Geoscientific Model Development Discussions

Geosci. Model Dev. Discuss., referee comment RC1
https://doi.org/10.5194/gmd-2022-177-RC1, 2022

Comment on gmd-2022-177

Anonymous Referee #1

Referee comment on "ForamEcoGEnIE 2.0: incorporating symbiosis and spine traits into a trait-based global planktic foraminiferal model" by Rui Ying et al., Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2022-177-RC1, 2022

Summary

The authors have expanded an existing BGC model (ForamEcoGenie) which includes a single calcifying foraminifera (foram) group into one which distinguishes between 4 separate foram functional groups (ForamEcoGenie 2). Each group is differentiated based on the presence/absence of a spine and/or the implicit presence of symbiotic photosynthetic organisms. Each functional groups relative physiologies are parameterized with various trade-off in metabolic cost, protection from grazing and disease, and grazing efficiency among other traits. The new model (w/ ~12 new parameters) is then tuned to observations of the relative distribution (from sediment cores), absolute biomass (from net tows), and export production (from traps) across all four functional groups. The model shows a strong ability to recreate the relative distribution of foram functional groups as inferred from the sediment cores, but a is unable to recreate absolute biomass as observed in net tows. However, I believe the later may have as much to do with limitations of tow data as it does with those of the model. Model skill to reproduce foram POC export lays somewhere in between. Overall, this is an interesting study and potentially worthy contribution to the field, however, I have two major concerns that should be addressed, in addition to a variety of minor comments.

First, the authors must provide a much more detailed description of the cost function (or M-score) they use to tune the model and assessment of the limitations of the various data sets they are tuning it to. For example, it is not clear how they deal with substantial space-time patchiness in tow data and how they control for seasonal biases therein. It appears to me they may be comparing annual means form the model to relatively instantaneous

tow observations which presumably occurred at different times in different places. Such observations would be unlikely to represent the annual mean anywhere with any seasonality. If this is true, I would strongly recommend conceiving of a more robust way to control for seasonality in the sparse obs. If it is not true then what they did do needs to be much more clearly explained. Similar concerns are detailed in Major Comment 1.

Second, a much richer analysis/discussion is warranted justifying why the inclusion of increasing foram complexity is useful for resolving large BGC cycles (beyond just being able to resolve foram diversity for its own sake). What large scale BGC processes/mechanisms might be getting missed by not resolving this level of complexity? Are there any metrics by which ForamEcoGenie 2 performs better than its predecessor which can be contributed to improve fidelity of foram diversity?

Third, some additional discussion point warrant consideration. For example, I am curious if the model, and sensitivity study in particular can provide insights into the seemingly stark dichotomy between observed foram biomass (very low) and observed foram export production (very high). It would also nice to add a more focused discussion of the mechanisms (i.e. parameterized physiologic trade-offs) that lead to the emergent foram distributions .

Finally, the authors need to be much clearer in their nomenclature throughout, both in figures and text it is often unclear what set of observations and sometimes what metrics are being referred to. Moreover, it is often unclear what dimensions/scales variables are being average over.

Major Comments

Model Evaluation and Cost Function.

I think a lot more focus on the methodology explaining the assessment of model skill is needed. As currently described, I was left with many questions about what, exactly, was being computed. For Example:

- **Primarily, it is not clear how the time dimension is incorporated into the evaluation of model skill and/or if model and obs are being compared on consistent time scales.**
  - How do you deal with the different time scales of different obs? The cores samples are presumably treated as averaged across a much longer time scale (certainly averaged across any seasonal signature). However, the tows and traps are measureing things on much shorted time scales. Depending on the method you could probably pretty easily get annual averaged in the trap data but the tows would certainly carry a seasonal signature which could bias the comparison with model means. How can/do you control for this? Although as noted below, it is not actually clear the model metric is the annual mean.
  - Are all M-scores computed on annual averages? If yes, then are they climatologies? And then, is there any accounting for the fidelity of the seasonal cycle?
  - For the obs that don't average out the seasonal cycle is the M-score somehow paired in time between model and obs? Or are there enough obs in all cases for a robust annual mean to emerge (this seems unlikely for the tows)?
  - What assumptions justify comparing pre-industrial paleo data for one metric (relative abundance) to very recent anthropogenically forced data for the others (absolute concentration and export)?

- Why is the trap data in units of count/m3/d rather than count/m2/day. Presumably the trap POC starts as a volume, but shouldn't that be divided by the height of the trap container to get a flux?

- **Line 312:** What is the 'time slice comparison' for which you regridded? I couldn't find the term 'time slice comparison' mentioned anywhere else in the manuscript? Was there any re-gridding for the other comparisons?

- Describe a little more specifically how the median absolute deviation measurement ensures 'close to reality data'.

- Is it necessary to discard species with less than 3% abundance when you are aggregating species into function groups anyway? Considering there are 50 some species I would assume there are quite a few beneath that threshold in each functional group and thus integrate to a non-trivial proportion of the group. It would be good to quantify how many were discarded in each group (in some average sense), or perhaps see how including them influence the M-score of just the optimal parameter set.

- Is the POC flux just separated into that from just Foram groups are all POC?

- Cite the figures in which the distribution of each observation is included.

- It would be interesting if there was some discussion on the similarities and differences of the parameterization of each Ensemble cluster of Figure 2 (A-E).

- It is often ambiguous throughout when biomass and export is being integrated across the whole ecosystem or just forams. Please err on the side of redundant clarification for this as it gets a bit confusing as is. For instance in Figure 3 you look at 'ecosystem biomass' which I assume is integrate across all plankton but also look at POC export which I assume here is integrated across the ecosystem but there is no way to tell from looking at the figure label. Additionally, using consistent use of POC flux and POC export would help (unless you mean different things?) Similarly, it seems like biomass is sometimes referred to as 'living biomass' and sometimes just biomass. Does this mean I am to assume biomass = living biomass + POC?

- You should define the export depth horizon somewhere else other than the caption of Figure 6. Further there should be some mention of what depth horizon the traps are at. At least on average.

- **Section 4.3:** Again, it is not clear what time scale you are comparing these on. Are the model distributions global means? And the net tows relatively instantenous points in time? Why would we expect these values to be related as there is presumably some seasonal cycle? Presumably, there is something left unexplained that justifies the comparison, but if not I don't think this a particularly useful metric to assess model skill as it does not appear to be comparing the same thing.

- It would be useful to provide some context on what a good M-score is. In section 4.3 you argue the M-scores are close to zero thus demonstrate the models *inability* to recreate living biomass concentrations; however, every other metric is also closer to 0 than 1. Is that acceptable? Further, does a negative value indicate an inverse correlation or just a worse overall bias? It seems odd that the biomass score is always 0 and never negative unless 0 is some fundamental limit which models with poor skill approach? But then what does a negative value indicate? A strong inverse correlation between model and obs?

Comparison to Prior Model Iterations:

- The second paragraph of Section 4.1 and Figure 3 touch on how the optimal foram parameter set for EcoGenie 2 compares to previous iterations of the model, but I think this matter warrants considerably more attention.

- Presumably, the reason for increasing the complexity of a BGC model is to include mechanisms necessary to accurately resolve larger scale carbon and nutrient cycling such that they respond realistically to environmental/climatic perturbations. That is to get things *right* for the right reasons rather than overtunning models without the right mechanisms. So I am curious how this addition improves the performance of the model w/r/t global bgc cycles that might lead us to believe it can offer more accurate predictions that justify its higher cost (computationally and in terms of parsimony). I am thinking about questions like what conditions favour foram groups that transfer carbon to depth or into higher trophic level more efficiently and do we expect climate change to shift that underlying balance in a meaningful way? At a minimum I think some discussion on this front is warranted. But preferably, it would be nice to see some further quantitative comparison of what aspect of global BGC cycling are improved relative to prior, simpler, but computationally cheaper, runs.

- At a minimum I would like to see what happens to NPP relative to previous iterations? It is somewhat surprizing that you could achieve similar model skill after adding 3 new tracers without having to tune the parameters of the original model.

- Structurally, with this expanded analysis I think it would flow better if you first describe the skill with which the optimal parameterization of ForamEcoGenie 2 recreates the obs (i.e. Sec. 4.2-4.4 and Figs. 4-6). Then go on to discuss how include accurately resolved foram PFTs changes the overall ecosystem variables in the broader bgc model compared to previous iterations of the model (i.e. Fig 3 and the end of Section 4.1).

Additional Discussion

- **Discussion of model utility:** Per above, can you quantify, or at least more deeply consider, how the added complexity of four foram groups could help BGC models improve large scale nutrient and carbon cycling?

- **Discussion of low biomass and high export:** The observations of such low biomass and high export are striking. Especially since the model seems to need much higher biomass to match observed export. A deeper discussion of this could be quite interesting. Could it be a bias in the obs? Nets and traps (especially those that are decoupled in space and time) have plenty of sources of error. Alternatively, what can we learn from the model about how this might be possible from an inverse modelling perspective. Can you identify parameter sets that lead to similar results? What are those parameters? I would assume very low vulnerability to grazing and very high mortality could create such an outcome by preventing recycling and increasing export efficiency. It might also be interesting to look at export efficeincy for forams explicitly. Depending on if there are any interesting findings this may be more suited for a subsection of Results.

- **Discussion of physiological trade-offs:** More discussion of how the assumed (ie parameterized) advantages and disadvantages of each group lead to their emergent distribution would be interesting and warranted.

Figures and Tables

Figure 1.

- This is redundant with Figure 4, column 2, no? I see how it is useful in an introductory context and definitely needed in Figure 4 for comparison, however, I think you could remove it here and just reference Figure 4 where required. Especially if you are tight on space.

Figure 2.

- Is the export production shown on the right the globally integrated total foram value used to calculate to the M-score for POC flux? Or is it the total ecosystem POC flux and the former just forams? This is an example of where carefuly labelling on what is actually being integrated/averaged is so important.
- Clarify if each column is the sum of M-scores for all 4 groups with a maximum of 4 (rather than 1) to be transparent that even bright red values are really quite low.
- Column three should be labelled 'Relative Abundance', not 'Abundance', no?
- Can you add a column showing the total M-score?
- Can you highlight the parameter set you chose as optimal?

Figure 3:

- Is there any reason not to show columns 2 and 3 as percent deviation from EcoGENIE such that the bias (relative to EcoGENIE) can be compared across all metrics consistently.
- Regardless, it would be useful narrow the colorbar for biomass and POC export as to discern the distribution.
- I encourage adding an additional row for NPP.
- Clarify in labels and caption that these are ecosystem integrated values, not foram integrated. For example, there is hard to tell if there is a difference between 'POC export' here and 'POC flux' in Figure 2, but I believe they are very different variables. I also can't figure out if the you mean something different between 'flux' and 'export'? If not, pick one and stick with it. Otherwise please clarify throughout.
- Potentially move to after Figs 4-6 following my suggestion to shift discussion of model-model ecosystem level comparison to after the model-obs foram level comparison

Figure 4:

- "Model relative abundance of each group are calculated based on POC flux rates" – Huh? Is this a typo?
- Here and elsewhere, I think the column 1 header should be ForamEcoGENIE **2** to distinguish it from the previous iteration (as in Fig. 3)
- Change 'mean' to 'global mean' for clarity.
- Consider moving the M-score to the row heading on the left, just after the functional group. I think this would be clearer as it is a function of both model and obs and then the heading for each column would be identical (the global mean)

Figure 5:

- I understand why you have overlaid the obs as there are many less data points than in the case of Figure 4. However, I think it would be clearer to present Figes 4-6 in a consistent way, with the model output on the left and obs on the right. Even though there are sparse obs for the other metrics I think this would be easier to compare and better communicate that the obs are in fact sparse (which is an important point). Further, it would help the reader get their head around all three if they were organized consistently.
- Include units and labels for what I assume is the global mean in the header.
- Include M-score here too, as in Fig 4. Ideally in the row headers as suggested above

Figure 6:

- Same comments as Figure 5.
- Are these units right? Shouldn't export (a flux) be /m2 not /m3 as in Figure 3 and 8?

Figure 7

- Are the units of panel b) correct? Shouldn't a flux be /m2 not /m3. Or is there some distinction in the flux, flux rate, and production rate I'm missing?
- Headings for c) and d) appear wrong. I think c) should be 'globally integrated biomass' not 'production' and d) something like 'globally integrated export production' not POC production rate. I'm positive what 'POC production rate' means (NPP I suppose?) but I think you are talking about export, no?

Figure 9/10

- Be clear about what obs are being used in each. Presumably tows in 9 and traps in 10, but mention this explicitly in the caption.
- What do multiple obs data points for the same functional group at the same site during the same month mean?? If these are different species I would integrate them into their corresponding functional groups as done for the M-scores.
- Minor, but maybe make the model v obs legend in grey rather than blue so that it isn't visually associated with a specific functional group.

Tables

Table 3

- Why is Biomass zeros across the board? I understand it is poorly resolved but being all uniformly 0 and never negative seems odd? See comment above on clarifying interpretation of M-Scores.
- Caption should read 'M-Score from best model run (or optimal parameter set preferably, per other comment).'
- Why not include the total M-score (col sum + row sum) as this is ultimately used to decide which parameter set was optimal, no?

Minor Comments

Trait Based Model Description.

I think it would be useful to have some more introductory discussion on the difference between species-based, PFT-based, and trait-based models, as you often reference species-based models as a foil. I may be wrong, but my understanding is that they are fundamentally very similar in that they all prognostically integrate some finite amount of discrete plankton tracers but differ in how they are parameterized. That is, species-based models use parameter values derived empirically for specific species, PFT-based models use parameter values intended to average across species but capture key functional differences and physiological trade-offs between groups, then trait-based allometric models connect PFTs which are distinguished nominally by size using allometric relationships that vary parameters with size. The later then typically allows for more tracers by reducing the number of independent parameters required. However, I am not clear if, without the allometric parameterization, there is anything fundamentally different between PFT and trait-based BGC models. Both seems to cluster myriad species into functional (or trait-based) groups and resolve them separately. The difference seems to be just the resolution of the groups (e.g. how many size classes) and how their parameters are related. Further, I think it could be argued that very few BGC models are truly species-specific, but rather, at least implicitly, are averaging over many particularly species. Is there something else essential I am missing? Either way, it would be useful to include a paragraph introducing the differences (similar to the broader intro to BGC model in Ward et al).

Line 1-35: Do coccolithophores and pteropods perform worse as paleoproxies? Mostly, I'm just curious.

Line 60: You have a sentence introducing the 'trait' of 'symbionts' and its prevelance. It would be useful to do the same for 'spines' up top here. Perhaps both following the next sentences. i.e. 'foremost trait is calcification… but spines and symbionts are two more important ones… then sentence on prevalence and definition of symbionts… and sentence on prevalence and definition of spines"

Line 64: Define what 'core-top' data is.

Line 68: "spines extruding from the test". Define what the test is?

Line 111: Describe the cell quata/carbon quota here a little more explicitly. You focus on how it varies with size but its fundamental role (to vary stoichiometry I think?) is not clear.

Eq 5. Does V stand for Volume and nutrient uptake? If so, change one.

Line 150 (and elsewhere): It is a bit confusing to use epsilon in the grazing formulation as the common disk parametrization uses the prey capture rate (typically referred to w/ epsilon) instead of the half saturation coefficient (K) to describe a mathematically identical version of the type II response curve. If there isn't a strong reason to use epsilon for the spine effect, I'd suggest changing it to avoid the confusion.

Line 299-301: Can you make this either 1 or 3 sentences. As currently written it sound like there is some inherent reason tows and traps a grouped together separate from cores. But as I understand they are three independent data sets each used to evaluate a different aspect of model skill.

Line 342: What is the difference between "POC export scores" and "showing the closest export rate to observations"?

Line 344: Above you say the relative abundance M-score reaches as high 1.2 but here you say the highest is 0.29. I think up top you're referring to the sum all scores for each group, but this could be clearer.

Line 345: Does this prioritization mean that the selected parameter set doesn't actually have the highest integrated M-score. Can you quantify this decision by assigning a weighting metric to each variable?

Throughout: I think 'Optimal Parameterization' would be more descriptive than 'best run' which could refer to differences in forcing, initial conditions, etc.

Line 389: "Although the general distribution pattern of foraminifera living biomass agrees with the observations from plankton nets:" --- Does it really? I would qualify this a little more.

Line 395: Export or net primary production? Or primary + secondary production for mixotrophs?

Line 409: Here and elsewhere it would help to be really specific if you are talking global POC export of all foram groups, one foram group or all POC. Additionally, it is not clear if you mean something different between POC flux and POC export. Presumably no, in which case use consistent language where possible.

Line 414: You use two different references to cite the same range of $CaCO_3$ export. Was that intentional? If so, why?

Line 437: Clarify what you mean by species-species discrepancy.

Lines 404: Agreed. But how does this all influence your M-scores?

Line 433: "The model successfully reproduces the first-order seasonal patterns observed by sediment trap data at a basin scale". Does it? Looking at Figure 9 I cant find one panel with a particularly convincing match.

Section 5: This section on limitations focuses entirely on increasingly complex traits that are not resolved but mentions nothing of uncertainty associated with the parameterization of those included or in the observations to which they are tuned. I think some discussion of the latter two limitations is warranted.

Line 486: Be specific here: Foram C export or all all C export? Also when you say global mean do you mean globally integrated? Or are you referring to an inter-annual time average?

Typos and Other

Throughout there is a lot of inconsistent/poor grammar that should be improved for clarity.

For example, in the **abstract**:

- "increasing functional trait diversity and expand**ing** their ecological niches
- "focusing on functional trait**s** rather than individual species" should
- "observations from global core-top**s**, sediment trap**s**, and plankton nets"
- "Our model approximates…, account**s**"
- "19% of the global pelagic marine calcite budget **which is** within the lower"

And **Intro:**

- "built an ecophysiology based dynamic model" ->" built ecophysiology based dynamic model**s**"

I've tried to stopped flagging these (although list a few more below) but the grammar warrants a careful review throughout.

Line 44: This sentence is structured as if the model 'reconstructed' the future scenarios int the second clause. Perhaps revise to "…and **simulated** potential…"

Line 70: "traits…lay down the foundation of a trait based model" is a bit of tautology ï☐☐

Line 95: extra 'and'

Line 174: Section title?

Line 404: is a flux rate’ different then a flux?

Line 445: Should this be a header?