Reply to reviewer 2

“The Baltic Sea model inter-comparison project BMIP - a platform for model development, evaluation, and uncertainty assessment” by Gröger et al. 2022
The manuscript presents a MIP for Baltic Sea models, which to this date has not existed before. So far, 4 models are participating, but 2 come in different resolutions which gives 6 in total. There is also a 7th model but the data availability from this model seems to be very limited as it is not presented in most plots. The models are forced by the same surface fluxes (atm. forcing and river input) and this surface forcing is presented as well. All models are compared to available observations and reanalysis products and some striking differences are found.
I think the paper is overall very interesting and well written. I would recommend it for publication after some minor comments below are addressed to make it more readable.

We thank the reviewer for his thorough reading of the manuscript and the specific comments which we will consider and which will help to improve the manuscript significantly.

Overall:

The authors rarely use the word “bias” in when discussing the differences between obs and model results. For example, line 309 says “positive anomalies in comparison with BSH climatology”, but this occurs several other times in the manuscript. I would use the word “bias” more often to make the text easier to read.

We agree and will use the word “bias” to replace the other terms.

The resolution and/or size of almost all figures (3,5,6,11,14 seem the worst) is pretty poor. Perhaps the final layout will be different, but I struggled to even find the results in some figures when on printed paper. Model names are hard to read in Fig 5, and the lack of a coastline or filled continents in Fig 11 is odd.
I would recommend making the figures larger (maybe taking up a full page), and that the authors add coastlines or filled land in Fig 11.

We agree figures have to be improved and we will definitely revise them (also in accordance with suggestions of rev#1). By it’s nature, model comparisons with many models require much place and so individual maps are somehow constrained in size. We have however ensured a quality of the pictures of 300 dpi resolution so that at least on a screen individual maps can be adequately assessed.

Detailed comments:
Page 3, Line 75: The additional reference to Myrberg 2010 is superfluous in my mind since it was given already in the previous sentence.
We agree and will remove the reference.

Page 5, line 158: I am very curious to know more details about the atmospheric forcing that is used here. I’m not familiar with UERRA. The website listed is not actually a set of instructions, but instead just a website for the project. Also, I would prefer if the authors spent some time in the main text of the manuscript to describe the data rather than point the reader to an external website. The authors should describe:

1) What the shortcomings are and what the corrections are.
2) What radiation was used? 2M temperature etc can be taken from analysis fields, but radiation and other fluxes must come from forecasts. In my experience, one would typically use the difference between the +6 and +12h forecasts of radiation, but I’d like to know what the authors do here.
3) Are any corrections needed for radiation? The commonly used DRAKKAR forcing set 5.2 (https://www.drakkar-ocean.eu/publications/reports/report_DFS5v3_April2016.pdf) had to do quite some corrections to the radiation fields. 4) Is there any effort in the data set to ensure that the surface water budget is closed, i.e. E-P-R = 0 over some time scale, or is

this done by the individual models?

General reply:
We appreciate that the reviewer shows particular interest in the applied forcing data and we are happy to share more information on it. In the article, several websites are listed that are related to the forcing data, e.g.:
- line 143/144: a link to homepage of the service
- line 155: a link to a GitHub page is given, which includes source code explaining how UERRA-HARMONIE data can be prepared for NEMO-Nordic. When following the link from line 155, please choose “create_forcing_for_NEMO”. There, you will find Python and shell scripts that were used to prepare the forcing data.
In case the reviewer is interested to learn more about the most recent regional reanalysis for Europe, the reviewer might take a look at CERRA (Copernicus European Regional ReAnalysis). CERRA was released in August 2022 and has a horizontal resolution of 5.5km and the same domain as UERRA-HARMONIE. Data are available in the CDS, here:
https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-cerra-single-levels?tab=overview

Specific replies:

Reply 1)
Major shortcomings of the dataset are related to parameters based on the forecast model only. Hence parameters, which are not assimilated. A prominent example here is the total precipitation. The total precipitation is overestimated in the UERRA-HARMONIE dataset and therefore it is reduced by 20% for BMIP. The UERRA-HARMONIE cloudiness was corrupted in the post-processing step before archiving. Unfortunately, a cloud cover of 100% was archived as cloud free (0%). Therefore, it is suggested to use coastDat-2 cloudiness in the BMIP-context. Otherwise, no corrections were made to the UERRA-HARMONIE data. This information is included in the Suppl. Mat. S6. However, we will give this information also in the main document in a revised version.

Reply 2)
Analyses are only available at 00 UTC, 06 UTC, 12 UTC and 18 UTC but the forcing frequency is hourly. Hence, data at time stamps without analyses are from the forecast model. For analyzed parameters (e.g. temperature) the forcing data is a blended set of analyses and forecasts as follows:
00 UTC (analysis), 01 UTC (forecast), 02 UTC (fc), 03 UTC (fc), 04 UTC (fc), 05 UTC (fc), 06 UTC (an), 07 UTC (fc), …
Parameters, which are not analyzed, are taken from the forecast model only.
To avoid spin-up effects after the data assimilation/analyses, longer forecasts can be used as mentioned by the reviewer. For the BMIP forcing, that is also done for precipitation. Here, we subtract the 12h forecast from the 24h forecast to avoid the model spin-up. For radiation, the BMIP forcing uses hourly data from the forecast model. The parameters “Time-intergrated surface solar radiation downwards” and “Time-intergrated surface thermal radiation downwards” were used.

Reply 3)
No correction was applied to the radiation parameters.

Reply 4)
The surface water budget in UERRA-HARMONIE is not closed. As explained above, the model precipitation is reduced by 20% and the runoff is based on observations.

Page 6, line 171: “The GETM_1nm and GETM_2nm domain is limited to the southern Kattegat” this makes it sound like the model domain only covers the Kattegat, which I’m sure it does not. The sentence should rather be “The GETM_1nm and GETM_2nm domains cover the Baltic Sea including the Kattegat while the two MOM domains also include parts of the Skagerrak. Both the NEMO and HBM domains encompass the Baltic and the North Sea, for which they also use tidal forcing on the lateral boundary condition.”

Thank you very much for this correction which we will include in the revised manuscript.

Page 8, line 256: “it is limited to regions where the coastline is mainly oriented along an east/west axis as in the Gulf of Finland”. Does this mean the method is only applicable there? I think maybe you mean that the method is most applicable when the coastline is north/south, and not so well applicable where it’s east/west?

We agree that our formulation is ambiguous. This method calculates the difference with the zonal mean, so the method is most applicable when the coastline is oriented north/south.
We will modify this sentence as follows to avoid this confusion: "This method is less reliable in regions where the coastline is mainly oriented along an east/west axis as in the Gulf of Finland".

The authors discuss the biases in upwelling along the Swedish coast (mainly meridional) and the Gulf of Riga (zonal and meridional) so if the method is less reliable for a specific direction, it could explain some of the larger biases they find.

We agree with the reviewer and this is clearly a limitation of this method. Nevertheless, as mentioned, upwelling occurs mainly along the southern Swedish coast where this limitation does not occur. For the Gulf of Riga, this could be an explanation.
We will add this sentence to discuss this point: "Along the zonal coasts, we are not able to disentangle whether the bias is due to the model or to the limitation of the upwelling detection method.

Fig 3. Perhaps the figure could be made to take up a full A4 page. It is very small and labels are difficult to read.
We agree and will try make figure 3 to fill a full A4 page.

Fig 5: This figure would also benefit from being larger.
Will be done.

Page 14, line 380: "NEMO_2nm showed that salinities were lowest in the deep later but highest in the upper layers". This makes it sound like NEMO and GETM simulate fresh bottom and salty upper ocean, which is surely not the case. I think the authors mean to say that NEMO and GETM are fresher at depth and saltier in the upper ocean compared to the other models.

Thank you for the correction! That is indeed what we meant. We will correct the sentence in the revised version.

Fig 6: This figure needs to be made larger. It is at times difficult to see the differences authors are referring to in the text.
We agree and will make the figure larger.

Fig 9: Please make the model names larger (can hardly be read when on printed A4 paper). Also please add a vertical line in the top-left subfigure to indicate in what year the reanalysis ends, and please explain this in the figure caption as well.
Thank you. We will change Fig. 9 and the caption accordingly.

Fig 11: Why is MOM_1nm and GETM_2nm not in this plot? Was the data not available? I think the authors computed the upwelling themselves using the temperature, i.e. it is not an online diagnostic, so it should be possible to do for both MOM and GETM as well. Or do those model runs, which differ only in resolution from their twins, produce the same result? I would think upwelling can be sensitive to the horizontal resolution.
Also, I would strongly recommend adding filled land or coastlines in this plot to make it easier to view.

Due to the delays in the production of the simulations there was an offset between analysis and data availability from the respective models. However, meanwhile all analysis is complete and we will complete the analysis. Hence, MOM_1nm and GETM_2nm will be included in the revised version. Also we will include coastlines in this plot.

Figs 11,12. It strikes me that NEMO simulates a very different pattern of biases, and much smaller biases in upwelling overall. I understand the authors do not want to deep dive into why this is, but I think some speculation on why NEMO is so different could be warranted. The final answer could be left for future work.

We agree and will more emphasize the low bias of NEMO and hypothesize about the reason.

Fig 13: Why is GETM_2nm not in this plot?

Will be include in the revised version. See also our reply to Fig. 11.