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**

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.

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.

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)?
2. Line 19: the references are not in alphabetical or chronological order.
3. Line 44-45: too many comma's 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.
4. Line 50: "rereferred to *as* cyanobacteria" and I miss a reference for the statement that cyanobacteria do not grow in the North Sea.
5. Line 78: bad grammar, I suggest ", which is particularly true in"
6. Line 80: "but *contain* biases for"
7. Line 87-90: Not sure why this text is here, not relevant to the subject of this manuscript.
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.
9. Line 131: please refer to the figure before the textual explanation, to make it easier on the reader.
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?
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?
12. Line 167: please provide a reference for the applied reduction factor, assuming this is a generally available dataset. If it is not I don't quite see why the authors would use this particular product.
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?

   - 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?

   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?

14. Line 233: the reference year was chosen because of high nutrient values. Where those simulated values or observational values?
15. Line 245: explanation of the applied seasonal delineation (meteorological? astronomical?) is only given in the caption of figure 10. Please provide this here.
16. Line 280: this has been stated before.
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).
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.
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).
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.
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 hydrodynamical 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.
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!
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.
24. Line 463: "in the Baltic Sea, four HELCOM-OSPAR assessment areas". Surely these are HELCOM assessment areas?

25. Line 477-4780: surely you can see in your simulation results if accumulation happens or not?
26. Line 483-485: this is an important message for the monitoring organisations, please make it stronger.
27. Line 492-493: grammatically incorrect sentence and it doesn't make much sense.
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.
29. Line 505: sentence is too long and loses it grammatical structure by the end. Please rephrase.
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?
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?
32. Line 526: you mean the Rhine, arguably the largest river to exit into the North Sea, has no influence here? Surely not.
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?
34. Line 542-544: maybe, but a comparison with satellite observations could verify this point better spatially.
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.
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 model's inner workings. A difference of 3 months in spring bloom timing is a lot, even for a large scale biogeochemical model.
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.
38. Line 580: "allows *for a study of the* North Sea"
39. Line 584: *", rather than prescribed boundary conditions*"
40. Line 585: again, this should not be a model contest on who performs best.
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.

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.
43. Line 613: how about suspended sediment?
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).