Title: Quantifying biological carbon pump pathways with a data-constrained mechanistic model ensemble approach.

by Stukel et al.,

General comments:

The study investigates the pathways of biological carbon pump, performing ensemble simulations of biogeochemical model parameterizations, constrained by data assimilation with the use of several data types obtained from Lagrangian experiments.

The ms. is well written and well structured, being very informative for the processes controlling BCP pathways. The idea of using an ensemble-based approach to quantify model parameter uncertainties and constrain them by data assimilation is innovative and the general approach is meaningful.

I am not an expert on the various aspects of biogeochemical model parameterizations, but I understand the most important feedbacks between the different compartments of the BGC model and the importance of the physical forcing. In this work, there are some assumptions that can be considered as simplifications (e.g., 1D model, physical forcing, length of simulations etc.), but in my opinion there are all justifiable and there are other novelties that compensate for the study approximations.

Overall, I find the ms. worthy of publication in Biogeosciences after minor revisions. Please find below a list of comments that I would like the authors to address.

Specific comments:

1) P6, L183 and L188. Vertical eddy diffusivity is varying with depth or is set constant? Please clarify.

Vertical eddy diffusivity varies with depth. We will make this clear in the revision.

2) P6, L196. Which model variables, in addition to the euphotic zone, you could have simulated? Please clarify why those variables were excluded from the simulation (e.g., computational cost?) and explain how this may affect model uncertainty in relation to other error assumptions.

This question is not entirely clear to us. We did not exclude any variables from the model. We simulated all model variables in the depth range from the surface to the base of the euphotic zone as defined by the 0.1% light level (which was typically at depths ranging from 40 – 100 m and hence included approximately 10 – 20 model layers). We made the decision to only model from the surface to the base of the euphotic zone (rather than from the surface down to an arbtitrary deeper depth, such as 1000 m) for two reasons: First, it does substantially decrease computational cost. Second, the vast majority of our field measurements were made in the euphotic zone. Thus simulating the twilight zone would not give us substantial improvements in model fit (due to a lack of validation data below the euphotic zone) and also would have added problems associated with defining boundary conditions (i.e., determining what the concentrations of each state variable should be at 1000 m depth would not have been feasible since most of them were not measured beneath the euphotic zone).

3) P7, L237 and P8, L252-264. In the context of data assimilation, observational errors are often considered as a combination of instrument and representativity errors, the latter usually being the most important of all. The authors here quantify observational errors as the standard deviation of their measurements and/or the instrument error; if I understood correctly, representativity errors are not considered here. Are these errors relevant in terms of magnitude with observation representativity errors?

We completely agree with the reviewer that representativity errors are often the dominant source of error. However, we believe that the standard deviation of multiple distinct measurements inherently accounts for the most significant form of representativity error in our analyses. Following Janjic et al. (2017) we consider three types of representatitivity error: I) error due to unresolved scales and processes, II) observation-operator error or forward model error, and III) errors associated with pre-processing or quality-control. We note that since our data is mostly derived from direct *in situ* measurements, II and III are much less significant than they tend to be with, for instance remote sensing measurements. We thus believe that error due to unresolved scales and processes is the dominant component in representativity error for our study. These unresolved scales and processes include such phenomena as temporal variability in vertical mixing or surface irradiance (i.e., inaccuracy of our steady-state physical forcing), diel variability in phytoplankton carbon:chlorophyll ratios, internal waves that displace communities upwards or downwards, etc. When we state that we used the standard deviation of our measurements, these are measurements from different sampling points within a model layer during the Lagrangian experiment (i.e., different times and depths). This variability

from one measurement to another thus incorporates representativity error (or at least the portion of this due to unresolved scales and processes) along with measurement error. Typically, this standard deviation (which incorporates representativity error + instrument error) is the error that we used. However, in the rare cases where the standard deviation was less than expected instrument error (which can happen, for instance if four nitrate measurements all returned a value of 0.4 mmol m^{-3}), we used the instrument error.

4) I am confused with the threshold limit "detlim" referred as "experimental detection limit". How this threshold is defined? I see that the "detlim" depends on indeces i,j,k and that k-index is not an option for the observations; why? I think the authors should provide more explanations regarding the "detlim" threshold, because the cost function decrease (after several iterations) largely depends on this (at least this is what I understand from the definitions of J(p) and error_i,j,k at the end of page 7).

The experimental detection limit varies for each measurement type. For instance, for particulate nitrogen the observational detection limit was 0.2 mmol N m^{-3}. This means that when values are below this (i.e., a measurement of 0 mmol N m^{-3}, we have no knowledge of whether the actual value was 0.001, 0.01, or 0.1 mmol N m^{-3}). Thus we cannot penalize the model if it returns any value less than the detection limit when the observation is also less than the detection limit. So if, for instance, the observation was 0.1, but the model returned a result of 0.02 we cannot say that there is any model mismatch at all (since both are less than the detection limit). In practice, the actual value of detlim for each measurement was not very important to our results, because observations were seldom less than detlim. However, this formal definition is necessary with log-normally distributed errors, because occasionally the reported observation value was zero (or even negative, in the case of NPP) and since the model can never take on values less than or equal to zero, this would lead to an infinite cost.

5) Overall, in the data assimilation Section 2.4, it is not clear to me which model variables consist the control vector e.g., is it the same with the model state vector described in Table 1 (or not)? Please clarify.

Just to be clear, since terminology can vary across disciplines, we assume that "control vector" is used here to denote the adjustable variables or parameters that determine the model's predictions and hence model-data misfit. As such our control vector is the set of 102 model parameters that we allow to vary (given in Supp. Table 1). This is essentially all parameters *except* for TLIM (the temperature dependence of growth, grazing, and respiration rates), which we chose not to allow to vary because it is both fairly well constrained from

measurements and because allowing it to vary would obfuscate interpretations of variability in other parameters. We note, however, that model results also depend on the initial conditions, boundary conditions (at the base of the euphotic zone), and physical forcing (temperature, vertical diffusivity, and surface irradiance), which we prescribe directly from field measurements and hence do not allow to vary.

6) P18, L669-670 "our work shows that very different parameter sets can result in similar cost function values, despite generating distinctly different model outputs". This is an interesting result, but what does it means exactly (especially here where the cost function is different wrt variational approaches)? Please elaborate.

Most medium- to high-complexity biogeochemical models still utilize an approach a single biogeochemical parameter set to reach their conclusions. Sometimes this parameter set is determined by manually "tuning" the model to approximately match a set of observations, while other times the parameter set is determined through formal data assimilation that seeks to find the parameter set that produces a global minimum in a cost function relating model output to observations. Both of these approaches, however, seek to find a single "best" set of model parameters that can then be used for a model run, which will be used to interrogate aspects of the marine system (e.g., in our case to understand the different pathways of the biological carbon pump). Our study shows that in a high-dimensional system (as all medium- to high-complexity biogeochemical models are) distinctly different sets of parameters can match the observations equally well but produce very different model results. Indeed, all of the parameter sets identified by our OEP_{MCMC} approach had approximately identical values of the cost function, but some produced model ecosystems in which mixing was the dominant pathway of vertical carbon transport while others produced ecosystems with sinking particles as the dominant pathway. With either a typical "tuning" procedure or a more formal variational data assimilation approach, investigators would arrive at a single parameter set that would predict either that the ecosystem was dominated by mixing or by sinking particles (or perhaps a 50/50 split) that would give them a false certainty about the behavior of the ecosystem. An ensemble approach, using different biogeochemical parameter sets, is necessary to diagnose this model uncertainty.

7) P19, L690-692 "A further study (Anugerahanti et al., 2020) simultaneously perturbed physical circulation fields and the biogeochemical model and found that results were most sensitive to variability in the biological model". Vervatis et al., (2021a) and (2021b)

performed ensemble simulations, using a 3D high-resolution ocean physics and biogeochemical coupled model, to investigate unresolved scales and processes, perturbing (1) only ocean physics, (2) only BGC sources and sinks, and (3) both physics and BGC simultaneously, and found that uncertainties in physical forcing and parameterizations have larger impact on chlorophyll spread (and other BGC variables) than uncertainties in ecosystem sources and sinks. Moreover, this had an impact on increment analysis correction and on empirical consistency between model-data misfits, using various datasets (e.g., SST, SLA, total CHL and/or class-based PFTs). I think part of this information would improve the quality of the paper. This is merely a suggestion and I leave it up to the authors to decide if it is relevant to their work.

Thank you for pointing us to these recent studies. We plan to add their results to our discussion.

Minor comments:

1) P1, L26. Please avoid acronyms in the abstract e.g., CCE.

Thank you for noting this. We will correct in the revised version.

2) P7, L250. Do you mean N_O,i,j instead of N_M,i,j?

Yes, thank you for catching this. We had originally used 'M' for measurement and changed to 'O' for observation, but clearly missed one spot.

Best regards,

V. Vervatis

References:

Vervatis, D. V., P. De Mey-Frémaux, N. Ayoub, J. Karagiorgos, M. Ghantous, M. Kailas, C.-E. Testut and S. Sofianos, 2021: Assessment of a regional physical-biogeochemical stochastic ocean model. Part 1: Ensemble generation, Ocean Modelling, 160, 101781, https://doi.org/10.1016/j.ocemod.2021.101781.

Vervatis, D. V., P. De Mey-Frémaux, N. Ayoub, J. Karagiorgos, S. Ciavatta, R.J.W. Brewin and S. Sofianos, 2021: Assessment of a regional physical-biogeochemical stochastic ocean model. Part 2: Empirical consistency, Ocean Modelling, 160, 101770, https://doi.org/10.1016/j.ocemod.2021.101770.