Response to Reviewer 2

Tuomas Kärnä et al.

August 18, 2021

In the following, all the comments raised by the Reviewer (in blue font) are addressed (in normal font). In the revised manuscript, all changes have been marked with red font color.

Revision of Nemo-Nordic 2.0 by Kärnä et al.

Summary:

The authors present and evaluate Nemo-Nordic 2.0, an operational marine forecast model of the Baltic Sea, which is based on Nemo-Nordic 1.0. The evaluation is sound and covers the most important aspects of the Baltic Sea physical oceanography, and the paper has overall a good structure and is in general well written. This is an important paper that documents the development of the Nemo-Nordic configuration, and deserves being published after revision.

Major comments:

-The introduction is not very coherent and needs to be revised. Maybe you could divide it into two subsections, with one about the modelling, and one more detailed about the physical oceanography of the Baltic and the North Sea? Alternatively, you could shorten the part on physical oceanography and only keep the most important aspects. At the moment there is a lot of information on the physical oceanography in there without references (lines 25-34), please add appropriate references that support your description if you want to keep this text.

We have revised the introduction. It is now shorter, focusing on numerical modeling with less discussion about the physical oceanography of the system.

-The model covers the Baltic Sea and the North Sea, but you only evaluate its performance in the Baltic Sea, and thus only a part of the model. Please describe why you do this. Still, I think that it is important, and that would be of value, if you also evaluate the model in the North Sea. If you want to focus on the Baltic Sea only in the main manuscript, you could maybe put some figures in supplementary information?

The model has been developed in the CMEMS BAL MFC project, where the goal is to model the Baltic Sea. As stated in the introduction, modeling the Baltic Sea requires a large portion of the North Sea as well due to the tight coupling of these seas. This is especially important for the modeling of tides and weather-induced water elevation variability across the Danish Straits region. We have added a comparison against 4 tide gauges in the North Sea to the SSH Taylor diagram (Fig. 3). The results show that the tides in the North Sea are captured with similar accuracy as in Skagerrak and Kattegat.

We agree that the Nemo-Nordic 2.0 model could potentially be used for operational purposes in the North Sea as well, and should therefore be thoroughly evaluated there as well. This is a laborious task, however, and out of the scope of the present paper.

-You write that in this paper you evaluate the Nemo-Nordic forecast system. But, doesn't a forecast system also include data assimilation and forecasts? (or hindcasts). Indeed, you call your simulation a hindcast, but you do not evaluate its ability to "predict" the past,i.e. for how long the model manage to reproduce observations if starting from initial conditions created with data assimilation. I do not think that you need to do this in the paper, and that it would be a paper of its own. I think that this is just a question of adding some extra text in the discussion/introduction about this, and/or revising the choice of words.

We have revised the manuscript in this regard: the focus is now on Baltic Sea modeling in general, not only on the forecast task.

-Wouldn't it be interesting to show how Nemo-Nordic 2.0 performs in comparison to Nemo-Nordic 1.0?

It would. This is, however, not straightforward as the Nemo-Nordic 2.0 configuration is a completely new setup implemented on top of the NEMO 4.0 version. Thus, we do not have comparable model configurations: The previous Nemo-Nordic 1.0 runs used different grids, forcings, time periods, and had different model parameters. In order to carry out a reliable comparison, we should essentially port the presented 2.0 configuration (bathymetry, forcings, and model parameters) back to NEMO 3.6 which implies a substantial amount of work. Despite its usefulness, we therefore chose not to pursue such a comparison.

Minor comments:

lines 4-5: the 1 nautical mile horizontal resolution is an update as well no?

Hordoir et al. (2019) present both 1 and 2 nautical mile versions of Nemo-Nordic 1.0.

lines 15-17: These models do not only simulate the circulation... maybe it is better to write: "several ocean circulation models have been set up for the Baltic Sea", or something similar

We have reformulated the sentence as suggested.

lines 84-85: you repeat "as well as" twice in one sentence, please revise

The text has been revised.

In section 2 it is not very clear what settings that are updates since Nemo-Nordic 1.0, and what you have kept the same, please clarify this.

We have added a paragraph summarizing the main differences to the beginning of Section 2.1.

line 110: do you resolve baroclinic eddies in your configuration?

Only in the larger basins, where the internal Rossby radius of deformation is approximately 3 to 7 km. Our 1.8 km grid can resolve the larger eddies in this range. Nevertheless, the numerical schemes that decrease numerical mixing in the eddying regime are likely to perform well in the case of unresolved eddies as well.

line 174: 14 months are not enough to spin up the deep Baltic Sea. What did you use to initialize the spin-up run?

The description of the spin-up run is described in more detail in Section 2.3.

line 244, and also elsewhere in the manuscript: how do you define good, relatively good and other, similar, qualitative words?

That is a good question. The notion of "good" skill in the manuscript is based on a) the previous modeling expertise of the authors, and b) the skill metrics. We do not present quantitative comparisons against previous modeling efforts, because the modeled time periods and model configurations are different and hence a direct comparison is not meaningful. In terms of the metrics, the normalized Taylor diagram gives a good overview of the statistics of the (centered) deviation. Also, as stated in Section 2.5, NRMSD is a particularly useful metric for defining "bad" skill in general terms (i.e. NRMSD¿1). But as the complex model behaviour can never be expressed with a single metric, we consider several of them simultaneously (e.g. in figs. 2-5 and 8).

figure captions and in the text discussing these figures: please describe if it is based on daily or monthly output. It should also be mentioned in the methods what time-frequency your analyses are based on. A bit in a similar manner as you do for the comparison with the Ferrybox data.

The model output was stored at 1 h resolution (now mentioned in Section 2.1). The tide gauge data was used in its original temporal resolution, i.e. 10 min, 15 min, or 1 h (added to Section 2.4). Prior to computing the error metrics, the model outputs were interpolated to the observation time stamps (as already mentioned in Section 2.4).

Could you put some lines in figure 1 showing the routes of TransPaper and Finnmaid ferries? As it is now it is difficult to relate figures 6-7 to a geographic location.

Added.

I like figure 8, but I have some questions related to it; Why did you choose the upper 10 m for the surface layer? The summer thermocline is generally located deeper. When showing the skill as you do for the surface and deep layer, it does not tell us if the bias is due an eventual mis-placement of the thermocline/halocline, or if it is the modelled temperature/salinity that is off. It would be valuable if you could evolve the text around this and discuss it, a bit like you do for the salinity at BY5.

We agree that comparing the thermo- and halocline structure between the model and observations is very valuable. However, this is a complicated task as the criteria for detecting the clines depend on the region (e.g. the Bothnian Bay and the Gotland Deep), and in many cases the water column is stratified with several possible thermo/halocline locations. The assessment, then, is highly dependent on the chosen metrics. We agree that such an analysis would be of interest, but it should be done separately for each sub-basin. This is a topic for future publication.

In the present paper, we have taken a simpler approach: The RMSD of the whole column indicates general agreement, and the top/bottom bins aim to assess the similarity of the surface/bottom waters.

figure 9 and 10: please write that the difference shows the model-observations

This was already mentioned in the caption of Figure 9. Now it reads: *[...] c) Difference (model minus observations)*. Figure 10 does not have a difference signal.

section 3.6: please write why you have chosen these specific locations for your analysis

We have added a sentence:

Figure 13 shows SSH time series at four stations in different parts of the Baltic Sea to illustrate the propagation of seiche oscillations.

figure 12: describe in the caption what the blue lines show.

We've added a sentence in the caption:

The blue vertical lines indicate events discussed in the text.

from figure 9 it looks like the model is too diffuse in the vertical. Maybe you could add this to your discussion on lines 408-416.

We have emphasized the role of vertical mixing in the discussion.

lines 429-430: do you have some references to other baltic sea operational models that you could put here?

We have added references to other models in this paragraph:

From an operational modeling perspective, the presented model configuration delivers sufficient performance, generally comparable to other models (e.g., Meier et al., 1999; Burchard et al., 2009; Dietze et al., 2014; Hordoir et al., 2019).