Biogeosciences Discussions

Biogeosciences Discuss., referee comment RC2
https://doi.org/10.5194/bg-2021-2-RC2, 2021

Comment on bg-2021-2

Marcello Vichi (Referee)

Referee comment on "Evaluation of biogeochemical models performance and recommendation on observing system design using an unsupervised machine learning algorithm, BGC-Argo floats and assessment metrics" by Alexandre Mignot et al., Biogeosciences Discuss., https://doi.org/10.5194/bg-2021-2-RC2, 2021

General comments

This manuscript is indeed a valid compendium of diagnostics for assessing global ocean ecosystem models, which has been prepared with the aim to demonstrate the use of the multi-disciplinary dataset made available by the BGC-Argo array. The authors should thus be praised for their intention to bring together the community and follow the steps taken by Russel et al. (2018). However, that paper had different entry points, since it was specifically dedicated to a poorly sampled oceanic region and offered a multi-model analysis. This manuscript is well written and constructed, but only conveys a demonstrative message. I am thus not fully convinced by the scope of this present version of the manuscript, as well as by its effective novelty, since it does not add further knowledge to the existing literature.

I have chosen to sign this review because I deem important to fully disclose my intentions and avoid any misunderstanding. Over the past 20 years, I have personally been one of the modellers who felt the need to engage with the thorough validation of global ocean models and the related limitations. My review may sound over-critical, and I would like the authors to appreciate my intention to give a constructive critique, which is meant to assist with model improvements.

Hence, I have carefully thought about how to write this review, and realised that the most relevant point of clarity would be to illustrate some cases of how readers could approach it. From a point of view of someone approaching modelling validation as a student or early career researcher, this manuscript offers a limited perspective, and one would gain more theoretical and methodological background in the 2009 JMS special issue (Lynch et al., 2009, and all the other papers in the issue), if not from earlier papers in the ecological modelling literature (Oreskes et a', 1994; Rykiel, 1996). If a reader is interested in the validation of the global version of PISCES, this manuscript is insufficient, because it provides a series of figures with few comments and discussions. It is surely of interest to

the PISCES developers who are knowledgeable of the model details and possible deficiencies, but then an internal report would suffice. Finally, for experienced global ocean modellers, this manuscript is an illustration of the minimum set of assessments (which I prefer to the term "validation") that serious modellers have been doing in the last ten years when evaluating their model results. In terms of "metrics", it gives indications to compare the model output against the state variables that can be measured by the array of floats and to add derived state variables from applications of artificial intelligence. Ultimately, the assessment is based on visual comparisons of coarsely gridded spatial maps and time series, or through the use of basic univariate scores (bias and RMSD) and cumulative diagrams that combine the same skill scores (e.g. the Taylor diagram, which also includes linear correlation).

The BGC-Argo data are certainly invaluable, and this is the reason why the community has strived to develop the technology and the financial support to deploy them. The authors did not however succeed in showing their enhanced value for model assessment, beyond the obvious consideration that this increases the number of data, which would be much more evident if this same assessment was done by comparing datasets with and without the contribution of the BGC-Argo.

In summary I have found two major issues with this manuscript that the authors have not considered to a satisfactory extent:

- The loose definition of metrics and the absence of uncertainties' treatment. The authors use the term metrics in a rather ambiguous way. They also do not differentiate between measured data and artificially generated data. This implies that the evaluation process does not necessarily lead to an improvement of the model(s).
- The unconvincing enhancement of the effective role of BGC-Argo data in model assessment. Basically, the question I have is: why BGC-Argo are good enough and should be used separately and not as part of a global compilation of data such as the World Ocean Atlas? (which incidentally includes or will include the BGC-ARgo data) Since BGC-Argos are ultimately increasing the availability of data that are usually collected by means of traditional oceanographic cruises, what is indeed their value in model validation? The authors state that their aim was to demonstrate the invaluable opportunities offered by the BGC-Argo observations for evaluating global BGC models. I'm afraid this intent cannot be met unless some of the above questions are addressed

For clarity, I would like to elaborate more on the first concept above, while the second point is mostly derived from the specific comments detailed in the next section. Russel et al (2018) also use the concept of metrics in a wider sense, although they define metrics as "any quantity or quantifiable pattern that summarizes a particular process or the response in a model to known forcings". The strength of the ACC transport at Drake Passage or the latitude of the maximum zonal mean winds over the Southern Ocean are "metrics" in this context. They are combinations of state variables, or values of state variables at specific locations.

In this context, all the surface state variables listed in Table 2, are indeed components of the biological carbon pump, but they are not metrics. They are simply state variables. Only when considered together to evidence emergent patterns they may give indications of proper process functionality (e.g. the ratio of particulate organic carbon to total chlorophyll, de Mora et al, 2016). I agree that the DCM and the "nutricline" (which would deserve a more appropriate definition, see specific points below) are "metrics", as well as the depth of the hypoxic layer. Mixing together indicators of processes with state variables is confusing, unless a rigorous link between a single state variable and the process is established.

This manuscript increases the risk of misinterpretation by mixing together "metrics" and skill scores. Neither Russel et al (2018) and this manuscript expand on the concept of metrics performance and objective assessment (performance indicators, skill scores, cost functions, are all synonyms that depend on the specific discipline), which was instead done by Allen et al. (2007), Friedrichs et al. (2009), Vichi and Masina (2009) and others in the JMS special issue. For ease of simplicity, I will use the term skill score, which is the one used in the more mature field of weather forecasting. State variables can be assessed using univariate skill scores, and this is a necessary exercise for any modeller to ensure the model has some grip with reality. Figure 3 and the other density plots in the Appendix give a visual indication of the skill score, but they do not quantify it (e.g. Smith and Rose, 1995; Rose and Smith, 1998). I also have another question linked to my Point 2 (and further detailed in the specific comments): why should this exercise be done only with the BGC-Argo and not also including the other existing data? Since BGC-Argo are evaluated against cruise cast benchmarks, then those data are usually considered always superior, and should be used. Again, the real value of the BGC-Argo would have been shown if the score had been substantially modified with the inclusion of the Argo data.

Specific comments

P2L1 - Earlier work has specifically addressed the impact of assimilation on the carbonate system (Visinelli et al., 2017)

P2L26-29 - This sentence is mixing together sensor accuracy, which has been assessed by Johnson et al and Mignot et al, in two specific regions of the world ocean) and temporal/vertical resolutions, which have not been assessed as far as I am aware. This is misleading. 10 days may not be sufficient for all variables, as well as the vertical binning that is done. The comparisons have assessed the equivalence between rosette casts and the floats, but they say nothing about the temporal and vertical resolution. For certain processes, such as carbon exchange and phytoplankton biomass through chlorophyll and backscattering proxies, a resolution of 10 days would lead to sampling aliases either of the mean or of the variability (Monteiro et al., 2015, Little et al., 2018). These are examples from the Southern Ocean, where there is the highest density of buoys.

P2L32-34 - The authors should be more specific. Other datasets, such as for instance remote sensing, are less limited in terms of temporal and spatial resolutions. This is connected to the concerns expressed in Point 1 above.

P4L3-5 This sentence seems to imply that one can only perform point-by-point comparisons when there are few floats, which is odd. Again linked to my main Point 1 above. The authors should explain why given the current computing capability, they only suggest to perform diagnostics for few selected tracks and not for the overall dataset (Section 5.d).

P4L12-16 The connection between the variables and the ocean health/ecosystem functioning is not made explicit in the text. Taking as an example the ocean health index (http://www.oceanhealthindex.org/), establishing ocean health is obtained as a multivariate analysis of several data layers, forming a selected set of drivers and their associated thresholds. The authors should be more explicit about their intent here.

P5L12-13 This is not an objective criterion. What is an acceptable level of compromise?
P5L22 There are many other relationships, and they have been shown to give different results (e.g. Thomalla et a., 2017l). The authors should explain why they are recommending this one.

P6L12-15 It appears that this method of linear resampling would artificially increase the number of data, and hence bias the statistical results, especially in conditions where there are not enough data.

P7L10-12 The authors do not discuss what would happen if the MLD is different between the observations and the model.

P7L29-30 Related to my point 1 above. The relationship between the state variables and the ecosystem functions is not made explicit. The term “useful” should be motivated.

P8L7-8 Same as above, the value of DCM as an indicator should be contextualized. Why are BGC-Argo data providing a better estimate of this metric than other data?

P8L13 Please explain what H is.

P8L14-16 This may be confusing for some readers, since it's not technically a gradient. The cited paper uses and justifies this definition. I'd suggest the authors to be more precise and give their definition and how this is an effective metric of the carbon pump. Also, there is a difference in sampling between argo and the layers of discrete models. How is this taken into account?

P8l28-30 At P4L11 it is reported “depth of the OMZ”. This the depth of the oxygen minimum. It should be explained how and why this is a good indicator, and why the BGC-Argo data are superior in its identification.

P9L26 This statement about non-linearity is odd in the context of model goodness-of-fit (Smith and Rose, 1995; Pineiro et al, 2008; Vichi and Masina, 2009). If it’s non-linear, then the assessment is failed.

P10-8-12 The choice of the binning interval should be discussed. What is the advantage of losing the variability measured by the floats? Why not using the standard deviation as an indicator of the model skill to reproduce the proper scales? These are enhanced features that only the BGC-Argo data would allow to compute.

P10L22-24 Allen et al (2007) warned against the visual comparison of time series. This sentence is generic and should be explained in the context of the augmented data provided by the BGC-Argo.

P11L11-14 The results are not presented according to the concept of the biological carbon pump “metric”. It is evident that the nutrients are correlated while all carbon flux variables are not performing. Which ultimately questions the use of surface nutrients as indicators of carbon cycling.

P11L31 I cannot see the data “around” the line. I rather see an overestimation. (it is either Cape Verde or Cap Vert)

P12-L2-17 Linked to Point 2 above. The authors seem to imply that BGC-Argo data are more suitable than ocean colour for model assessment. I acknowledge that this is not explicitly written, but there is no clear rationale. This kind of map would certainly be superior in terms of spatial and temporal resolution when using that product as benchmark.

P12-section-d This is the section that mostly led to the inclusion of Point 2 above. The shown time series is 2 years long, which is an invaluable source of data from a region that has been influential in shaping our understanding of the spring bloom. I am missing the point why the authors are writing the term spring bloom in quotes. The advantage of time series from floats that remained in a given province of the global ocean is of huge potential in model validation. The offered description is quite generic, which could have been done even using monthly climatological time series obtained from the WOA, or from the existing long-term observational ocean sites (BATS, PAPA, HOT). The BGC-Argo floats are an unprecedented source of multiple opportunities to do validation in several regions of the world ocean (with some limitations), but this present form of the manuscript does not offer any specific recommendation of what numerical modellers should do to unleash

this potential. I would be very interested in seeing an exploitation of the multivariate nature of BGC-Argo, while I only see multi-panel plots.

P13L4-5 The authors should do more than simply say "correctly represented". This is a subjective statement, which is based on a visual comparison, exactly what the community challenged in the last 10-15 years. The advantage is that now we can use a frequency of 10 days, when initially phenology analysis was based on monthly data. Again, the authors are missing an opportunity to demonstrate the intrinsic value of this new data set.

P13-L13-20 This is a more detailed analysis of this specific model, which indeed brings in some of the advantages of a multivariate data set. However, there is a combination of measured and derived variables, which are treated as if they were equivalent. Quite a few questions come to mind: Is there a possibility that there is artificial correlation in the derivation of the phosphate and silicate concentration? What is the error associated with the CANYON-B method? Which is the effective (measured) variable mostly responsible for the response of the other estimated nutrients? The reduced consumption occurs during the spring period, and is continued during summertime. Hence, there is a factor at play during the late spring period, which is less likely to be reduced uptake from smaller phytoplankton during summer as suggested. It may thus be a delayed onset of the phytoplankton succession, or maybe a faster remineralization occurring in the upper layers, which retain more inorganic nutrients closer to the surface. This may indeed be beyond the scope of the manuscript, but it has been the authors' decision to propose some mechanistic explanations of this discrepancy. Showing a complete example of how the use of multivariate data allows modellers to investigate model deficiencies would offer guidelines to other modellers.

P13-L22-23 This sentence bears lots of assumptions. This is really where BGC-Argo can make a difference. The related uncertainties should however be highlighted, together with recommendations to other modellers on how to best approach the assessment of the carbon cycle metrics.

P13L26-29 This argument is flawed. If the occurrence of the peak is matched in the mesopelagic layer rather than at the surface, it is a clear indication of vertical mismatches in the export. I would thus argue that POC concentration is a proper metric for the export component of the carbon cycle. I would again encourage the authors to replace the use of subjective terms such as "consistent" with objective indicators (see Allen et al., 2007). For instance the comparison of the skill score computed in two consecutive years would give indication if there is some variability or if the model tends to repeat the same pattern.

P14L16-19 I would recommend more clarity on this statement. Are these sensors not available on the global ocean floats? It is not clear why this example is presented for Mediterranean floats, and not introduced earlier as one major advantage of the BGC-Argo floats.

P14L26-28 This sentence is similar to the statements done in the earlier sections. This is not technically a perspective statement.

P15L1-6 The question is whether these data should be used "on their own" or in conjunction with the other existing datasets. The authors should clearly explain in the conclusion why this dataset should be exploited as a separate unit.

P15L32-P16L3 I would thus recommend the authors to thoroughly address the issue of how the uncertainties should be treated. This is particularly important in the case of mixing measured and derived variables. If BGC-Argo are capable, within their limits, to reduce uncertainties in model assessment exercise, this should be adequately argumented. The fact that there are more data available is undoubtedly of relevance, but I wonder if it does help to reduce uncertainties in model states.

P16L15-18 Please highlight in which part of the results this is shown.

P17L2 Please add in the caption the meaning of the codes (or a link to where they are explained more in detail). Also, in the heading of the 3rd column, correct Date with Data.

Figure 2 Taylor diagrams are based on geometric properties of the circle. Hence they should be presented using equal axes.

References

Allen, J.I., Somerfield, P.J., Gilbert, F.J., 2007. Quantifying uncertainty in high-resolution coupled hydrodynamic-ecosystem models. Journal of Marine Systems 64, 3–14.

de Mora, L., Butenschön, M., and Allen, J. I.: The assessment of a global marine ecosystem model on the basis of emergent properties and ecosystem function: a case study with ERSEM, Geosci. Model Dev., 9, 59–76, https://doi.org/10.5194/gmd-9-59-2016, 2016.

Friedrichs, M.A.M., Carr, M.-E., Barber, R.T., Scardi, M., Antoine, D., Armstrong, R.A., Asanuma, I., Behrenfeld, M.J., Buitenhuis, E.T., Chai, F., Christian, J.R., Ciotti, A.M., Doney, S.C., Dowell, M., Dunne, J., Gentili, B., Gregg, W., Hoepffner, N., Ishizaka, J., Kameda, T., Lima, I., Marra, J., Melin, F., Moore, J.K., Morel, A., O'Malley, R.T., O'Reilly, J., Saba, V.S., Schmeltz, M., Smyth, T.J., Tjiputra, J., Waters, K., Westberry, T.K., Winguth, A., 2009. Assessing the uncertainties of model estimates of primary productivity

in the tropical Pacific Ocean. Journal of Marine Systems 76, 113–133.

Little, H.J., Vichi, M., Thomalla, S.J., Swart, S., 2018. Spatial and temporal scales of chlorophyll variability using high-resolution glider data. Journal of Marine Systems 187, 1–12. https://doi.org/10.1016/j.jmarsys.2018.06.011

Lynch, D.R., McGillicuddy, D.J., Werner, F.E., 2009. Skill assessment for coupled biological/physical models of marine systems. Journal of Marine Systems, Skill assessment for coupled biological/physical models of marine systems 76, 1–3. https://doi.org/10.1016/j.jmarsys.2008.05.002

Monteiro, P.M.S., Gregor, L., Lévy, M., Maenner, S., Sabine, C.L., Swart, S., 2015. Intraseasonal variability linked to sampling alias in air-sea CO2 fluxes in the Southern Ocean. Geophysical Research Letters 42, 8507–8514. https://doi.org/10.1002/2015GL066009

Oreskes, N., Shrader-Frechette, K., Belitz, K., 1994. Verification, Validation, and Confirmation of Numerical Models in the Earth Sciences. Science 263, 641–646.

Pineiro, G., Perelman, S., Guerschman, J.P., Paruelo, J.M., 2008. How to evaluate models: Observed vs. predicted or predicted vs. observed? Ecological Modelling 216, 316–322.

Rykiel, E.J., 1996. Testing ecological models: the meaning of validation. Ecological Modelling 90, 229–244.

Rose, K.A., Roth, B.M., Smith, E.P., 2009. Skill assessment of spatial maps for oceanographic modeling. Journal of Marine Systems 76, 34–48.

Rose, K.A., Smith, E.P., 1998. Statistical assessment of model goodness-of-fit using permutation tests. Ecological Modelling 106, 129–139.

Russell, J.L., Kamenkovich, I., Bitz, C., Ferrari, R., Gille, S.T., Goodman, P.J., Hallberg, R., Johnson, K., Khazmutdinova, K., Marinov, I., Mazloff, M., Riser, S., Sarmiento, J.L., Speer, K., Talley, L.D., Wanninkhof, R., 2018. Metrics for the Evaluation of the Southern Ocean in Coupled Climate Models and Earth System Models. Journal of Geophysical Research: Oceans 123, 3120–3143. https://doi.org/10.1002/2017JC013461

Smith, E.P., Rose, K.A., 1995. Model goodness-of-fit analysis using regression and related techniques. Ecological Modelling 77, 49–64.

Thomalla, S.J., Ogunkoya, A.G., Vichi, M., Swart, S., 2017. Using Optical Sensors on Gliders to Estimate Phytoplankton Carbon Concentrations and Chlorophyll-to-Carbon Ratios in the Southern Ocean. Frontiers in Marine Science 4, 34. https://doi.org/10.3389/fmars.2017.00034

Vichi, M., Masina, S., 2009. Skill assessment of the PELAGOS global ocean biogeochemistry model over the period 1980-2000. Biogeosciences 6, 2333–2353.

Visinelli, L., Masina, S., Vichi, M., Storto, A., Lovato, T., 2016. Impacts of data assimilation on the global ocean carbonate system. Journal of Marine Systems 158, 106–119. https://doi.org/10.1016/j.jmarsys.2016.02.011