Answer to reviewer 3: Review of the manuscript Validation of the coupled physical-biogeochemical ocean model NEMO- SCOBI for the North Sea-Baltic Sea system, by Ruvalcaba Baroni et al, submitted to Biogeosciences

Note: Authors' answers are given in bold

Reviewer #3

Manuscript overview

This manuscript provides an extensive overview of the validation performed on a new coupled model set-up. This set-up existed for the hydrodynamic model (NEMO-Nordic), but not previously for the biogeochemical model (SCOBI). The validation therefore focusses mainly on the biogeochemical part. Model performance is not bad for such a model, and model extensions are also reported fully in the manuscript and appendix. The manuscript has no other focus than presenting the validation, and concludes this model set-up is not better or worse than others previously published by other research groups. As such, it is deemed to be a valuable addition to model ensemble studies.

Review overview

The manuscript is in general well written (some grammatical errors remain) and includes an exhaustive validation exercise. The authors focus a lot of attention on the fact that this set up includes both the wider North Sea and the Baltic Sea: most model set-ups do either one or the other due to the different governing mechanisms. For this set up the biogeochemical model was extended, from its natural domain the Baltic, to cover the North Sea and Channel areas. Therefore is seems strange that the validation is mainly focused on the Baltic and the Kattegat/Skagerrak area (for which SCOBI was designed) and hardly on the new areas it now has to represent. The authors truthfully cite a lack of observational evidence in the newly covered areas, but some station validation is surely possible. The biogeochemical model tends to capture the phosphorous and oxygen dynamics pretty well, but this is what I would expect from a model designed for and tuned to the Baltic Sea. The North Sea has very different dynamics, and although hypoxia and anoxia can occur there as well they are not a defining feature of the modern system. The authors themselves state that the model performs better in P-limited areas than N-limited areas, and the North Sea is mainly the latter. Modelled phytoplankton consists of diatoms, flagellates and cyanobacteria, but the latter hardly play a role in the salty North Sea and Channel: what groups could be added for a better representation of phytoplankton in the North Sea? Would addition of Phaeocystis spp. be an option (the North Sea nuisance species), and how good is the model at representing primary production at pycnocline depth? Figures 7 and 9 seem to indicate an overestimation of the top mixed layer depth, which should be discussed more. Overall, the model misses the seasonal dynamics of both systems (e.g. timing of spring bloom, autumn bloom), which could be due to the light climate, silica dynamics, phytoplankton parametrization or the nutrient inputs (temperature is usually easy to get right). Without additional analysis it is hard to say what is the main cause, particularly as this might differ per region. But some light attenuation validation could be added, as could a comparison with continuous Chla observations (few locations, usually short temporal coverage) or comparison with Chla satellite observations to get a better grip on this issue. I miss an in-depth analysis and discussions on these topics in the current manuscript. But the manuscript itself is worth publishing, as the model represents a valuable addition to both North Sea and Baltic modelling efforts.

We thank reviewer #3 for the interest and the thorough review of the manuscript. The reviewer pin points already here important questions that we dissect and address one by one below:

"Therefore is seems strange that the validation is mainly focused on the Baltic and the Kattegat/Skagerrak area (for which SCOBI was designed) and hardly on the new areas it now has to represent. The authors truthfully cite a lack of observational evidence in the newly covered areas, but some station validation is surely possible."

0.1) The reviewer has a good point. However, it should be noted that monitoring programs are handled very differently in the North Sea than in the Baltic Sea. In the North Sea, single stations with long time series are greatly lacking or not publicly available, mainly confined to coastal areas and have less observations per year than in the Baltic Sea. The latter means that monthly or even seasonal profiles cannot always be obtained, and therefore, plots such as 7 or 9 are not that useful for the North Sea and statistically not representative. Many observations in the North Sea come from cruises limited in time that do not always cover the same exact transect from year to year. In addition, from our point of view, we showed that the evaluation per area is also very informative for detecting regional model biases. Following earlier studies (e.g. Pätsch et al., 2017) in the North Sea, we have prioritized the spatially compiled observations instead of isolated station data. Also note that both the Kattegat and the Skagerrak are new areas with respect to models that cover either the Baltic or the North Sea. In such models, both these areas constitute the boundaries of the models, which implies having simplified dynamics in these areas (this is written in lines 579-585 and also mentioned in our conclusions, line 659). We will add more discussion on this in section 3.5, to make it more clear that both Kattegat and Skagerrak are better represented in models covering both the North Sea and the Baltic Sea. In this respect, the North Sea, even though it can now be studied with NEMO-SCOBI, is mainly used here to better capture the complex dynamics of the Skagerrak-Kattegat transition zone. As a Swedish governmental institution, covering the entire Swedish coast including both the Skagerrak and Kattegat areas is important. Expanding the model to cover the North Sea allows for better dynamics in both these complex areas. This will be better emphasize in the text. Importantly, we have analyzed results from several stations in the Skagerrak and Kattegat areas (see figure 1 in the main text), therefore covering also both of these new areas. Please also see our reply to comment 17.

The biogeochemical model tends to capture the phosphorous and oxygen dynamics pretty well, but this is what I would expect from a model designed for and tuned to the Baltic Sea. The North Sea has very different dynamics, and although hypoxia and anoxia can occur there as well they are not a defining feature of the modern system. The authors themselves state that the model performs better in P-limited areas than N-limited areas, and the North Sea is mainly the latter.

0.2) Our analysis indicates that chlorophyll-a is best captured in areas where the same nutrient is limiting in both model and observations (lines 498-503). The reviewer could have been confused by line 417 (in section 3.2): "Because phosphate is better captured in the model, chlorophyll-a is also best captured at sites where phytoplankton is limited by phosphate ...". This phrase refers to stations in the Baltic proper (and not the North Sea). However, we agree that this is confusing and misleading. Line 417 will be rephrased as follows:

"In the Baltic proper, good model skill for chlorophyll-a is found at 6 stations (i.e. BY4, BSCIII-10, HANOBUKTEN, BY10, BY15 and BY20), which also show good or acceptable model skill for both nitrate and phosphate".

0.3) It is actually the other way around. It is surprising that by expanding the model to a such different dynamic region, and with very little additional tuning, most processes in the North Sea are still well captured. This point will be better highlighted in the text. It is only specific coastal areas (mentioned in line 457) which are not well represented in NEMO-SCOBI and in other models as well. Such areas are affected by multiple uncertain factors (e.g., riverine input, non-Redfield ratios for phytoplankton growth, light penetration depth, etc). The fact that phosphorus is better captured in the model in both the North Sea and the Baltic Sea could be indeed linked to a better model representation/parametrization of the phosphorus cycle than that for the nitrogen cycle. It can also be related to a better phosphorus forcing than that for nitrogen. It is true that the oxygen and phosphorus cycle in the Baltic Sea are key study points to understand this system, but this knowledge also applies to the North Sea.

Modelled phytoplankton consists of diatoms, flagellates and cyanobacteria, but the latter hardly play a role in the salty North Sea and Channel: what groups could be added for a better representation of phytoplankton in the North Sea? Would addition of Phaeocystis spp. be an option (the North Sea nuisance species)

0.4) Yes, the cyanobacteria species included in SCOBI do not grow in the North Sea, but they do grow in other areas of the domain. As in other biochemical models covering a similar model domain (e.g., ERGOM Maar et al. (2011)), we consider diatoms and flagellates to be the dominant species in the North Sea and do not plan to add additional species for the North Sea in this version of the model. Note that the category PHY2 in SCOBI accounts for "flagellates and others" (line 126) and therefore implicitly accounts for other species such as Phaeocystis spp. However, explicitly including a phytoplankton group that reacts on changes in the N to P ratios, as seen in Phaeocystis spp in Dutch coastal waters Riegman et al. (1992), could indeed improve results in the southern coast of the North Sea. The results presented here are results from the first version of NEMO-SCOBI, which is under constant development. The reviewer suggestion will be certainly considered for future improvements.

how good is the model at representing primary production at pycnocline depth? Figures 7 and 9 seem to indicate an overestimation of the top mixed layer depth, which should be discussed more.

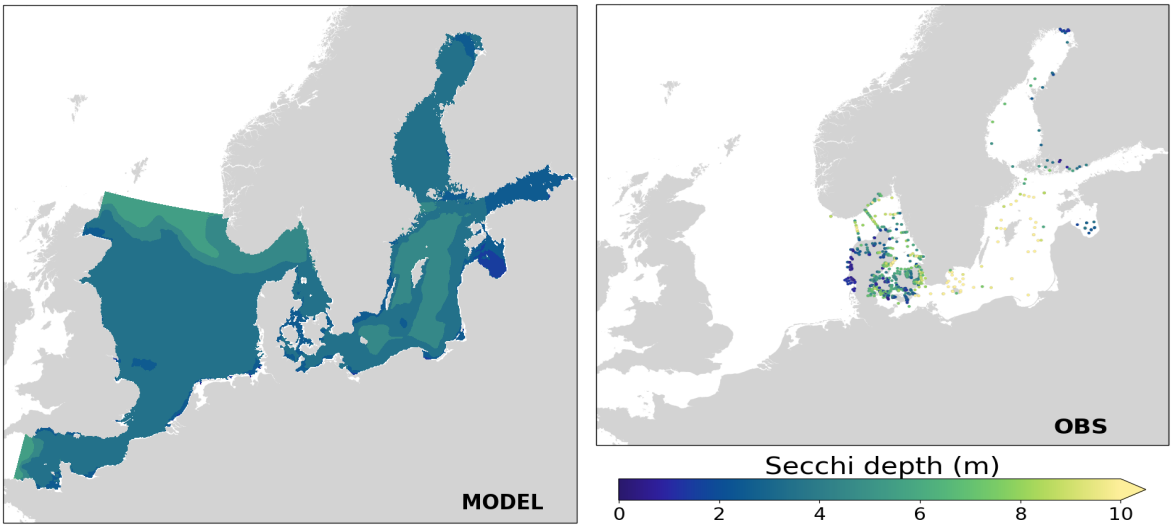
0.5) The chlorophyll-a values at the pycnocline are in very good agreement with observations (both seen in the Hovmöller diagrams and the seasonal and period averaged profiles of Figures 7, 9, B1, B2 to panels). The overestimation above the pycnocline seen in the summer months in the Hovmöller diagrams comes from the delay in the timing of the phytoplankton bloom. The maximum chlorophyll-a concentrations is generally well captured by the model at all stations, but comes too late in the year at several station in the Skagerrak-Kattegat and the Baltic proper. This is mentioned in section 3, more specifically in line 335 for the Skagerrak-Kattegat (ANHOLT) and in line 516 for the Baltic proper (BY15). Station-specific differences in chlorophyll-a are also explained in detail for ANHOLT and BY15 and discussed in lines 367-377. However, we could add a line summarizing the fact that the overestimation above the pycnocline for the summer months comes from the bloom delay in lines 570-575.

Overall, the model misses the seasonal dynamics of both systems (e.g. timing of spring bloom, autumn bloom), which could be due to the light climate, silica dynamics, phytoplankton parametrization or the nutrient inputs (temperature is usually easy to get right). Without additional analysis it is hard to say what is the main cause, particularly as this might differ per region. But some light attenuation validation could be added, as could a comparison with continuous Chla observations (few locations, usually short temporal coverage) or comparison with Chla satellite observations to get a better grip on this issue. I miss an in-depth analysis and discussions on these topics in the current manuscript. But the manuscript itself is worth publishing, as the model represents a valuable addition to both North Sea and Baltic modelling efforts.

0.6) To capture all aspects of the seasonal cycle is challenging. We have, however, shown that the seasonal cycle for all assessed parameters are well captured at least in several regions of the model (e.g. all dots falling within the inner and outer quarter circle in Figure 10 and written in lines 418-420). In the text, we do highlight the specific parameters that are not well captured for specific months and areas in order to understand what to improve in the model. We also acknowledge that the reviewer consider the manuscript worth publishing.

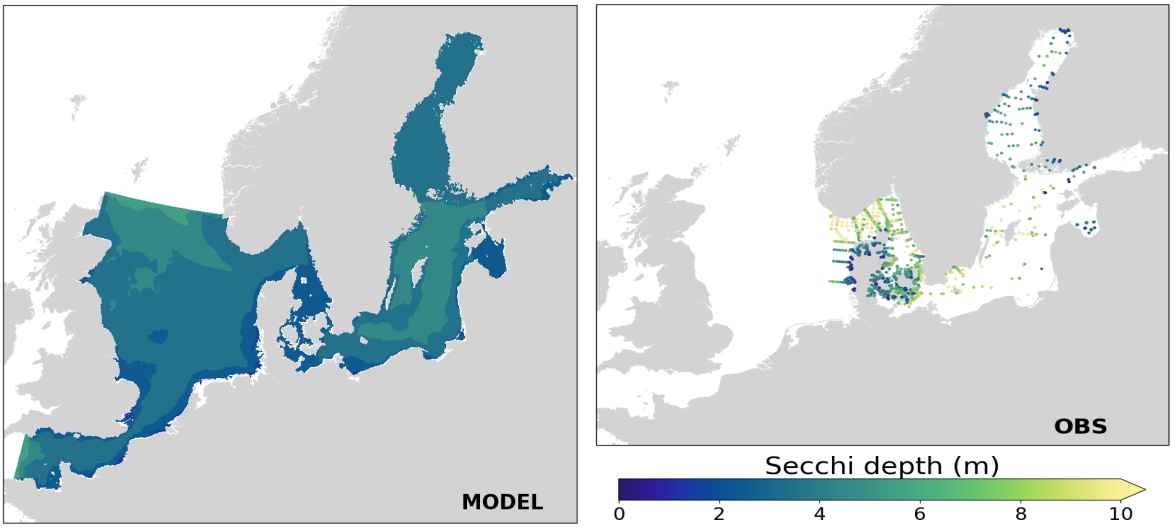
0.7) Regarding the phytoplankton bloom delay, it is mainly occurring at some stations in the Skagerrak-Kattegat transition zone and the Baltic proper, as highlighted in lines 421-424. To our knowledge, no current biogeochemical model of the Baltic Sea and/or the North Sea have good representation of chlorophyll-a everywhere in their domain. Much of the modelling community work is focusing in improving this, but it is not trivial work. NEMO-SCOBI is not an exception and therefore we are currently working, in a parallel study, in improving chlorophyll-a concentrations, where we indeed aim to include high temporal resolution data and satellite data sets. For the present study, this is not feasible, as it requires much more additional information that would make the paper indigestible. For example, satellite data should come with additional analysis on its own, especially for chlorophyll-a, and additional evaluation periods as well as additional stations would be required. Importantly, in this paper we focus on long-temporal trends. In addition, the model has been partially validated against chlorophyll-a satellite data in the North Sea in a recent study (van Leeuwen et al. (2023)). We will add the latter information.

0.8) Regarding the light attenuation validation, the reviewer is correct and this is indeed an important parameter for phytoplankton growth. In fact, as mentioned in line 562, we have analyzed the light attenuation in the model and it is overestimated in the Baltic proper and the Skagerrak-Kattegat transition zone. The figure for observations and model results averaged over 1975 to 2017 for winter and spring are shown below, as an example. We will specifically add a line summarizing this information and specify that the model overestimation of the light attenuation in the Skagerrak-Kattegat is likely linked to the delay of phytoplankton bloom in lines 574-577. Please see also our reply to reviewer #1.



(a) Modelled secchi depth in surface waters for winter months averaged over 1975 to 2017.

(b) ICES secchi depth in surface waters for winter months averaged over 1975 to 2017. Note that the scale is the same for both model and observations.



(a) Modelled secchi depth in surface waters for spring months averaged over 1975 to 2017.

(b) ICES secchi depth in surface waters for spring months averaged over 1975 to 2017. Note that the scale is the same for both model and observations.

Recommendation

Moderate revision

I like the manuscript but would like to see further validation results added to it for 1. the North Sea area and/or Channel area (main) and 2. the riverine forcing used (appendix). This would require no new simulations but new post-processing. I would also like to see figures 6 and 8 reorganized and figure B3 moved from the appendix to the main article. If necessary, figure 5 could be banned to the appendix to make way for figure B3.

0.9) In our reply to the reviewer's first comment (reply 0.1), we have detailed the reasons why we have chosen not to validate model results for the North Sea at single stations. Please note, however, that we are aware of higher temporal resolution observational data for recent years and will be included in future work. We would like to pin point again that validation for stations in both the Kattegat and Skagerrak (new areas for this model) have been done and shown (Figures 1,6,7,10 and B1).

0.10) Regarding figure B3, we are not sure that moving figure B3 to the main text will be better, as it shows similar (almost repeating) information as figure 11, due to the fact that observations are mainly confined to surface waters. However, we will consider this within the author team (please also see our more detail answer to comment 23, page 11). Regarding figures 6 and 8, please see our reply to comment 19.

0.11) We would prefer to not ban figure 5, as it clearly shows the lack of observations in the North Sea with respect to the Baltic Sea. This means that extra care must be taken when analysing trends in the North Sea, that some analysis cannot be performed and that analysis for the North Sea will not have the same statistical significance as those for the Baltic Sea. This is one of the main reasons we did not show observations for the North Sea at single stations. Regarding the river forcing, please see our more detailed answer below (comments 12 and 13).

Detailed Comments

1) Line 14-16: The validation is in agreement with assessment areas ... ? Do you mean that you are using assessment areas for the validation (i.e. method), or that validation within these assessment areas confirms with reported values in those same areas (i.e. validation result)?

This will be rephrased as follows:

Hence, these observations represent only a snapshot of nature and there are no guarantees that the measurements actually captured the chlorophyll peaks and may ...

2) Line 19: the references are not in alphabetical or chronological order.

This will be corrected to be in chronological order

3) Line 44-45: too many commas and bad grammar. Not sure what is meant here: such areas refers to the deeper parts of the North Sea (previous line), but those are not coastal.

This will be rephrased

4) Line 50: rereferred to as cyanobacteria and I miss a reference for the statement that cyanobacteria do not grow in the North Sea.

This will be rephrased (also in response to reviewer #1, comment 1), but the overall message is that the filamentous (and sometimes toxic) cyanobacteria that we parameterize in the model in the Baltic Sea do not grow in the North Sea, because the latter is too salty for this species. We will add the following reference:

Olofsson, M., Suikkanen, S., Kobos, J., Wasmund, N., Karlson, B., 2020. Basin-specific changes in filamentous cyanobacteria community composition across four decades in the Baltic Sea. Harmful Algae 91, 101685. https://doi.org/10.1016/j.hal.2019.101685

5) Line 78: bad grammar, I suggest , which is particularly true in"

This will be rephrased according to reviewer's suggestion

6) Line 80: but contain biases for

This will be corrected

7) Line 87-90: Not sure why this text is here, not relevant to the subject of this manuscript.

We agree with the reviewer that these lines are not relevant in this section. We will therefore move the sentence to the section "relevance of this study" (section 3.5).

8) Fig 1: I would say observational SHARK stations. The acronym is later used with capitals, without those it is rather confusing here. Indeed, SHARK is only explained in line 205, so some explanation is due here.

Yes, this is a typo and shark should be capitalized (also in response to reviewer 1).

9) Line 131: please refer to the figure before the textual explanation, to make it easier on the reader.

We agree with the reviewer that it would be better and will therefore move line 131 to line 121.

10) Line 148: do you mean that the phytoplankton parameters were tuned to represent both the Baltic and North Sea areas? That is to say, the parametrization the model had previously was tuned to the Baltic and these parameters did not fit with North Sea simulations and so needed adjustment? If so, can you speculate why this was necessary? What processes/groups/functionality difference is there between these areas that make this adjustment necessary?

The explanation for these adjustments are given in appendix A1 (lines 674-678):

"In the SCOBI version coupled to NEMO, rates and dependencies for phytoplankton growth were modified with respect to previous SCOBI versions in order to account for silica limitation of diatoms (not included in earlier versions), to improve the occurrence of dominant groups in both the North Sea and the Baltic Sea and to limit cyanobacteria growth in the Skagerrak-Kattegat transition zone and in the North Sea."

For more clarity, this information will be moved to the main text, instead.

11) Fig 2 : this is a spaghetti diagram, hard to read for the many arrows. I think it is a bad idea to include so much detail that the model visual abstract (which is what this is) becomes visually unattractive. I would leave out the coloured arrows explanation, readers can see for them selves if a flux stays in the pelagic or not. Or use different line styles. Maybe group it a bit more, with all pelagic nutrients together in a circle and 1 arrow going in and out if all nutrients are needed? And why are all the fluxes named in the caption rather than in a separate table?

This is a matter of taste, but we agree that to meet the color blind test, the usage of colors for the different arrows are not the best choice. We will try different line styles instead. Gathering nutrients in one circle would simplify too much the figure and hide relevant information as all nutrients do not have the exact same pathway or dependency. Regarding the fluxes, we could add a table with all fluxes but we do not think it will be relevant for the main text as it would be too much detail. It can then be added in appendix, but the fluxes still would need to be mentioned in the main text in relation with the figure. If the reviewer really wants the table, we can add it, but we think our choice to mention them in the caption, even if we agree that it is not perfect, it is still the best choice in this case.

12) Line 167: please provide a reference for the applied reduction factor, assuming this is a generally available dataset. If it is not I dont quite see why the authors would use this particular product.

The runoff data used here (as mentioned in line 165) comes from a dedicated run performed with the EHYPE model, thus especially done for this study. The EHYPE model was forced with an atmospheric database called Hydro-GFD (Berg et al., 2020), which overestimated precipitation by about 10 percent, according to a validation done by the meteorological group at our institute. Hence the river discharge correction factor of 0.9. This will be better clarified in the text. Therefore, there is no published reference for this particular EHYPE run, but it has been tested against standard check ups within the EHYPE team before delivering it for this work. It also accounts for more realistic number of river outlets than those in observed runoff data sets., which will also be emphasized in the text. In addition, the EHYPE model Donnelly et al. (2009), is a hydrological model that has been well validated and broadly used in many different settings (e.g. Donnelly et al., 2016; Hundecha et al., 2016; Nijzink et al., 2018; Macian-Sorribes et al., 2020. In fact, together with LISFLOOD and VIC, EHYPE is one of the most popular models applied for large-regional scale Piniewski et al. (2022)).

13) I am getting a bit confused about the riverine forcing applied. Am I correct in thinking that you used

- discharge values calculated by a hydrological model, which were adjusted evenly across the domain for a known, uneven model error?

Yes. A constant reduction factor was applied over the entire model domain, as mentioned in our reply to comment 12.

- nutrient values based on the same hydrological model, but adjusted for each year and basin to observational values based on two different observational data sets?

Yes, one data set is used for the Baltic Sea Gustafsson et al. (2012) and the other for the North Sea Lenhart et al. (2010). Note that the ICG-EMO data, even if it contains data for the Baltic Sea, it neglects important river outlets in this sea.

If so, then I think it unlikely that any modeller could replicate your efforts, as they cannot replicate this forcing set. And it makes me wonder why the observational sets were not used directly. This mix up of 3 different sources complicates interpretation of results, which are reported in eutrophication-relevant variables. Please provide more detail on this forcing set (in an appendix), as well as a comparison for a few selected rivers (e.g. some of the larger dots in figure 3) of the applied discharge and nutrient loads compared to the observations that you state you also have. Can some of your mismatches in coastal zones be related to this forcing data?

This is indeed a large part of the work we did. It remains, however, a side part for this manuscript. There are many possible ways to compile river forcing and we selected the method that we think best fits our purpose.

The EHYPE data set covers all relevant rivers and models physically consistent and well calibrated changes in river input according to meteorological forcing. Therefore, as written in lines 188-190, captures well the interannual variability (both for runoff and nutrients). However, it does not yet account for the anthropogenic effect on nutrients (detailed in lines 189-191). The latter is extremely relevant for our study or else our model results for nutrients will be meaningless. Importantly, the ICG-EMO and the Baltic Sea riverine data sets are not directly applied in this study because the EHYPE data includes much more river outlets than in both of these data sets combined (636 *vs* 208 outlet points, respectively). Another relevant point, is that the ICG-EMO and the Baltic Sea riverine data sets are not pure observations per say. They have been derived from observations, meaning that several assumption had to be applied to those data sets. Real observations for nutrients (for example those reported per country) lack not only spacial resolution, but also temporal resolution (mostly available per year). Also, the Baltic Sea data (Gustafsson et al., 2012) provides monthly loads per basin, which need to be redistributed in the different rivers of the corresponding basin in the Baltic Sea. Thus, both the ICG-EMO and the Baltic Sea data sets come with large uncertainties. We will add information on this in the main text.

Regarding the effect of the river forcing on our model results, the nutrient trends in the model closely follow those seen in the riverine data set in many regions of the model domain, especially in the Skagerrak-Kattegat area and coastal areas. Thus, improving the riverine data set will certainly lead to improved model results. In fact, improving the riverine data sets in both the Baltic Sea and the North Sea is an active field of research. For example, the ICG-EMO data set is constantly being updated and discussed within a large group of experts from which we are part of (see acknowledgements). More information of the importance of the river forcing will be added in the main text. Note, however, that we will not necessarily show results for one or few specific rivers, as this implies validating riverine data sets which is out of the scope of this manuscript.

Currently, the riverine data applied in this study is available on demand but we will look for a place to make it freely downloadable, transparent and citable. Note that, due to its high spatio-temporal resolution, our riverine data set is quite heavy, even if provided in netCDF files and it

may not be possible to store it in an open platform. If this is the case, we will add that the dataset is available on demand and provided with a full description.

The reviewer has a good point regarding the re-productivity of our riverine data set and, as also mentioned in comment 12, we will look for a platform that allows for large data sets to be stored, explained, freely downloadable and citable. Note that due to high spatio-temporal resolution this data set is quite heavy, even if provided as netCDF files. If storage in an open platform is not possible, we will add that the dataset is available on demand and provided with a full description.

14) Line 233: the reference year was chosen because of high nutrient values. Where those simulated values or observational values?

The period was chosen when nutrients where high in both the model and the observations. This will be clarified in the text.

15) Line 245: explanation of the applied seasonal delineation (meteorological? astronomical?) is only given in the caption of figure 10. Please provide this here.

We will add this information in line 245 as well.

16) Line 280: this has been stated before.

We understand that the phrasing is similar, however, line 255 refers to the fact that we do not show areas with less than 100 observations in figure 5. In line 280, we refer to the method for the evaluation of the model, where we do not evaluate such areas. Thus, the information, even if similar, do not refer to the exact same thing.

17) Section 3.1: the authors should include a North Sea station here, maybe one on the Danish or Dutch transects or an individual research station from the UK. It may not have everything the authors want but an extension of the SCOBI model into the North Sea and Channel areas should be validated in detail there. Stations like the Oystergrounds (NL), West Gabbard (UK) or L4 (UK) spring to mind, though the latter is I think just outside of the domain. These may have standard surface monitoring and limited at depth monitoring, but it is better than nothing. They also have high resolution observational data for a few years, generally. In the very least a North Sea station comparison will provide more detail on the local Chl-a seasonal signal and bloom timing capacity of the model there (difficult to derive from figure 12).

We thank the reviewer for providing recommendations of further stations. In our study we used the ICES data, which should include all publicly available data for the North Sea. Figure 12 together with figure B4 already show that the general spatial pattern of the phytoplankton bloom in the North Sea is actually well captured, but that the model mainly underestimates it in the southern coast of the North Sea in winter, spring and autumn, while in summer it overestimates it. They also give information of the timing of the bloom and show that chlorophyll-a values in the model do increase when observation values also increase in most areas in the North Sea, except at very coastal areas during winter and autumn. This indicates that the timing is generally well captured in the North Sea, except for these very coastal areas (so a very small part of the domain) where the blooming lasts longer in observations than in the model. We therefore argue that single stations in the North Sea will not provide much more new information in terms of long-term

trends and general patterns. However, we will further look into specific stations, keeping in mind that this is an ocean model and not a coastal model, and add text if new relevant information is found.

Regarding the high resolution chlorophyll-a, we are aware that these exist for recent years (this is also the case for the Baltic Sea), but have chosen not to include them in this particular study as they do not cover the complete studied period (2001 to 2017). Again, the focus of this work is to validate the model performance on long-temporal trends and general biogeochemical patterns, which we have explained in detail in the text. We agree with the reviewer that high resolution data is extremely valuable and relevant for model improvements and we are currently including such observations (as well as satellite data) for further analysis. However, we think it is too much information to add in this study, as such observations should come with additional detail description. Please also see our reply above (page 4, reply 0.7).

18) Line 315-317: can you provide an overview of the trends in table form in the appendix? Now it is hard to see and compare trends.

We will summarize the trends shown in figures 6 and 8 in a readable table and add it where relevant.

19) Figures 6 and 8: I would suggest restructuring these. A label over results in a graph is a no-go, in any case. Suggestion: make a two column graph (which these are already). Top left: the legend for surface values. Rest of left: surface graphs for T, S, NO3, PO4, Chl-a. Top right: legend for bottom values. Rest of right: bottom graphs for T, S, NO3, PO4, O2. The legend for O2 can be removed and explained in the caption. This would make the graph more accessible as surface or bottom processes can be viewed at a glance (vertically) while top and bottom values can still be compared easily (horizontally).

The figures will be improved (also in reply to reviewer # 2, comment 14).

20) Line 330-333: no guarantees that the measurements did not fail to, the double negative here makes this sentence hard to read. I presume your point is that observational evidence is discrete in time and so can easily miss the peak of the spring bloom. This is a very valid and important point to make, which merits unambiguous text.

The sentence will be rephrased as mentioned in our reply to comment 1:

"no guarantees that the measurements actually captured the chlorophyll peaks..."

21) Line 350-354: the model correctly predicts inflow of North Sea waters into the Baltic proper, though bottom temperature and salinity values are too low compared to observations. But this is a feature of the existing hydrodynamics model, NEMO-Nordic, which was already used in the presented domain before, and calibrated and validated there. I would not expect the extension of the SCOBI model to influence these dynamics.

The reviewer is correct: the code for the dynamics in NEMO-Nordic has, indeed not been changed and the SCOBI model does not affect the hydrology of NEMO-Nordic. However, the applied physical forcing has changed, especially that for runoff. As mentioned in lines 378, the

new river forcing significantly improved the surface salinity results in the Baltic Sea when compared to results in Hordoir et al. (2019), but increased the existing negative bias in intermediate and deep waters of the Baltic proper. The Baltic Sea in NEMO-Nordic is, fairly sensitive to changes in fresh water input and, while we are currently working on improving this bias, here we have chosen to compromise the known bias in deep waters in the Baltic proper for improved surface results.

22) Figures 7 and 9: the little cyan plusses (not crosses as it says in the caption, that would be x) are very hard to see. Can this be done by shading instead? I do love the surface values on top of the depth graphs, very nicely done!

Thanks! We have already tried the shading and it is not great as it hides the standard deviation for the model (or make it less visible). However, we will further look on how to improve the averaged profile figures.

23) Line 435: figure B3 is mentioned here, I would prefer to see figure B3 in the main text rather than in the appendix. If there is a limited number of figures allowed, I would suggest moving figure 5 to the appendix instead, as it does not show simulated results. Within B3 the markers are very hard to see, can you make then larger? Some of the colours are quite light, resulting in a number without a visible marker in my printed version: enlargement might help with this too.

We will indeed enlarge the markers of figure B3. Regarding it's position in the manuscript, we had the same discussion within co-authors before submission and decided that it was too detailed information for the main text. If adding it to the main text, a detailed discussion on it should be added as well, describing main differences between areas, parameters and position in the water column. This will significantly complicated the flow and readability of the manuscript. The main relevance of this figure is to show where the model performs good or not. For example, if wanting to use the model results for a special variable in a particular area, one can look at this figure and have an idea if such results can be trusted or not. It also helps the developers of NEMO-SCOBI to see if changes in the code or forcing improves both these model performance figures (namely figure 10 and B3). Importantly, the figure results for the North Sea below surface waters are uncertain due to the great lack of observations (as clearly seen in figure 5). Moreover, we have summarized the relevant patterns in section 3.4 and discussed the main processes that could have an effect on the seen model biases in section 3.6, which is the main focus of this manuscript. It is also somewhat repetitive information as many of the area points show similar positions than those of Figure 10, especially when looking at surface values.

24) Line 463: in the Baltic Sea, four HELCOM-OSPAR assessment areas. Surely these are HELCOM assessment areas?

We agree that this is confusing, this will be changed.

25) Line 477-4780: surely you can see in your simulation results if accumulation happens or not?

Yes. We do see an accumulation over time in the time series at stations in the Gulf of Bothnia when compared to observations. However, there are only a few data points for nitrate at for example F3 and F9 in the entire time series and even if they are clearly below model results in recent years, it is difficult to say much more than this.

26) Line 483-485: this is an important message for the monitoring organisations, please make it stronger.

It is indeed a relevant point and we will emphasize it more in the text.

27) Line 492-493: grammatically incorrect sentence and it doesnt make much sense.

We agree with the reviewer that this line is confusing. "The Northern North Sea" is the official name for the central area in the OSPAR assessment areas. We refer to "the HELCOM-OSPAR assessment areas" to the areas adapted to our model domain, which means that they are not exactly the same, notably for OSPAR. The OSPAR assessment areas include also the North Atlantic and therefore some of the areas around the boundaries are cut in our domain. What we mean in this sentence is that the central part of the North Sea is the largest assessment area in figure 1 but also the least measured, which makes the model skill evaluation unreliable. We will rephrase this sentence.

28) Line 494-495: surely this is not about which model is better? Grammatically also incorrect, I assume the model by Daewel et al captures the southern coast of the North Sea just fine. Maybe not in biogeochemical terms, but the coastline itself is in the model so it captures it.

Correct, it is certainly not a competition between models. We are only highlighting the areas where most models fail to capture the nutrient dynamics. This will be rephrased as follows:

"Nonetheless, our biogeochemical results are comparable to those of Daewel and Schrum (2013): an overall good agreement with observations in the North Sea but with biases in the southern coast of the North Sea."

29) Line 505: sentence is too long and loses it grammatical structure by the end. Please rephrase.

We will split the sentence in two.

30) Line 513-516: please speculate on what the missing process for phytoplankton growth could be. And add riverine nutrient validation to the appendix (e.g. figure 3 but with an applied forcing-observational evidence focus), to better quantify the nutrient input issue. How well does your input capture suspended matter from fluvial sources?

The missing processes we refer to in this sentence are later on explained in section 3.5. For clarity, we will add at the end of the sentence "(further discussed in section 3.5)". Regarding suspended matter from riverine input, again, here we do not validate the riverine data set. In addition, we have modified EHYPE nutrient inputs to match those of observations and therefore we do not really understand the question. This said, we do mention (in lines 614) that there are large uncertainties regarding organic matter (detritus) in both seas. This also concerns all riverine data sets, as suspended and particulate organic matter is not necessarily automatically measured in all countries.

31) Line 524: have you considered the following works?

Capuzzo, E., Stephens, D., Silva, T., Barry, J., & Forster, R. M. (2015). Decrease in water clarity of the southern and central North Sea during the 20th century. Global change biology, 21(6), 2206-2214.

Capuzzo, E., Painting, S. J., Forster, R. M., Greenwood, N., Stephens, D. T., & Mikkelsen, O. A. (2013). Variability in the sub-surface light climate at ecohydrodynamically distinct sites in the North Sea. Biogeochemistry, 113, 85-103.

And how does this work relate to your findings?

We thank the reviewer for providing additional interesting references. We will look into both references and add findings if relevant for this study.

32) Line 526: you mean the Rhine, arguably the largest river to exit into the North Sea, has no influence here? Surely not.

Thank you for catching this. It is a typo and the Rhine will be added.

33) Line 532-535: figure 12 shows no observational support for this. How do you know your model is not simply overestimating the local light climate?

Lines 532-535 do not refer to our results (no reference to Fig. 12 there), but discuss findings in the literature. The sentence will be rephrased as follows:

"The elevated chlorophyll-a concentrations along the eastern UK coast has been shown to be likely related to frequent upwelling events under a predominant westerly wind regime ...

Please note that we do mention in line 562 that our model does overestimate the light attenuation depth. See also our reply 0.8 on light limitation and attached figures.

34) Line 542-544: maybe, but a comparison with satellite observations could verify this point better spatially.

Please sea our reapply in page 4, comment 0.7.

35) Line 547: in figure B4 the matching points are hard to see as they are white, overemphasizing the discrepancies. Can you use a blue-yellow-red colourbar here to emphasize where model and observational evidence do agree, and where there is simply an observational dessert? The same applied to figure 13, where observational points with a N:P ratio of (near) Redfield values are invisible.

We will see how to improve the visibility of matching points. The matching points are now in gray (not white), so we agree that they are difficult to distinguish not from missing data but between an island/land point from an actual observation point.

36) Line 570-575: spring bloom timing is mainly driven by temperature and light conditions in the North Sea, so a discussion on the simulated light climate is due here. Diatoms have evolved to be more light receptive than most other phytoplankton species, so they lead the spring bloom. Does the biogeochemical model allow for a proper succession of species? Figure B5 indicates it does, but a general seasonal succession graph (daily resolution, maybe for the different basins) would be better to display the models inner workings. A difference of 3 months in spring bloom timing is a lot, even for a large scale biogeochemical model.

Yes, the growing succession of species have been plotted and analyzed as these were tuned for the North Sea. It is not the main issue causing the bloom delay in the Skagerrak-Kattegat. In the text, we have highlighted the main factors that are likely the cause: the light limitation and the fact that its seasonality is not well captured in the Skagerrak-Kattegat area. We will make this point more clear in the text (in lines 562).

37) Figure 12: I love this graph but at the current size results are hard to compare to observational evidence. Can these graphs be enlarged? The colourbars are also hard to read.

Thanks! The figure is full page (A4), so enlarging it may be difficult. This may depend on how the journal handles it in the final version. However, we can enlarge the fonts.

38) Line 580: allows for a study of the North Sea

This will be changed according to reviewer's suggestion.

39) Line 584: , rather than prescribed boundary conditions

This will be changed according to reviewer's suggestion.

40) Line 585: again, this should not be a model contest on who performs best.

We fully agree, but still comparison is required to understand model discrepancies. We leave the sentence as it is also in response to reviewer #2, who is keen on knowing the performance of this model in comparison with others.

41) Line 595-597: you have shown that your model is capable of simulations from which you can derive relevant indicators for HELCOM and OSPAR, taking into account model performance and bias. Certainly for Chl-a there would be caveats, but most models have these. But you have not shown that the model can be used for climate projections with specific relevant improvements, as you have not made these improvements yet. And there is no detailed information in the manuscript on what these improvements would be: several ideas have been floated but there was no priority list of things to implement in the model. I would remove the latter part of this statement. For example, line 650 list improving the seasonal cycle of benthic denitrification, but contains no statement to how important this is with regard to other suggested improvement (e.g. cyanobacteria life cycle inclusion), or how this will be achieved.

A priority list is difficult to provide, because all mentioned future changes have a relevant role and can lead to significant improvements. However, solving the timing of the chlorophyll-a bloom is the first thing we will consider for further improvements. Therefore, it is the first point mentioned in the discussions of the section 'Future plan and knowledge gaps'. For this to be clear, we will rephrase the first sentence of this section as follows:

"Solving the phytoplankton bloom timing in the southeastern coastal North Sea and the Skagerrak-Kattegat transition zone in NEMO-SCOBI is a priority as it would significantly improve the model results, ..."

Note that this work necessarily implies improving the light penetration depth (as mentioned in line 612). In addition, we have shown that the main spatio-temporal patterns are well captured

and with the evaluation made here, we have a very good idea of where results can be trusted and where not. Of course, we do want to improve our model results and further understand them. However, this is not a major problem for the future projections, because that work will not focus on seasonal patterns, but on yearly averages (or even period averages). We have shown that these are in good agreement with observations and therefore, the model can be used as it is. For climate projection the aim is to keep the model as simple as possible, but ensuring that the main processes are captured. This is to be able to perform long-term runs at a reasonable computational cost.

42) Line 596-597: this is why I want to see a validation of the applied riverine forcing. The atmospheric deposition bias was discussed, but the reader lacks information on the riverine input bias.

Again, the focus of the paper is not on river validation. However, we will add more discussion on the river's effect. Please see our detailed reply to comment 12 and 13.

43) Line 613: how about suspended sediment?

The model does not take into account suspended inorganic sediments. However, it accounts for resuspended material from the sediments and it is taken into account for the light penetration depth as they go back to the corresponding nutrient pool in the water column (see for example N1 and P5 fluxes in figure 1).

44) Line 633: please provide references for the claim that the model compares well with previously published estimates (assuming you mean other publications than Dalsgaard et al, 2013).

In line 633, we indeed compare the model values to the values mentioned in Dalsgaard et al, 2013, which are mentioned just above (line 631). We agree with the reviewer that this is confusing and will rephrase these lines as follows:

line 630 "The total nitrogen removal from water column denitrification in the Baltic proper with persistent large hypoxic areas has been estimated to be..."

and

line 633 "This compares well with previous published estimates" will be removed from the text.

References

Donnelly, C., Andersson, J., and Arheimer, B. (2016). Using flow signatures and catchment similarities to evaluate a multi-basin model (E-HYPE) across Europe. *Hydr. Sciences Journal*, 61(2):255–273.

Donnelly, C., Dahne, J., Lindström, G., Rosberg, J., Strömqvist, J., Pers, C., Yang, W., Arheimer, B., et al. (2009). An evaluation of multi-basin hydrological modelling for predictions in ungauged basins. *IAHS publication*, 333:112.

Gustafsson, B. G., Schenk, F., Blenckner, T., Eilola, K., Meier, H., Müller-Karulis, B., Neumann, T., Ruoho-Airola, T., Savchuk, O. P., and Zorita, E. (2012). Reconstructing the development of Baltic Sea eutrophication 1850–2006. *Ambio*, 41(6):534–548.

Hordoir, R., Axell, L., Höglund, A., Dieterich, C., Fransner, F., Gröger, M., Liu, Y., Pemberton, P., Schimanke, S., Andersson, H., et al. (2019). Nemo-Nordic 1.0: a NEMO-based ocean model for the Baltic and North seas–research and operational applications. *Geoscientific Model Development*, 12(1):363–386.

Hundecha, Y., Arheimer, B., Donnelly, C., and Pechlivanidis, I. (2016). A regional parameter estimation scheme for a pan-European multi-basin model. *Journal of Hydrology: Regional Studies*, 6:90–111.

Lenhart, H.-J., Mills, D. K., Baretta-Bekker, H., van Leeuwen, S. M., van der Molen, J., Baretta, J. W., Blaas, M., Desmit, X., Khn, W., Lacroix, G., Los, H. J., Mnesguen, A., Neves, R., Proctor, R., Ruardij, P., Skogen, M. D., Vanhoutte-Brunier, A., Villars, M. T., and Wakelin, S. L. (2010). Predicting the consequences of nutrient reduction on the eutrophication status of the North Sea. *Journal of Marine Systems*, 81(1):148–170.

Maar, M., Möller, E. F., Larsen, J., Madsen, K. S., Wan, Z., She, J., Jonasson, L., and Neumann, T. (2011). Ecosystem modelling across a salinity gradient from the North Sea to the Baltic Sea. *Ecological Modelling*, 222(10):1696–1711.

Macian-Sorribes, H., Pechlivanidis, I., Crochemore, L., and Pulido-Velazquez, M. (2020). Fuzzy postprocessing to advance the quality of continental seasonal hydrological forecasts for river basin management. *Journal of Hydrometeorology*, 21(10):2375–2389.

Nijzink, R., Almeida, S., Pechlivanidis, I., Capell, R., Gustafssons, D., Arheimer, B., Parajka, J., Freer, J., Han, D., Wagener, T., et al. (2018). Constraining conceptual hydrological models with multiple information sources. *Water Resources Research*, 54(10):8332–8362.

Pätsch, J., Burchard, H., Dieterich, C., Gräwe, U., Gröger, M., Mathis, M., Kapitza, H., Bersch, M., Moll, A., Pohlmann, T., et al. (2017). An evaluation of the North Sea circulation in global and regional models relevant for ecosystem simulations. *Ocean Modelling*, 116:70–95.

Piniewski, M., Eini, M. R., Chattopadhyay, S., Okruszko, T., and Kundzewicz, Z. W. (2022). Is there a coherence in observed and projected changes in riverine low flow indices across Central Europe? *Earth-Science Reviews*, page 104187.

Riegman, R., Noordeloos, A. A., and Cadée, G. C. (1992). Phaeocystis blooms and eutrophication of the continental coastal zones of the North Sea. *Marine Biology*, 112:479–484.

van Leeuwen, S. M., Lenhart, H.-J., Prins, T. C., Blauw, A., Desmit, X., Fernand, L., Friedland, R., Kerimoglu, O., Lacroix, G., van der Linden, A., Lefebvre, A., van der Molen, J., Plus, M., Ruvalcaba Baroni, I., Silva, T., Stegert, C., Troost, T. A., and Vilmin, L. (2023). Deriving pre-eutrophic conditions from an ensemble model approach for the North-West European seas. *Frontiers in Marine Science*, 10.