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.

Specific comment

For these comments I list them in the same order as in the publication.

Abstract

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

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

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?

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.

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

Specifically, the upwellings detection method used in this study is that of Lehmann et al, 2012 despite the biai 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.

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.

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.

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.

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.

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?

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.

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?

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.

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.

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

Technical corrections

Reference error

This is not an exhaustive list

Name written differently:

Meier HEM, Döscher R, Coward AC, Nycander J, Döös K: 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