Reply to reviewer 1

General comments
The paper addresses relevant scientific modeling issues using a never-before-used and undeniably rigorous protocol for comparing numerical models of the Baltic Sea dynamics. This method is useful in determining the discrepancies between the models because the same forcings were applied to each of the models, thus allowing the discrepancies between the models to be interpreted as being due to the model's own configuration or grid resolution. The authors have well introduced the lack of inter-comparison studies of regional models against other (global) models and the need to do so for the Baltic Sea and the North Sea. To justify the importance of this exercise, they highlighted the complexity of the site and the diversity of dynamic models used to simulate, among others, the general circulation. A state of the art of the models used in the study area is quite complete, it is presented at the beginning of the article. The detailed method is available on the project's website and allows for identical replication of the experiments. The added value is in the potential reproducibility of the method to other marginal seas. Thus, the technical approach is clearly explained with some exceptions that will be mentioned in the "Specific Comment" section.
The overall structure of the paper and its presentation make it clear and easy to read. However, some parts of the results need revision which I detail in the "Specific Comment". In addition, many of the figures need to be reworked. However, the main messages of the publication are clear but the results lack discussion and consideration of related work.

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

Abstract
It is detailed and includes the main results. The absence of information about salinity is deplored.

That’s a good idea. We will include a short statement regarding the main salinity results.

Methods:
This part is very clear, which makes it easy to notice the few missing information.

Runoff used for this study are clearly referenced but the way they are implemented is not explained, if the runoff is added to the first mesh from the coast or diffused over several meshes, if they are applied only on the surface or on the whole water column?

The implementation is not specified in the protocol because the models manage this differently, as it is constrained by internal model configuration such as the vertical and horizontal grid spacing and options in the model code.

We will note the implementation for every model (GETM,MOM,HBM,NEMO) in the revised manuscript:

HBM, MOM: runoff is added to one coastal grid cell in the top layer
GETM: always in the top cell
discharge < 500 m^3/s - one cell discharge near the coast.
discharge < 1000 m^3/s - spread evenly over two horizontal cells near the coast
discharge > 1000 m^3/s - spread evenly over three horizontal cells near the coast
NEMO: one cell discharge near the coast. Distributed over the whole water column.

Moreover, contrary to the choice of atmospheric forcing which is justified, the choice of runoff is not explained.

As there are no homogeneous river discharge datasets for the entire period 1961-2018 available and because the last years were only covered by the E-HYPE model forecast product, we merged the two E-HYPE model datasets and the observational records. At least for the basin averages the BMIP dataset is homogeneous and consistent.

Spin-up : There are many references to model stability in the article, however, in the supplementary material there is no figure showing the stability of each of the models. Although the recommendations are clear, they are not explained. Why is it recommended to run the simulations again in July 2004 and not another month? What initialization was used to start the spin up runs?

The BMIP protocoll provides no initial data for the start of the spinup. As the Baltic Sea has an overturning time of about 30 years, BMIP gives the conservative recommendation for a 1961-2004 (44 years, thus > than the Baltic Sea overturning time) spinup to reach an equlibrium where potential drifts can be minimized. BMIP recommends to start the production runs in mid-summer as in this season Major Baltic Sea salt inflows (MBI) from the North Sea are extremely unlikely.
We will make this more clear in the revised version.

Also implied by these comments is the issue of applying the same spin-up for all models despite differences in grid resolution and turbulence schemes.

This is correct. We will include a comment that model internal turbulence schemes and resolution may influence the time the model reaches equilibrium. That's why we recommend an at least 44 year spinup duration (>overturning time) which is a compromise between costs to drive the model and the minimization of potential drifts.

We will make this more clear in a revised version.

Analysis methods : It may be of interest to indicate the error associated with the postprocessing of AVHRR data.

Thank you for your comment. To address it, we downloaded the raw AVHRR dataset and compared it with the post-processed dataset (Fig R1). This figure shows that the raw AHRR dataset underestimates the upwelling frequency by 0.9% but overestimates the spatial variability because it overestimates the frequency in Bothnian Bay. This result is consistent with the principle of post-processing as it unmasks regions misidentified by the cloud detection algorithm. We added this paragraph in the ms to discuss this point: "A comparison between raw AVHRR dataset and the post-processed dataset reveals an underestimation of annual upwelling frequency of ~1% (not shown) which is of the same magnitude order as the models error. Therefore it is important to note that in order to assess the ability of the regional model to simulate coastal upwelling, the choice of the satellite data set is crucial."

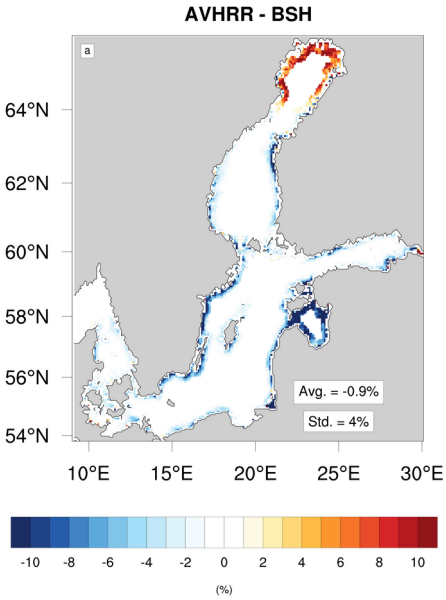

Figure R1: Difference of annual upwelling frequency between the raw AVHRR dataset and the post-processed dataset. The difference in average and standard deviation are shown in the bottom-right corner.

Specifically, the upwellings detection method used in this study is that of Lehmann et al, 2012 despite the biais from the position of the coastlines whose axis is different from the East/West axis. Why this choice of method? Why not use another method as described in Schlegetk & Smit; 2018 and Abraham, Schlegetk & Smit; 2021.

The method we chosed is easy to implement and has been tested and applied many times in the Baltic Sea (e.g. Lehmann et al., 2012; Gurova et al., 2013 Dutheil et al., 2021) contrary to the suggested methods. Nevertheless we acknowledge that the suggested method can indeed used to avoid the bias related to the orientation of the coast. However, in the original study of

Abrahams A, Schlegel RW, Smit AJ (2021) A novel approach to quantify metrics of upwelling intensity, frequency, and duration. PLoS ONE 16(7): e0254026. https://doi.org/10.1371/journal.pone.0254026

the suggested method was adapted to the coast of South Africa and requires the choice of certain thresholds and includes also the wind field evaluation. Hence, more investigation and intense analysis will be necessary to adapt this method to the Baltic Sea which is beyond the scope of this study. However, we are encouraged to to the work and adapt the method for the Baltic Sea in a followup study with specific focus on upwelling.

References

Gurova, E., Lehmann, A., Ivanov, A.: Upwelling dynamics in the Baltic Sea studied by a combined SAR/infrared satellite data and circulation model analysis, Oceanologia, 55(3), 687-707.DOI: 10.5697/oc.55-3.687

Dutheil, C., Meier, H.E.M., Gröger, M. and Boergel, Understanding past and future sea surface temperature trends in the Baltic Sea. Clim Dyn **58**, 3021–3039 (2022). https://doi.org/10.1007/s00382-021-06084-1

Results:

In the introduction of the results, it is stated that different runoffs were for the HBM model. This part should be in the material and method explaining the reason for this choice and specifying which runoff were used.

As HBM is an operational setup, it is straight forward for its implementation to utilize the respective runoff data set for this purpose. Nonetheless, the hydrological dataset is derived from the same source as for other models, i.e. E-HYPE forecasts (Donnelly et al., 2016).

Donnelly, C., Andersson, J.C., Arheimer, B., 2016. Using flow signatures and catchment similarities to evaluate the E-HYPE multi-basin model across Europe. Hydrol. Sci. J. 61 (2), 255–273. http://dx.doi.org/10.1080/02626667.2015.1027710.

We will include this note in the methods section of the revised version.

In the first part of the results the role of thermoclyne formation in the sensitivity of the SST to variations in meteorological forcing is stated but sparsely discussed. This lacks discussion and bibliographic references.

We agree. We will add a short paragraph on the role of thermocline formation and add bibliographic references.

The section dealing with seasonality needs to be restructured. Suggestion: Discuss the divergences of the models, station by station, with respect to temperature and then do the same for salinity, in the same way as the introduction to Figure 5. Indeed, the paragraphs

introducing the stations describe sometimes the variability of temperature, sometimes that of salinity.

The discussion of temperature variability for the Nemo model is missing.

Thank you for the suggestion. We will think about the structure and revise it accordingly. We will also include NEMO temperature variability in the discussion.

Long term variability: In this part we still refer to the stability of the models. It is therefore necessary to put the figures that illustrate these remarks in the publication.

Yes, the section 3.4 “Long term variability of temperature and salinity” shows deep water time series which are related also to stability. The long-term development of salinity is a good indicator for this. The salinity at the deep stations BY15 and F9 show that for all models but HBM there are no significant drifts. We will include a remark about this in the revised version. However, we want to avoid any further analysis and production of new figures on this issue as this is beyond the scope of the manuscript.

Also, in this and several times, it is referred to divergences of models because of their different management of ice modules, what about turbidity that can limit the heat flux?

It is true water turbidity and of course the individual models’ light penetration scheme will also influence the heat fluxes in addition to sea ice. We will make a remark about this in the revised manuscript.

Marine heat waves: Figure 8 with Table 1 again confirm what was explained in section 3.5 without adding additional information. It would be interesting to compare the models with the data in Figure 8 to see which model is closer to the observed extreme values and not just that the models diverge more for extreme temperatures.

Figure 8 shows the annual mean surface and bottom temperatures averaged over the entire Baltic Sea. Table 1 is an overview showing model setup characteristics.

Maybe you mean Table 2 which lists yearly mean and maximum surface and bottom temperature trends in the spatial averages over the Baltic Sea? We agree a comparison with observed extreme values would be very interesting but to our knowledge no observational data sets exist that would allow the calculation of such long term trends in spatial averages over the entire Baltic Sea and over such a long time. Thus, this would require additional intense processing of observational data to allow a reasonable comparison with the models. This work is however, beyond the scope of our study which aims to highlight model differences (and thus uncertainty) despite one and the same forcing.

Upwelling: In figure 11, the GMT_1nm model is analyzed, while in figure 12, GMT_2nm is analyzed. Why this choice and why not treat the outputs of the MOM_1nm model in the upwellings analysis?

We agree. 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 include MOM_1nm in the upwelling analysis.

Water column stratification: This section ends with “Further detailed analyses of model output may reveal the reasons underlying the difference in the timing of thermocline formation

despite identical atmospheric forcing." What do you suggest? This section should be discussed with references.

Thank you for the comment. We will include a short note on what could be investigated in further studies to elaborate on the timing of thermocline formation, such as vertical turbulence schemes, the momentum transfer from wind into the sea, or different schemes for the light penetration into the water column. We will also include references for this.

Summary :

In the conclusion, taking Hordoir et al., 2019 as an example of non-validated models in long-term simulations is not accurate because, in the first instance, the HBM model was chosen in the experiments as an example of an operational model. Furthermore, in Hordoir et al. 2019, the model is described as one that allows research on long-term simulations as much as on operational applications and whose simulations are devoid of data assimilation.

We agree. The Hordoir et al., 2019 model is validated for long term multi-decadal simulations. We will remove Hordoir et al., 2019 in this context and refer solely to HBM. Thank you for the correction.

Finally, salinity has once again been little discussed even though it is strongly impacted by runoffs, MBI…

That's true. This reflects also that salinity dynamics is very complex in the Baltic Sea. In this first BMIP introduction paper, however, we can not go to deep into the details. Definitely this interesting topic will be taken up in follow-up studies.

Technical corrections

We thank the reviewer for the technical suggestions given below to improve the figures. We will revise the figures accordingly to facilitate the interpretation for the readers. We also thank for the correction of the reference list.

Reference error. This is not an exhaustive list

Name written differently:

Meier HEM, Döscher R, Coward AC, Nycander J, DöösK: RCO—Rossby Centre regional Ocean climate model: model description (version 1.0) and first results from the hindcast period 1992/93. Reports Oceanography No. 26, SMHI, Norrköping, Sweden, p 102, 1999.

Meier, H. E. M., and S. Saraiva : Projected Oceanographical Changes in the Baltic Sea until 2100. Oxford Research Encyclopedia of Climate Science, online publication date:. DOI: 10.1093/acrefore/9780190228620.013.69, 2020.

Meier, H.E.M., Dieterich, C., Gröger, M.: Natural variability is a large source of uncertainty in future projections of hypoxia in the Baltic Sea. Commun Earth Environ 2, 50 (2021). https://www.nature.com/articles/s43247-021-00115-9, 2021a.

Listed as duplicates:

Meier, H. M., Höglund, A., Döscher, R., Andersson, H., Löptien, U., & Kjellström, E. (2011). Quality assessment of atmospheric surface fields over the Baltic Sea from an ensemble of regional climate model simulations with respect to ocean dynamics. Oceanologia, 53, 193-227

Figures

Fig.3 Use a different color palette for absolute values and differences for better readability.
Figure 3.e does not seem to have a colorbar with such a layout. Correct the extends of the colorbars that look truncated.

Fig.5 :Negative temperatures referred to in the text are not displayed on the scale
-Fig.5.a put the colorbar at the end of the figure horizontally
-Use the same width for all colorbars
-Center the station names-
Fig.5.b set the colorbar below each figure concerned and horizontally

Fig.10 : Reorganize the colorbars, the choice of palettes is not appropriate, the
Fig10.c and Fig.10.d seem to have the same color palette