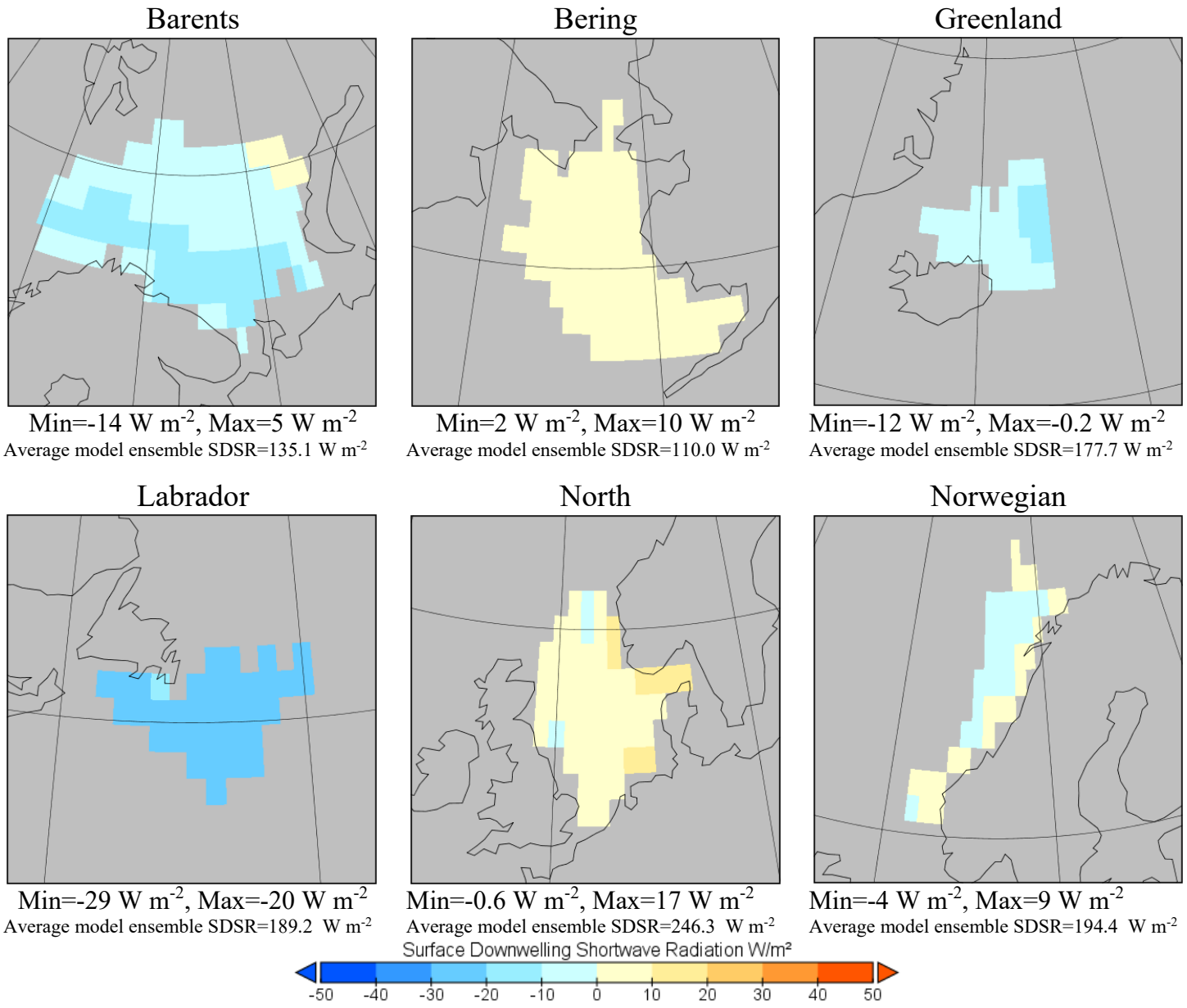

Figure 5b. Spatial distribution of biases in surface downwelling solar radiation between models and reanalysis in six target seas averaged over the vegetation season and the time period 1979-2005.

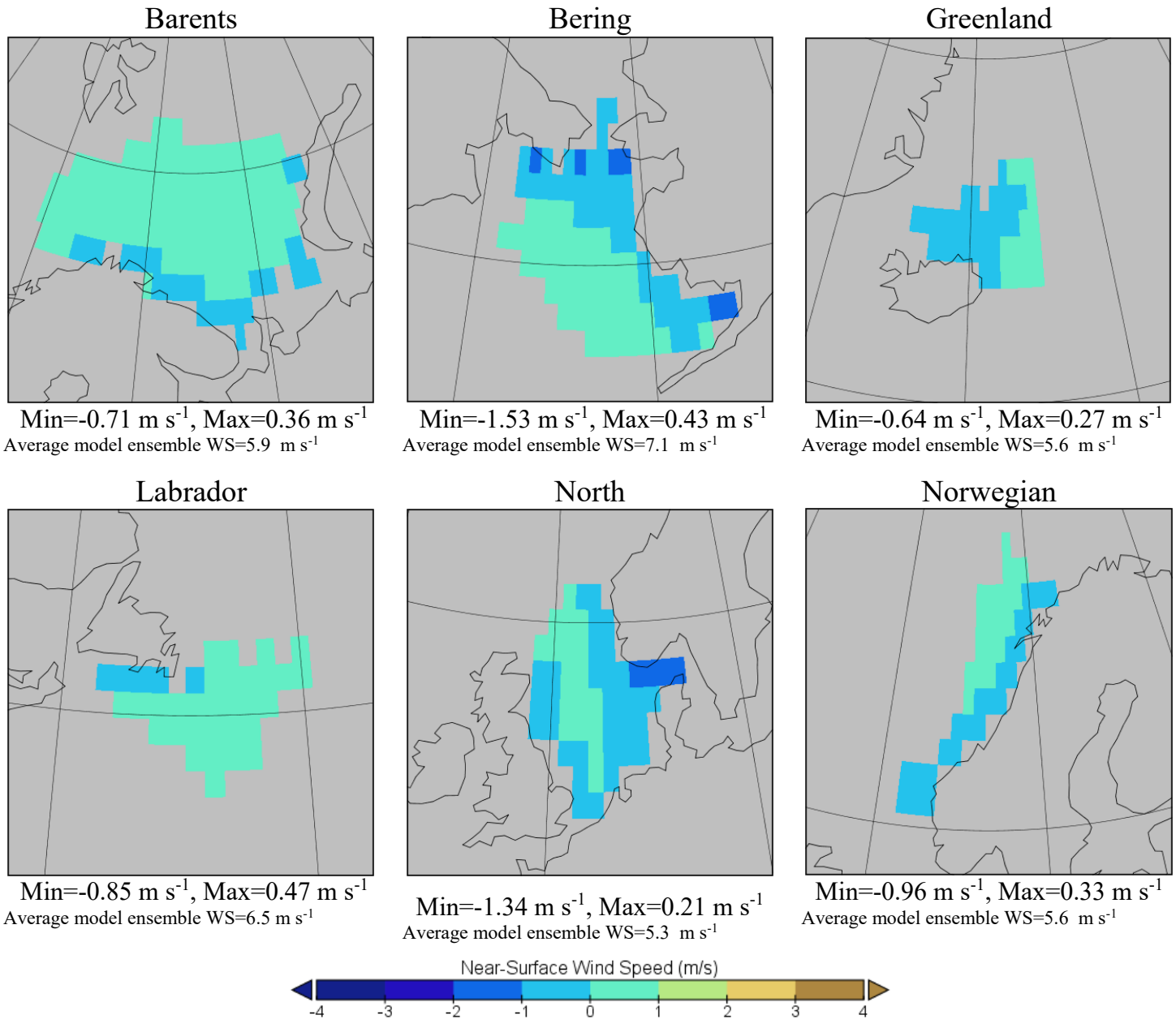

Figure 5c. Spatial distribution of biases in near-surface wind speed between selected model ensemble and reanalysis in six target seas averaged over the vegetation season and the time period 1979-2005.

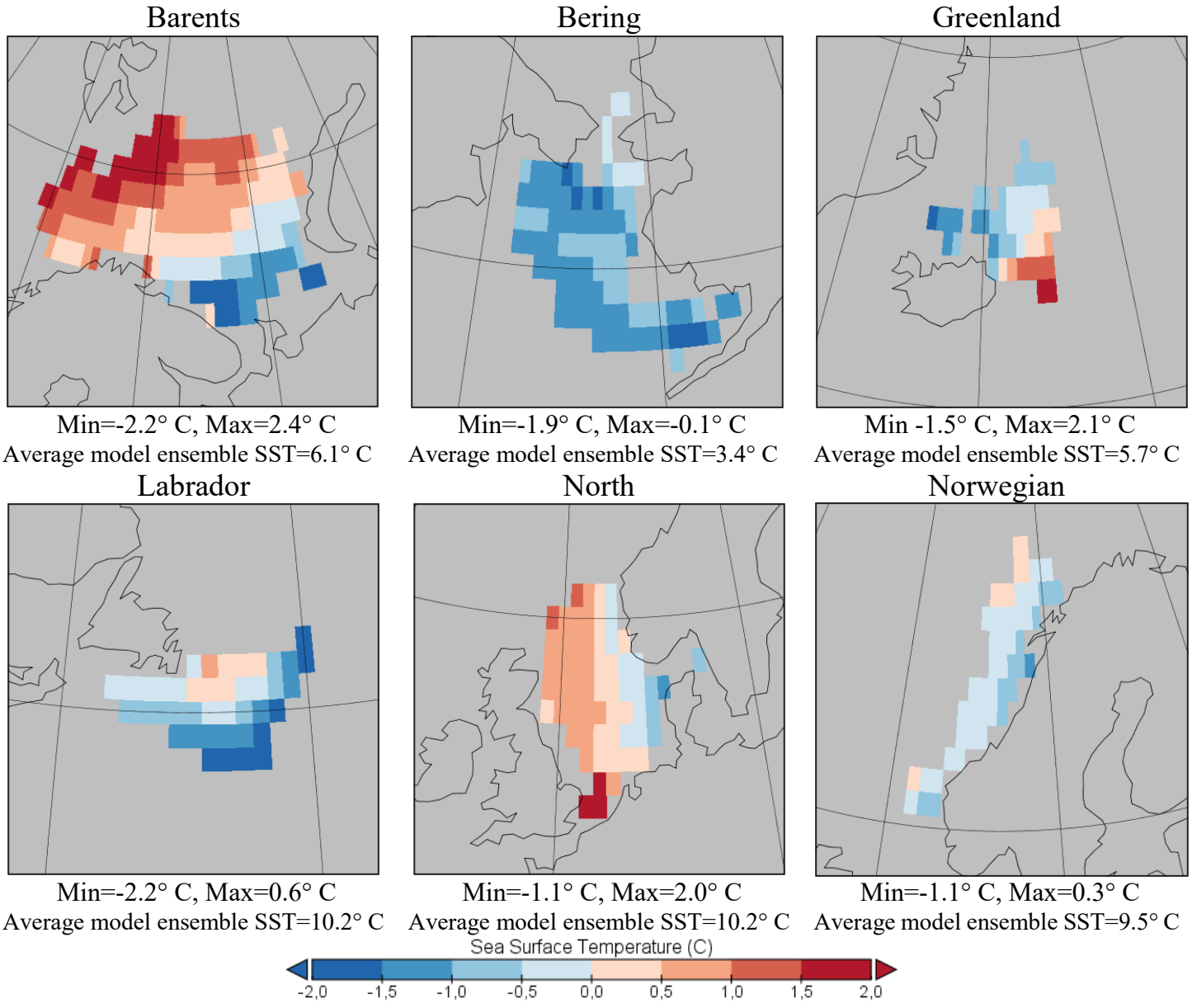

Figure 5d. Spatial distribution of biases in sea surface temperature models and reanalysis in six target seas averaged over the vegetation season and the time period 1979-2005.

5

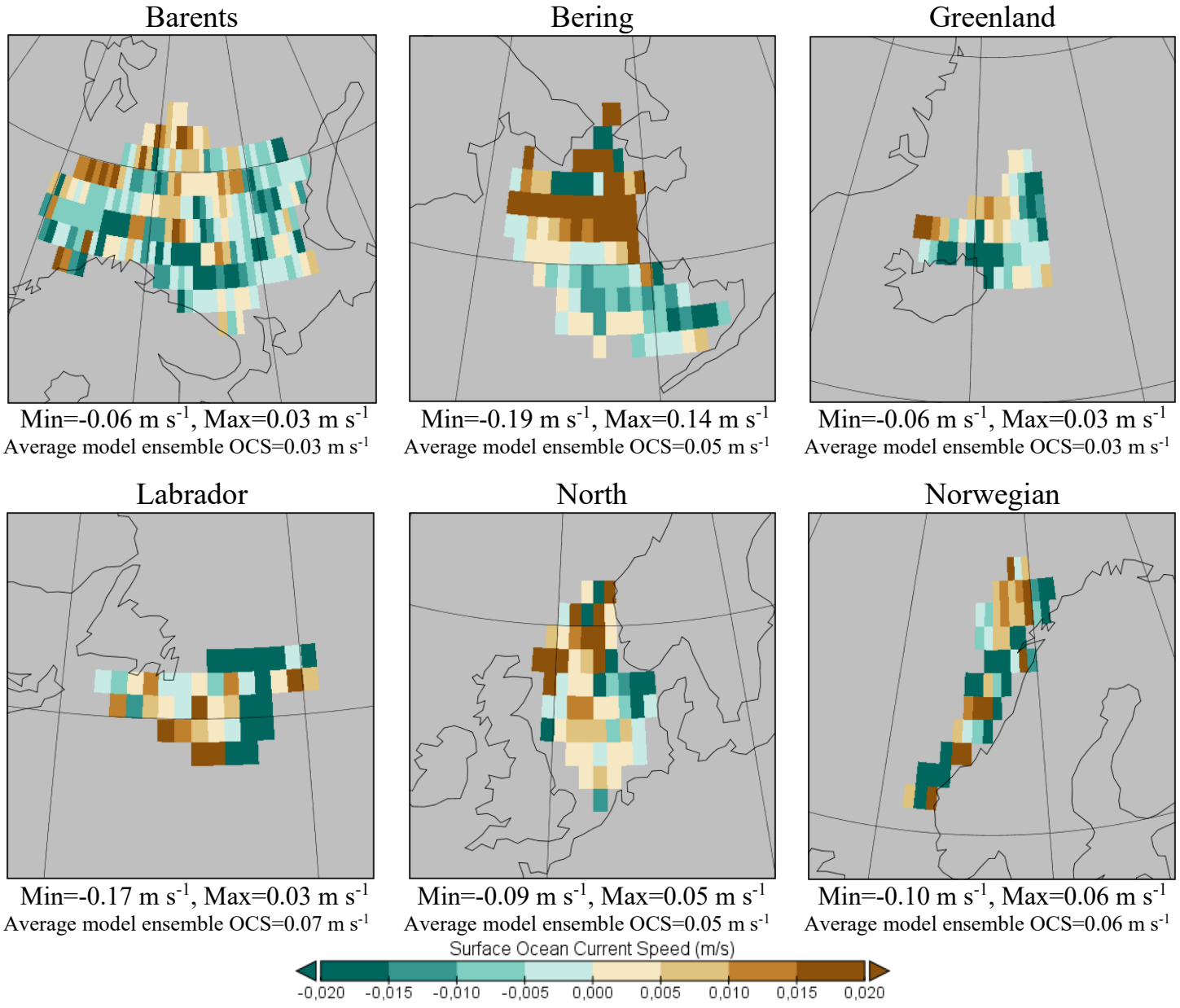

Figure 5e. Spatial distribution of biases in surface ocean current speed models and reanalysis in six target seas averaged over the vegetation season and the time period 1993-2005.

5

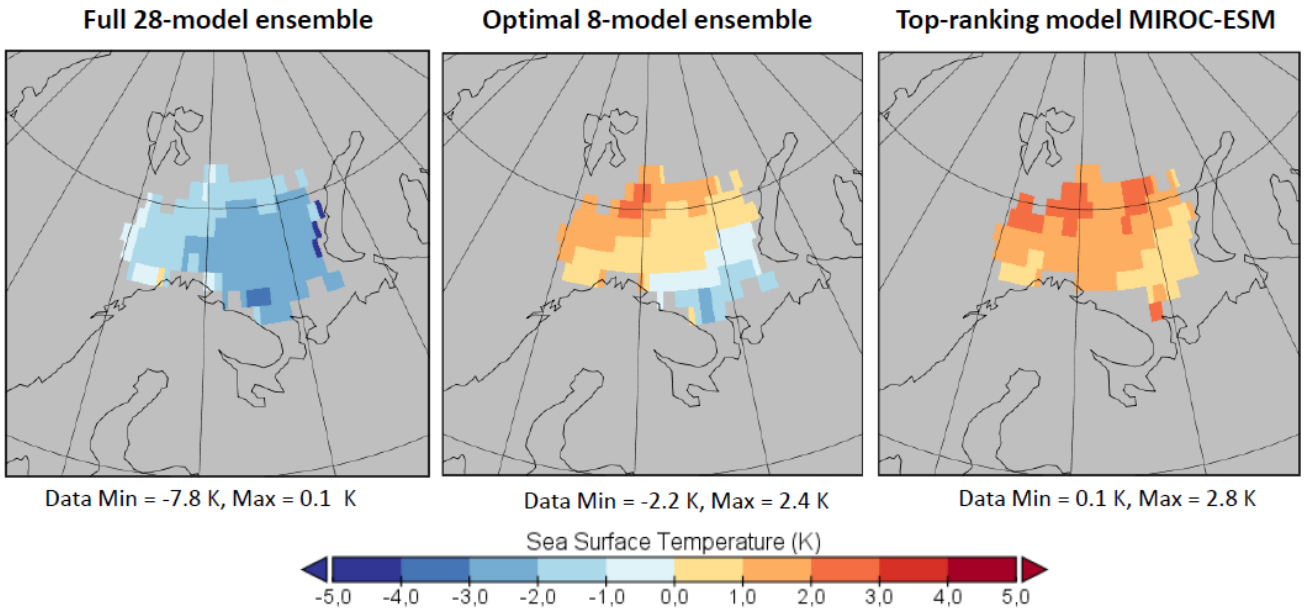

Figure 6a: Spatial distribution of biases in SST (K) between models and reanalysis data in the Barents Sea; the biases are averaged over June-September.

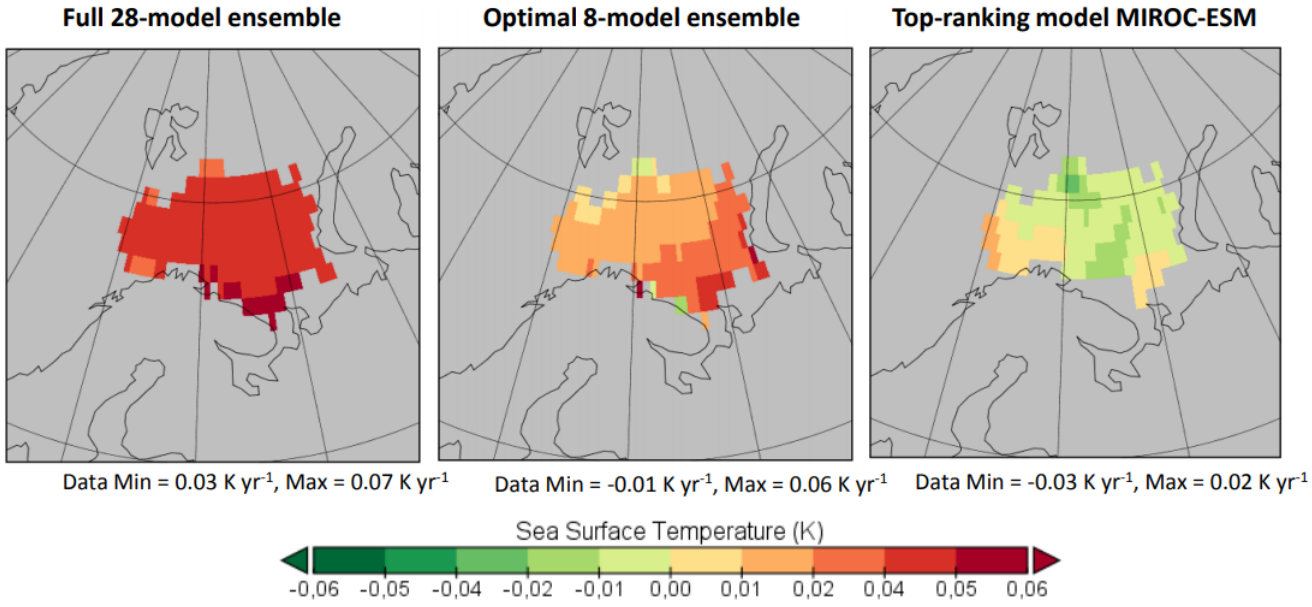

Figure 6b: Spatial distribution of errors, which are calculated as the difference between model and reanalysis values of annual SST trends (K yr^{-1}) in the Barents Sea (June-September)