Dear Prof. Heinze,

Thank you very much for your assistance with our manuscript: "**Simulation of factors affecting *E.huxleyi* blooms in arctic and subarctic seas by CMIP5 climate models: model validation and selection**" by N. Gnatiuk, I. Radchenko, R. Davy, E. Morozov, and L. Bobylev.

We are very thankful to the anonymous reviewers and greatly appreciate their valuable comments and suggestions that considerably improve our manuscript. Regarding the comments from both reviewers, we decided to refuse from focusing on the *E. huxleyi* problematics per se and concentrate on the main goal of the manuscript, which is a selection of those climate models that simulate most efficiently the state of abiotic parameters relevant to living conditions of the phytoplankton communities inherent in a number of seas at subpolar and polar latitudes. Accordingly, the Introduction is thoroughly recast. As both reviewers suggested us to improve the "Results and Discussion" section, we moved the text related to the description of the methodology and two figures to section "Materials and Methods". Also, Figures 3, 5, and 8 are deleted as they are either a mere modification of presentations of some other akin figures or their presence in the manuscript is not so important. We also decided to add a new figure (#Figure 5a-e) to section "Results and Discussion" as both reviewers suggested to cover in discussions all studied parameters and seas. So, Figure 5 displays a spatial distribution of biases in five parameters between models and reanalyses in six target seas. The biases are averaged over the vegetation season and 1979/1993-2005 period. We improved all sections following the comments from both reviewers.

Please kindly find attached the responses to the reviewers and effected revisions, as well as a detailed specification of the changes we introduced.

We are looking forward to hearing from you considering these changes and await further instructions.

On behalf of the paper's co-authors

Best regards,
Natalia Gnatiuk
(and co-authors)

Response to Reviewers and Proposed Revisions

We are very grateful to both reviewers for their constructive and valuable comments and very pointful suggestions, which helped greatly improve our manuscript.

Concerning the comments of both reviewers on the choice of factors controlling phytoplankton blooms in general, and coccolithophore in particular, and other reviewers' comments directly related to the coccolithophore blooms, we fully agree with the arguments provided by reviewer #2 in section "general comments":

"By not having a primary focus on E.huxleyi blooms in the Introduction, the reader will be able to recognize the wider implications of this extensive intercomparison of climate models – it will also alleviate some of the major issues of neglecting "what else" underpins coccolithophore blooms and their occurrences."

Actually, the main objective of the study was to analyze how CMIP5 climate models reproduce different oceanographic and meteorological parameters in the arctic and subarctic seas as well as to form a methodology for validation and selection of the optimal model sub-set. To have practical use of the results we have chosen for case study oceanographic and meteorological parameters that influence coccolithophores blooms in studied arctic and subarctic seas. Due to the fact that we did not consider in the article all the factors (including biotic ones) that influence coccolithophores bloom, we mistakenly paid too much attention to coccolithophores and the factors affecting their blooms. This resulted in shifting the paper's focus away from the main goal of the study, i.e. to develop a methodology of validation and selection of climate models that simulate most accurately the abiotic conditions within the target marine areas.

To mend the situation, we decided to refuse from focusing specifically upon the issue of coccolithophore blooms and put at the forefront the methodology of validation and selection of climate models.

In the absence of a close connection to coccolithophores, the article indeed gains greater clarity and becomes focused on the substance of the research done on the comparative effectiveness of global climate models for specific marine objects. We corrected the manuscript according to the recommendations of the reviewers and tried to make the goal of our research clear and precise. Below we have presented all the answers to the comments and all text changes.

We earnestly thank the reviewers for their critical comments.

Reviewer #1

General comments

- **G.C.1:** In my opinion, the choice of the variables to be validated in the models is incomplete and the chosen subset is not obvious when thinking about the factors controlling phytoplankton blooms. First, all possible drivers should be carefully introduced in the introduction (it is nowhere clearly stated in the manuscript what factors impact phytoplankton dynamics in general). Important variables such as (macro/micro) nutrients and carbonate chemistry modeling at all and it is not even thoroughly discussed why. Motivating the choice of drivers to be analyzed in this paper simply by referring to an individual study (Kondrik et al., 2019) which is currently in review for publication in Biogeosciences is not sufficient in my opinion. Additionally, reviewers of the manuscript by Kondrik et al. (2019) have raised similar concerns with regard to the chosen drivers considered in the analysis. In my view, the choice made for the manuscript at hand is a missed opportunity as the evaluated models can provide more comprehensive information on factors impacting phytoplankton/coccolithophore growth than the variables the authors chose here.

  G.C.1 answer: As mentioned above, we believe that focusing on factors affecting coccolithophore blooms will take the readers away from the true purpose of the publication. Also, as we provide an incomplete list of factors controlling phytoplankton blooms, we have decided not to use in the manuscript formulations like "factors controlling/affecting coccolithophore blooms". Instead, in the article, we will emphasize that we analyze the models for a number of meteorological and oceanographic parameters. We believe that in this case the results may be of interest to a wider readership.

  We would like to add that at the beginning of the research we also planned to consider different biogeochemical variables. However, CMIP5 models have only monthly outputs for ocean biogeochemical variables, whereas, in this study to develop a methodology for selecting climate models, we employed daily data.

- **G.C.2**: In the current version of the manuscript, a discussion of the results is completely missing. While section 3 is called “Results & Discussion”, it currently only represents a description of the Figures, without putting the results into the context of previous literature or how the results impact the overall motivation for the study (assessing the potential future development of E. huxleyi blooms). It is not clear to me e.g. what modellers should take away from their analysis. In a revised version of the manuscript, I suggest to include a thorough discussion on e.g. the sensitivity of the resulting model combinations on chosen thresholds in the ranking, the impact of the identified model biases on coccolithophore blooms, the impact of neglecting important forcing factors (nutrients, carbonate chemistry) or biotic interactions (which cannot be assessed with this approach as opposed to when coccolithophores are included as an explicit functional type to the model).

G.C.2 answer: We agree and we improved section Results and Discussion.

- **G.C.3**: Regarding choices of the presentation of the results, I would personally like to see more than just temperature in the Barents Sea to be included in the main text (the choice of the figures could be reconsidered, especially given the title of the manuscript). The current choice makes it very hard for the reader to assess how the representation of present-day coccolithophore blooms in these models is potentially affected by biases of all variables impacting phytoplankton dynamics (and not just temperature). Including more detail in the study at hand will also make the assessment in a follow-up paper on future changes easier.

  G.C.3 answer: Based on one example – the temperature in the Barents Sea, we aimed to describe the methodology of climate model validation and selection in detail. Of course, from the point of view of assessing how climate models represent present-day coccolithophore blooms, such choices of presentation of the results is very uninformative. However, we have tried to illustrate with this example each step of the model's validation and the selection and to show the spread of the model's values for each selection criterion. Using this approach, we intended to prove the need for a comprehensive analysis that is not confined solely either to the seasonal cycle or inter-annual variability or trends or spatial errors.

- **G.C.4**: Overall, I think that the literature review in the introduction on factors controlling coccolithophore blooms in the arctic/subarctic (or North Atlantic) and possible drivers for observed changes in their distributions is not comprehensive enough in its current form. In my detailed comments below, I suggested a few papers that could be considered in my view – a result of a very brief literature search I have done (as I am not 100% familiar with the literature of the arctic/subarctic), but this list is by no means exhaustive. The authors should revise the manuscript accordingly, as this might also help to motivate why certain variables are (or are not) considered in their study.

  G.C.4 answer: We agree and we improved the Introduction part.

- **G.C.5**: Throughout the manuscript, the writing needs to be more concise and to the point. Often, it is not clear to the reader why certain information is given, i.e. what the relevance is for the study at hand or what the take-away message is (see detailed comments below of e.g. the introduction). The authors should especially revise the result section, which is currently a list of brief descriptions of the Figures without making it clear enough why they were chosen to be included and what the key message for each Figure is, which makes this section quite hard to read in its current form. Ultimately, all figure captions are currently incomplete as they do not describe what is actually shown in the respective figures.

G.C.5 answer: We improved the manuscript and the Results and Discussion section accordingly.

Detailed comments of Reviewer #1:

Abstract:

D.C.1 p. 1, L. 9: Please add ocean acidification here. It is not only the effect of global warming that can be expected to impact future coccolithophore blooms.

D.C.2 90 p. 1, L. 10: I find this statement on the aim of the paper misleading, as none of these models (or very few) includes an explicit parametrization of coccolithophores to the best of my knowledge. I think it you need to state more clearly what exactly you do here. You don't actually assess the blooms, but only how well the models reproduce the present-day environmental conditions that favour coccolithophore blooms. Also, please be precise here what you mean by "optimum combination".

D.C.3 p. 1, L. 11: This last statement of the paragraph is misleading because you don't actually address this in the study. Please point out that this is future work that can/will be done following this study. Additionally, please add a "potential" in the last part of the sentence 100 "[…] potential future changes […] can be assessed."

D.C.1-3 Answer:

We thank the reviewer for the comments and re-wrote the paragraph:

"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."

Specific comments of Reviewer #1:

p. 1, L. 14: Please delete the "complex" or describe what methodology you're using. I suggest to rephrase to something along these lines "Here, we present the validation of 34 CMIP models over the historical period. Furthermore, we present the procedure for model selection, which is based on their skill to represent important forcing factors for coccolithophore blooms."

Answer: We agree and rephrased the sentence in the following form:

"Here we describe the complex methodology used for the validation of 34 CMIP5 climate models, and the selection of models that best represent the regional features of the oceanographic and meteorological factors affecting E.huxleyi blooms in arctic 15 and subarctic seas: sea surface (i) temperature and (ii) salinity; (iii) wind speed at a height of 10 m above the surface; (iv) ocean surface current speed; and (v) surface downwelling shortwave radiation."

p. 1, L. 15: Are these five factors really known to be the dominant factors impacting coccolithophore dynamics in the arctic and subarctic?
Answer: We consider these five factors as one of the important abiotic controlling factors in the studied seas without a direct reference to coccolithophore ecology. Since we changed the focus of the article, we no longer use the word "factors".

p. 1, L. 16: The chosen set of environmental factors to be validated in the models is not obvious to me. There is no rationale from my point of view as to why one would completely neglect nutrient fields and carbonate chemistry in the validation of the models (see general comments and detailed comments further down). Furthermore, I suggest to include a brief description what the environmental conditions wind speed, current speed, and salinity are proxies for as phytoplankton growth in these models is not a direct function of these variables.
Answer: We agree that nutrients and carbonate chemistry are important for phytoplankton. However, we aim to select the appropriate CMIP5 models for one of the important abiotic factors. We added the following sentence:
"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."

p. 1, L. 16: Please check throughout the text: are you using sea surface salinity (as stated e.g. in the abstract and introduction) or the salinity averaged over the top 30m (as e.g. stated in section 2.1 or the caption of Figure 9)?
Answer: We thank the reviewer for this comment. We added this information in the Abstract:
"...the sea surface (i) water temperature and (ii) salinity (averaged over the top 30 m)..."

p. 1, L. 20: "best models" in what respect? Please be precise here.
Answer: In Introduction we add text
"… 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."

p. 1, L. 22: I don't understand this the statement about "30 combinations of most-skillful models were selected". Selected for what? Additionally, I suggest to state how many models are considered within each combination.

p. 1, L. 23: “common” is used in what sense here? I don’t understand this. How do you define “high skill”? This is rather subjective. I suggest to rephrase.

Answer: We thank the reviewer for the above two comments. The sentences are now modified as follows: “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.”

p. 1, L. 25: What should e.g. modelers conclude from your analysis? I miss a statement on the broader implications of your study in the abstract.

Answer: We decided to remove this paragraph as Reviewer #2 suggested us to do it.

Introduction:

Note: The generalized answer will be given for the listed below Reviewer’s comments:

p. 2, L. 2-5: Please include a brief description on how exactly coccolithophores impact the carbon cycle (as done for the sulfur cycle). Additionally, I don’t think Rivero-Calle et al. (2015) and Winter et al. (2013) are appropriate references for the biogeochemical impact of coccolithophores here, as these only describe changes in the biogeography and occurrence over time. Check e.g. Iglesias-Rodriguez et al. (2002) or Balch (2018) (and references therein) for the biogeochemical imprint of this phytoplankton group.

p. 2, L. 3: Please delete the “additionally”. You describe the impact on the sulphur cycle here.

p. 2, L. 6: It is not only essential to study E. hux. blooms, but coccolithophore blooms in general. E.huxleyi has not yet been introduced in this line of the text. Please change to “coccolithophores” instead of “E. huxleyi”

p. 2, L. 7: Please introduce the abbreviation “E. huxleyi” here.

p. 2, L. 9: Please add a reference to the temperature and salinity tolerance.

p. 2, L. 10: I suggest to add the more recent reference “Krumhardt et al. (2017)” here, as they provide the most recent compilation of the global present-day distribution of coccolithophores (to my knowledge).

p. 2, L. 11: Have coccolithophores really expanded because of ecosystem changes in the Arctic? Don’t you mean “as a result of recent changes in environmental conditions, coccolithophores have expanded poleward”? Please revise the logic in this sentence.

p. 2, L. 12: Henson et al. (2018) is not an appropriate reference here (they don’t talk about the changes in E. huxleyi blooms in the cited paper). Please consider adding e.g. Rivero-Calle et al. (2015) here.

p. 2, L. 12: Winter et al. (2013) suggest that that the poleward expansion is mainly driven by temperature, salinity, or nutrients, but Rivero-Calle et al. (2015) and Krumhardt et al. (2016) suggest that carbonate chemistry matters as well. Please be more comprehensive in the discussion of possible drivers for the expansion.

p. 2, L. 14: Please be more precise here: When you say "E. huxleyi blooms have a high positive correlation with […]", do you mean the occurrence, their size, their duration…?

p. 2, L. 14: The description of controls on E. huxleyi blooms (and causes for its changes) is not comprehensive enough. I only did a very brief 10-minute search in the literature and found a number of papers that could be relevant for the introduction of this paper (only focusing on those of the northern hemisphere, i.e. disregarding the wealth of recent literature on Southern Ocean coccolithophore dynamics, see e.g. Balch et al., 2016, Nissen et al., 2018 and references therein): please have a look at e.g. Daniels et al. (2015), Harada et al. (2012), Oziel et al. (2017), and Smyth et al. (2004) (and references therein). I suggest to first describe the factors that contribute to phytoplankton/coccolithophore blooms in general (these are currently not introduced) and to then discuss what has been suggested for coccolithophores in general and in the (sub)arctic in particular. Please motivate why you think nutrients and carbonate chemistry are not important as this is not at all obvious.

p. 2, L. 12-20: Please clearly differentiate between discussing drivers of present-day coccolithophore blooms as opposed to possible drivers of observed/future changes in coccolithophore distributions and bloom dynamics.

p. 2, L. 20-32: I find it problematic to focus so much on a single paper here, especially as the discussed paper by Kondrik et al. (2019) has not yet been accepted. One of the main criticisms by the reviewers of that paper was the neglect of important variables as potential drivers of coccolithophore blooms (such as e.g. carbonate chemistry). I think the study at hand can be much more generally motivated, without going into the details of this specific one. To that aim, and similarly to the points raised in the review of Kondrik et al. (2019), the analysis in the manuscript by Gnatiuk et al. should be more comprehensive in the assessment of potential drivers of coccolithophore blooms, especially because the output from models is assessed here, which can provide information on all environmental variables impacting phytoplankton growth. There should not be a a-priori-restriction to the drivers assessed here without giving a good reason to do so. Please revise the introduction and the analysis in that respect.

p. 2, L. 2-32: The whole first part of the introduction does not provide a comprehensive summary of what is known about drivers of coccolithophore bloom dynamics and does not naturally result in the knowledge gap that will be assessed in this study. From my point of view, it should be substantially revised following the comments made above. Additionally, the models are not properly introduced. E.g. no reference to the CMIP is given. Furthermore, it should be clearly stated that none (or maximum a few, to be doublechecked) of the CMIP5 models includes an explicit parametrization of coccolithophores, which is why it is currently only possible to project potential changes of their blooms based on changes in environmental conditions (but note the recent paper by Krumhardt et al., 2019). This comes with the limitation that biotic interactions cannot be assessed, which should be clearly stated in the discussion section (see also Krumhardt et al., 2017).

Answer:
We thank the reviewer for the valuable and constructive comments for Introduction (page 2) of the manuscript. In view of the reviewer's comments, we decided to remove the first three paragraphs in the Introduction and we concentrated on the main goal of this manuscript, i.e. a selection and validation of the CMIP5 climate models against the available reanalysis data for the important abiotic parameters that influence phytoplankton blooms in six sub-arctic and arctic seas (Barents, Bering, Greenland, Norwegian, North and Labrador Seas). We changed it to the following:

"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."

p. 3: The first paragraph does not link well with the above. Please work on your flow in the introduction. Additionally, this whole page reads like it should be in the method section. Please revise and consider moving at least parts of it to the method section.
Answer: Thank you for the comment. Since we changed the focus of the paper, we decided to leave this information in the Introduction. This part reviews the previous experience in the field of climate models validation and selection, that leads to formulating of a scientific task.

p. 3, L. 7: How are the "best models" defined here? Please be precise what you mean.
Answer: We modified it as follows:
"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; Knutti et al., 2010b; Reichler and Kim, 2008); 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; Knutti et al., 2010b; Taylor, 2001)."

p. 3, L. 8: Please revise the sentence “These two approaches usually give a good result”. Good in what respect? Please add references.

Answer: We improved the sentence as follows:

“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.”

p. 3, L.11: Choosing this method implies the assumption that whatever model is representing present-day conditions best will also do the best job in projecting these into the future, doesn’t it? I think this is important to state here.

Answer: Thank you for the comment, we modified it as follows:

“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.”

p. 3, L. 16-21: It is not clear to me what the take-away message of this paragraph is. How does the first approach, assessing how well models do in representing air temperature, sea level pressure, and precipitation help in the assessment of environmental factors impacting phytoplankton/coccolithophore growth? Please work on this paragraph and make it more specific to the goal of your work. Consider combining it with the next to avoid having a 2-sentence paragraph with no clear take-away message.

p. 3, L. 21 – p. 4, L.2: Again, the take-away message in context of your specific goals for the paper are not clear. Please re-write.

Answer for above 2 comments: We think that now the information p. 3, L. 16-21 (version 1) suits to the new version of the manuscript. We modified the paragraphs as follows:

“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."

p. 4, L. 3: Why do you conclude that? This is not clear to me from what you have presented so far in the introduction.

Answer: We deleted this sentence. We describe more clearly which available ideas we adopted in the developing of our methodology of climate model validation and selection. We added the following information:

"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 (25th,50th, 75th) that is widely used in statistical analysis, and, finally, (iv) selecting the top 25% ranked CMIP5 models following Ruan et al. (2019)."

p. 4, L. 3-5: You have not presented the differences in environmental conditions of the different focus areas. Please revise the introduction accordingly. I don't understand what the second half of the sentence means: How can areas have a wide range of parameters?

Answer: We agree that the formulation is not very good, and we modified it. Here is the new version of the whole paragraph:

"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."

Methods

p. 4, L. 17-19: As mentioned before, I don't understand based on what grounds you neglect the assessment of nutrients and the carbonate chemistry in the models.

Answer: We agree that nutrients and the carbonate chemistry are important for coccolithophores bloom. Since the focus of the manuscript is changed, we do not analyze all factors influencing coccolithophores bloom, we analyze only abiotic parameters. We added additional sentences and modified the sentence as follows:

"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)."

p. 4, L. 23-25: Did you include regional models, e.g. CORDEX? I can't find it in Table 1. If you didn't include those models, don't make that statement here. I am bit confused. Please distinguish between regional and global models and state which kind you considered.

p. 4, L. 25: Did you only consider global models in the end? This is not clear from your description in this section. Please clarify.

Answer for two above comments: We did not include regional models, we used only global models. We deleted this sentence.

p. 4, L. 25-26: I suggest to give the range of models available: number available for FFs ranged from X1 for variable Y1 to X2 for variable Y2 (see Table 1). What do you mean by "main characteristics"? Please rephrase.

Answer: We added the following sentence:

"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."

Also we modified the sentence that included the phrase "main characteristics" as follows: "The list of climate models used is presented in Table 1."

p. 5, L. 4: replace "has been shown" by "have been shown".

Answer: We changed it.

p. 5, L. 6: Please choose a better description in the title than simply “methods”, maybe something like “model evaluation metrics”?

Answer: We changed it to “Study regions” and “Model evaluation metrics”

p. 5, L. 7: Please rewrite “regions under the study”. Add “Sea” behind “Norwegian”.

Answer: We changed it: “The target regions are six arctic and subarctic seas: the Barents, Bering, Greenland, Labrador, North and Norwegian seas.”

p. 5, L. 7-8: How do you define a bloom here? Please state this and add references. Additionally, you don’t state what data you base Fig. 1 on to define the blooms. Please clarify in the text and the figure caption. I suggest to draw the study regions into Fig. 1. To help the reader localize the different subregions.

Answer: We changed the sentence as follows:

“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.”

We changed the figure caption accordingly.

p. 5, L. 11: Do you mean model output here when you say “data? Please clarify.

Answer: We clarified it at the beginning of the section as follows:

“Thirty-four CMIP5 GCMs outputs for the historical period 1979-2005 were used in this study.”

p. 5, L. 12: I have a hard time believing that the blooming period lasts from January-December in the Bering Sea. What bloom definition is used for this?

Answer: We modified the sentences and added some information:

“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).”

p. 5, L. 10-14: I don’t fully understand why you’re restricting the analysis to the times and locations of identified E. huxleyi blooms under present-day environmental conditions for each sea (if the models do not necessarily reproduce the environmental conditions at these exact locations and times). Don’t you want to restrict the analysis to the observed environmental conditions at the times/locations of the blooms (i.e. the observed environmental niche)? As a consequence, I am wondering why don’t you define each subregion as a slightly larger area than currently done.

Answer: First, we conducted an analysis for the entire territory of the seas and found a significant difference if we consider the whole sea and the area we are interested in. Trying to answer the question that the reviewer has raised here, we entered the following sentences into the text:

"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."

p. 5, L. 17: The interannual variability of what exactly? The seasonal cycle/amplitude, summer average, average over blooming period, …? Please be precise here.

Answer: We modified the sentence here:

"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."

p. 5, L. 18: The seasonal cycles […]

Answer: We changed it.

p. 5, L. 19: "but the interannual variability " of what?

p. 5, L. 19: Replace "sea" by "subregion"

p. 5, L. 25-26: Rephrase to something like "For the assessment/evaluation of the interannual variability […]"

Answer for the two above comments: We changed it to:

"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."

p. 5, L. 23: Can you rephrase "RMSD-observations standard deviation ratio"? I have a hard time understanding what you mean here. Please consider to add the formulas to make it very clear.

Answer: We changed the sentence:

"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."

p. 5, L. 26: Do you mean the difference in the spatial distribution of temporal trends between the model output and the reanalysis data? This sentence is not clear to me. Please rephrase to clarify.

Answer: We changed it to the following sentence:

"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."

p. 5, L. 26-27: What exactly is "your percentile score-based model ranking method"? This method is defined nowhere in the method section up to this point. In particular, the description of this ranking method should be very clear (e.g. by including an overview listing the metrics are included in the ranking), as the main result of your study is based on this ranking.

315 p. 5, L. 31: Less than 25% of what? Please be precise. What do you base these thresholds on? It seems rather subjective to me. What is the effect of the choice on the outcome? This needs to be discussed somewhere in the text.

p. 6, L. 4: Again, choosing 25% seems random to me (see previous comment).

Answer for the three above comments: We modified the text as follows:

"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."

Results & Discussion

p. 6, L. 6: Personally, I find it a bit unfortunate that only results for temperature and the Barents Sea are presented in the main text. Isn't there a better way to synthesize the results and present more than just one tiny subarea and one forcing factor?

Answer: We agree with the reviewer, and added an additional figure to sections Results and Discussion with the spatial distribution of biases in five parameters between models and reanalysis data in six target seas. The biases are averaged over the vegetation season and the time period 1979/1993-2005. We added new Figure 5 a-e and following description:

"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."

Below we present Figure 5 (a-e):

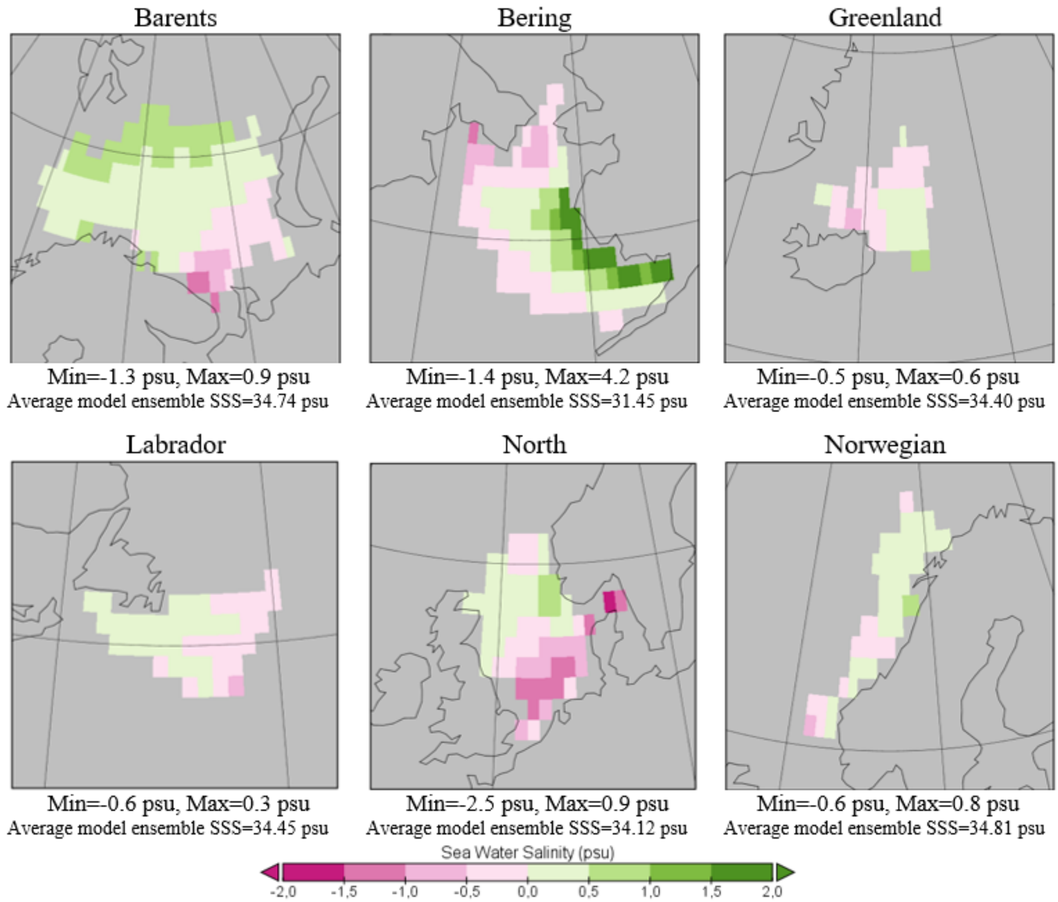


Figure 5a. Spatial distribution of biases in sea surface salinity models and reanalysis in six target seas averaged over the vegetation season and the time period 1993-2005.

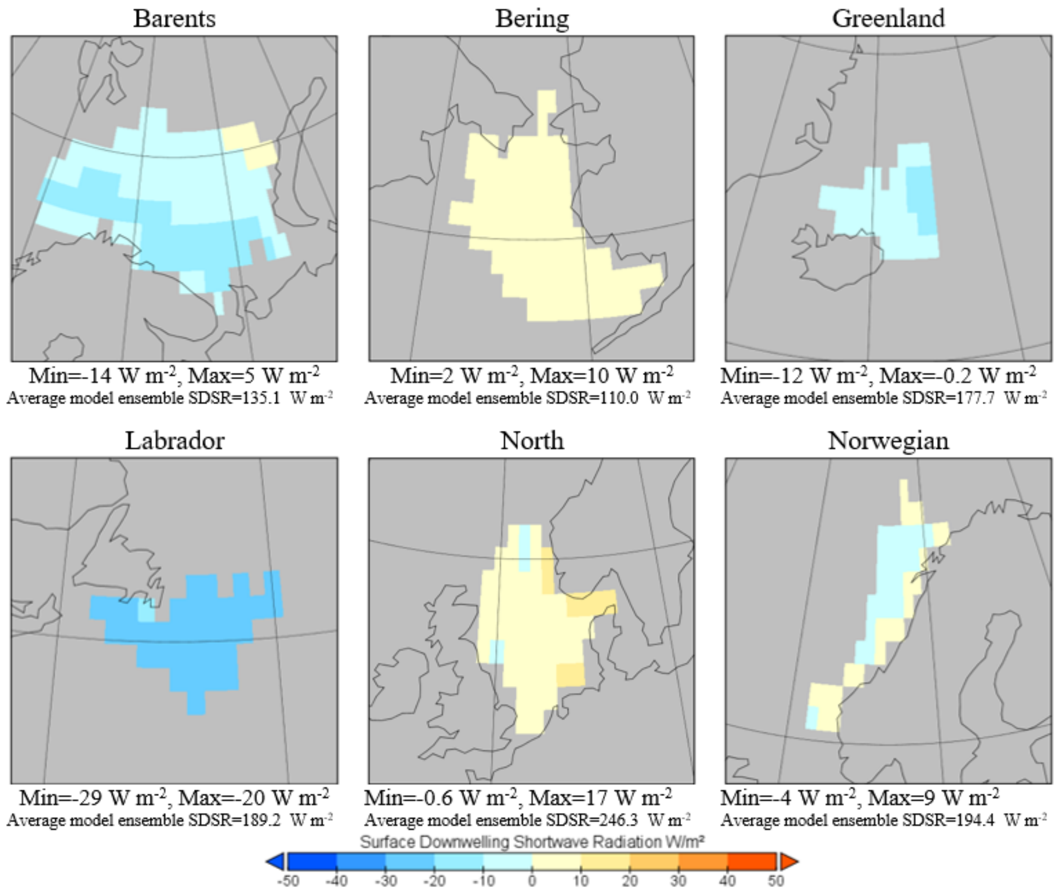


Figure 5b. Spatial distribution of biases in surface downwelling solar radiation between models and reanalysis in six target seas averaged over the vegetation season and the time period 1979-2005.

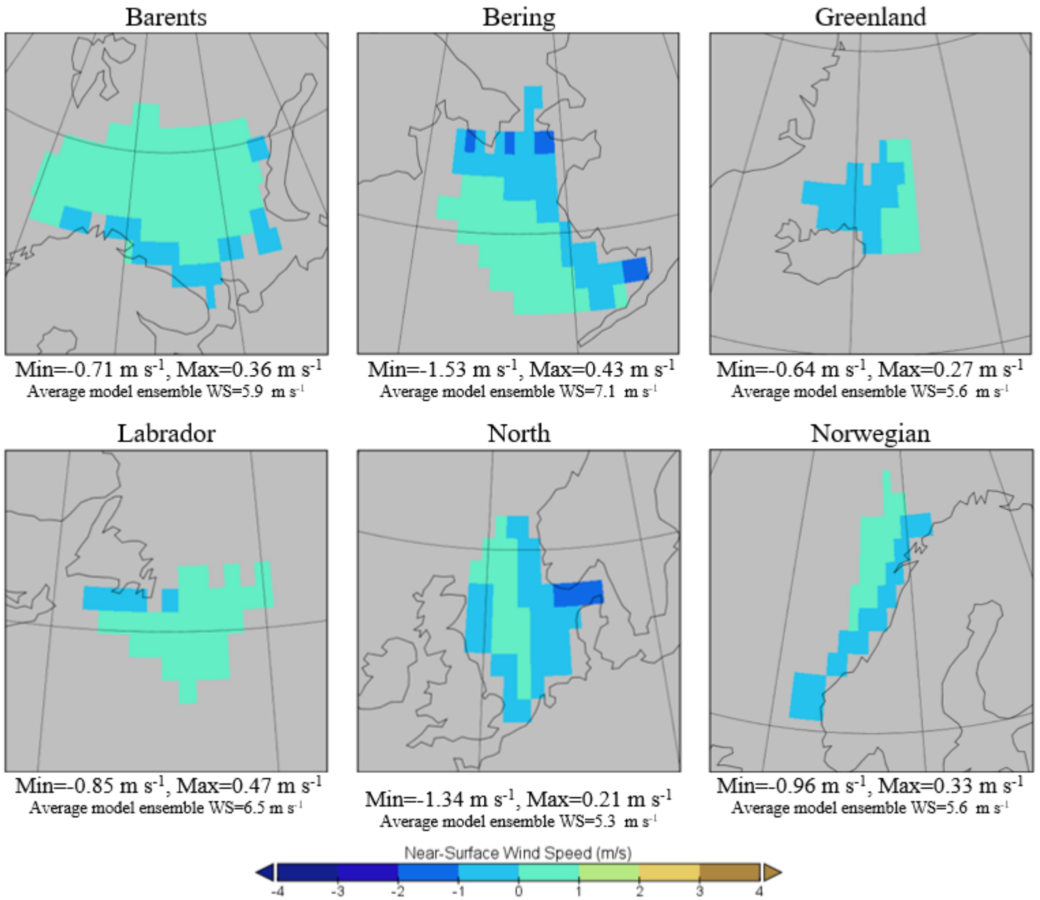


Figure 5c. Spatial distribution of biases in near-surface wind speed between selected model ensemble and reanalysis in six target seas averaged over the vegetation season and the time period 1979-2005.

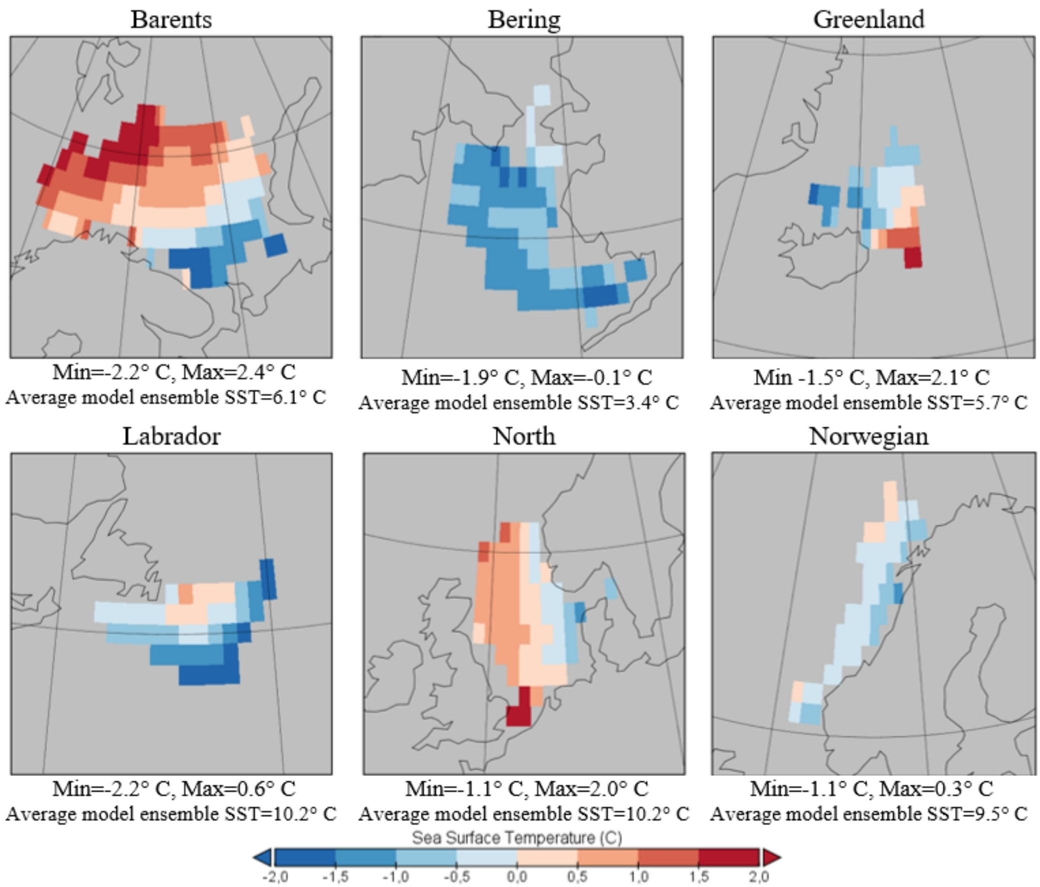


Figure 5d. Spatial distribution of biases in sea surface temperature models and reanalysis in six target seas averaged over the vegetation season and the time period 1979-2005.

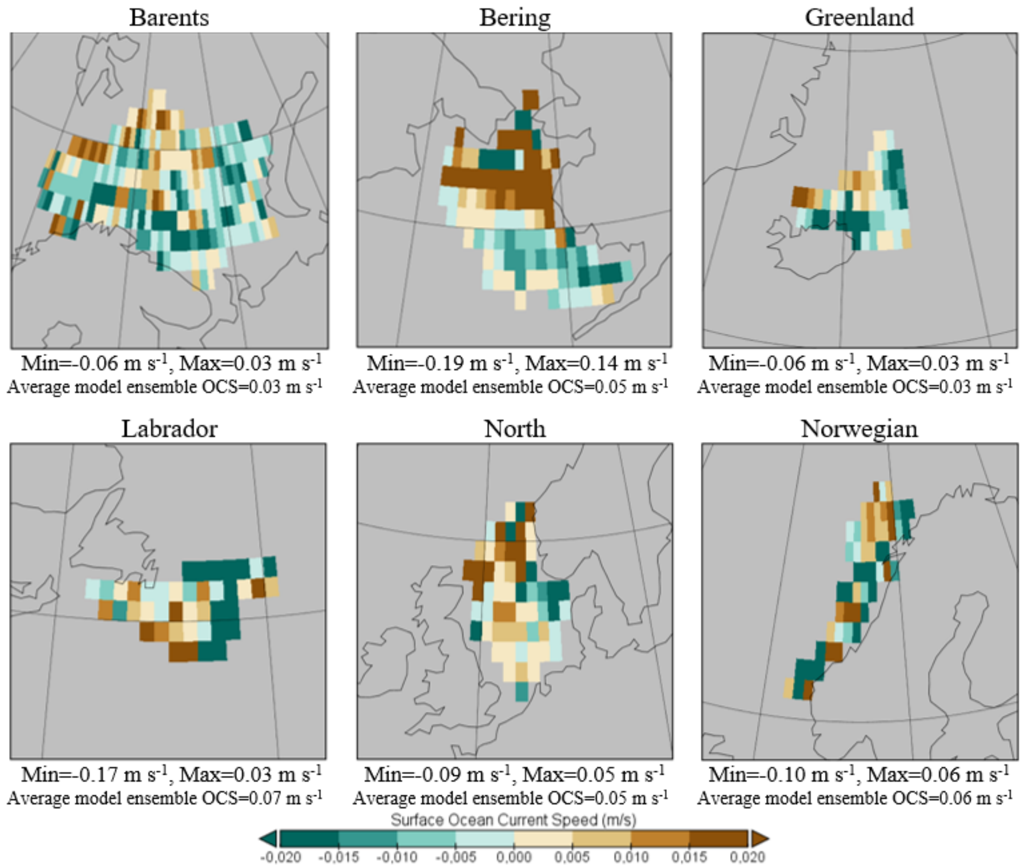


Figure 5e. Spatial distribution of biases in surface ocean current speed models and reanalysis in six target seas averaged over the vegetation season and the time period 1993-2005.

p. 6, L. 9-11: I don't think a Taylor diagram needs to be explained in the result section. I suggest to rather briefly explain when the agreement between model and reference data set is good (i.e. how to interpret the plot) instead of simply stating what can be seen (see comment on L. 15-17).

p. 6, L. 11/12: Please add "[…] capture the climatological seasonal cycle […]". Furthermore, please explain how it can be seen from the plot that the seasonal cycle is represented better than the interannual variability (see previous comment).

p. 6, L. 14: Are these numbers really unitless? If so, define somewhere that you plot normalized SD and RSMD (method section, consider adding formula there) and state that here by saying e.g. "the SD and RMSD normalized by XX are between …". This will help the reader to follow.

p. 6, L. 15-17: This is the information you should start your paragraph with (see previous comments). First explain to the reader how to interpret the plot. However, the statement that "the closer the model data is to the x-axis, the better the correlation coefficient" is not entirely correct, as the correlation coefficient is shown on the radial axis. A point with RMSD/SD/CorrCoeff of 0.1/0.1/0.1 is closer to the x-axis than a point with 1.0/0.8/0.9 (note that this is under the assumption that RMSD is on the x-axis, SD on the y-axis and the correlation coefficient on the radial axis, see comment on the Figure further down)– but the correlation coefficient of the second point is higher. Please be precise in the description. Also, a correlation coefficient is high/low and not good/bad.

p. 6, L. 17: Replace “climate parameter” by “e.g. SST”.

p.6, L. 17-18: Please revise the grammar of this sentence.

p. 6, L. 18-20: This statement is redundant with the method section.

p. 6, L. 9-20: For the whole description of the Taylor diagram, please add the names of models here that show the highest/lowest correlation coefficients, RMSD etc. to make it easier for the reader to extract the information from the plot.

Answer to above 8 comments:

Our intention was to describe our methodology using Figures 2-8. We decided to delete the Taylor diagram as it only illustrates the root mean square deviation, standard deviation and correlation together; but we analyzed these statistical metrics separately in the form of a table. Thus, we deleted the second paragraph from page 6 and moved Figures 4 to section Methodology section. In addition, we deleted Figure 5 as well, since the analysis is very similar to Figure 4, and Figure 8.

p. 6, L. 21-23: If you say you show the “spatial distribution”, I expect maps. Do you mean the spatial variability of the climatological SST bias across the subregion? Please be more precise throughout the description.

Answer: We modified the sentence as follows: “Figure 3 illustrates the box plots of the spatial variability of SST biases in the selected area of the Barents Sea for the vegetation season (June-September) during 1979-2005 and the time period 1979-2005.”

p. 6, L. 22: I see median biases that are >0 (e.g. for the model 2). Please double-check.

p. 6, L. 24: Do you mean the maximum bias? I don’t understand “amplitude bias” (throughout paragraph). Similar to above, please add the names of the models showing the numbers you’re stating to make it easier for the reader to find the information you’re stating in the plot.

Answer to the above 2 comments: Thank you for this correction. We changed it as follows:

“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).”

p. 6, L. 24-25: Please revise this statement, e.g. simply stating that the comparison shows large variability across the models.

Answer: We changed it as follows:

"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)."

p. 6, L. 27-29: Where is this seen? This is not included in Fig. 4. If you're referring to a different plot here, please add the reference.

Answer: We changed it as follows:

"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)."

p. 6, L. 29: Similar to above, be more precise in your description. From just the wording "spatial distribution of annual trends") the reader expects maps here, not box plots.

p. 6, L. 31: How are "significant trends" defined here? How can that be seen in the plot? Please be precise. What models show a significant trend? What is an "unrealistic trend" for you here?

p. 7, L. 1: How do you know that? As mentioned above, I think it is important to state in the method section that this is the assumption you make (a model that reproduces the observations best over the historical period (however you define "best"), also gives the "best" projections for the future).

Answer to the above 3 comments: We deleted the figure plotting the spatial variability of the trends as the procedure of the analysis is similar to the figure displaying the spatial variability of biases. Also, we will move Figure 4 to section Method.

p. 7, L. 9-12: Is the +/- 1K the average over the domain? Currently, the reader at this point has totally forgotten why you're doing this exercise. I suggest to always relate your analysis back to your goal of projecting potential future changes in coccolithophore blooms. I understand that this will be a follow-up paper, but this paper would gain a lot if you speculated at least. How can these biases be expected to impact these estimates? You could do some basic calculations using a Q10 function (see e.g. Nissen et al., 2018) or a temperature optimum function (see e.g. Krumhardt et al., 2017) describing the impact of temperature on phytoplankton growth.

Answer: We modified it as follows:

"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"

p. 7, L. 13: What error do you mean here? Please be precise and make sure that all the metrics you present are carefully introduced in the method section.

p. 7, L. 15-16: Please revise the grammar of this sentence.

p. 7, L. 13-17: Similar to above, I don't understand from the current presentation of the results what these mean.

Answer: We modified the sentence as follows:

"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."

p. 7, L. 24-28: This is repetitive with the method section and what should be in the figure caption. There is no need to state it this detailed in the main text.

Answer: We modified the sentence as follows:

"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."

p. 7, L. 28-30: Does it surprise you that the model combinations vary?

Answer: In many studies that use climate model data, vast regions are considered, in particular, the entire Arctic. Most studies also use the approach when one set of models is selected for different parameters. Our results confirm that the same model does not properly reproduce the distribution features of all the parameters we examined and is not suitable for the analysis of large regions. It is one of our messages to readers to be more careful when to choose climate models at the study.

p. 8, L. 3-5: How is "better performance" defined here? Is not clear to me how you conclude this.

Answer: We modified the sentence as follows:

"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."

Conclusions

p. 8, L. 13-16: The statement that the Arctic is often considered as a single region in other studies is never made in the introduction, but should be included there as a motivation to look at subregions. Furthermore, you don't actually assess the whole area, so I suggest to revise this statement, as you don't actually compare the performance over the whole area to the smaller subareas.

Answer: We agree that we do not show results for all Arctic, but we study 6 seas, and we didn't find models same good for such a large area. We modified the sentence as follows:

"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."

p. 8, L. 18: What about the temporal development of the environmental conditions?

Answer: We suppose that trends are responsible for the temporal development of environmental conditions.

p. 8, L. 18-21: Are more important than what? Please be precise. I cannot follow your logic here. Please revise to clarify, taking also into account the comments I made in the result Section.

Answer: We modified the sentence as follows: "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."

p. 8, L. 24: And the time series is even shorter for SSS and ocean currents, isn't it? What is "out-of-sample" testing? Please try to avoid introducing concepts in the conclusion section which were not discussed before. Why did you not test by excluding certain time periods from the analysis?

Answer: We agree with the reviewer and deleted the following sentences: "Due to the short sample period of reanalysis data (1979-2005), we did this evaluation without out-of-sample testing. Definitely, it is better to test any model ranking method on another historical period. It will be possible to consider the period 1950-2014 with the release of new data, e.g., CMIP6, ERA5."

We consider the period is very short to be divided into two independent periods for the analysis.

p. 8, L. 27: important for what? Please be precise.

Answer: We modified the sentence as follows: "We can conclude that the range of different factors is important for model selection, including the spatial pattern of model biases, and the proposed methodology is a way of enhancing the model selection procedures sophistication that promises a better chance to identify more skilful models for the features we are interested in."

p. 8, L. 31: Why only at regional scales?
Answer: We modified the sentence as follows: "Thus, the proposed method can be used for analyses regarding other regions with the purpose to evaluate climate model performance with respect to various atmospheric and oceanic parameters at different scales."

Figures/Tables
Fig. 1: You don't simply show the "locations of the blooming areas" here, but the spatial distribution of the frequency of blooms. Please be more precise. I suggest to show the subregions in the plot directly (put names and add e.g. a black contour to show the extent). Please add to the caption what data this map is based on and how you define a bloom.
Answer: We corrected it as follows: "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."

Fig. 2: Be more precise in caption, a lot of information on what is seen in the plot is missing. What is the unit of the RMSD?
Answer: We corrected the caption as follows: "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)."

Fig. 3: The way I know it, a Taylor diagram shows the RMSD (normalized by the standard deviation of the reference data set) on the x-axis, the standard deviation (normalized) on the y-axis and the correlation coefficient on the radial axis. It is not clear to me what exactly you're showing. Please add labels to the plot (y-axis, grey circles) and also say what you're showing in the caption (including units or state if you normalize by something). Also, please add panel labels to the plot and the caption.
Answer: We decided to delete the Taylor diagram as it only illustrates root mean square deviation, standard deviation and correlation together; however, we analyzed these statistical metrics separately in the form of a table.

Fig. 4 & 5: Possibly replace “distribution” by “variability”? Be precise in what you show. What are the orange line and the whiskers? How is the bias defined (Fig. 4)? What trend is shown Fig. 5; trend in average over blooming period averaged over subregion?)?

Answer: We corrected the caption of Figure 4 as follows: “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.”

We decided to delete Fig. 5, since it illustrates a similar analysis procedure as that in Figure 4. After revision, it is Figure 3 that we moved to the section Method.

Fig. 6: Restate blooming period in caption, add unit of SST bias.

Answer: We changed the caption as follows: “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.”

Fig. 7: What error are you showing here? Please add the unit of the SST trend in the caption.

The colorbar currently states that you’re showing SST (K) – please double-check. Please restate the blooming period.

Answer: We corrected the caption as follows: “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)”

Fig. 8: In my view, it is not really common to plot SST in Kelvin, consider changing it to °C. Please add the units of the variables in the figure caption. Explain what the fit is, exchange “x” and “y” by the actual variables you fitted. Please don’t use black/dashed for all fits, I suggest to change the color of each fit to the color of the respective full time series.

Answer: We deleted this figure to avoid overloading of the paper with figures as it is not that much important in the manuscript.

Fig. 9: Please briefly summarize what the numbers for each model-variable combination represent and refer back to the method section and Fig. 2. Please also explain in the caption what the white areas are and refer back to Table 1. Please add the units to the variables in the Figure caption.

Answer: We modified the caption as follows: “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..”

Table 1: Replace “concrete” by “respective. Please define all abbreviations in the Figure caption (e.g. SST, WS…) and add units.
Answer: We corrected accordingly: “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 “+”)”

Table 2: Please add units in the Figure caption. What is SDdif? This is never explained in the text (method section). Please be consistent with the use of underscores in caption and Table (e.g. Trm vs Trm). What does “modulus of standard deviation difference” mean? I don’t understand this. Please use the exact same names as introduced in the method section.
Answer: We improved this part in the section Method. We corrected the caption as follows:
“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; |SDdif| is the modulus of the standard deviation difference (model minus reanalysis), K; |Trm| is the modulus of spatial trend median difference (the model minus reanalysis), K yr-1; |Tra| is the modulus of spatial trend amplitude difference (the model minus reanalysis), K yr-1; |Brm| is the modulus of spatial bias median difference (the model minus reanalysis), K; |Bra| is the modulus of spatial biases amplitude difference (the model minus reanalysis), K).”

General comments – Reviewer #2

This study evaluates (and ranks) the performance of 34 climate models in simulating 5 physical parameters [namely, 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); shortwave downwelling solar radiation (SDSR)] on a sub-regional scale in the Arctic and Subarctic regions. These 5 parameters are selected as "forcing factors (FFs) controlling E. huxleyi blooms in arctic and subarctic seas" (p. 2, line 21) and tested in six seas (Barents, Bering, Greenland, Labrador, North and Norwegian seas) where the coccolithophore Emiliania huxleyi is known to form blooms.

I believe the core of the study is interesting and merits publication, but I think the authors could and should do a better job at discussing (all) the results. By not having a primary focus on E. huxleyi blooms in the Introduction, the reader will be able to recognize the wider implications of this extensive intercomparison of climate models – it will also alleviate some of the major issues of neglecting "what else" underpins coccolithophore blooms and their occurrences. Nothing wrong with mentioning your motivation for selecting the regions of interest and the potentially relevant abiotic parameters, but as is, the reader is expecting more than is actually presented re. coccolithophore blooms (see below).

Specific comments

The authors explain that this study is a precursor for another study (in prep/planned by Kondrik et al.), in which these FFs will be applied to "model the future dynamics of E.huxleyi blooms" – so in fact, the current study has very little to do with E. huxleyi blooms apart from being the motivation for the presented set-up. There is no objection to test the model performance of the selected parameters, but the authors should do a better job at explaining why these factors were selected, and others ignored (i.e. because they cannot be assessed in the models? I wonder). Because it could be easily argued that the authors miss a crucial parameter in their line-up of FFs – nutrient availability – that arguably underpins any phytoplankton bloom (i.e. sustained exponential growth). Any biotic factors (e.g. grazing pressure) are ignored herein.

Indeed, it is unclear what correlations are sought between the various FFs and E. huxleyi blooms – what do you mean with "affecting"? – e.g., the onset (triggers), the duration/maintenance of blooms, other affects?

After reading the ms, I felt that the study is a valid and interesting intercomparison of climate models, raising important issues in simulating abiotic parameters, but that the initial focus on E. huxleyi seems too specific here; i.e. without it, all results could be presented just as well – or even better as these parameters surely affect more than just E. huxleyi, thus giving the study a wider relevance. In fact the authors conclude rather generally, without discussing specific implications for the next/planned study by Kondrik et al. – so that also gives the impression that the initial motivation need not take central stage in this ms (or the title). Still, I don't fully understand what the strategy would be in "selecting the best models" for such follow-up study, given the multivariate outcomes, this could/should be better explained in the final discussion and conclusion.

Figure 9 ("heat map") is a good visual representation of the amount of work performed and the complexity of the outcomes; not only does it show the range in performance between the listed models (1-34), but also how within one model the chosen parameters are simulated at different strengths – and, possibly even more intriguing (disconcerting?) that a model that performs very well for one sea, does not in another (for example, compare model 1, ACCESS1-3 in Barents and Bering Seas). Indeed, the authors conclude that the results "show that there is no optimal model ensemble or one top-model which could best simulate all factors across all of the study regions. 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" (p.8, L. 12-15).

What I miss, is an in-depth discussion why these inter-model, inter-parameter and inter-subregional differences exist – is this due to issues of spatial resolution, initial parameterization of each model (what it was built for) or real physical differences between the seas that models cannot address/capture? Again, I don't know, but would be interested to learn what factors could underpin the results in Fig. 9. Currently, the "results and discussion" section reads as a list of figure descriptions rather than highlighting the main take-home messages (while figure captions could do with more information). Moreover, only one of the 5 factors is highlighted (SST) as "an example" – I believe the paper would have a much greater impact if the other parameters get equal treatment or at least their highlights mentioned and discussed in the main text, not only in a supplement.

Reply to Reviewer # 2:

We thank reviewer #2 for very helpful comments. We fully agree with the arguments regarding the suitability of application of the proposed climate models selection methodology based on parameters that impact blooms not only E. huxleyi but also other phytoplankton in the study areas. Therefore, we decided to change the focus from the coccolithophore and concentrate on the methodology of choosing climate models. This step is fully justified, since this work, in its essence, is certainly not connected with the work performed by Kondrik et al. (2019) of simulating conditions that modulate the intensity of coccolithophore blooms. In the absence of a close connection to coccolithophores, the article indeed gains greater clarity and becomes focused on the substance of the research done on the comparative effectiveness of global climate models for specific marine objects.

We improved all manuscript sections. Also we consider to move Fig. 4 and related text from the Results section to Materials and method, and delete Figures 3, 5, 8 due to either their resemblance to some akin figures or because their presence is not so important in the manuscript. In addition, we decided to add to the section Results and Discussion a new figure displaying the spatial distribution of biases in five parameters between the models and reanalysis data in target studied seas; the biases are averaged over the vegetation season and the time period 1979/1993-2005. We added Figure 5a-e, and following description:

"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."

Below we present Figure 5 (a-e):

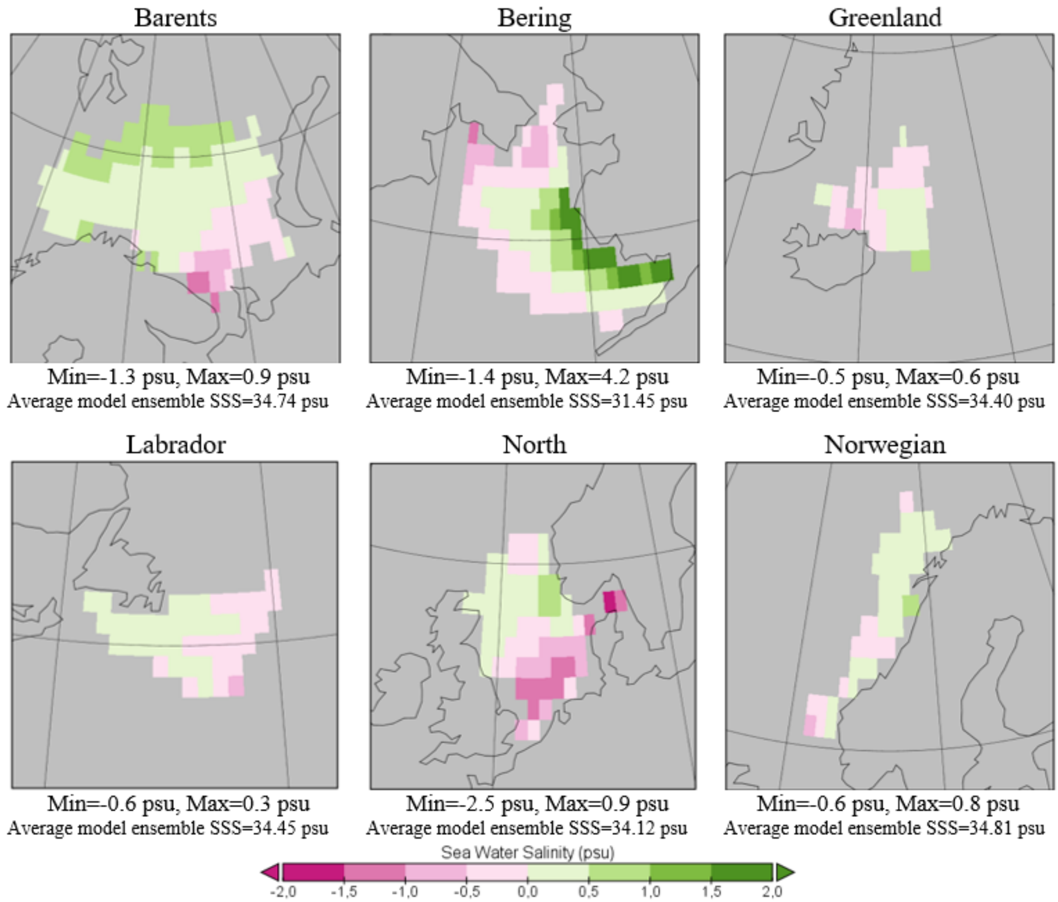


Figure 5a. Spatial distribution of biases in sea surface salinity models and reanalysis in six target seas averaged over the vegetation season and the time period 1993-2005.

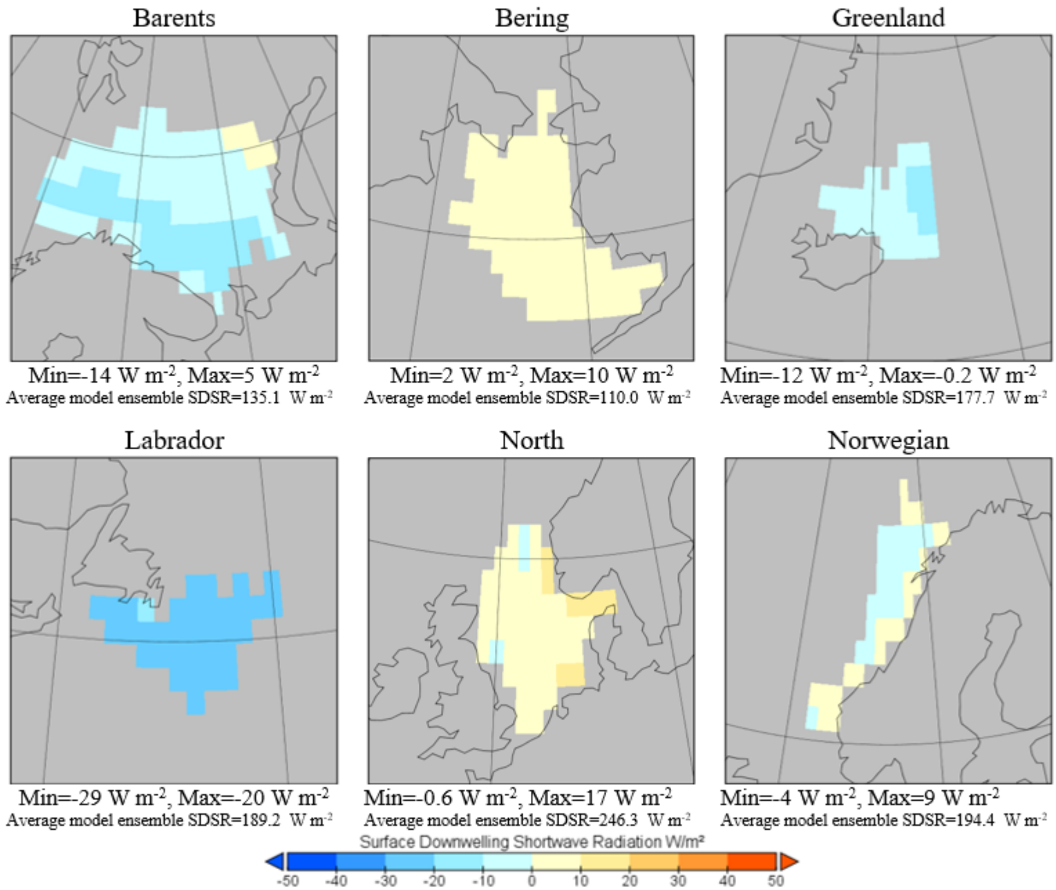


Figure 5b. Spatial distribution of biases in surface downwelling solar radiation between models and reanalysis in six target seas averaged over the vegetation season and the time period 1979-2005.

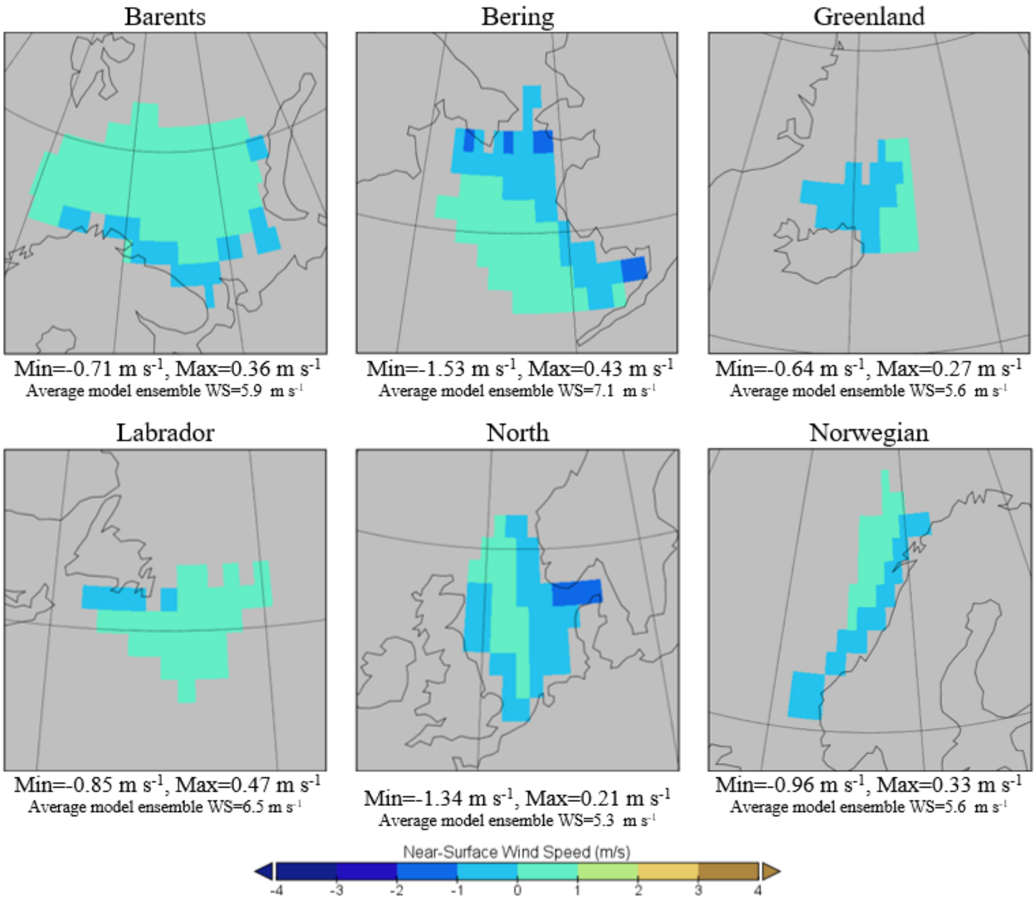


Figure 5c. Spatial distribution of biases in near-surface wind speed between selected model ensemble and reanalysis in six target seas averaged over the vegetation season and the time period 1979-2005.

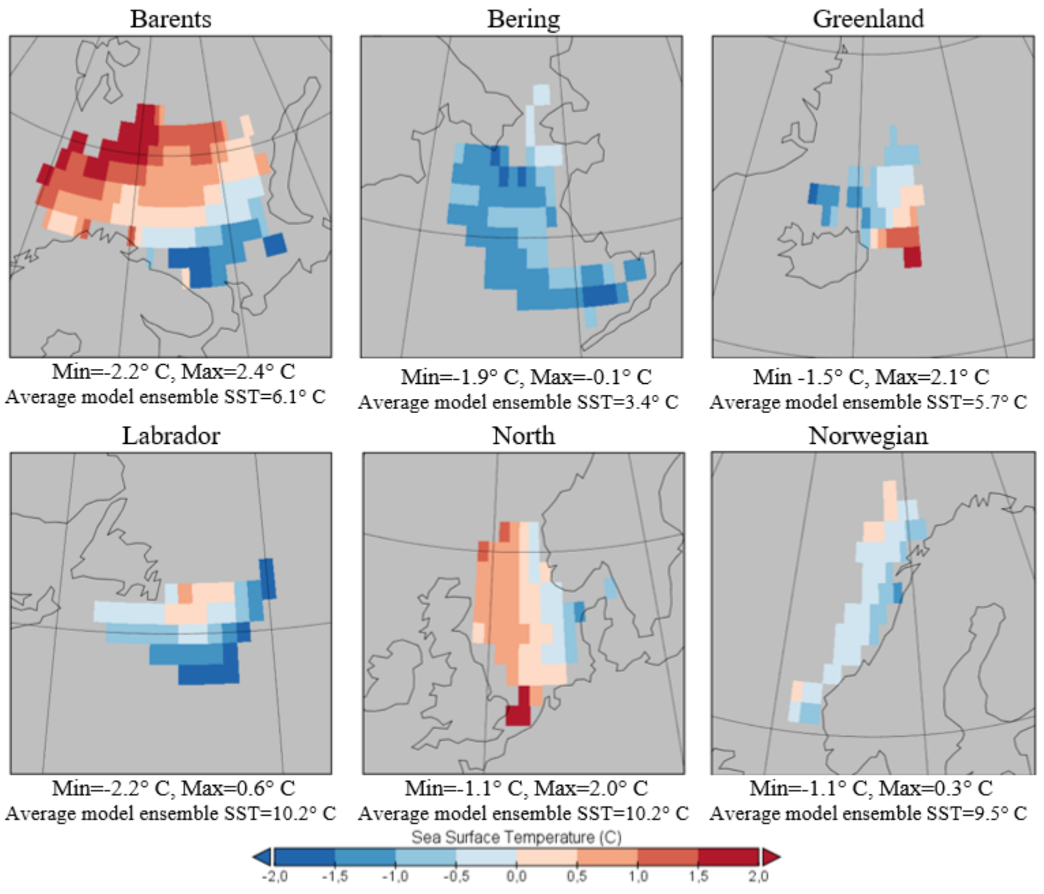


Figure 5d. Spatial distribution of biases in sea surface temperature models and reanalysis in six target seas averaged over the vegetation season and the time period 1979-2005.

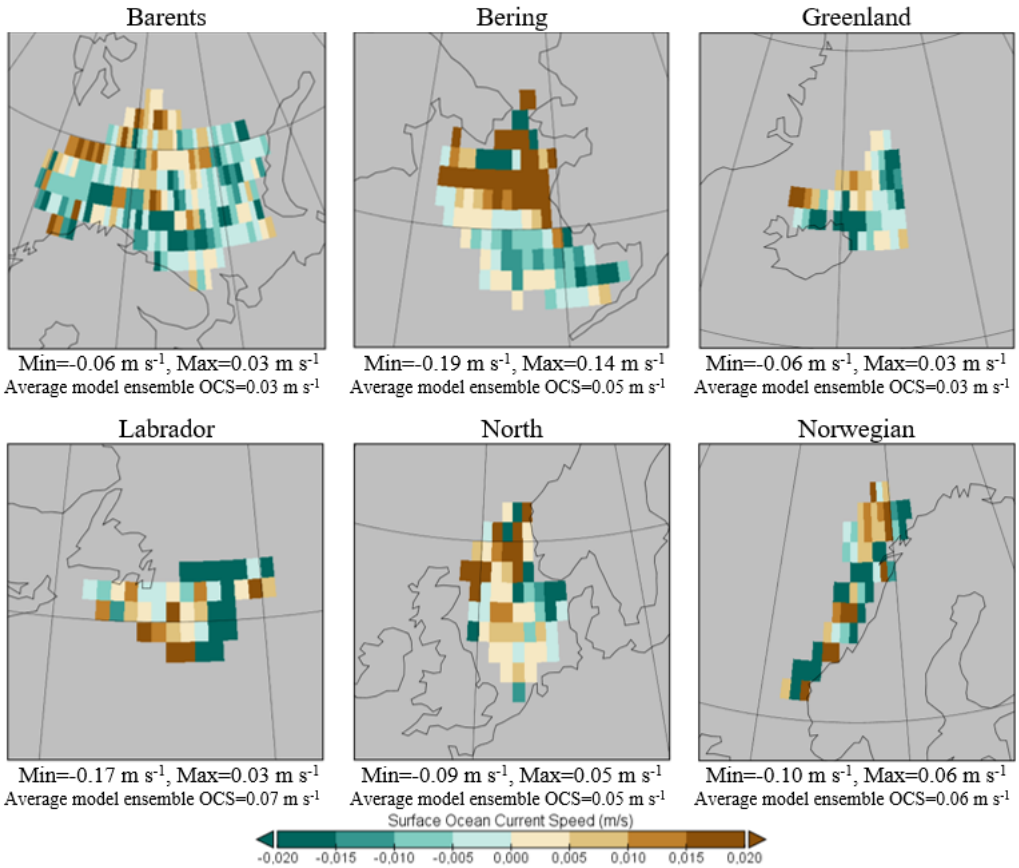


Figure 5e. Spatial distribution of biases in surface ocean current speed models and reanalysis in six target seas averaged over the vegetation season and the time period 1993-2005.

Abstract: first sentence, shortly name the reasons; why only carbon cycle mentioned here, as opposed to carbon and sulphur cycles in first sentence of Intro?

Answer: We removed the first paragraph of Abstract in order to concentrate on the main goal of our manuscript. We changed the first paragraph as follows:

"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."

Line 25 (last paragraph): too much information (and acronyms) for abstract. Remove.
Answer: We removed the last paragraph.

Intro, p. 4. Lines 8 -14 - this paragraph "goes without saying"; what follows is generic order of methods, results, discussion.
Answer: We agree with the Reviewer and we deleted this paragraph.

Intro/Methods: What is a CMIP5 climate model / the CMIP5 project? Define and describe – currently not done anywhere.
Answer: We introduced "CMIP5" in the first paragraph of the Introduction as follows:
"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)."

Figure 1 (if kept as motivation for selected regions), please state what type of data are shown and cite data sources in caption.
Answer: We corrected it as follows:
"Figure 1: Spatial distribution of E. 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."

General: Many figure captions need more details for reader to follow or identify data Sources.
Answer: We improved the figures captions as follows:
"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).

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 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)"

Technical corrections

If you decide to keep Emiliania huxleyi in, know to write the full species name the first time the species is introduced in the text, as well as any time you start a sentence with "E. huxleyi" (change to "Emiliania huxleyi"). Also put space between E. and huxleyi. Alternatively, as motivation you could mention "coccolithophore blooms" as a more generic way – and comment that many of the blooms in the Arctic and Subarctic are indeed formed by one species.

Check: Winter et al., publication year is 2014?

Answer: We thank the Reviewer; we have corrected 'Emiliania' word writing in the text; the correct form is Winter et. al (2013)

p. 5, Line 7: delete "the" between "under" and "study" / and consider replacing as "under investigation". Add "seas" after list of sea names.

Answer: We corrected the sentence as follows:

"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."

Line 18: add: "The" seasonal cycle

p. 7, Line 32: models (plural)

p. 8, Line 28: add "the" before proposed methodology

Answer: We corrected it.

Below we present the proposed changes and modifications in the manuscript, and we will improve it further.

Simulation of ~~factors affecting *E.huxleyi* blooms~~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. ~~The coccolithophore *E.huxleyi* plays an essential role in the global carbon cycle. Therefore, considering the ongoing global warming, the assessment of future changes in coccolithophore blooms is very important. Our paper aims to provide a framework for selecting the optimum combination of global climate models to conduct such an assessment. To do this we analyse the forcing factors influencing present and future blooms using climate model projections. Then, based on the projected changes in the forcing factors, future changes in the dynamics of coccolithophore *E.huxleyi* blooms can be determined.~~

~~Here we describe~~ Currently, there are a large number of climate models that give projections for various oceanic and meteorological parameters in the ~~complex~~ 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 ~~used for~~ 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 ~~climate models, and the selection of models that best~~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 ~~the regional features of the~~ several important oceanographic and meteorological ~~factors affecting *E.huxleyi* blooms in~~ 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 ~~CMIP5 Atmosphere-Ocean General Circulation Models~~GCMs against reanalysis data includes analysis of the interannual variability, seasonal cycle, spatial biases and temporal trends of the simulated ~~forcing factors. Here we propose a percentile score-based model ranking method for the selection of the best models from the CMIP5 ensemble. The selection of the best models was performed separately for each study area in the Barents, Bering, Greenland, Labrador, North and Norwegian Seas and for each of the five forcing factors affecting the coccolithophore blooms.~~parameters. In total, 30 combinations of ~~most~~high-skilful models were selected~~.~~ for 5 variables over 6 study regions. The results show that there is no ~~common~~mutually optimal combination of models, nor is there a one top-model, that has ~~high~~a skill in reproducing either the regional climatic-relevant features ~~across the combination~~of the whole Arctic region or all combinations of the ~~five~~ considered ~~forcing factors and all arctic and~~

~~subarctic~~parameters in target seas. ~~However, some climate models consistently show good skill for many of these combinations e.g. ACCESS1-3; ACCESS1-0; HadGEM2-AO; HadGEM2-CC; HadGEM2-ES; GFDL-CM3; INMCM4; GISS-E2-R; GISS-E2-R-CC. The models that have the smallest skill for the majority of the study regions are CMCC-CM; FGOALS-g2; IPSL-CM5A-LR; IPSL-CM5A-MR; IPSL-CM5B-LR; MIROC5; MRI-ESM1.~~ 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

~~The coccolithophores contribute significantly to the global carbon and sulphur cycles (Malin et al., 1993; Matrai et al., 1993; Rivero-Calle et al., 2015; Winter et al., 2013). Additionally, they contribute to the generation of sulfate aerosols which scatter solar radiation in the atmosphere and act as cloud condensation nuclei, enabling cloud formation; therefore, coccolithophores are responsible for both warming and cooling effects on the global climate (Charlson et al., 1987; Wang et al., 2018a, 2018b). Consequently, it is essential to study the dynamics of *E.huxleyi* blooms.~~

~~*E.huxleyi* belongs to the group of small phytoplankton with a size of a few µm, and can form extensive blooms covering more than 100,000 km^2 (Brown and Yoder, 1994; Kondrik et al., 2017; Raitsos et al., 2006). *E.huxleyi* is a temperature and salinity tolerant alga; therefore, it is distributed in the waters of the arctic and subarctic, as well as in equatorial and subtropical waters (Fernandes, 2012; Flores et al., 2010; Kondrik et al., 2017; Okada and McIntyre, 1979; Winter, 1994).~~

~~Since changes to the regional climate have influenced the ecosystems of the Arctic seas, coccolithophores, particularly *E.huxleyi,* have increasingly expanded their range into polar waters (Henson et al., 2018; Winter et al., 2013). Winter et al. (2013) suggest that this poleward expansion of *E.huxleyi* is driven by changes in water temperature, salinity, or nutrients. *E.huxleyi* blooms have high positive correlation with the following set of parameters: temperature (3-15 ˚C), high light intensity (25-150 µmol quanta m^{2} s^{-1} or 5-33 W m^{-2}) (Iglesias-Rodríguez et al., 2002) and higher N:P ratio (Lavender et al., 2008). Raitsos et al. (2006) noted that the combination of high solar radiation, increased sea surface temperature, and shallow mixed layer depth contributed to the increase in coccolithophores in the North Atlantic region. In addition, wind is one of the environmental factors that influences a bloom of coccolithophores by controlling the amount of vertical mixing in the subsurface layer; therefore, the decrease in wind stress during summer months results in a decrease in the mixed layer depth, and consequently it has a positive effect on *E.huxleyi* growth (Raitsos et al., 2006).~~

~~A detailed analysis of a wide range of *forcing factors* (FFs) controlling *E.huxleyi* blooms in arctic and subarctic seas was performed using a machine learning method (Kondrik et al., 2019). Kondrik et al. (2019) identified that sea surface temperature~~

~~and salinity, near-surface wind speed at a height of 10 m, shortwave downwelling solar radiation, and ocean surface current speed are the most important oceanographic and meteorological factors for the blooming of coccolithophores and ranked the degree of their influence for each of the Barents, Bering, Greenland, Labrador, North and Norwegian Seas. Further, Kondrik et al plan to model the future dynamics of *E.huxleyi* blooms using the statistical models developed in the course of implementing a multifactorial statistical algorithm based on machine learning techniques, described in detail in Kondrik et al. (2019). The model-selection procedure and the resultant optimum combinations of CMIP5 models presented in this paper will be used as input data for Kondrik et al to model the dynamics of blooms in future scenarios. Therefore, the main goal of our study is to validate the ability of CMIP5 climate models to reproduce regional features of the FFs and then to select the best combination of CMIP5 model ensembles to be used in the modelling of the future dynamics of *E.huxleyi* blooms in six arctic and subarctic seas.~~

~~It is well established that the ensemble averaging method can be used to reduce the errors, biases and uncertainties in the individual climate models (Flato et al., 2013; Gleckler et al., 2008; Knutti et al., 2010a, 2010b; Pierce et al., 2009; Reichler and Kim, 2008). The main recommendation from climate model developers, in case it is not possible to calibrate a model for a selected region, is to take into consideration more than one climate model (Flato et al., 2013; Gleckler et al., 2008; Knutti et al., 2010b; Pierce et al., 2009). There are two main approaches for the use of climate model ensembles: (i) use of the full-ensemble average data (Flato et al., 2013; Gleckler et al., 2008; Knutti et al., 2010b; Reichler and Kim, 2008); and (ii) selection of an ensemble of the best models from the full set of available climate models (Herger et al., 2018; Knutti et al., 2010b; Taylor, 2001). These two approaches usually give a good result. 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 a very strong smoothing of the data, the interannual variability will be poorly reproduced, and only the long-term trend of a given variable will be well captured. Therefore, we chose the second approach – selection of climate models that properly simulate the regional features (spatial distribution) of the influencing factors 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 an appropriate methodology for the selection of the best model ensembles.~~

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 ~~based~~with focus only on ~~the~~some key climatological parameters, ~~e.g.,~~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)~~, believing~~. This approach assumes that if the models skillfully reproduce these key parameters, ~~then~~ they also ~~have skill in~~must be good at reproducing the regional climate in general. Another approach, which is ~~often~~ used in this study, is to select a unique combination of models for each ~~parameter~~study variable (Agosta et al., 2015; Anav et al., 2013; Fu et al., 2013; Gleckler et al., 2008).

~~There are many publications which address the selection of climate models, including the application of a ranking method. For example, Agosta et al. (2015) ranked the CMIP5 models using only one statistical metric – a climate prediction index (ratio of~~

~~root mean square error to standard deviation of observations). Gleckler et al. (2008) evaluated climate models and ranked them analyzing the climatology of the annual cycle, inter-annual variability, and relative errors. They noted that the performance of the climate models varies for different parameters. Das et al. (2018) assessed 34 CMIP5 models using three criteria: mean seasonal cycle, temporal trends, and spatial correlation, and selected the models using a cumulative ranking approach. Fu et al. (2013) and Ruan et al. (2019)~~ 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 the creation of~~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 models using a weight criterion from 0.5 to 1.0. Ruan et al. (2019) reported that introducing multiple criteria gives less uncertainties in the models' performance in comparison with observation data~~(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 consider that applying a score-based method using multiple criteria is most appropriate for our study. Since we deal with six arctic and subarctic seas with rather different environmental conditions and a wide range of parameters, it was decided to individually analyze each sea. Moreover, we analyzed the model data only for the areas where *E.huxleyi* blooms were observed to occur, because we highly prioritized the capability of the models to properly capture the local features in the areas of the blooms.~~

~~In the following section, we give a description of the data used and methodology applied for the validation analysis and selection of the climate models based on a percentile score-based model ranking for each sea and factor. Section 3 gives a detailed example of the applied methodology and contains results and discussions of the selection of the best climate model ensemble for sea surface temperature in the Barents Sea, along with the overall ranking of models for each considered sea and factor. Analysis of the results from our percentile score-based model ranking method are given in the Supplementary material~~

~~(spatial distribution of biases and errors in trends for each sea & FF combination). And finally, Section 4 presents the conclusions.~~

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

~~As mentioned above, the FFs influencing the blooms of *E.huxleyi* in arctic and subarctic seas are:~~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). ~~For the selection of the best climate models for reproducing the regional features of the distribution of these factors, we used historical projections from Atmosphere-Ocean General Circulation Models (GCMs) that were carried out in the framework of the CMIP5 project, available from the ESGF portal (https://esgf-node.llnl.gov). On the one hand, these models have low resolution (on average it is 150 km), but on the other hand, they include both the atmospheric and oceanic components, and cover all studied regions. Whereas the regional models have high resolution of 11-25 km (e.g., CORDEX) but simulate only atmosphere or ocean separately, and do not cover all six seas within the same model run. In total, we considered 34 GCMs for the historical experiment, but the number of models available for concrete forcing factors varies. The list of climate models used and their main characteristics are presented in Table 1. The atmospheric and oceanic reanalyses data were used for the evaluation and verification of the climate model performance: (i) Era-Interim – for surface wind speed at 10 m, sea surface~~

~~temperature and shortwave downwelling solar radiation for the period from 1979 to 2005; (ii) GLORYS2V4 - for sea surface salinity and ocean surface current speed for the period from 1993 to 2005. The period for verification of the climate models was chosen based on the length of the reanalysis data and the limitations from the "historical" runs of the climate models, which usually end in 2005. We used the Era-Interim Reanalysis with the resolution 0.75° x 0.75° from the European Centre for Medium-Range Weather Forecasts (https://apps.ecmwf.int)~~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)~~. The~~ 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 ~~(Global Ocean Reanalysis and Simulation version 4) Reanalysis is available at a global scale (with resolution 1° x 1°)~~ from the European Copernicus Marine Environment Monitoring Service ~~(http://marine.copernicus.eu). Selected~~(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 ~~has~~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 ~~and Methods~~

~~Regions under the study are six arctic and subarctic seas: Barents, Bering, Greenland, Labrador, North and Norwegian. The areas where *E.huxleyi* blooms occurred in these regions presented in Fig. 1, showing the number of 8-day periods when blooms were observed. Before conducting a selection of climate models, we applied a spatial-coverage mask to confine the territories of the study seas where blooms occurred more than one 8-day period during 1998-2016. We focused only on the periods of *E.huxleyi* blooms and analysed data for a specific blooming period for each sea with respect to the seasonal distribution of the coccolithophore blooms: June-September for the Barents and Labrador seas, January-December for the Bering Sea, June-August for the Greenland Sea, May-July for the North Sea, and May-August for the Norwegian Sea (Kazakov et al., 2018). Therefore, the selection of the climate models was carried out individually for each sea.~~

~~In~~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 ~~CMIP5 climate models~~ they reproduce the regional features of ~~the FFs~~ distribution ~~they were validated by means of comparison of model simulations with the reanalysis data.~~the selected parameters/variables. The validation methodology ~~of~~for the ~~validation of~~ GCMs outputs included the analysis of the climatological-mean seasonal cycle~~ and~~, interannual variability ~~of FFs for~~and trends, and analysis of spatial distributions of climatological-mean biases and trends for selected parameters averaged over the blooming period in each sea.~~ Seasonal~~

a) *The seasonal cycle* was ~~analysed~~analyzed using the multi-year averaged monthly variables for all months of ~~the~~ year (i.e., a sample size of 12~~), but interannual variability was analysed based on monthly variables for the blooming periods only (sample size varied according to sea and FF combination, e.g., a sample size for SST in the Barents Sea was 108 – monthly variables from June to September during 1979-2005).~~). Basic statistical measures were ~~used for both analyses:~~calculated, such as the root-mean-square deviation (RMSD), the correlation coefficient ~~between GCMs and reanalysis~~ (r), ~~root mean square deviation (RMSD), and~~and the standard deviation (SD) (Fu et al., 2013; Gleckler et al., 2008; Kumar et al., 2015; Ruan et al., 2019). ~~Additionally, we calculated RMSD-observations standard deviation ratio (RSR) – one of the model evaluation statistics that weighs the simulated data against the observations~~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).~~ For the interannual variability analysis we also calculated the spatial distribution of temporal trends and spatial bias between the model data and reanalysis (Anav et al., 2013; Das et al., 2018; Fu et al., 2013; Gleckler et al., 2008; Kumar et al., 2015; Ruan et al., 2019). Further, we applied our percentile score-based model ranking method of giving a score to each statistical measure for each model. Figure 2 shows an example of this approach applied to RMSD of sea surface temperature in the Barents Sea. We divided the statistical measures into 4 groups based on the amplitude of the calculated metrics~~

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 ~~25%)~~25th percentile of the statistical metrics distribution) were ~~assigned~~given a score of 3; (ii) good models (between ~~50%~~50th and ~~25%)~~25th percentile) were ~~assigned~~given a score of 2; (iii) satisfactory models (between ~~75%~~75th and ~~50%)~~50th percentile) were ~~assigned~~given a score of 1; and (iv) unsatisfactory models (more than ~~75%)~~75th percentile) were ~~assigned~~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 ~~ranked with a score of 3, and so forth. Finally, we summed up the total score for each GCM and selected the optimal ensemble of climate models which we take to be the top 25% of GCMs ranked according to their total score. This procedure was applied to each factor and study region~~were ranked with a score of 3, and so forth.

~~3 Results and discussion~~

~~In this section, we describe the selection of the best GCM ensemble for SST in the Barents Sea as an example of our methodology. In addition, we provide final results of the model~~ For ranking ~~– the best model ensembles for each considered factor and sea. The rest of~~ models based on the obtained ~~results are summarised~~differences in the ~~supplementary material.~~ ~~Figure 3 shows two Taylor diagrams, which combine the following statistical measures: correlation coefficient (r), standard deviation (SD)~~ spatial distributions of biases and ~~root-mean-square deviation (RMSD) on one graph~~trends between model outputs and ~~show how well the simulated data fit the observed patterns (Taylor, 2001). Figure 3 illustrates that all GCMs capture the seasonal cycle (left) much better than the interannual variability (right). The obtained correlation coefficient for all models is more than 0.95 for the seasonal cycle, whereas, for the interannual variability it varies from 0.28 to 0.83. Simultaneously, the SD and RMSD have a wide spread of values in both cases – SD for the seasonal cycle varies from 0.27 to 2.67 and for the interannual variability – from 0.75 to 2.55, RMSD ranges from 0.26 to 5.15 for the seasonal cycle and from 0.98 to 7.06 for the interannual variability. The closer the model data to the x-axis, the better the correlation coefficient, and the closer the model data to the dotted line (that represents SD of the~~ reanalysis~~), the better the model reproduces the variability of the climate parameter. The closer the model data to the reanalysis point, the smaller RMSD that is represented by semicircle lines in Fig. 3. The ranking method was carried out for the correlation coefficient in the range from 0 to 1, where the group with the maximum values in the range 0.75-1 was assigned the score of 3.~~

~~Figure 4 presents the box plots of spatial distribution of SST biases in the bloom area of the Barents Sea for the blooming period (June-September) during 1979-2005. For the model ranking~~, we analysed the absolute values of ~~both~~the median ~~bias~~ 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 ~~0 to~~ -6.6 ~~K~~(model #20) to 1.5 K (model #24) among the models, whereas the amplitude bias has a wide spread of values from ~~10.8~~7.3 (model #21) to ~~19.8~~16.5 K~~. We~~ (model #3). Thus it can ~~conclude~~be concluded from Fig. ~~4~~3 that the analysis of spatial distribution of biases is very important, e.g., if we compare ~~the~~ model #2 (ACCESS1-3) with ~~the~~ 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 ~~the~~ model #3 is much higher than that for model #2. ~~After the application~~Application of the percentile score-based method~~, the model~~ to modes #2 (ACCESS1-3) ~~was included into the optimal ensemble, whereas and the model~~ and #3 (CanESM2) ~~was not included.~~

~~Box plots of the spatial distribution of annual trends of SST in the Barents Sea are shown in Fig. 5. The median for SST does not reveal any trend in the Era-Interim reanalysis, while for the models~~resulted in inclusion of the former in the best-model sub-set, whereas the latter was placed beyond it ~~varies from -0.02 K yr^{-1} to 0.18 K yr^{-1}. From Fig. 5 we can conclude that some models show significant trends. Therefore, if these models show an unrealistic trend during the historical period, then they could give higher errors in the projections for the future period. For the ranking of models, we analysed the absolute values of~~

~~differences between model and reanalysis trends, specifically median of the trends and the amplitude of the spatial variation in the trends.~~(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-~~AO; 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

~~Figure 6 shows~~The selected best CMIP5 model sub-sets for five oceanographic and meteorological variables, viz. the ~~spatial distribution~~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., ~~biases for SST~~ 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 ~~models and reanalysis in the bloom~~ 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~~ in the Barents Sea~~, 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~~ 28~~-model ensemble, selected ~~8~~best-model ~~ensemble,~~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 ~~we can see~~seen in Fig. 6a, the full 28-model ~~set~~ensemble underestimates the SST in the

~~study~~target region while the top-model, MIROC-ESM, overestimates it. The selected 8-model ensemble shows smaller biases (± 1 K) in SST for the ~~majority~~most part of the ~~bloom area~~sea. Illustrated in Fig. 5b, the ~~Barents Sea.~~

~~The~~ spatial distribution of ~~errors in~~ SST trends (the difference between ~~models~~model data and reanalysis ~~in~~data) indicates that the ~~study region is presented in Fig. 7. The~~ full 28-model ensemble overestimates the trends for the whole ~~study region~~sea (model-reanalysis errors are 0.03-0.07 K yr^{-1}), the top-model MIROC-ESM partly underestimates the SST trend, but ~~mainly~~for the larger area it reveals ~~similar to Era-Interim~~ reanalysis ~~insignificant~~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 ~~for~~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}.

~~SST variability and spatially-averaged annual trends are presented in Fig. 8. As we can see, the full 28-model ensemble notably underestimates the interannual variability of SST and~~Analysis of comparison of all selected model sub-sets (see Supplements) shows ~~a significant positive trend (0.04 K yr^{-1} – statistically significant at the level of $p < 0.05$), while the top-model MIROC-ESM overestimates SST and shows a non-significant negative trend (-0.005 K yr^{-1}). The optimal 8-model ensemble has better performance in the SST, even though it shows a slight positive trend (0.02 K yr^{-1} – statistically significant at the level of $p < 0.05$) compared with a non-significant trend in the Era-Interim reanalysis (0.002 K yr^{-1}).~~

~~The selected optimal CMIP5 model ensembles for the other seas and FFs are presented in Fig. 9. The heat map shows the final model scores, which represent the results of our percentile score-based model ranking approach. The map summarises scores for five FFs that influence blooms of *E.huxleyi* (OCS, SSS, SST, WS, and SDSR) in six arctic and subarctic seas (Barents, Bering, Greenland, Labrador, North, and Norwegian). The top 25% of GCMs were selected as the optimal model ensemble for each sea and forcing factor combination (total 30 model ensembles: six seas multiplied by five factors). From the heat map we can conclude, that there is no optimal model ensemble, or one top-model, which could properly simulate all factors over all study regions.~~ that, in~~However, some climate models show good results for many cases, e.g., ACCESS1-3; ACCESS1-0; HadGEM2-AO; HadGEM2-CC; HadGEM2-ES; GFDL-CM3; INMCM4; GISS-E2-R; GISS-E2-R-CC. The model that have higher biases across the majority of the study regions are CMCC-CM; FGOALS-g2; IPSL-CM5A-LR; IPSL-CM5A-MR; IPSL-CM5B-LR; MIROC5; MRI-ESM1.~~

~~To examine our percentile score-based model ranking method we analysed the spatial distribution of biases and errors in trends for the full-model ensemble, selected best-model ensemble, and top-model vs. reanalysis data for each sea & FF combination (see Supplements). In~~ general, the selected best-model ensemble ~~shows~~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~~. The~~ 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 ~~better~~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 ~~worse~~they are less skilful in the simulations in 13% ~~and 7% of cases than the full-model ensemble or/and top-model for biases and trends respectively~~(biases) and 7% (trends) cases.

4 Conclusions

A percentile score-based model ranking method has been presented for ~~the ~~selection of ~~the~~ optimal model ensembles~~,~~ from a total of 34 CMIP5 models, for five different climate-relevant variables ~~that have previously been identified as influencing *E.huxleyi* blooms ~~(SST, WS, SSS, OCS, SDSR) in six arctic and subarctic seas ~~(~~, viz. the Barents, Bering, Labrador, Greenland, North, and Norwegian~~).~~ seas. The ~~optimal~~best model ensembles for each ~~factor~~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 ~~factors~~parameters across all ~~of the study regions~~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 ~~optimal~~best model ensembles show quite good results with lesser biases in smaller study regions, i.e., ~~individual Arctic~~some specific arctic seas.

~~Since we plan to apply CMIP5 model projections for the modelling~~To assure best implementation of the ~~dynamics of *E.huxleyi* blooms in the future~~model selection results, it is essential to select climate models that properly simulate the spatial distribution of the ~~FFs.~~chosen variables. Therefore, we suppose that the spatial distribution of biases and trends in ~~FFs ~~the considered parameters are ~~more~~as well important ~~in ~~as other statistical metrics within the framework of the model selection procedure~~. From ~~ performed. Based on our results, we can also conclude that it is essential ~~to ~~not only to analyse spatially averaged values, but also the ~~amplitude of the ~~spatial distribution of their amplitudes.

The results of ~~examining~~examination of ~~our~~the percentile score-based model ranking method proposed in ~~the~~this paper generally ~~show~~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.~~ Due to the short sample period of reanalysis data (1979-2005), we did this evaluation without out-of-sample testing. Definitely, it is better to test any model ranking method on another historical period. It will be possible to consider the period 1950-2014 with the release of new data, e.g., CMIP6, ERA5.~~

We can conclude that ~~a~~the range of different factors ~~are~~is important for model selection, including the spatial pattern of model biases, and that the proposed methodology is ~~one~~a way ~~we could increase~~of enhancing the ~~sophistication of ~~model selection procedures ~~to give us~~sophistication that promises a better chance ~~at selecting~~to identify more skillful models for ~~those~~the features ~~in which ~~we are interested in. Thus, the proposed method can be ~~applied~~used for ~~the analysis in~~analyses to be done for other seas/regions with the purpose to evaluate ~~climate model~~the performance ~~for~~climate models in terms of various atmospheric and oceanic parameters at ~~regional~~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 ~~gratitude~~thank Dmitry Pozdnyakov and Dmitry Kondrik for the invitation to participate in the project ~~and thank him and Dmitry Kondrik for discussion~~as well as for very useful discussions of the ~~manuscript~~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 ~~(mentioned~~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.

~~Brown, C. W. and Yoder, J. A.: Coccolithophorid blooms in the global ocean, J. Geophys. Res., 99(C4), 7467, doi:10.1029/93JC02156, 1994.~~

~~Charlson, R. J., Lovelock, J. E., Andreae, M. O. and Warren, S. G.: Oceanic phytoplankton, atmospheric sulphur, cloud albedo and climate, Nature, 326, 655–661, 1987.~~

~~Das, L., Dutta, M., Mezghani, A. and Benestad, R. E.: Use of observed temperature statistics in ranking CMIP5 model performance over the Western Himalayan Region of India, Int. J. Climatol., 38(2), 554–570, doi:10.1002/joc.5193, 2018.~~

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.

~~Fernandes, M.: The Influence of Stress Conditions on Intracellular Dimethylsulphoniopropionate (DMSP) and Dimethylsulphide (DMS) Release in Emiliania huxleyi., 2012.~~

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.

~~Flores, J. A., Colmenero-Hidalgo, E., Mejía-Molina, A. E., Baumann, K. H., H., J., L. and Rodrigues, T.: Distribution of large Emiliania huxleyi in the Central and Northeast Atlantic as a tracer of surface ocean dynamics during the last 25,000 years, Mar. Micropaleontol., 76, 53–66, 2010.~~

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.

~~Henson, S. A., Cole, H. S., Hopkins, J., Martin, A. P. and Yool, A.: Detection of climate change-driven trends in phytoplankton phenology, Glob. Chang. Biol., 24(1), e101–e111, doi:10.1111/gcb.13886, 2018.~~

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, ~~2010a~~2010.

~~Knutti, R., Abramowitz, G., Collins, M., Eyring, V., Gleckler, P. J., Hewitson, B. and Mearns, L.: Good practice guidance paper on assessing and combining multi model climate projections., 2010b.~~

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.

~~Kondrik, D., Kazakov, E., Chepikova, S. and Pozdnyakov, D.: Prioritization of the vector factors controlling Emiliania huxleyi blooms in subarctic and arctic seas: A multidimensional statistical approach, Biogeosciences Discuss., 1–24, doi:10.5194/bg-2019-104, 2019.~~

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.

~~Lavender, S. J., Raitsos, D. E. and Pradhan, Y.: Variations in the Phytoplankton of the North-Eastern Atlantic Ocean: From the Irish Sea to the Bay of Biscay, in Remote Sensing of the European Seas, pp. 67–78, Springer Netherlands, Dordrecht., 2008.~~

~~Malin, G., Turner, S., Liss, P., Holligan, P. and Harbour, D.: Dimethylsulphide and dimethylsulphoniopropionate in the Northeast atlantic during the summer coccolithophore bloom, Deep Sea Res. Part I Oceanogr. Res. Pap., 40(7), 1487–1508, doi:10.1016/0967-0637(93)90125-M, 1993.~~

~~Matrai, P. A. and Keller, M. D.: Dimethylsulfide in a large-scale coccolithophore bloom in the Gulf of Maine, Cont. Shelf Res., 13(8–9), 831–843, 1993.~~

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.

~~Okada, H. and McIntyre, A.: Seasonal distribution of modern coccolithophores in the western North Atlantic Ocean, Mar. Biol., 54(4), 319–328, doi:10.1007/BF00395438, 1979.~~

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.

~~Rivero-Calle, S., Gnanadesikan, A., Castillo, C. E. Del, Balch, W. M. and Guikema, S. D.: Multidecadal increase in North Atlantic coccolithophores and the potential role of rising CO2, Science (80-.)., 350(6267), 2015.~~

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.

~~Wang, S., Maltrud, M. E., Burrows, S. M., Elliott, S. M. and Cameron-Smith, P.: Impacts of Shifts in Phytoplankton Community on Clouds and Climate via the Sulfur Cycle, Global Biogeochem. Cycles, 32(6), 1005–1026, doi:10.1029/2017GB005862, 2018a.~~

~~Wang, S., Maltrud, M., Elliott, S., Cameron-Smith, P. and Jonko, A.: Influence of dimethyl sulfide on the carbon cycle and biological production, Biogeochemistry, 138(1), 49–68, doi:10.1007/s10533-018-0430-5, 2018b.~~

~~Winter, A.: Biogeography of living coccolithophores in ocean waters, Coccolithophores, 1994.~~

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~~., Beaufort, L., Rickaby, R. E. M. and Brown, C. W~~. and -, B.: Poleward expansion of the coccolithophore Emiliania huxleyi, J. ~~Plankton Res~~Plankt., 36(2), 316–325, ~~doi:10.1093/plankt/fbt110,~~ 2013.

Table 1. CMIP5 models used for simulation of ~~forcing factors influencing E.huxleyi blooms~~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 ~~concrete factor~~respective variable are marked as "+")

Model	**ID**	**Modelling Center** (acronym, full name, and country)	**Resolution** (°lon x °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~~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 ~~trend~~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 ~~Interannual variability~~ (averaged over the territory)				Interannual variability ~~Seasonal cycle~~ (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

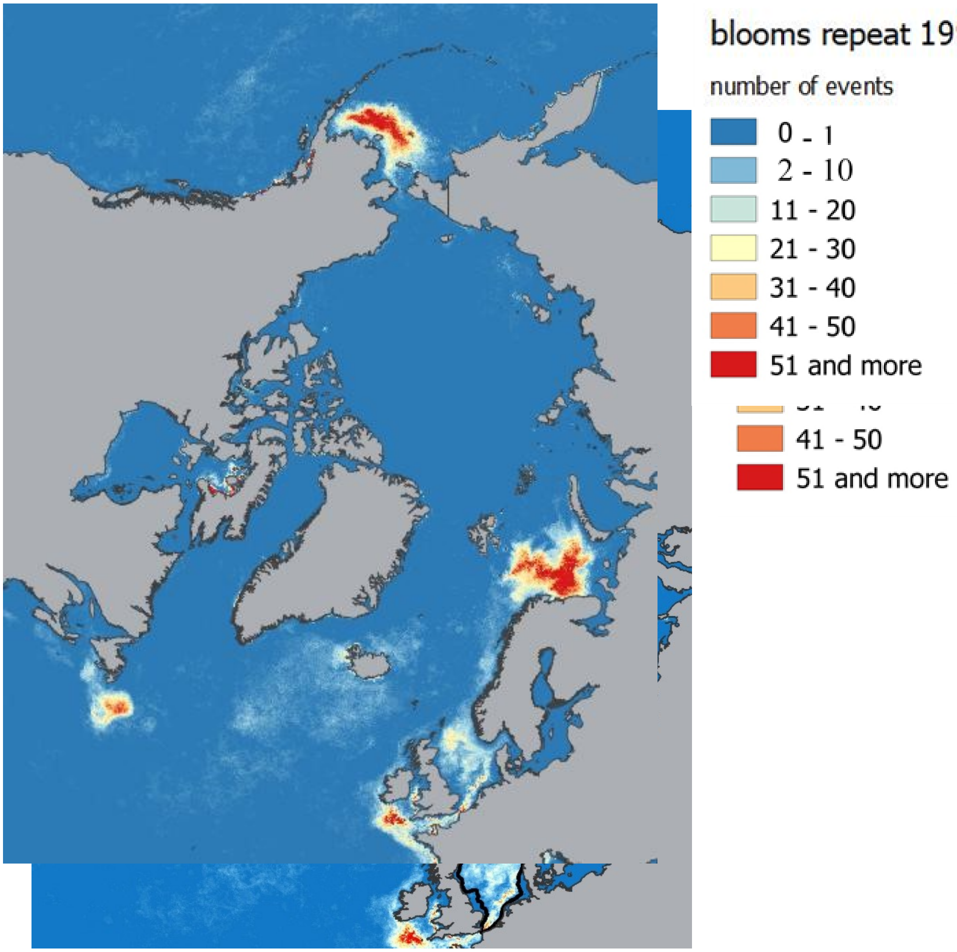


Figure 1~~**: Location of the E.huxleyi blooming areas in the study regions.**~~

: 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.**

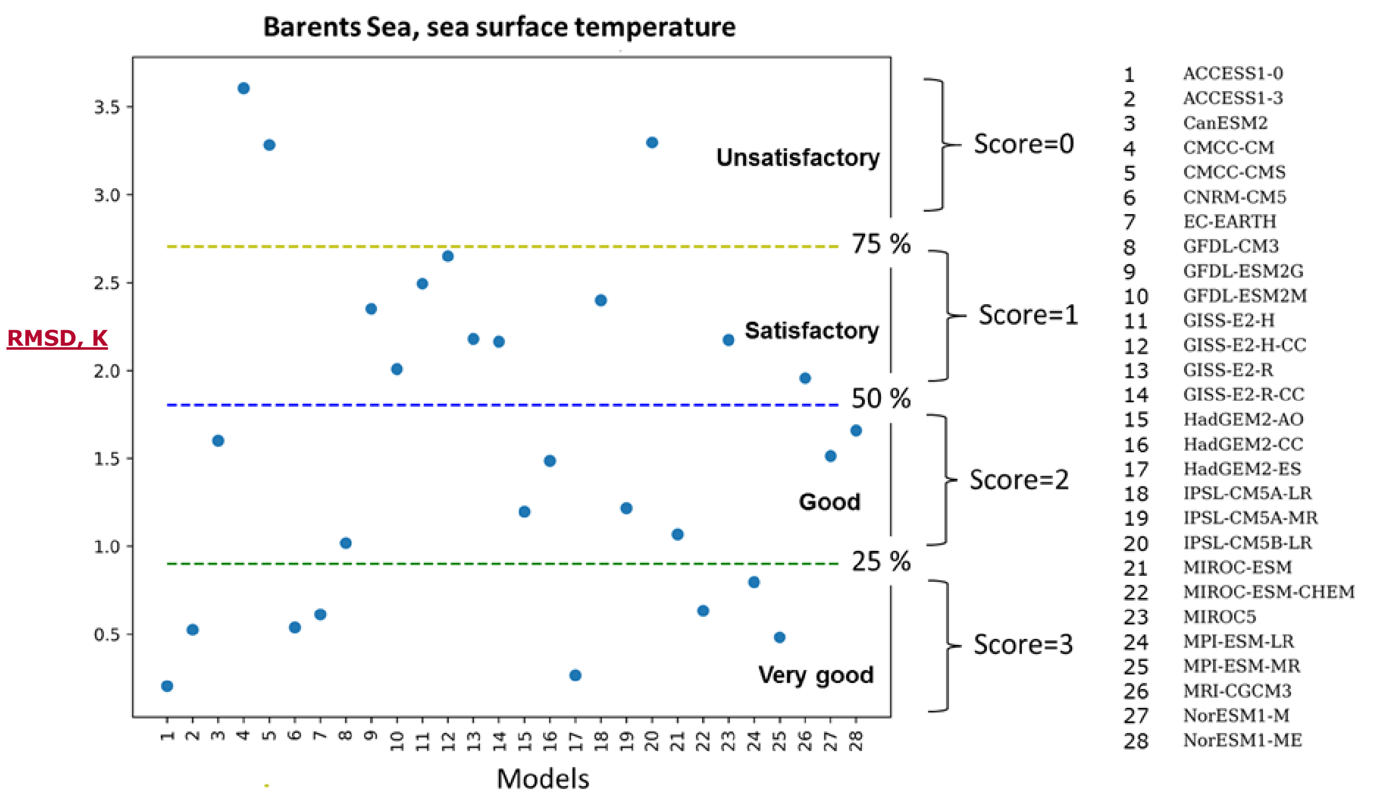
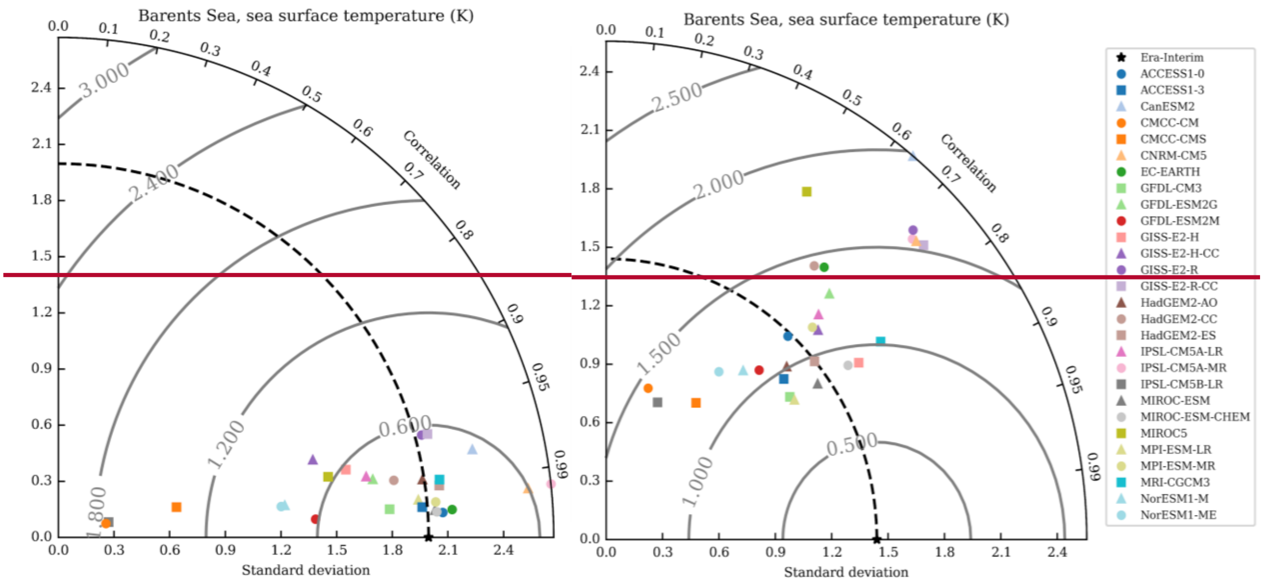


Figure 2: A schematic representation of the percentile score-based model ranking method~~.~~ (Division of RMSD values distribution of 28 models into four groups that are limited by 25th, 50th and 75th percentiles and the relative assignment of scores from 3 to 0 to each group accordingly - very good,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



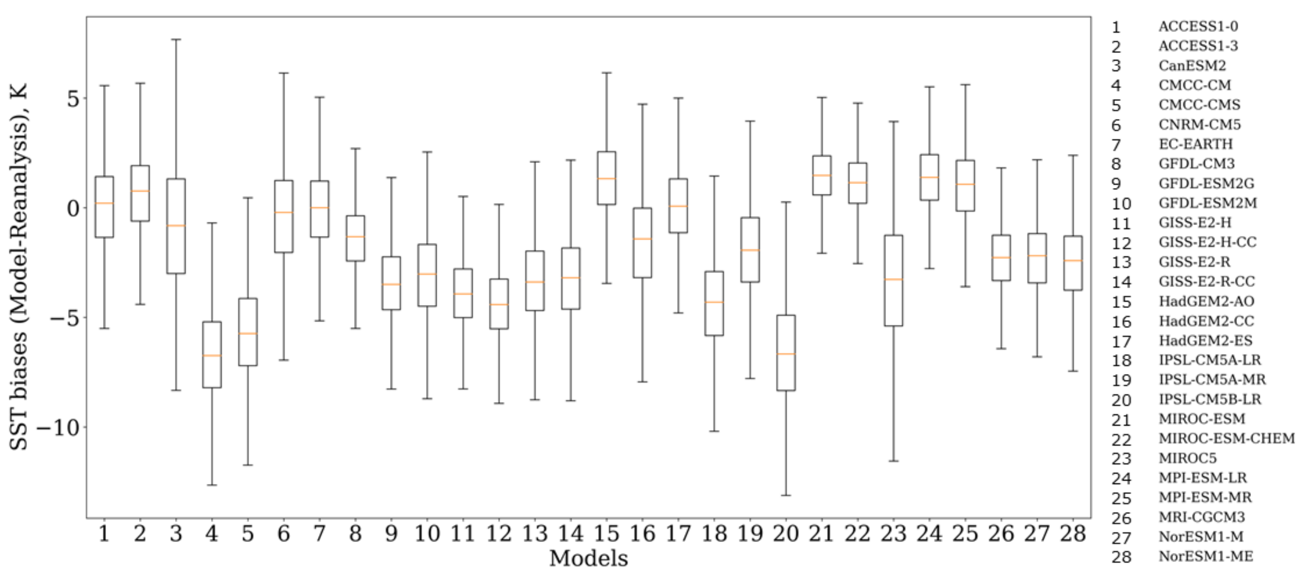
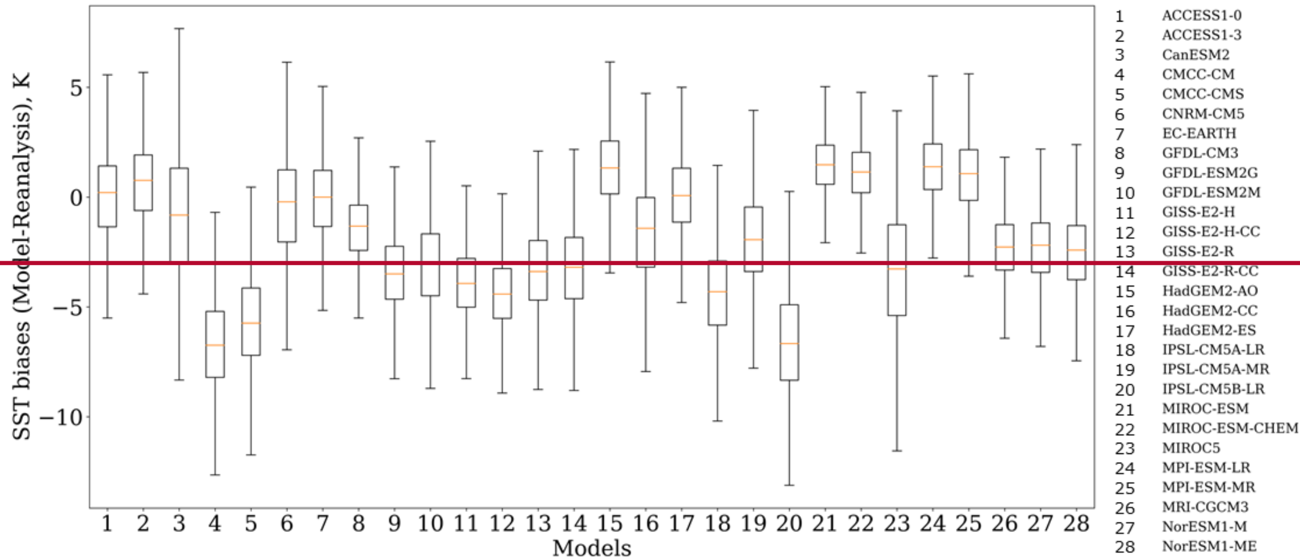
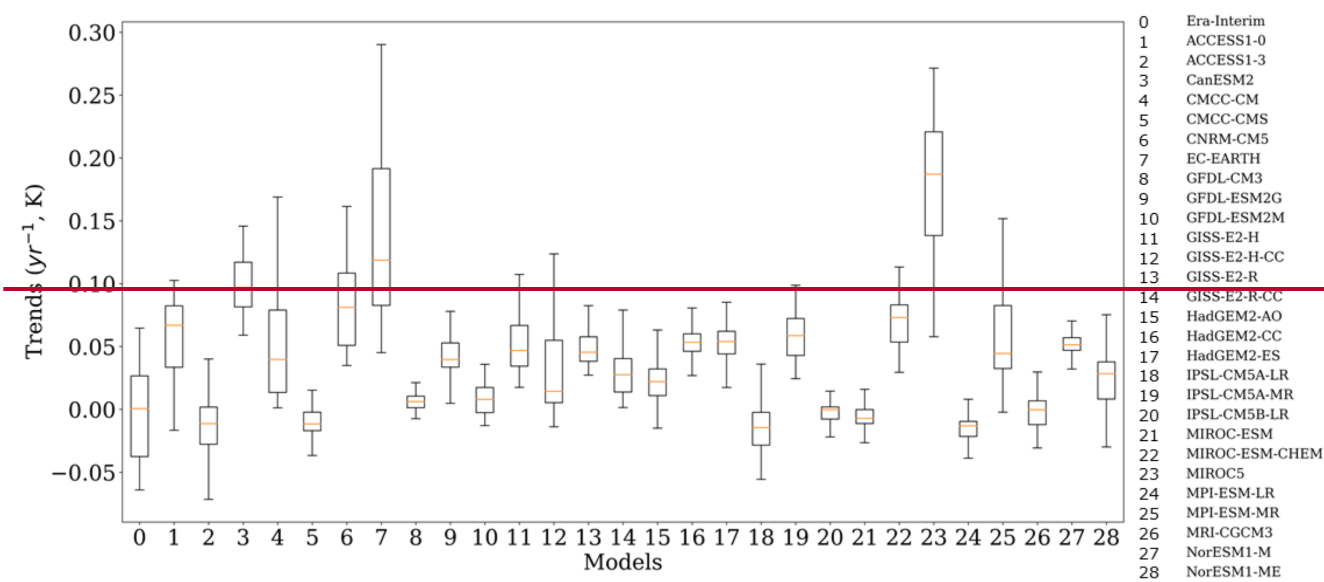


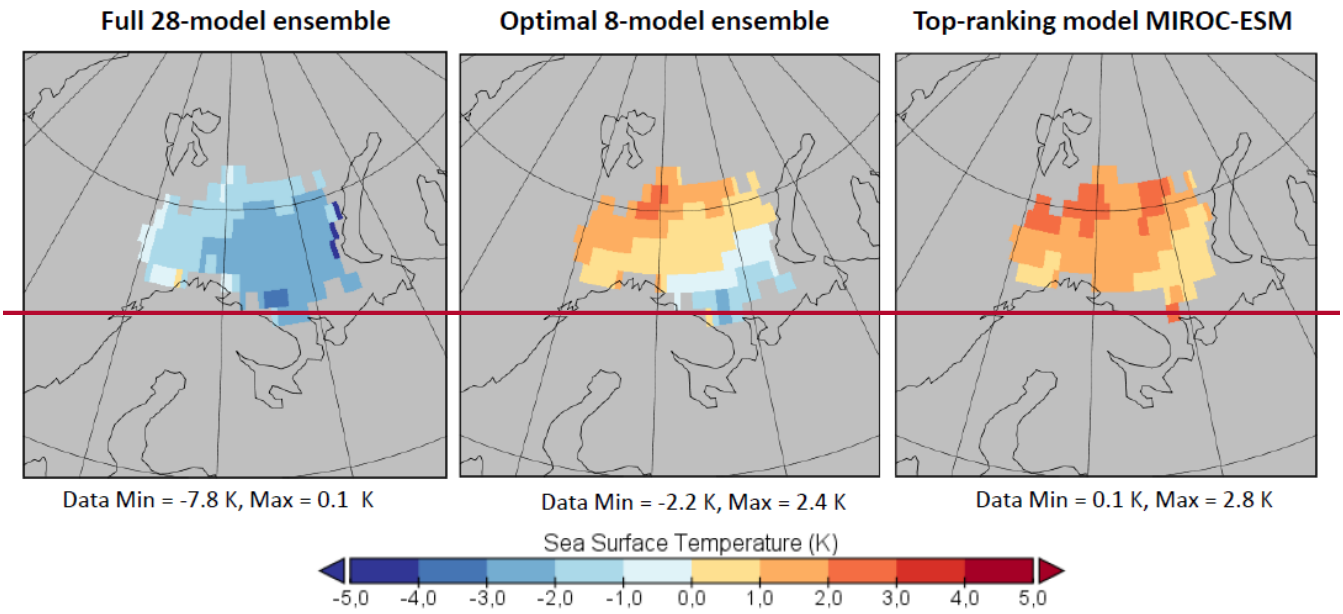
Figure 3: ~~Taylor diagrams for~~ Box plots of the ~~seasonal cycle (left) and interannual~~ spatial variability ~~(right) of SST in the Barents Sea.~~



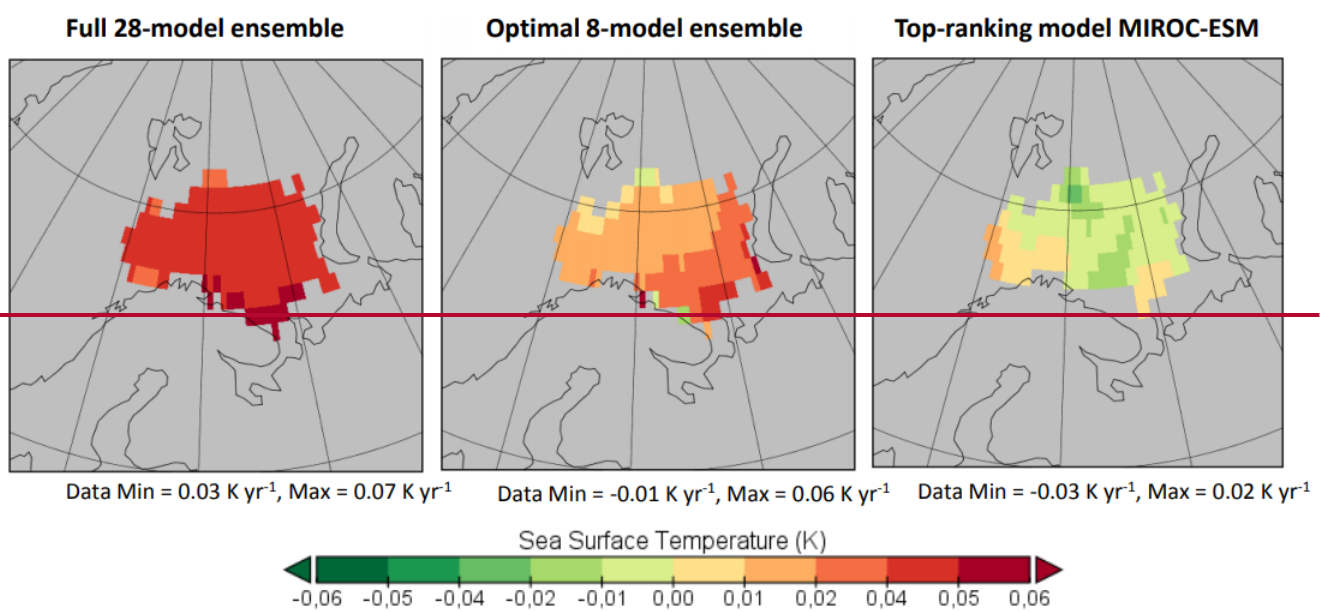
~~**Figure 4: Box plots of spatial distribution**~~ **of SST biases** ~~**in**~~ **, which are calculated as the difference between the** ~~**Barents Sea.**~~



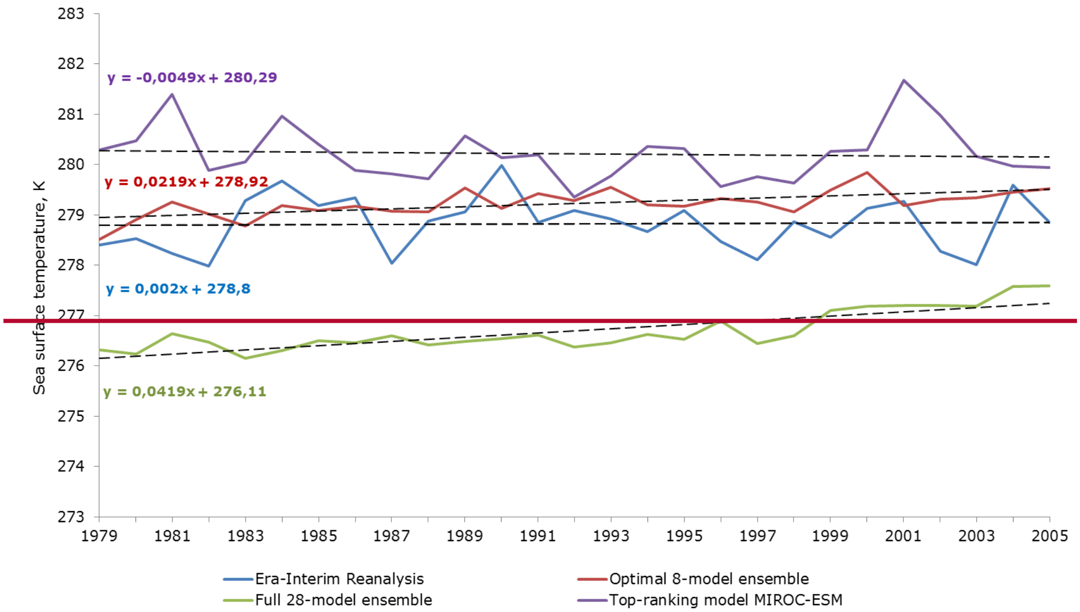
~~Figure 5: Box plots of spatial distribution of SST trends~~model and reanalysis data in the Barents Sea~~.~~



~~Figure 6: Spatial distribution of biases in SST between models~~ over the vegetation season and ~~reanalysis in the bloom area of~~ the ~~Barents Sea for the blooming~~time period~~.~~



~~Figure 7: Spatial distribution of errors in annual SST trends (model minus reanalysis) in the bloom area of the Barents Sea for the blooming period~~



~~Figure 8: SST variability and trends in the Barents Sea for the blooming period over~~ 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.

ID	CMIP5 models	Barents Sea					Bering Sea					Greenland Sea					Labrador Sea					North Sea					Norwegian Sea				
		OCS	SSS	SST	WS	SDSR	OCS	SSS	SST	WS	SDSR	OCS	SSS	SST	WS	SDSR	OCS	SSS	SST	WS	SDSR	OCS	SSS	SST	WS	SDSR	OCS	SSS	SST	WS	SDSR
1	ACCESS1-3	**23**	**34**	**33**	28	**27**	30	23	17	27	24	22	**31**	26	**29**	**31**	**27**	29	18	**30**	13	**29**	**30**	32	23	27	23	32	**36**	24	25
2	ACCESS1-0	**26**	**33**	**34**	28	**27**	27	24	26	26	**29**	18	**31**	27	27	**33**	**27**	26	22	26	20	**31**	**30**	30	23	25	**28**	31	**35**	25	24
3	CanESM2	**25**	26	24	**29**		27	24	26	14		19	15	**30**	19		16	29	**33**	9		**26**	22	**34**	18		**29**	22	**35**	21	
4	CMCC-CM	7	26	9	23	21	29	22	25	27	14	21	28	16	27	21	23	**30**	18	20	14	**27**	23	25	24	8	13	**33**	22	**30**	8
5	CMCC-CMS	16	22	14	24	23	29	23	25	28	15	**25**	**33**	**32**	22	16	**25**	**35**	15	21	15	24	19	30	25	13	24	31	**36**	**28**	14
6	CNRM-CM5	18	31	29	28	13	**31**	25	26	30	26	21	**32**	23	26	19	**29**	**30**	**30**	26	**29**	23	**31**	30	**28**	**29**	25	**34**	31	**27**	25
7	CSIRO-Mk3-6-0	20	23		19	21	21	**26**		**31**	14	20	**35**		26	10	21	27		**30**	17	23	25		24	16	19	**33**		15	13
8	EC-EARTH			27					27					**35**					28					30					**36**		
9	FGOALS-g2	17					4					8					24					11					12				
10	GFDL-CM3	20	32	**34**	27	23	**32**	20	**32**	**32**	26	19	30	**32**	21	**28**	**27**	25	25	**28**	**28**	23	19	31	**29**	22	**26**	**33**	**36**	**27**	24
11	GFDL-ESM2G	21	30	23	26	26	29	25	20	30	14	**24**	27	22	**30**	24	20	27	**29**	**27**	21	22	27	32	27	26	**26**	**33**	30	26	25
12	GFDL-ESM2M	15	**33**	26	27	25	**32**	20	24	29	20	**23**	**33**	23	23	18	**27**	**32**	24	**27**	**27**	24	18	29	**28**	28	25	**33**	33	23	27
13	GISS-E2-H	10	29	25	**29**	12	26	19	29	30	**28**	16	**32**	28	**28**	25	15	15	14	19	**28**	20	**30**	32	**28**	**31**	17	**33**	**36**	19	**34**
14	GISS-E2-H-CC	14	24	24	**30**	12	25	21	**32**	**32**	26	13	24	25	**28**	17	18	23	23	18	19	19	**31**	32	26	**29**	20	27	**35**	26	**32**
15	GISS-E2-R	19	8	20	26	12	28	25	25	**32**	**29**	**25**	29	28	**30**	22	22	26	27	26	**29**	23	28	31	**29**	**30**	23	32	33	**27**	**34**
16	GISS-E2-R-CC	20	9	22	27	11	29	**27**	28	**32**	**30**	**24**	28	26	**30**	25	22	22	**30**	**28**	**28**	22	25	30	**30**	**29**	24	**35**	29	**27**	29
17	HadCM3				16					28					25					**27**					27					19	
~~18~~	~~HadGEM2-AO~~	~~26~~	~~32~~	~~31~~	~~30~~	~~29~~	~~30~~	~~28~~	~~29~~	~~32~~	~~30~~	~~17~~	~~23~~	~~27~~	~~31~~	~~33~~	~~19~~	~~11~~	~~30~~	~~28~~	~~13~~	~~28~~	~~30~~	~~35~~	~~20~~	~~28~~	~~26~~	~~31~~	~~34~~	~~21~~	~~31~~
19	HadGEM2-CC	**22**	32	25	**30**	25	29	**26**	**32**	30	**29**	20	19	**31**	**29**	**33**	22	20	**30**	**30**	16	**29**	**31**	**33**	**28**	**31**	**27**	32	**35**	25	**32**
20	HadGEM2-ES	21	**33**	**33**	27	**30**	25	24	28	30	27	17	25	28	**28**	**33**	25	17	26	**29**	13	**28**	26	32	**29**	**30**	**28**	30	33	23	**32**
21	INMCM4				**30**	**32**				26	**32**				16	**33**				18	**30**				23	**31**				24	28
22	IPSL-CM5A-LR	18	12	22	23	**29**	30	25	**34**	27	26	18	29	25	19	25	19	**31**	23	24	26	22	12	21	13	20	17	29	28	17	25
23	IPSL-CM5A-MR	20	18	23	24	**29**	**33**	22	**32**	**31**	24	17	28	**32**	27	**27**	21	27	25	24	23	25	7	26	23	28	25	31	31	18	27
24	IPSL-CM5B-LR	11	9	11	15	**27**	**33**	**27**	22	**31**	26	15	11	12	18	13	14	21	**31**	23	19	21	13	18	14	16	12	13	25	14	22
25	MIROC4h				**32**					18					**28**					21					27					**28**	
26	MIROC5		31	14	28	22		14	16	24	**31**		**32**	**33**	28	**32**		**31**	19	21	**27**		25	20	**28**	25		24	17	25	**32**
27	MIROC-ESM		31	**35**	15	26		13	**31**	**33**	20		29	22	26	20		**30**	**29**	26	9		26	**34**	16	13		30	34	16	25
28	MIROC-ESM-CHEM		30	**34**	19	23		15	**31**	**31**	21		29	20	25	18		**34**	28	21	10		28	**34**	15	18		28	33	16	25
29	MPI-ESM-LR	21	31	**34**	25	21	**32**	**29**	24	**31**	11	12	**33**	29	21	19	16	22	21	21	10	**26**	**31**	**33**	27	19	13	31	34	**28**	23
30	MPI-ESM-MR	17	**33**	**32**	24	19	**31**	**28**	21	29	15	17	**31**	**31**	25	18	12	24	28	20	15	23	**31**	**35**	25	18	13	25	**35**	**27**	23
31	MRI-CGCM3	**26**	20	27	13	25	28	**28**	**30**	10	26	**26**	13	25	16	19	21	16	26	14	18	20	29	32	12	28	**28**	20	33	15	**33**
32	MRI-ESM1				12					9					11					14					8					16	
33	NorESM1-M		**33**	25		20		17	24		13		30	26		10		23	23		14		**30**	**34**		25		31	33		25
34	NorESM1-ME	**23**	**33**	27		23	28	23	23		15	**23**	**31**	20		14	**27**	21	28		10	25	**30**	31		28	24	**35**	32		23
	total selected models	7	7	8	7	8	7	8	8	11	8	7	11	8	10	9	7	8	8	10	8	8	11	8	9	8	8	9	10	9	8

30 - selected optimal model ensemble

- score < 25% "very good"

- 25% < score < 75% "good" & "satisfactory"

- score > 75% "unsatisfactory"

ID	CMIP5 models	Barents Sea					Bering Sea					Greenland Sea					Labrador Sea					North Sea					Norwegian Sea				
		OCS	SSS	SST	WS	SDSR	OCS	SSS	SST	WS	SDSR	OCS	SSS	SST	WS	SDSR	OCS	SSS	SST	WS	SDSR	OCS	SSS	SST	WS	SDSR	OCS	SSS	SST	WS	SDSR
1	ACCESS1-3	**23**	**34**	**33**	28	**27**	30	23	17	27	24	22	**31**	26	**29**	**31**	**27**	29	18	**30**	13	**29**	**30**	32	23	27	23	32	**36**	24	25
2	ACCESS1-0	**26**	**33**	**34**	28	**27**	27	24	26	26	**29**	18	**31**	27	27	**33**	**27**	26	22	26	20	**31**	**30**	30	23	25	**28**	31	**35**	25	24
3	CanESM2	**25**	26	24	**29**		27	24	26	14		19	15	**30**	19		16	29	**33**	9		**26**	22	**34**	18		**29**	22	**35**	21	
4	CMCC-CM	7	26	9	23	21	29	22	25	27	14	21	28	16	27	21	23	**30**	18	20	14	**27**	23	25	24	8	13	**33**	22	**30**	8
5	CMCC-CMS	16	22	14	24	23	29	23	25	28	15	**25**	**33**	**32**	22	16	**25**	**35**	15	21	15	24	19	30	25	13	24	31	**36**	**28**	14
6	CNRM-CM5	18	31	29	28	13	**31**	25	26	30	26	21	**32**	23	26	19	**29**	**30**	**30**	26	**29**	23	**31**	30	**28**	**29**	25	**34**	31	**27**	25
7	CSIRO-Mk3-6-0	20	23		19	21	21	**26**		**31**	14	20	**35**		26	10	21	27		**30**	17	23	25		24	16	19	**33**		15	13
8	EC-EARTH			27					27					**35**					28					30					**36**		
9	FGOALS-g2	17					4					8					24					11					12				
10	GFDL-CM3	20	32	**34**	27	23	**32**	20	**32**	**32**	26	19	30	**32**	21	**28**	**27**	25	25	**28**	**28**	23	19	31	**29**	22	**26**	**33**	**36**	**27**	24
11	GFDL-ESM2G	21	30	23	26	26	29	25	20	30	14	**24**	27	22	**30**	24	20	27	**29**	**27**	21	22	27	32	27	26	**26**	**33**	30	26	25
12	GFDL-ESM2M	15	**33**	26	27	25	**32**	20	24	29	20	**23**	**33**	23	23	18	**27**	**32**	24	**27**	**27**	24	18	29	**28**	28	25	**33**	33	23	27
13	GISS-E2-H	10	29	25	**29**	12	26	19	29	30	**28**	16	**32**	28	**28**	25	15	15	14	19	**28**	20	**30**	32	**28**	**31**	17	**33**	**36**	19	**34**
14	GISS-E2-H-CC	14	24	24	**30**	12	25	21	**32**	**32**	26	13	24	25	**28**	17	18	23	23	18	19	19	**31**	32	26	**29**	20	27	**35**	26	**32**
15	GISS-E2-R	19	8	20	26	12	28	25	25	**32**	**29**	**25**	29	28	**30**	22	22	26	27	26	**29**	23	28	31	**29**	**30**	23	32	33	**27**	**34**
16	GISS-E2-R-CC	20	9	22	27	11	29	**27**	28	**32**	**30**	**24**	28	26	**30**	25	22	22	**30**	**28**	**28**	22	25	30	**30**	**29**	24	**35**	29	**27**	29
17	HadCM3				16					28					25					**27**					27					19	
18	HadGEM2-AO	**26**	32	31	**30**	**29**	30	**28**	29	**32**	**30**	17	23	27	**31**	**33**	19	11	**30**	**28**	13	**28**	**30**	**35**	20	28	**26**	31	34	21	**31**
19	HadGEM2-CC	**22**	32	25	**30**	25	29	**26**	**32**	30	**29**	20	19	**31**	**29**	**33**	22	20	**30**	**30**	16	**29**	**31**	**33**	**28**	**31**	**27**	32	**35**	25	**32**
20	HadGEM2-ES	21	**33**	**33**	27	**30**	25	24	28	30	27	17	25	28	**28**	**33**	**25**	17	26	**29**	13	**28**	26	32	**29**	**30**	**28**	30	33	23	**32**
21	INMCM4				**30**	**32**				26	**32**				16	**33**				18	**30**				23	**31**				24	28
22	IPSL-CM5A-LR	18	12	22	23	**29**	30	25	**34**	27	26	18	29	25	19	25	19	**31**	23	24	26	22	12	21	13	20	17	29	28	17	25
23	IPSL-CM5A-MR	20	18	23	24	**29**	**33**	22	**32**	**31**	24	17	28	**32**	27	27	21	27	25	24	23	25	7	26	23	28	25	31	31	18	27
24	IPSL-CM5B-LR	11	9	11	15	**27**	**33**	**27**	22	**31**	26	15	11	12	18	13	14	21	**31**	23	19	21	13	18	14	16	12	13	25	14	22
25	MIROC4h				**32**					18					**28**					21					27					**28**	
26	MIROC5		31	14	28	22		14	16	24	**31**		**32**	**33**	**28**	**32**		**31**	19	21	**27**		25	20	**28**	25		24	17	25	**32**
27	MIROC-ESM		31	**35**	15	26		13	**31**	**33**	20		29	22	26	20		**30**	**29**	26	9		26	**34**	16	13		30	34	16	25
28	MIROC-ESM-CHEM		30	**34**	19	23		15	**31**	**31**	21		29	20	25	18		**34**	28	21	10		28	**34**	15	18		28	33	16	25
29	MPI-ESM-LR	21	31	**34**	25	21	**32**	**29**	24	**31**	11	12	**33**	29	21	19	16	22	21	21	10	**26**	**31**	**33**	27	19	13	31	34	**28**	23
30	MPI-ESM-MR	17	**33**	**32**	24	19	**31**	**28**	21	29	15	17	**31**	**31**	25	18	12	24	28	20	15	23	**31**	**35**	25	18	13	25	**35**	**27**	23
31	MRI-CGCM3	**26**	20	27	13	25	28	**28**	**30**	10	26	**26**	13	25	16	19	21	16	26	14	18	20	29	32	12	28	**28**	20	33	15	**33**
32	MRI-ESM1				12					9					11					14					8					16	
33	NorESM1-M		**33**	25		20		17	24		13		30	26		10		23	23		14		**30**	**34**		25		31	33		25
34	NorESM1-ME	**23**	**33**	27		23	28	23	23		15	**23**	**31**	20		14	**27**	21	28		10	25	**30**	31		28	24	**35**	32		23
	total selected models	7	7	8	7	8	7	8	8	11	8	7	11	8	11	8	8	8	8	10	8	8	11	8	9	8	8	9	10	9	8

30 - selected optimal model ensemble

- score < 25% "very good"

- 25% < score < 75% "good" & "satisfactory"

- score > 75% "unsatisfactory"

Figure ~~9~~4: Heat map with the final model scores obtained using the percentile score-based model ranking method for the five ~~forcing factors~~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.

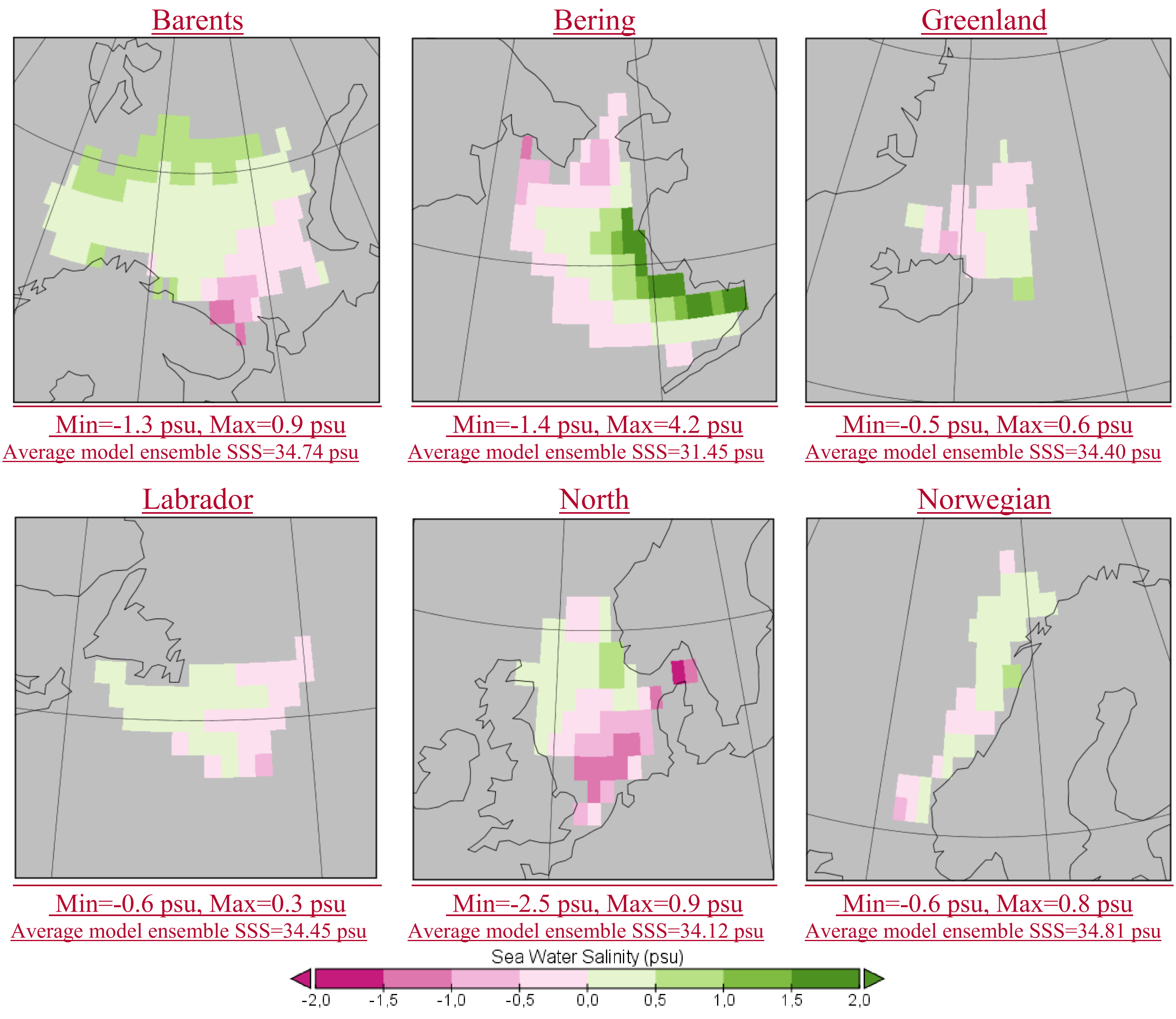


Figure 5a. Spatial distribution of biases in sea surface salinity models and reanalysis in six target seas averaged over the vegetation season and the time period 1993-2005.

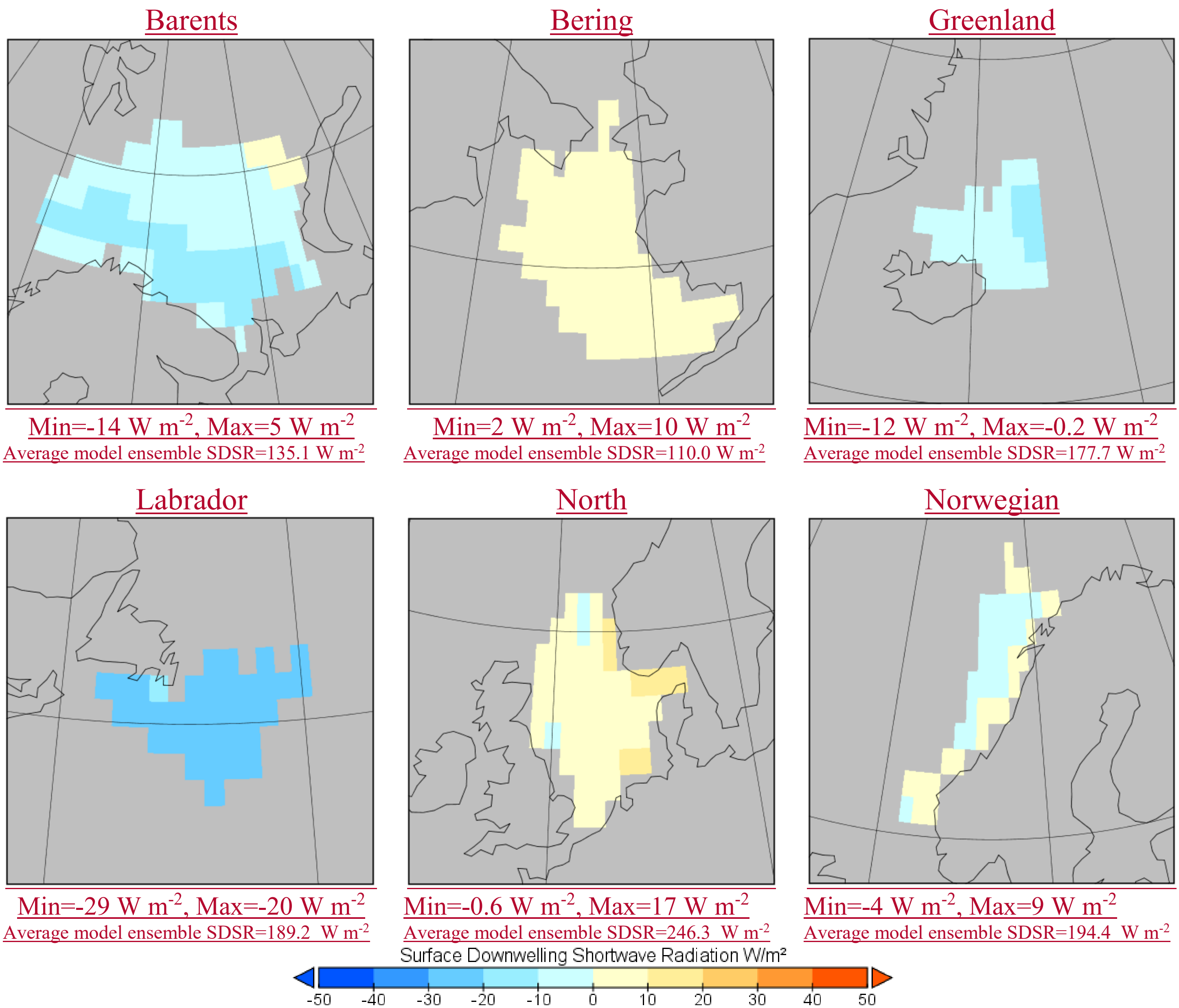


Figure 5b. Spatial distribution of biases in surface downwelling solar radiation between models and reanalysis in six target seas averaged over the vegetation season and the time period 1979-2005.

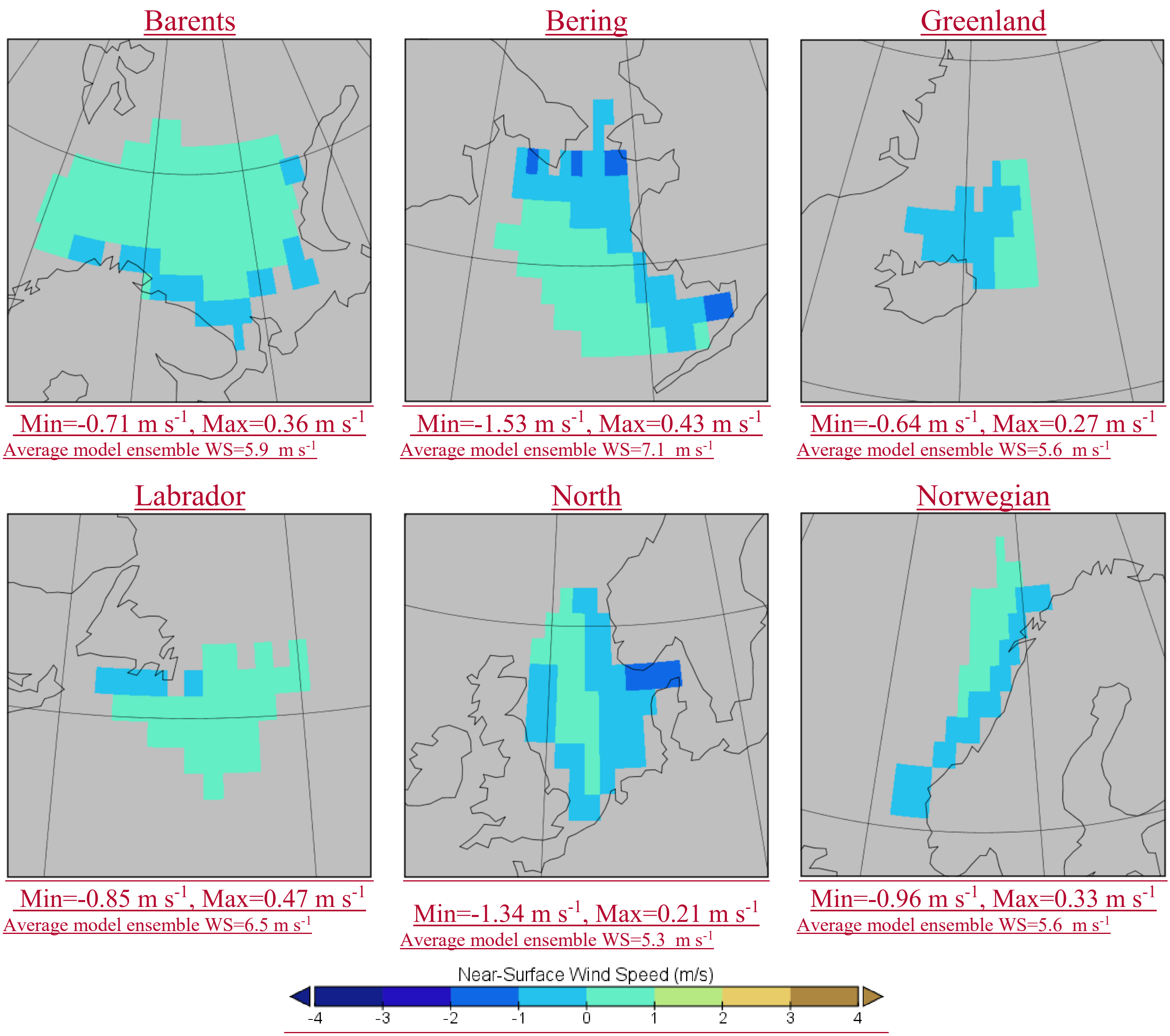
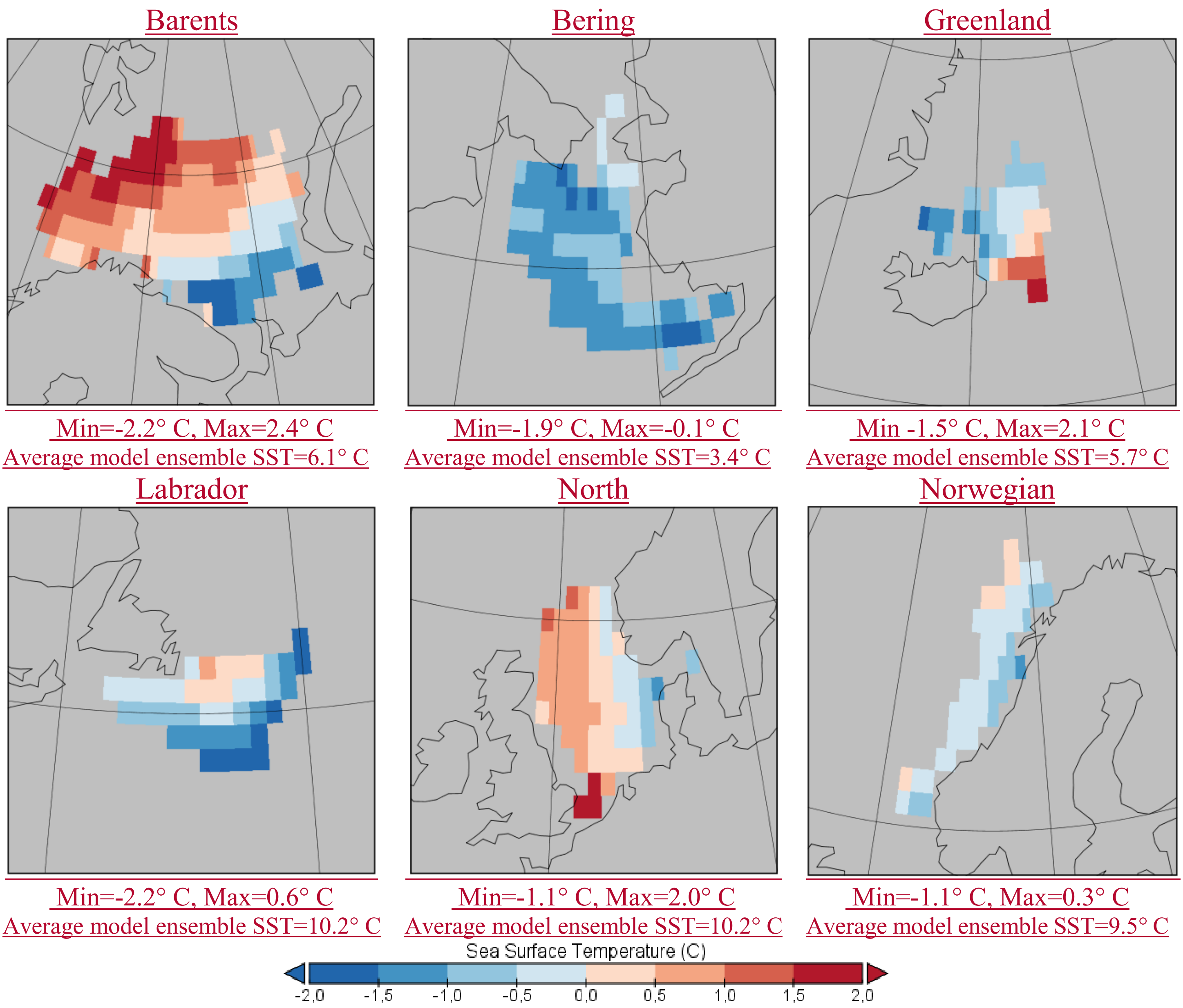

Barents
Min=-0.71 $m\ s^{-1}$, Max=0.36 $m\ s^{-1}$
Average model ensemble WS=5.9 $m\ s^{-1}$
Bering
Min=-1.53 $m\ s^{-1}$, Max=0.43 $m\ s^{-1}$
Average model ensemble WS=7.1 $m\ s^{-1}$
Greenland
Min=-0.64 $m\ s^{-1}$, Max=0.27 $m\ s^{-1}$
Average model ensemble WS=5.6 $m\ s^{-1}$
Labrador
Min=-0.85 $m\ s^{-1}$, Max=0.47 $m\ s^{-1}$
Average model ensemble WS=6.5 $m\ s^{-1}$
North
Min=-1.34 $m\ s^{-1}$, Max=0.21 $m\ s^{-1}$
Average model ensemble WS=5.3 $m\ s^{-1}$
Norwegian
Min=-0.96 $m\ s^{-1}$, Max=0.33 $m\ s^{-1}$
Average model ensemble WS=5.6 $m\ s^{-1}$
Near-Surface Wind Speed (m/s)
-4
-3
-2
-1
0
1
2
3
4

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.



5 **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.**

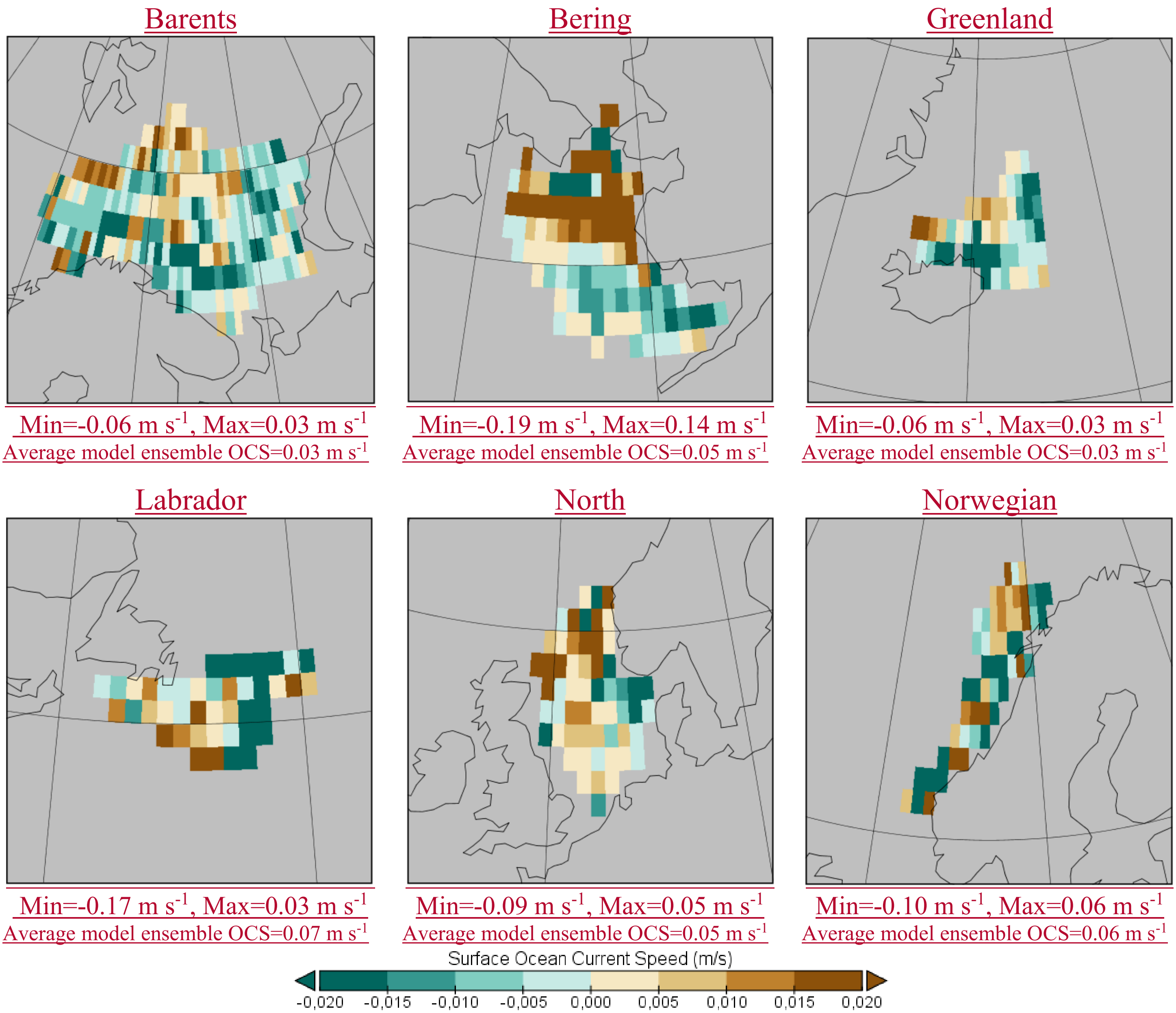


Figure 5e. Spatial distribution of biases in surface ocean current speed models and reanalysis in six target seas averaged over the vegetation season and the time period 1993-2005.

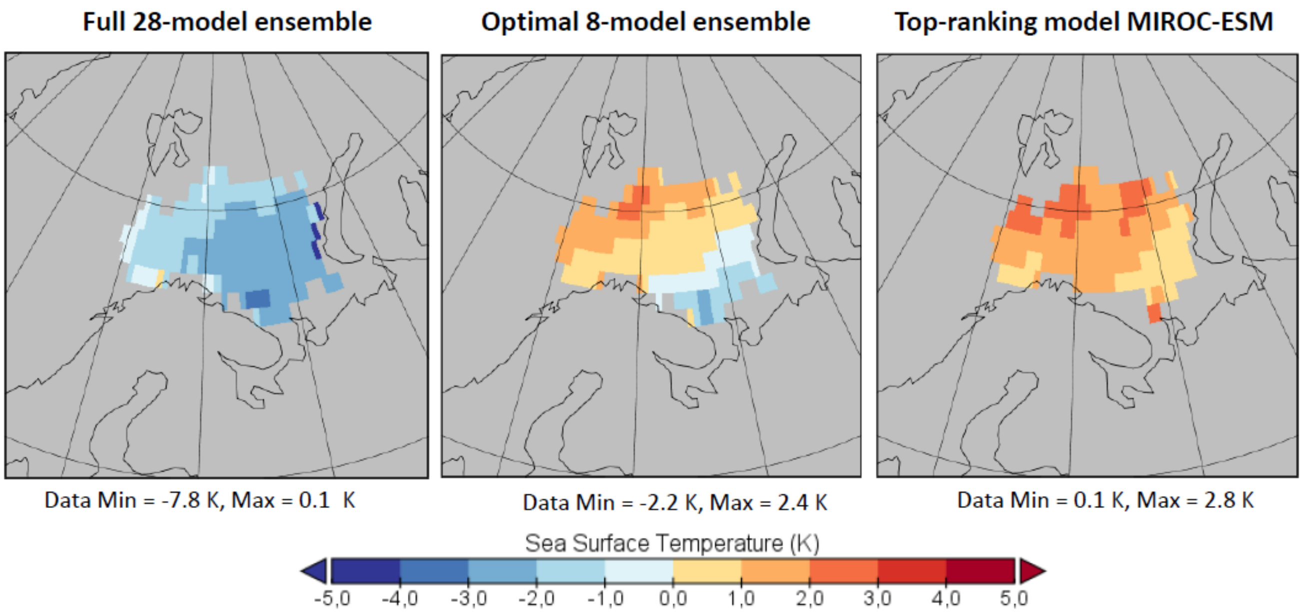


Figure 6a: Spatial distribution of biases in SST (K) between models and reanalysis data in the Barents Sea; the biases are averaged over June-September.

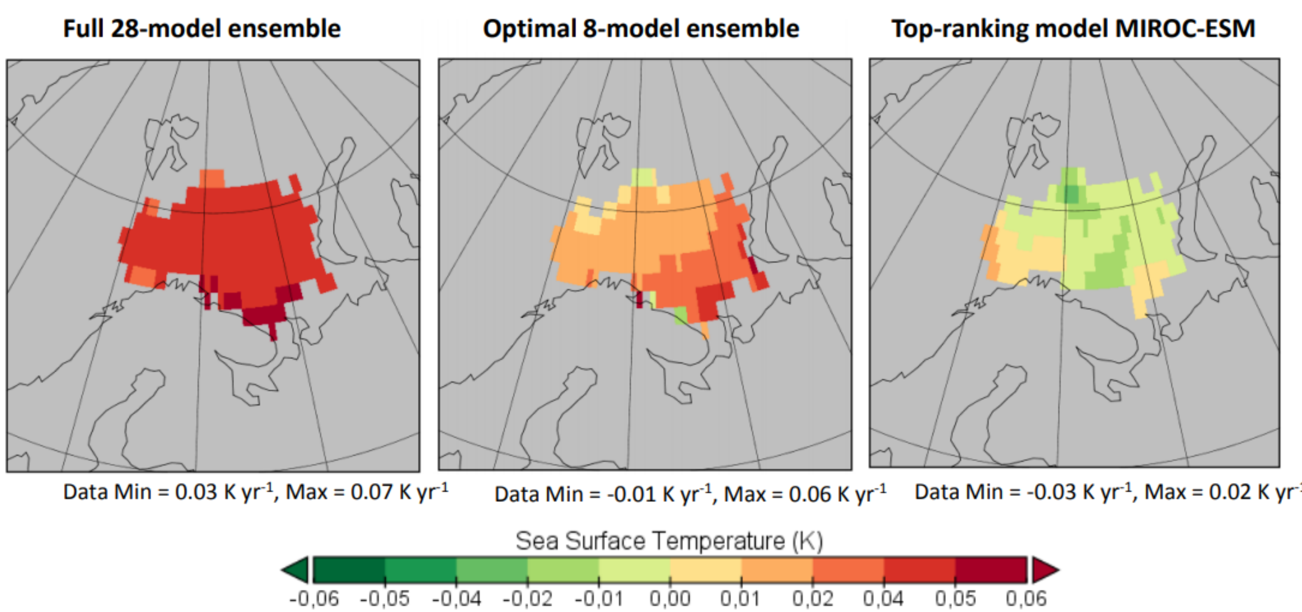


Figure 6b: Spatial distribution of errors, which are calculated as the difference between model and reanalysis values of annual SST trends (K yr^{-1}) in the Barents Sea (June-September)