We thank both reviewers for their constructive comments on the manuscript. We have made some major changes following the two reviewers' comments.

- Firstly, we changed the POC export unit conversion (Reviewer #1's 4th major comment) and slightly retuned the model to account for this change. This change generates better model-data comparison than the previous version.
- We also reformatted the text, improving the introduction and the model description as Reviewer #2 requested. The introduction now justifies the focus on the critical gap in foraminiferal model development and introduces better the traits of spine and symbiosis to the readers. The model description update now includes a new figure demonstrating the basic model structure.
- Lastly, we added additional discussions about why we need to increase the foraminifera complexity in a model and the possible limitations in the current parameterization.

These additions have increased the manuscript's length while improving its clarify. We hope the editor agrees with us that this was worthwhile.

We have responded to each comment below. Reviewer comments are shown in bold, our responses in blue and our actioned responses in red (with quoted text in Italics).

Reviewer #1

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.

We now provide more details on the cost function in the Section 6.3.

"We used the Mielke measure, or M-score (Watterson, 1996; Watterson et al., 2014) to quantify the model-data fit (Eqn. 21). This metric is essentially a non-dimensional transformed mean square error (Gregoire et al., 2011; Hemer and Trenham, 2016). The score spans from -1 to 1 with values closer to 1 representing better model performance, 0 representing no predicting skills, and negative values representing negative correlation."

It is correct that we calculated the cost function using model annual mean with observations, however, we estimated observed annual mean averaging all observations present in each model grid point in time (combining different seasons and years). To assess the observation temporal coverage, we compared the model with the observed seasonality (Figures 9, 10). Combining annual mean cost function with seasonal plots thus allow us to validate the model temporally and spatially. We will clarify this in the comparison section (Section 6.2).

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?

We agree with the reviewer that the long-term goal of implementing foraminiferal complexity is to better resolve large GBC cycles. However, this goal is beyond the scope of this current paper.

Foraminifera influence biogeochemistry mostly via the inorganic carbon cycle rather organic carbon cycle, as they have very low biomass. Modelling calcification of foraminifers is currently limited by the lack of understanding of the mechanistic drivers behind carbonate production. Instead, as is common in Earth System Models, we estimate calcification in the model using a fixed rain ratio to estimate the calcite export from the organic carbon export of the foraminifera.

Increasing the complexity of foraminiferal traits allows us to 1) capture their different niches, 2) improve the comparison to observations, which resolve this degree of ecological complexity, and 3) capture foraminifera biogeographic and ecogroup change in response to the environment. Having a more complex foram model also provides a powerful tool to

compare with past events and possible impacts of future climates. These additional benefits had relatively minor impacts on the performance of the ecosystem properties in EcoGENIE (Figure 12).

We have clarified our study's focus, resolving the ecology rather than the biogeochemical role of planktic foraminifera (Line 87). We have also added a more general discussion about the importance of model ecosystem complexity in Section 10, second paragraph.

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.

We thank the reviewer for questioning the cause of discrepancies between distribution patterns, biomass and export. To disentangle the cause, we plotted the histogram below showing the model parameter values that are associated with low foraminifera export production (links to the cluster with high M-scores). In brief, the foraminifera cell leading to high scores has higher probabilities falling in the 100-200 µm range. We interpret this as zooplankton in this size tend to resemble the observed foram's spatial distribution (Figure S5 below). As for the other parameters, the symbiont size is small (0-0.01 times the host size) so that their high nutrient affinity supports foraminiferal survival in oligotroph gyres. The calcification respiration peaks at 0.02 mmol C/d to achieve better comparison with the observed low standing stocks.

Thanks to this analysis, we were able to identify a mistake in the data processing, causing artificially low biomass to export production ratio. We previously converted observations of foraminifera export into the incorrect unit. Fixing the unit conversion of the foraminifera export and retuning the model now provides modelled magnitude for biomass and export more consistent with observations. Despite this improvement, our main concern of underestimating these two metrics because of the limited temporal and spatial coverage in the data still remains. The real biomass is likely higher than in the reported observations (if all seasons were equally sampled) to match the high export production.

We have added a histogram (Figure S3, shown below) to provide an overview of the optimal parameters (indicating the emergent distribution). We modified our unit conversion and present now the retuned model. We also provide a discussion in the 7.1 Section and a more distribution-relevant discussion in 7.2.

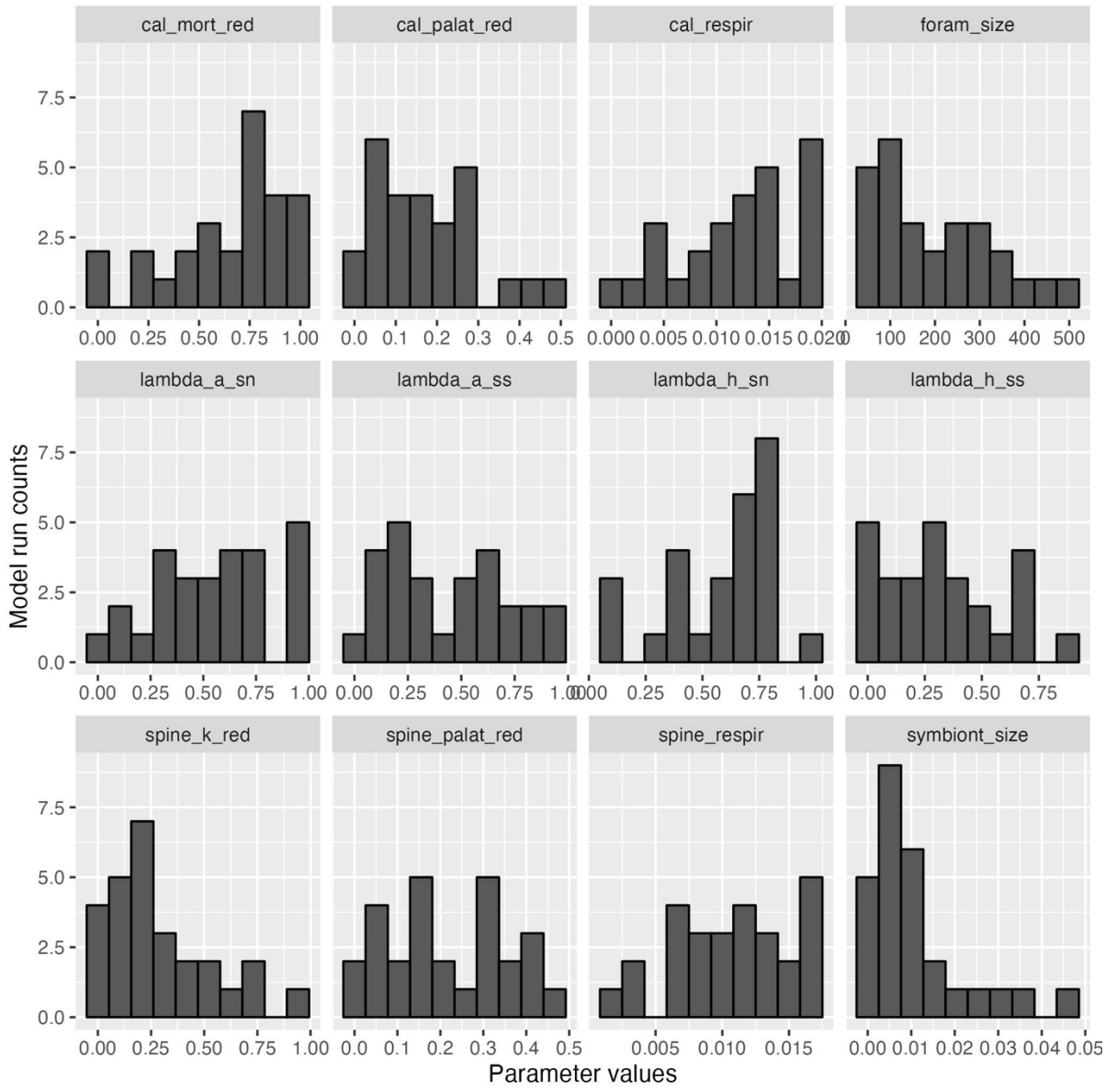


Figure S3. Histogram of the optimal parameters (linking to Cluster "A" in Figure 2) associated with low foraminifer annual mean export production (< 1 mmol C m^{-2} d^{-1}) and relatively high relative abundance M-score (>= 0.45). Parameter abbreviations are as follow. cal, calcification; mort, mortality; red, reduction strength; palat: palatability; respir, respiration; a, autotroph; h, heterotroph. ss, symbiont-obligate spinose foram, sn, symbiont-facultative non-spinose foram.

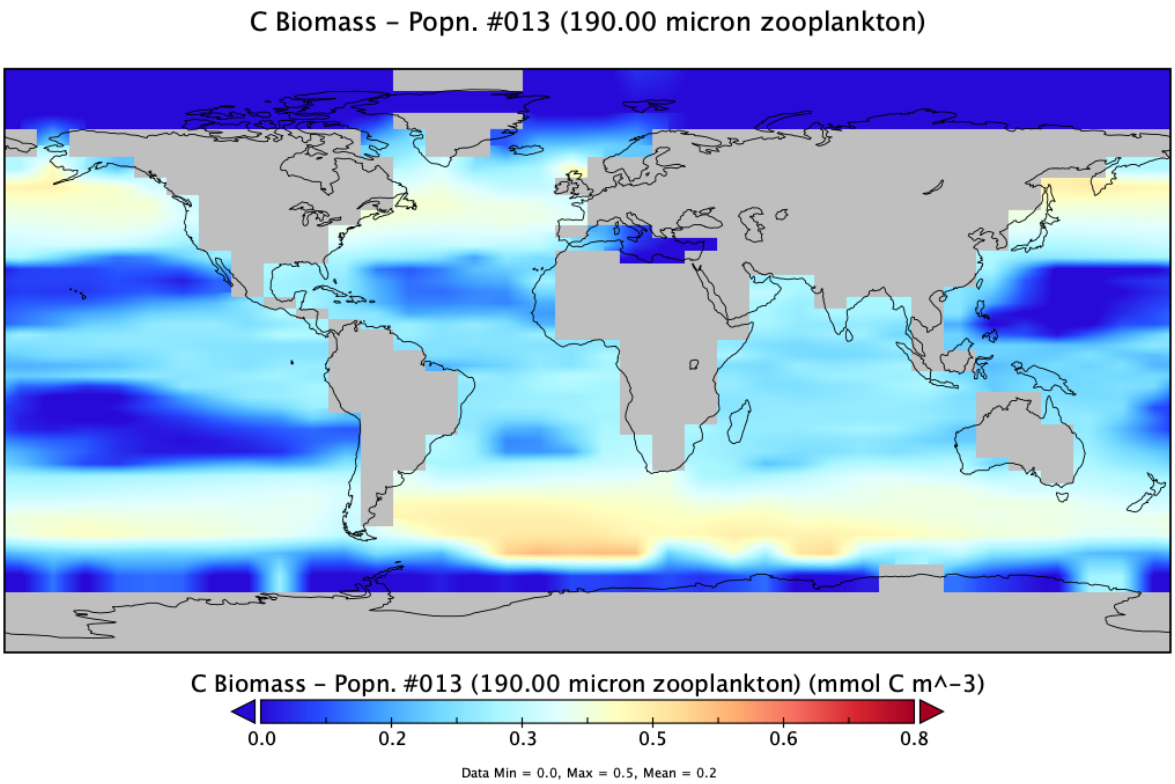


Figure S5. Biogeographical distribution of zooplankton biomass in 190 µm

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.

We thank the reviewer's careful observation on the caption/terminology. We have made corresponding changes in figures and text as the reviewer asked (see responses to the minor comments).

Major Comments

Model Evaluation and Cost Function.

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)?

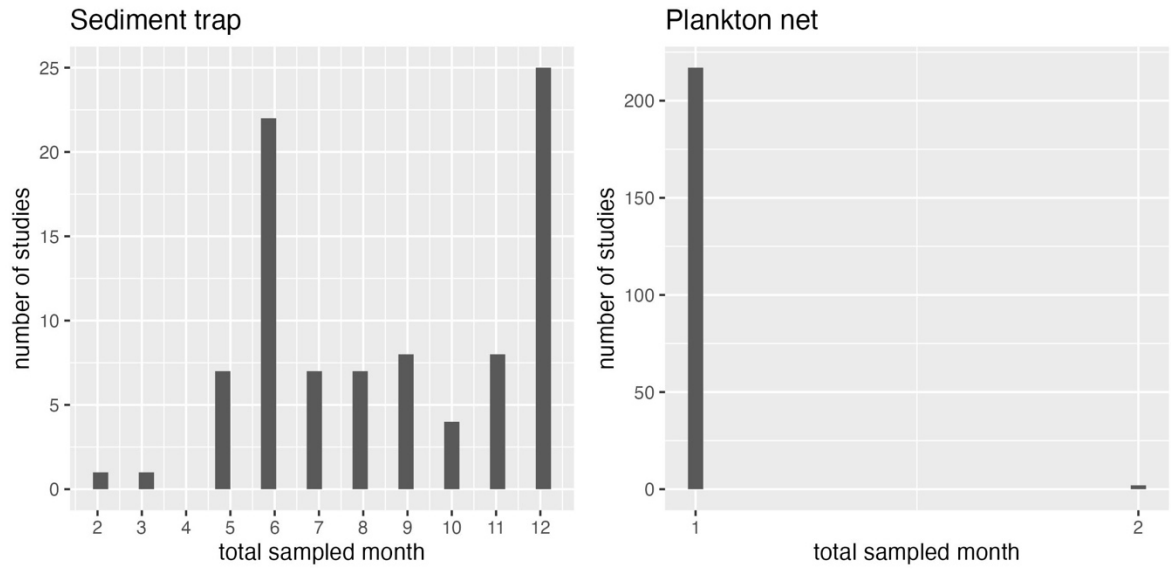
1) Time scale

While the core-top data are already averaged over several decades given bioturbation in surface sediments, we plotted a histogram below to show the sampled seasonality in plankton net and sediment trap studies for demonstration. As the reviewer points out, the plankton net studies barely consider seasonality, and this might be not robust for annual average comparison. We deal with this by directly comparing the model with seasonal time series in Figure 9 & 10. However, we will also clarify the limitation in annual average biomass comparison part (section 7.3).

2) M-score

The M-scores are based on annual averages, mimicking a climatology in the sense that we combine multi-year observations. We did not estimate a cost function for the time series comparison, because (1) the sampling data temporal coverage is too low at most locations, and (2) the number of available locations is insufficient (like those in Figure 10), creating large spatial bias towards specific oversampled locations, (3) the coarse model grid resolution isn't that well suited to resolving seasonal cycles in detail so an annual average is a more consistent comparison with the model.

As mentioned above, we now clearly specify the data processing and model-data comparison in the method section 6.2.



A histogram of sampled month in collected sediment trap and plankton net data. The sediment traps tend to carry seasonal signatures while plankton net not.

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)?

To clarify, we calculated a score for each metric, so did not quantitatively compare different observations. The model is forced with pre-industrial boundary conditions to match the core-top data (relative abundance). So there is possible inconsistency in comparing the top-core data (representative of the pre-industrial state) and water-column observations (plankton net and sediment trap, which represent the current climate). But we assume that such inconsistency is negligible at the first-order level considering (1) the scale of foraminifer living stocks is small and (2) the difficulties of tackling different time scales of those plankton net and sediment trap (from 1970s-2010s). This is clarified in the Line 316.

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?

We are grateful that the reviewer has found this inconsistency. The sediment trap data should indeed be in count/m2/d and not count/m3/d as used in the manuscript. We have now fixed the sediment trap data unit conversion and retuned our model. This updated version improves our results with POC export and biomass more consistent with observations.

We converted all modelled flux units into "mg/m2/d", retuned the model and have updated the text throughout to reflect this.

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?

We apology for this confusing term. We meant "annual average" and have rephrased it as is.

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

Using a median absolute deviation measurement improves the model-data comparison when the data are sparse. This uneven distribution results in a few data points with high biomass/export variability having a large effect on the overall scoring. Such high variability can be seasonal or caused by any other local changes to the environment such as storm events which is not resolved in the model.

Because we have now better model-data comparison of export and biomass, we do not need to include a median deviation and have removed its mention in the manuscript.

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.

We did not exclude the entire taxa because of rare abundance in some places. We only excluded the occurrence of taxa with less than 3 % of the total assemblage (traditionally 300 specimens would be counted) because the statistical and taxonomical accuracy of these rare occurrence counts is too low. Foraminiferal assemblages have very uneven distribution with a long tails of rare taxa with one or two counts, for which the taxonomy is often less certain than most of the over 30 more dominant taxa. Therefore most of the assemblage is represented by our approach.

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

Almost exclusively POC flux in the paper reflects the foraminifer-derived bulk POC flux. The only exception is the figure in comparing to prior model versions (Figure 12 in Section 8)

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

We assume that the Reviewer refers here to the relative abundance distribution figure.

We have added citations to the subplots in the result section.

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).

Similarly to Reviewer #1's third major comments, we picked the best cluster (Cluster A in Figure 2) to show the trends in the parameters values with model success in an histogram (Figure S3). We repeated the same analysis (Figure S4 showing below) for the parameters associated with negative M-score (Cluster D in Figure 2). Their distributions closely contrast with the one for the high-score parameters, especially for the first four general foraminifer parameters (i.e., calcification trade-offs and foraminifer size), indicating the important role of foraminifera size and calcification in scoring spatial distribution as they influence all the functional groups.

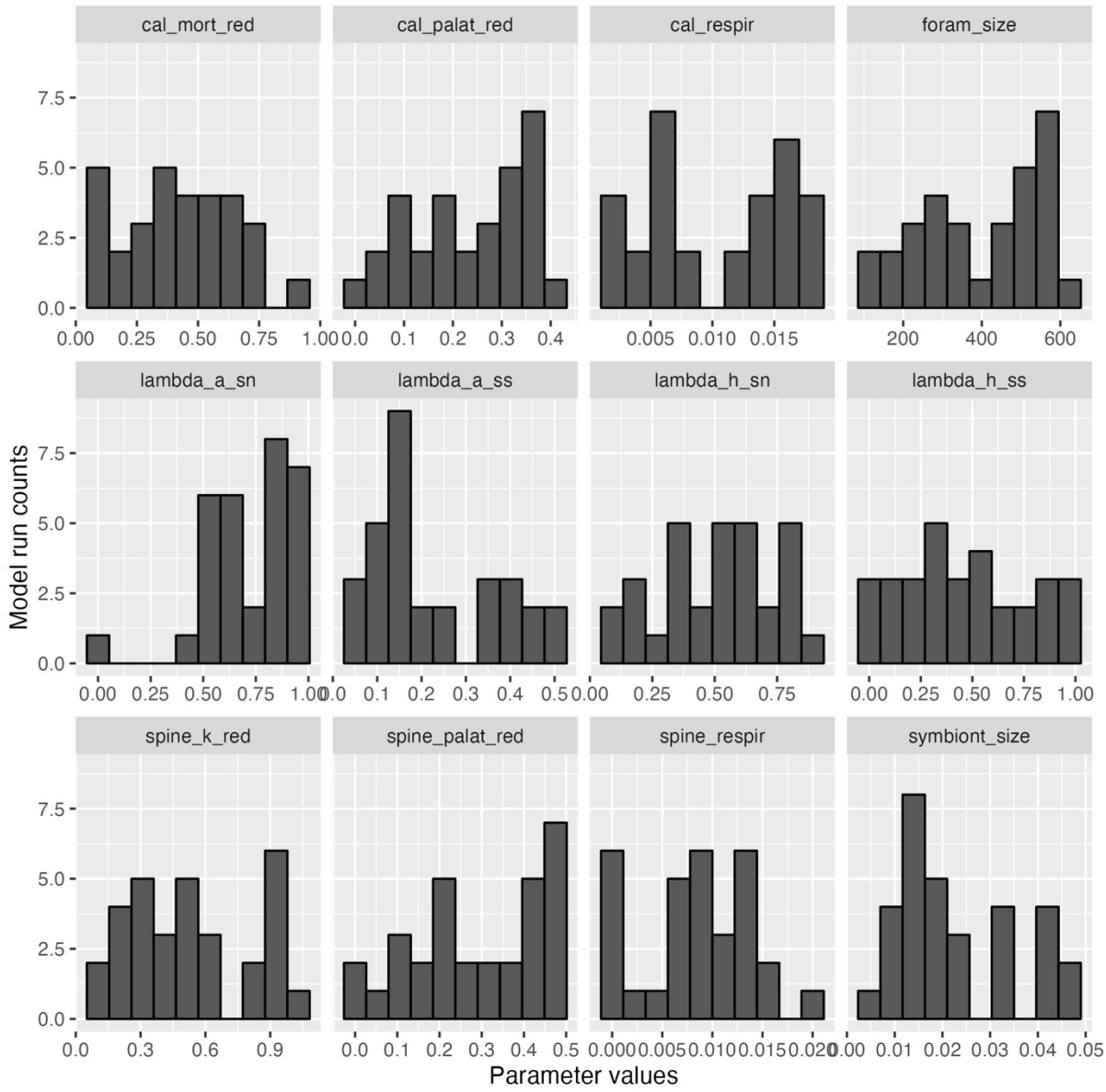


Figure S4. Same histogram as Figure S3 but associated with negative relative abundance M-score (<=-0.3, proxy of Cluster D in Figure 2).

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?

We apologise for the confusion. Here "POC export" and "POC flux" refer to the same quantity, and likewise "biomass" and "living biomass" are refer to the same quantity.

We now use these terms consistently throughout the revised manuscript.

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.

We agree it is not clear enough in the model-data comparison description.

We provide now more details about the export production depth (80.8m), the average trap deployment depth (-1960 m) in Line 325 and after.

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.

We addressed these issues in the above responses about net, trap and sediment samples. As mentioned above, the traps/net tows are relatively instantaneous, but we show both annual average and time series comparisons.

We changed the section headers for sections 7.3 & 7.4 to include "annual average biomass" and "annual average export".

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?

The M-Score for biomass, even with the changes we implemented following the reviewer's comments about unit conversion (see above), still has a relatively low value. This low M Score is despite a good agreement between global annual mean biomass data and model output, both in terms of geographical distribution and global mean range. We tested a method using the geospatial information of the observational points to match nearest model grid (i.e., point to grid) and calculated normalised root mean square. While this method results in a higher score than the grid-grid method we choose, such an approach is also giving undue weight to specific locations where the data is concentrated and therefore creating its own bias. We added text in section 7.3 to explain our approach and its limitations (copied here in Italic text).

We added more description and references to M-score in section 6.3 (see answer above).

Line 915: "The skill score, however, does not capture this good mode-data fit. This is mostly caused by regridding the data points into model grid resolution. The plankton net data are spatially concentrated in North Atlantic, North-western Pacific, Arabian Sea, and Indian

sector of Southern Ocean. Under such circumstance, re-griding causes sparser data and makes skill score sensitive to several outlier grids. Therefore, the insufficient data is likely the primary reason of low scoring in biomass."

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.

The reviewer raises some very interesting points on the link between increasing model complexity (i.e., functional ecology) and the fidelity of the modelled biogeochemistry. Adding foraminifera will impact two key biogeochemical fluxes: POC and $CaCO_3$ fluxes. We expect a minor impact on POC fluxes because foraminifera only contribute a small fraction of the total plankton biomass. However, associations with the dense $CaCO_3$ test, e.g., ballasting (Wilson et al., 2012) may alter this assumption.

The $CaCO_3$ fluxes of foraminifera tests is likely to impact biogeochemistry as foraminifera are estimated to contribute 23-56% of the total carbon flux (Schiebel, 2002). However, our model does not include an explicit representation of calcification as explained earlier. Secondly, our model does not include other major calcifying groups such as coccolithophores or pteropods (Daniels et al., 2018; Buitenhuis et al., 2019). Therefore, in our model the impact of $CaCO_3$ fluxes on biogeochemistry is limited to the dynamics of productivity via a fixed rain-ratio.

For these reasons, we have chosen not to expand on a quantitative comparison of biogeochemical variables. We have instead re-focussed the manuscript on the plankton ecosystem and associated fluxes such as productivity. We have removed text justifying the development of the model for improving biogeochemistry and have retained text reflecting the impact on biogeochemistry in the discussion (section 8) as a direction for future research.

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.

We have now added the NPP comparison in section 8 and figure 12. This analysis shows minor change of POC export with similar geospatial pattern and a small reduction mainly in

the subpolar regions. The similarity of model skill with prior models is likely because the biomass of foraminifera functional types is so small as mentioned before.

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).

Following the suggestion, we restructured the result section. Now it follows the "parameterisation result -> comparison with observations -> comparison with prior models" route. We also emphasize that we do not focus on how the inclusion of foraminifera diversity changes the nutrient cycles. This is an important direction but outside the remit of this study due to above mentioned limitations.

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?

We added a section specifically in comparing with prior models (section 8) and a more general discussion (section 10) to discuss how the increasing complexity of foraminifer/ecosystem is necessary.

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.

See our comments above in relation for many aspects raised here. As for the reason of low biomass and high export, we agree that the reviewer's suggestion of bias in observation is the source of error. In the modified manuscript, we use the correct unit conversion of sediment traps data as suggested by the reviewer and retuned the model as explained above. The biomass and export now compare well with net tow and traps in terms of annual average values. Because the model is calibrated against traps which were deployed in deep waters (average ~2km), and plankton net which likely did not capture the high production season (Line 545).

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.

We addressed this question above when it was raised before.

"We have added a histogram (Figure S3, pasted below) showing the optimal parameters (indicating the emergent distribution) and given a general discussion in the 7.1 Section and a more distribution-relevant discussion in 7.2. In brief, the foram size needs to peak in 100-200 um to resemble their prey distribution in high latitudes with abundant nutrients. The symbiont size is small (0-0.01 times the host size) so that their high nutrient affinity can help foraminifer survive in oligotroph gyres. The calcification respiration also peaks at highest bin to achieve better comparison with the observed low standing stocks."

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.

Yes, they are the same plot. We removed the first redundant one.

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?

The export production shown on the right of the figure is the global annual mean, which is averaged for four foraminifer groups. It shows the common feature of best cluster, i.e. the low export production.

We clarify the summed M-score in the figure caption, abundance is changed to relative abundance; we added the total M-score in the fourth column.

We find it hard to highlight the "best" parameter as this figure, as it summarises 1200 model runs in one plot.

We added some histograms in the supplementary information (Figure S3) showing our best parameter set.

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

Thanks for pointing out the importance of this figure.

We have added an NPP row; changed the anomaly to ratio, changed the "POC export" label to "ecosystem POC export"; moved the figure to the latter position and have some discussion about the potential reasons.

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)

We are sorry for confusion in terminology. Our model estimates relative abundance based on biomass or POC flux, and biomass and POC flux are highly correlated. We have removed this sentence in the caption and added one sentence in Line 378 to state this.

For the figure, the column title of this figure is now ForamECOGENIE 2 as suggested. We also changed the subplots title to global mean, added unit and moved the M-score to the left with functional group name.

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

We accepted the reviewer's suggestion to separate the model and data into two columns as the relative abundance figure. The others are same as Figure 4.

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?

The POC export unit is now converted as previously replied.

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?

The POC export unit is changed, and the heading is "globally integrated biomass" and "globally integrated EP".

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.

We changed the observations in the figure caption to "plankton net" and "sediment traps."

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?

We modified the table. However, we suggest that the low M-score is caused by the low-resolution data (Line 421 and after). The caption is changed to "The distribution of M-scores across foraminiferal groups from the optimal parameter set" akin to other parts of the paper where we now replace “best run" with "optimal parameter set”. We have provided the column and row sums.

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. .... 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).

We thank the reviewer for their discussion on the difference of multiple model types. In brief, the trait-based models allow the size spectrum to be continuous, using allometric relationship to determine physiological processes. Using this approach, higher size diversity can be resolved while keeping parameters at a minimum. In addition, PFT models rely on lab data to fit derived growth rates, while trait-based models can be developed and applied to taxa which are challenging to culture to derive physiological understanding like foraminifers. Finally, as pointed out by the other reviewer, PFT and trait-based models overlap, like here defining foraminifera 4 functional groups.

We added a new section (section 10) to introduce the different types of ecosystem models introducing NPZD model (not species-based models), PFT based models, and trait-based models and their strengths.

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

Yes. Isolating individual coccolithophores to monospecific analysis, given their size of a few microns, is challenging. Their organic remains are used as paleo proxies very successfully, though. Pteropods are rare and have a much lower preservation potential as their shell is formed by aragonite a form of calcite which is much less stable, resulting in a more limited use of this group for palaeoproxies.

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"

Thanks for the advice. We added a new sentence (Line 69) to connect the two parts of this paragraph.

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

We have changed the core top to "foraminifer sediment core-top census data" to make text clearer.

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

We changed this sentence to "spines extruding from the calcareous test" and add a note at the second sentences of introduction (Line 25) to explain that test is the synonym of shell.

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.

We have re-structured the paragraph to highlight the variable stoichiometry and how it influences the nutrient uptake rate (second paragraph in section 3.2)

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

Thanks for pointing out. We now use μ to represent uptake rate.

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.

Thanks for this advice. We now use τ to represent this coefficient.

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.

We have separated the sentence of sediment trap and plankton tows.

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

We wanted to express the higher score accompany lower absolute export values and have rephrased for clarity (Line 361)

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.

Yes, it is the summed M-score. We have clarified this in the text.

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?

We do not weight each parameter but choose better performance of relative abundance over the other two.

We have removed this sentence to avoid confusion.

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

We have replaced the "best run" with "optimal parameterisation" throughout the article.

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.

The modified result compares well with plankton net. The difference is same as previously stated.

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

The cited reference uses production for foraminifer biomass production.

We have removed this terminology.

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.

We apply "POC export" to avoid confusion and make a clear reference of carbon export from ecosystem or foraminifera only.

Line 414: You use two different references to cite the same range of CaCO3 export. Was that intentional? If so, why?

The reference should be both Schiebel 2002 Global Biogeochemical Cycle paper.

We have removed the wrong 2001 reference.

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

We removed the term as we rephrased the entire paragraph.

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

We changed the result as previously stated; consequently, this sentence was also rewritten.

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.

We rephrased the statement to clarify that for the seasonal time series part, we compared the peak time. While the model does not resemble the amplitudes of the sediment trap, the peak time in the model is largely consistent with data.

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.

We added an additional paragraph (Starts at line 540) to describe such limitations.

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?

We changed the global mean to global annual mean and added foraminifer-derived before the C export.

Typos and Other

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

abstract:

- **“increasing functional trait diversity and expanding their ecological niches**
- **“focusing on functional traits rather than individual species” should**
- **“observations from global core-tops, sediment traps, and plankton nets”**
- **“Our model approximates…, accounts”**
- **“19% of the global pelagic marine calcite budget which is within the lower”**

Intro:

- **“built an ecophysiology based dynamic model” ->” built ecophysiology based dynamic models”**

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

We thank the reviewer for taking the time to highlight these.

All suggestions have been implemented and the paper will be proof read before resubmission.

Line 44: This sentence is structured as if the model ‘reconstructed ’the future scenarios int the second clause. Perhaps revise to “…and simulated potential…”

Fixed

Line 70: “traits…lay down the foundation of a trait based model” is a bit of tautology

We changed "traits" to "observational studies" (as following).

"These observational studies of how functional traits affect biogeography and trophic activities lay the foundation of building a trait-based model."

Line 95: extra ‘and’

We have removed the 1st "and".

Line 174: Section title?

This is a formatting issue as the section title is shown in last page, not resolved.

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

We remove all the "flux rate" term.

Line 445: Should this be a header?

Fixed.

References

Buitenhuis, E. T., Quéré, C. L., Bednaršek, N., and Schiebel, R.: Large Contribution of Pteropods to Shallow CaCO3 Export, Global Biogeochemical Cycles, 33, 458–468, https://doi.org/10/gjpnzt, 2019.

Daniels, C. J., Poulton, A. J., Balch, W. M., Marañón, E., Adey, T., Bowler, B. C., Cermeño, P., Charalampopoulou, A., Crawford, D. W., Drapeau, D., Feng, Y., Fernández, A., Fernández, E., Fragoso, G. M., González, N., Graziano, L. M., Heslop, R., Holligan, P. M., Hopkins, J., Huete-Ortega, M., Hutchins, D. A., Lam, P. J., Lipsen, M. S., López-Sandoval, D. C., Loucaides, S., Marchetti, A., Mayers, K. M. J., Rees, A. P., Sobrino, C., Tynan, E., and Tyrrell, T.: A global compilation of coccolithophore calcification rates, Earth Syst. Sci. Data, 10, 1859–1876, https://doi.org/10.5194/essd-10-1859-2018, 2018.

Gregoire, L. J., Valdes, P. J., Payne, A. J., and Kahana, R.: Optimal tuning of a GCM using modern and glacial constraints, Clim Dyn, 37, 705–719, https://doi.org/10.1007/s00382-010-0934-8, 2011.

Hemer, M. A. and Trenham, C. E: Evaluation of a CMIP5 derived dynamical global wind wave climate model ensemble, Ocean Modelling, 103, 190–203, https://doi.org/10.1016/j.ocemod.2015.10.009, 2016.

Schiebel, R.: Planktic foraminiferal sedimentation and the marine calcite budget, Global Biogeochemical Cycles, 16, 3-1-3–21, https://doi.org/10/bdxfhs, 2002.

Watterson, I. G.: Non-Dimensional Measures of Climate Model Performance, International Journal of Climatology, 16, 379–391, https://doi.org/10.1002/(SICI)1097-0088(199604)16:4<379::AID-JOC18>3.0.CO;2-U, 1996.

Watterson, I. G., Bathols, J., and Heady, C.: What Influences the Skill of Climate Models over the Continents?, Bulletin of the American Meteorological Society, 95, 689–700, https://doi.org/10.1175/BAMS-D-12-00136.1, 2014.

Wilson, J. D., Barker, S., and Ridgwell, A.: Assessment of the spatial variability in particulate organic matter and mineral sinking fluxes in the ocean interior: Implications for the ballast hypothesis, Global Biogeochemical Cycles, 26, https://doi.org/10/gj35bn, 2012.