20/10/2015
Plymouth Marine Laboratory
Prospect Place
The Hoe
Plymouth
PL9 0AD

Dear anonymous referees #1 and #3, Dr. Robson, and Dr. Hargreaves,

Thank you for taking the time to read the paper and for your comments. After this introductory letter, a list follows where we address the referee comments in the order that they were received. Many of the comments relate to the overselling of the work in the title, so we've decided to revise the title to a more accurate description of the study:

> The assessment of a global marine ecosystem model on the basis of emergent properties and ecosystem function: a case study with ERSEM

We hope that this change will lead to less confusion for readers.

The majority of the requested changes from anonymous referee #1 were about toning down the optimism about the power of emergent property validation. We were happy to implement these changes. There were also some clarifications that were added. Unfortunately, we were unable to locate the appropriate Karl et al paper mentioned in point 7, but would be willing to amend our text if more details about this work can be shared with us.

We found the comments from the second referee, Dr. B. Robson, to be very positive. We're glad that you were happy that we could retain the current structure even though it is somewhat non-traditional. However, after going through the comments from anonymous referee #3, we decided that a restructuring would be of benefit anyway. Regarding the text seeming too dense, I feel that this is a regular issue with the Copernicus discussions format. The final GMD format has two columns per page and the images are placed appropriately near the text. To me, this makes it feels more spacious and clean, while reducing the page count. As the discussion section has also been expanded, sub-headings were added to the discussions section.

The comments of anonymous referee #3 were very in-depth, and we've done our best to address all of them, in the response section below. We're grateful for such a detailed review, and we feel that the resulting changes have resulted in a much stronger paper.

Once again, thanks for taking the time to read the paper, we are grateful for your honest opinions.

Sincerely,

L. de Mora
M. Butenschön
J. I. Allen

Anonymous referee #1:

General comments:

This manuscript presents several examples of how “emergent relationships” in model results (e.g., the fraction of chlorophyll contributed by diatoms, and the overall chl:C and POC:PON ratios) can be compared with data from observations to test the performance of plankton ecosystem models at the large scale. It argues that such comparisons provide a way of testing models that is more robust than the widely applied “point-by-point” direct comparisons of specific simulated values (e.g., nutrient and chlorophyll concentrations).
I find the comparisons meaningful and honest, in that they include cases in which the modeled emergent properties both agree and disagree with their counterparts from the observations. I also find that the authors make a compelling case that emergent properties should be used more as metrics for comparing model results against observation.

Thanks for your kind words, we're glad to hear that we've made a compelling case.

However, I do find that the authors are overly optimistic and tend to overstate the case that mere agreement of such emergent properties from the model with those from the observations constitutes “a strong indication that the model has a appropriate representation of the ecosystem functions that lead to the emergent relationship” (lines 6-8, in the Abstract). For example, the statement quoted above is at odds with the authors’ later statement that, “Many interacting parts of an ecosystem can affect the balance of diatoms chlorophyll against the total community chlorophyll:. . .” (p. 6098, lines 15-16). Given that so many factors may affect any such large-scale “emergent relationship”, it must therefore be possible to achieve similar emergent relationships using different model formulations, or different values of parameters for any given model formulation. In other words, some of the interacting parts may be made to counteract others in determining the overall emergent relationship. This is just another case of the well known problem that it is quite easy to get the correct modeled value for the wrong reasons, i.e., with many different incorrect parameterizations. Therefore, I recommend strongly that the wording be changed to not overstate the confidence that one might have that underlying model formulations are correct, merely based on the agreement of “emergent relationships”. I do think that testing models with such relationships is quite valuable and should be encouraged, but not over-hyped.

As mentioned above, we are happy to implemented these changes. The optimism of the first draft has been toned down, and the title has been changed. We have added a “getting it right for the wrong reasons” paragraph to the beginning of section 3:

> The use of emergent relationships as a tool to assess the quality of a marine biogeochemistry model is the central thesis presented in this work. However, there are some important caveats that should be stated first. Firstly, this method is not intended to replace current validation methods, but to complement them. The standard methods such as a point to point comparison should remain the first test for objectively determining model quality. Secondly, the emergence of a~coherent natural relationship in a~simulation is an indication that the model has an appropriate representation of the ecosystem functions that create the emergent relationship. However, the ability to reproduce an emergent property does not guarantee the accuracy of the model choices and parameterisation. Only through a thorough investigation of the model structure and behaviour can we know whether the model reproduces the emergent property for the right reasons. Nevertheless, we aim to demonstrate that the reproduction of emergent relationships by the model are a diagnostic tools which may be used to identify the origins of model strengths and weaknesses.

1. The wording in the abstract concerning the “strong indication”, as discussed above, should be revised not to overstate the case.

We changed:

> The emergence of a~coherent natural relationship in a~simulation is a~strong indication that the model has a~appropriate representation of the ecosystem functions that create the emergent relationship.

to:

> The emergence of a~coherent natural relationship in a~simulation is an indication that the model has a~appropriate representation of the ecosystem functions that create the emergent relationship.

We have also revised other uses of the word strong when describing the power of these tools.

2. The following sentence also overstates the case: “The ideal scenario regarding the model version of the ecosystem function is that that is is an emergent property of the model, and is not constrained or imposed in anyway.” (p. 6102, lines 2-4), in that the results of any deterministic model must in fact be constrained and imposed by the choice of equations and the values of parameters employed. I suggest changing to something like, “. . ., and is not purposefully and obviously imposed a priori by the choice of model parameterization.”

We've implemented the referees suggested text.

3. “in the absence of a causal relationship“ (p. 6102, line 27) should be changed to something like “a purposefully prescribed functional relationship”.

We've implemented the suggested text.

4. As in the abstract, the statement (p. 6103, lines 1-2) that, “The emergence of a coherent natural relationship in a simulation is a strong indication that the model has a appropriate representation of the ecosystem functions. . .”, specifically the word “strong” should be revised.

Removed the word "strong".

5. (p.6104, final paragraph): Was the model actually fitted (by tuning its parameters) to the data? I get the impression not. If this refers merely to the correlation from a model-data comparison, the wording should be changed to avoid confusion.

Changed wording from:

> Figure~\ref{fig:CSfits} shows the five fits of in situ community structure and the least squared fit of ERSEM to the three-population absorption model of \citet{Brewin2010}.

to:

> Figure~\ref{fig:CSfits} shows the five fits of in situ community structure \citep{Hirata2011,Devred2011,Brewin2012a,Brewin2014,Brewin2015a}. The fit of the ERSEM simulation to the three-population absorption model of \citet{Brewin2010} using a least squared fit is also shown. This fit was performed using model data after the simulation had been completed and did not influence the relationship of the plankton functional types during the simulation.

6. (p. 6111, line 19),“not a direct consequence” should be revised to something like “not an obvious and purposefully prescribed consequence of. . .”

We've implemented the referees suggested text.

7. Section 3.4: One possible reason for the discrepancy in the range of modeled vs. observed POC:PON ratios could be the contribution of dissolved organic matter (DOM) to observed POM, which as been documented by Dave Karl and co-authors in studies of the Hawaii Ocean Time-series.

We're unable to find the exact paper, could you please be more specific? Nevertheless, we're happy to add the following paragraph below p6120 L11, if we can find the appropriate Karl et al paper.

> In addition, it has been shown by \citet{Karl?} that measurements of particulate organic matter may be exaggerated by contributions from dissolved organic matter. Whereas the ratio calculated in the model does not include any contribution from dissolved organic matter.

8. (p. 6120, final paragraph). It does not make sense to state that the minimum dissolved inorganic phosphorus concentration is lower than the minimum of any ratio. This needs revision.

Good spot! We changed:

> Similarly to the nitrogen, the lower limit of dissolved inorganic phosphorus in the model is higher than the minimum inorganic phosphorus:carbon ratio observed in nature.

to:

> Similarly to the nitrogen case, the lower limit of the dissolved inorganic phosphorus to DIC ratio in the model is higher than the lower limit of the same ratio observed in nature.

9. (p. 6124, line 2). “strong” is an overstatement.

The entire discussions section has been revised, and this statement now appears as:

> The ability to reproduce this ratio from the combination of four phytoplankton functional types, three zooplankton functional types and 3 classes of particulate organic detritus, all of which have variable stoichiometry (except mesozooplankton), is a~sign of the validity of models parameterisation of the balance of carbon to nitrogen

10. (p. 6126, lines 11-12) “Most importantly, ecosystem functions are the only way to demonstrate the models capacity to represent ecosystem function. . .“ The tautology in the beginning of this sentence needs revision. Perhaps “. . .explicity consideration of ecosystem functions are. . .”.

We changed:

> Most importantly, ecosystem functions are the only way to demonstrate the models capacity to represent ecosystem function, as opposed to quantitative metrics of absolute ecosystem state.

to:

> Most importantly, an explicit analysis of the reproduction of ecosystem functions by the model is the only way to demonstrate the models capacity to represent ecosystem function, as opposed to a demonstration of quantitative metrics of absolute ecosystem state. None of the features shown here would be visible in a model against data comparison of static historical concentration measurements

Dr. B. Robson:

Paragraph 1: This is an excellent paper, representing, in my view, an important advance in assessment and evaluation of biogeochemical models. The authors have made a strong case for the utility of emergent properties and have provided carefully considered guidelines around how emergent should be considered in model assessment (some other examples that I have seen have fallen into the trap of calibrating the model to the properties being assessed, which can then no longer be considered "emergent" properties in the model). I am impressed also by the size of the datasets used to derive expected patterns in emergent properties: this has been achieved by drawing widely on the literature for the (very well studied) Atlantic ocean and beyond, and may provide a good starting point for those of us working in less intensively monitored parts of the world.

Thanks for praise, that is very kind. We are really glad to hear that this work may be of use for others that work in less studied regions.

Paragraph 2: The structure of the paper departs from the traditional Intro-Methods-Results-Discussion-Conclusion format and it is not clear that there was a need for this departure, however the present format is clear enough and given the major effort that would be required to restructure the paper, I am not recommending that this be done. The text is fairly dense, however, and could perhaps be made easier to read by including more subheadings. Also, while I know ERSEM is a well published model, it would save readers time if a diagram showing the conceptual structure of the model was included as an early figure.

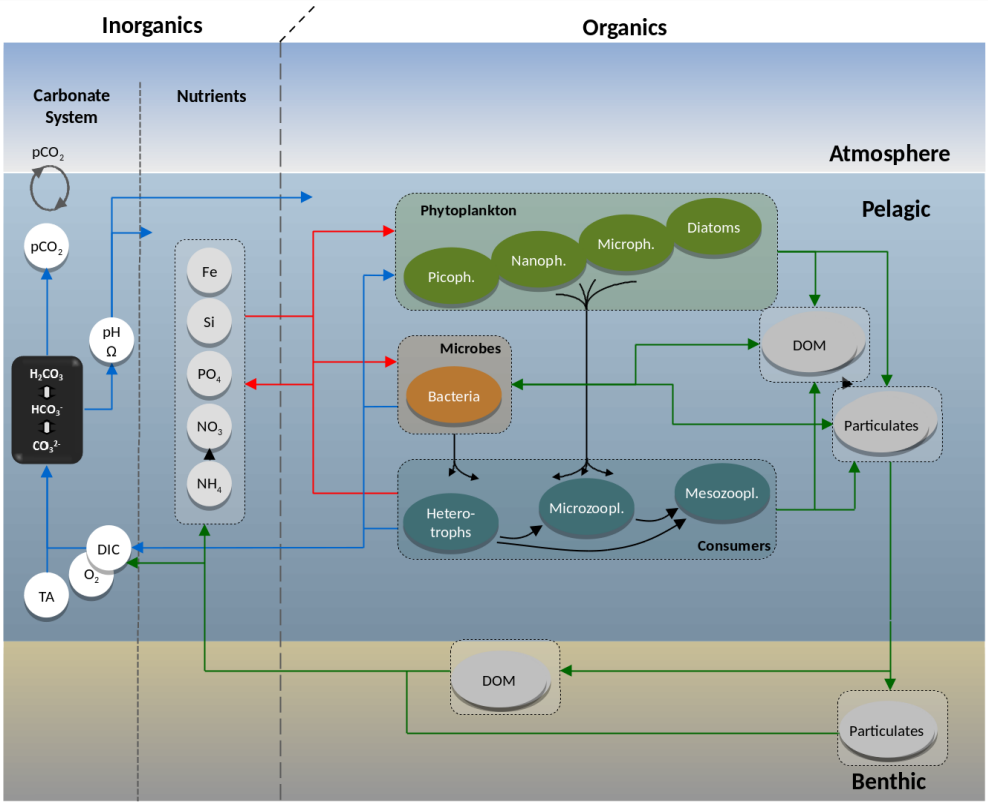
As a result of the changes requested by the third referee, we have changed the paper structure to the traditional structure with a more detailed discussions section.

Regarding the text seeming too dense, I feel that this is a regular issue with the Copernicus discussions format. The final GMD format has two columns per page and the images are placed appropriately near the text. To me, this makes it feels more spacious and clean, while reducing the page count. We also added sub-subheading to the discussions section.

We've added the following ERSEM conceptual structure image, and added the following to the model description:

> A diagram showing the major organisms, nutrients, chemical systems,
> organic matter and fluxes modelled in ERSEM ia shown in Fig.~\ref{fig:ERSEM}.

I also added a citation to the ERSEM 15.06 publication (currently in press with GMD) doi:10.5194/gmdd-8-7063-2015.



Caption: Schematic representing the major organisms, nutrients, chemical systems, organic matter and fluxes modelled in ERSEM. Blue connectors represent inorganic carbon, red represents nutrient fluxes, black represents predator prey interactions and green represents the non-living organics.

Paragraph 3: For future work, it would be good to see this sort of evaluation conducted in the context of testing hypotheses around bgc model structure. For instance, "X has suggested that bgc models need to include process Y to properly capture nutrient dynamics. If this is true, we would expect a model that did not include process Y to exhibit behaviour Z (e.g. systemically underestimating POC/DIN ratios)..."

This is a good example of future applications of this method, we've added the following text to the discussions section:

> In future works, it should be possible to use emergent properties to test hypotheses during biogeochemical model development. As an example, if some process X is expected to influence an emergent property Y, a comparison of the emergent property Y in the presence and absence of X may yield insight as to the value of that process in models. Similarly, a comparison of the emergent property Y in two models with alternative parameterisations for process X may facilitate the selection of a parameterisation.

Paragraph 4: Comments: p 6101, li 13-17: Is it appropriate to use the same parameters when presumably the models have slightly different structures, and hence the true biophysical meanings of even nominally identical parameters will be slightly different in the different models? (i.e. a parameter set that gives the best calibration for one model may not be the best for another, equally

valid, model).

I think some confusion was caused by our wording. The BGC models were free to use any parameterisations, only the physical model parameterisation was prescribed.
In order to make this point clearer, we're rewording

> " A unique requirement of this project was that the six biogeochemical models were run under identical physical conditions, using the same parameters and settings for NEMO, the same computing resource, the same coding framework and the same initial conditions for nutrients."

to:

> "A unique requirement of this project was that the six biogeochemical models were coupled to identical physical models. All six biogeochemical models were required to use identical model parameters and settings to run the physical component of the simulation, NEMO. Furthermore, all six models used the same computing resource, the same coding framework and the same initial conditions for nutrients."

Paragraph 5: p 6102, first paragraph: This is a very important point and should be highlighted, perhaps in conclusions or earlier in the introduction.

The "important point" in that line is:

> "As emergent relationships are large scale properties of the system, validation with emergence is possible even when there is very little or no appropriate data for the region under study, or when the physical circulation component of the model differs significantly from that observed in nature at the same location and time."

A summary of the lines 9-12 from page 6202 already appear in the abstract:

> Model validation with emergent properties is possible even when there is very little or no appropriate data for the region under study, or when the hydrodynamic component of the model differs significantly from that observed in nature at the same location and time.

And we added to the conclusion:

> As ecosystem functions arise independently of local physical conditions of the ocean, they can be used to demonstrate model quality in the case when the physical features of a sea are not co-located in the model and in nature. While these methods can not compensate for a catastrophic failure of the circulation model, these methods do overcome limited weaknesses in the ocean physics such as a displacement in ocean fronts or the mixed layer depth. These limited uncertainties have the potential of spoiling a point to point metric even though they may only be a minor error in the overall picture.

We also added the following sentence to the introduction (in context):

> In addition, the relationship that they observed was a widespread general response in all the plankton models that was independent of local hydrodynamic conditions. Furthermore, this relationship is a large scale property of the marine ecosystem, and is expected to hold true even in regions with few historical measurements. This relationship is important because it occurred independently of the hydrodynamic model, and because it reflected the functioning of the modelled ecosystem in a~way that would not be visible in a~simple point-to-point comparison of ecosystem state.

Paragraph 6: Minor technical edits: p 6101, li 16 (and elsewhere): "parametrisations" -> "parameterisations"

Changed all instances of "parametrisation" to "parameterisation".

Anonymous referee #3:

General comment: I would first like to apologise with the Authors for this late review and the late releaseof the comments on their manuscript. However, the reason of my delay is partly to attribute to the manuscript itself. I was very much captured by the title, which sounded definitely different from the typical works published by GMD and I thought there was really something new. Nevertheless, when I started reading, I was distracted by some overstatements, the several misuses of ecological and biogeochemical terms and by the lack of references to the most relevant literature on ecological functions, macroecology and stoichiometric concepts in biogeochemistry (see main comments below).

My initial perception was that the authors have learned about the functioning of marine ecosystems only through the model reality and that they only used references from the modelling literature, if not from their modelling group only. In addition, the English needs some care, words are doubled in many sentences ("is that that is is") and there are quite some repetitions of the same concept as well.

The choice of GMD implies that the presented methods or models are either linked to specific model versions of public availability, or have more general validity and should help other scientists in the application of new approaches or use innovative validation methodologies. I do not think it is the aim of the authors to demonstrate that this specific model (one of the various existing flavours) can do things that other models cannot, because this manuscript does not specifically address where this model is superior. Although I do realize that there exist two schools of thought, one that prefers to have the discussions done together with the presentation of the results, and one that prefers a dry presentation with a combined discussion of the presented results at the end, I do think that the way results are presented and discussed seem to be there only to state that the model is good and does what it should do.

But at last, I came to Sec. 3.4 and immediately the manuscript made sense to me; I perceived the underlying power of this work and the reason why it should be eventually published by GMD. That section is worth the whole paper and it is very well written. However, the entire manuscript in this current form is not ready and major changes are needed to streamline the concepts. I do hope that this review does not sound too negative, because it is indeed an encouragement to work further and contribute soundly to the science of biogeochemical modelling.

Thanks for your comments and especially thanks for the level of detail of your review. While it has taken some effort to re-work the paper, we feel that it has resulted in a stronger work. We have revised the title of the paper and hope that this will result in less over selling. These comments have allowed us to re-structure the paper such that it is not a simple "look at what our model can do" paper, but have refocused it so that it presents what the method can do.

Main point 1: The introduction is too much philosophical and much less methodological as it would be required by a journal like GMD. The authors do not make clear the distinction between biogeochemistry and ecosystem modelling; the use of concepts related to both disciplines is not always appropriate and the choice of literature is often limited to modelling papers from the same group of the authors. See detailed comments below for some examples of both.

We have re-worded the introduction to focus on biogeochemical modelling, as opposed to ecosystem modelling. Even though GMD is a methodological journal, we feel that a gentle introduction to emergent property validation by going over a previous example is still a valid way to begin the paper.

Main point 2: In particular, there is one main statement that the authors make that would deserve further (indeed philosophical) considerations. In Sec. 3 they state that “The emergence of a coherent natural relationship in a simulation is a strong indication that the model has a appropriate representation of the ecosystem functions that create the emergent relationship. If the emergent relationship is not seen in the model, this implies that the ecosystem functions that bring about the emergence are not correctly implemented in the model”. There is a kind of analogy to the Type I and Type II errors in statistics here. I think the authors should better justify why the emergence of an observed macro-ecological property in the model does guarantee that this is not a “false positive”, or maybe better expressed, that this is not “right for the wrong reasons”. This equivalence - same behaviour in ecosystem function means “all” underlying biogeochemical and ecological parameterizations are correct - is reported many times in the text, but I do not think the authors have provided enough evidence that this is a truism.

We have made specific changes to the text to highlight the exact conclusions that can be drawn, instead of the hazier wording used in the previous version. For instance, we changed:

> This indicates that the modelled phytoplankton may have an appropriate response to temperature, light and nutrients in much of the global ocean.

To:

> This indicates that the combination of temperature, light and nutrients in the modelled global ocean result in a reasonable and natural response in the modelled phytoplankton. This natural response in the phytoplankton indicates that the combination of model parameters describing the phytoplankton community in this simulation are a valid parameter set for the modelled physical and biogeochemical environment they inhabit.

In addition, we have added a caveats paragraph to introduce section 3.0, as well as a re-wording of much of the discussions section:

> The use of emergent relationships as a tool to assess the quality of a marine biogeochemistry model is the central thesis presented in this work. However, there are some important caveats that should be stated first. Firstly, this method is not intended to replace current validation methods, but to complement them. The standard methods such as a point to point comparison should remain valid tools for objectively determining model quality. Secondly, the emergence of a~coherent natural relationship in a~simulation is an indication that the model has an appropriate representation of the ecosystem functions that create the emergent relationship. However, the ability to reproduce an emergent property does not guarantee the accuracy of the model choices and parameterisation. Only through a thorough investigation of the model structure and behaviour can we know whether the model reproduces the emergent property for the right reasons. Nevertheless, we aim to demonstrate that the reproduction of emergent relationships by the model are a diagnostic tools which may be

used to identify the origins of model strengths and weaknesses.

Main point 3: A much more detailed discussion is required at the end, that collects all the points quickly mentioned when presenting the results. As it stands, the discussion is a summary of the same considerations made in the Results section but does not provide further insights or avenues for discussion. Some important points are not taken further (see detailed comments below).

We have beefed up the discussion section using the suggestions of the reviewer. It is now 9 pages long, instead of 4.

Main point 4: Are the emergent properties a concept of general applicability for the validation of any biogeochemical model or is it only specific to ERSEM? (as a sentence at page 6103, lines 12-13 seems to imply). The final section on code availability would become a more substantial added value to the manuscript if the data used to assess the ecosystem functions are made publicly available. I am not specifically referring to the tools to make the comparison (like the python scripts, which can actually be subject to a direct request to the authors as already stated in the manuscript). Most equations of the various statistical fits and ranges can actually be derived from the provided tables, but it would be very useful if the model data shown in the various figures and the Martiny et al distribution of Fig. 5 would also be provided in the author's website. This would allow other modelling groups to perform the same analyses and compare with the ERSEM model as a reference.

All tools have been applied to ERSEM, and some have also been applied to other models (MEDUSA, CSIRO-GBR), but the method itself should be applicable to any model, if a sufficient set of emergent properties can be identified.

The Martiny et al data is not mine to share and it was not included as part of their publication, but readers can contact the authors of that paper to request it.

I've packaged up the tools that I used and published them to our gitlab server:
https://gitlab.ecosystem-modelling.pml.ac.uk/ledm/gmd-2015-135

Also, I'd be happy to add the subsampled model data we used to the GMD website. However, as the full model data is a significant volume, so it might be better to share it upon request.

Main point 5: The concept of "independency from hydrodynamical models and physical conditions of the ocean" which is expressed in the introduction and in the conclusions should be better explained. The authors do see that some functionalities break in certain regions (both in model and observations) and, most of all, they have not demonstrated that the results are the same if single biogeochemical provinces are considered. It may be that the large spread is due to bad performances of the model in certain regions. I personally do not think that ecosystem functioning is independent of physical processes, though I do understand the concept that by pulling together data from various regions and only looking at macro-ecological properties we may overcome the limitations of global ocean models.

The independence from hydrodynamic models was mentioned multiple times in the manuscript. We interpret this as meaning that the emergent properties allow a validation of the biogeochemical model independently of the quality of the hydrodynamic model. We don't mean that the Biogeochemical model is independent of the ocean circulation model. Rather that these methods should allow a validation of the BGC parts of the model irrespective of the quality of the physical model. However, this can only compensate for limited weaknesses in the physical model. For instance, if the modelled ocean is entirely composed of storms and 10 degrees hotter across the board, then the problems with the physical model are probably going to be too great to use emergent properties to validate the BGC model. The point here is that these methods overcome limited weaknesses in the ocean physics like e.g. displacement of ocean fronts or the mixed layer depth, that have the potential of spoiling a point to point metric entirely while it's only a minor error in the overall picture. The following text was added to the conclusions:

> As ecosystem functions arise independently of local physical conditions of the ocean, they can be used to demonstrate model quality in the case when the physical features of a sea are not co-located in the model and in nature. While these methods can not compensate for a catastrophic failure of the circulation model, these methods do overcome limited weaknesses in the ocean physics such as a displacement in ocean fronts or the mixed layer depth. These limited uncertainties have the potential of spoiling a point to point metric even though they may only be a minor error in the overall picture.

We like the idea of investigating the independence of the BGC from the physics by comparing different regions (or physical models?). This is an ideal option for future extensions of this work and has been added to the conclusions.

> A second potential avenue for future research would be to test the independence of the biogeochemical model from the physical model by comparing the emergent properties from the same BGC model coupled against multiple physical models.

Details comments:

P6097_L27: Does this mean that matching “observables” does not imply a proper representation of ecosystem behaviour? Please explain

By itself, the point to point matching can not inform about the underlying machinery that causes the model to reproduce the observed data or about the interaction of the different parts of the model. Matches model and data informs only about how close the model is to the observables. To validate an entire model requires simultaneously investigating many parts and their interactions.

P6098_L4: The emergent property of “high-chlorophyll = diatom domination” should be backed up by references to the literature from real observations. The authors only mention Holt et al (2014), which is a modelling validation exercise, and then

say that this relationship is “seen in many in situ datasets” (L10). Some explicit references should be given. One may think that diatoms are the only functional group capable of large blooms in models because diatoms are among the most studied organisms; the behaviour of the other functional groups is less known and we are seeing diatoms emerging because the others are more inadequately parametrized.

There is some confusion here, as upon closer inspection, the Holt et al. Paper may indeed confuse the PFTs. They correctly compared modelled diatoms to the diatom component of community structure fit from Hirata2011, but they incorrectly included the microphytoplankton fit of Brewin2010 as a fit to diatoms, (and they didn't specify which line is which anyway).

However, the evidence that the fraction of the community chlorophyll that originates in the large phytoplankton size classes usually increases with total chlorophyll is uncontroversial, {Hirata2011, Devred2011,Brewin2012a,Brewin2014,Brewin2015a}.

We have changed the text to clarify these points. Furthermore, this introductory section is for illustrating the origin of this method by summarising the Holt et al results. In section 3.1, we investigate micro-phytoplankton, not diatoms.

p6099_L4-8: I would say complementary and not more valuable. Especially when data are scarce and scattered over large regions.

There was some confusion regard the meaning of this entire paragraph. We have changed it to make the idea more accessible:

> In addition, model validation with emergent properties can be more valuable as a model test than a~direct comparison of model to laboratory-based experiments, such as primary production bottle measurements. This is because emergent property validation can be a~large scale test of many combinations of factors, simultaneously testing the model in a~wide range of physical and biogeochemical environments, where as laboratory experiments can not simultaneously cover such a diverse range of settings.

P6100_L1: A web search of the ERSEM model gives the Baretta et al. (1995) paper as first reference. Why are the authors not mentioning this, especially if they say at the end that this shows the validity of a model initially meant for regional applications?

We have added Baretta1995 and Butenschon2015 to the ERSEM reference list.

Section 3: This is a key part of the manuscript and deserves some attention. I think the authors use quite some jargon and do not provide a clear definition of the terms they use. I do not understand if they implies a difference between “ecosystem functions” (the title of the section says function, used as singular) and “ecosystem functioning” as it was initially introduced in Sec. 1. Is there an

analogy with the state functions of thermodynamics? Moreover, they use properties of ecosystem and biogeochemical properties as interchangeable, but I think most of the properties they present are more related to biogeochemical considerations rather than ecosystem-based.

The text used the phrase ecosystem function when emergent property was meant, and used ecosystem model interchangeably with biogeochemical model. We have gone through section 3.0 and I hope that these misconceptions have been clarified. We changed the section title to :

> Emergent relationships in marine biogeochemical models

There have been many other changes to this section, and some of the changes listed below may have been moved, deleted or entirely re-written. Unfortunately, I'm not sure I understand the analogy with thermodynamics state functions.

P6103_L17: These sentences are full of errors or approximate concepts. I think phytoplankton cell sizes influence ecological and physiological processes, not ecosystem processes. Light absorption IS an internal process (I thought chloroplasts were in the cell) and nutrient uptake, metabolism and light harvesting are all physiological processes. Individual effects implies that they are different between individuals, which is not the case for unicellular organisms. What does it mean that phytoplankton function based classifications are also used? I thought this was a description of the properties considered in the functional group approach. This description mixes ecology, physiology and the specific choice of functional groups in ERSEM all in a bunch of sentences.

We rewrote the introduction to this section, with the aim of removing some of the approximate concepts.

> In modelling and observational marine biology, phytoplankton are often grouped into size or function based classifications Woodward et al. (2005). The size of a phytoplankter can directly influence a range of physiological processes, which combined together with other individuals can have an ecosystem-wide effect. Physiologically, a cells size may affect its nutrient uptake, metabolic rates, physiology and light absorption (Nelson et al., 1993; Stolte and Riegman, 1995; Huete-Ortega et al., 2012; Zhang et al., 2012). At the ecosystem scale, these individual-level effects combine to influence large scale observable properties such as the community primary production, export, the food web, and the light penetration depth into seawater (Riegman et al., 1993; Finkel et al., 2010; Huete-Ortega et al., 2012). In addition to size classes, phytoplankton function based classifications are also used in modelling and observational marine biology. Phytoplankton functional types allow a grouping of phytoplankton species by their role in the ecosystem, their preferred nutrient sources, and their production, export and sinking rates.

P6104_L4: Desirability? Do you mean palatability?

We change desirability for palatability.

P6104_L7-11: This sentence is a repetition (actually much better stated than the previous confused concepts).

This is hopefully less repetitive now, after cleaning up earlier section.

P104_L11-13:I presume none of the existing biogeochemical models has an explicit parametrisation of dominance.

The authors are not aware of any model which has an explicit parametrisation of dominance.

P6105_L1-4: This text is already in the figure caption

We have voluntarily included this information also in the text to improve readability of the section.

P6105_L17: This point recurs in the whole manuscript, related to Fig. 2, 3,4 and 5. The authors refer to density histogram, which is actually misleading. It is either a density distribution (i.e. normalized to 1) or a histogram distribution (i.e. counts). I presume they use 2-dimensional histograms as data numbers are shown for every bin. Also see the comment below for Fig. 5.

We changed density histogram to histogram distribution everywhere.

P6105_L18: Is this the same fit shown in Fig. 1?

Yes it is. We added the text:

> The fit to ERSEM, the Brewin 2015 and the Hirata 2011 lines are identical in Figs. 2 and 3.

P6106_L17: “more similar in shape”. Say in which part of the curve. I would say that it is more similar to Brewin

We added “At low chlorophyll concentrations,” to the start of the sentence: (now in the discussion section)

> In addition, the picophytoplankton chlorophyll as a fraction of the community behaves much more like the concave empirical fit to data seen in Hirata et al. (2011) than like the convex three-population model of Brewin et al. (2010) at low chlorophyll concentrations.

P6108_L3:I do not completely understand what the authors mean here. There is no need to discuss the shape of the model distribution as you only show the fits from the other satellite-based models.

It is a model-model comparison and I would limit the analysis to the fitted curves. I think the ERSEM fit follows the Brewin curve because of the constrain of fitting the picophytoplankton equation first.

These fits to data shown in this section are not satellite based. They are taken from in situ measurements, as such this is not a model-model comparison, but rather a fit-to-model vs. fit-to-data comparion.

What we're trying to say in lines 3-6 here is that the fits are performed with each point having equal weight, however, the underlying histogram distribution is plotted with a log scale. This visual difference between a linear-fit and a log scale sometimes makes the fit line looks like it doesn't match the model data very well, when in fact it does. I've changed the text to:

> In addition, the fit to ERSEM was performed such that each model point had equal weight, while the underlying histogram distribution is shown with a log scale. This means that the fits are logarithmically more influenced by the high data density regions in this figure. For these reasons, the fits may not appear to match the overall shape of the model distribution, while still being an acceptable fit.

P6108_L6: This final part of the section is a thorough discussion that should go in the Discussion section. It is linked to other results found in the next result sections and should be discussed together. What about the Southern Ocean where diatoms are usually dominating? Also, make clear if you discuss the fit or the distribution. I see no reason to comment the distribution as you only show the fit from Hirata and Brewin works.

Moved this section to the discussion section. Here is the new section 4.1:

> In the top panel of Fig. 3, there is a cluster of points where the diatom and large phytoplankton functional types unexpectedly dominate the community structure at low total chlorophyll. These points account for less than 0.2 % of the data, after the cuts described above were applied. Furthermore, they only appear adjacent to the excluded shallow and polar seas. While there is also some evidence of the proportion of large cell increased in the polar regions (Sosik and Olson, 2002), we postulate that it is a combination of multiple factors that creates this excess microphytoplankton. Firstly, in the polar regions there is an abundance of nutrients, and especially silicon, caused by excessive mixing in the physical model. Secondly, the model is parameterized to favour diatoms in low light regions. These factors collude to create an abundance of diatoms and large phytoplankton at low total chlorophyll. When the polar, shallow and inland sea regions are included in the model, the number of points included in these regions of the figure increases. As an example, the fit to the three population model was performed to all the model data from the top 40 m of the surface. The results of this fit are shown in the ERSEM (Top 40 m) column of Table 1. Relative to the Brewin et al. (2015) fit, the fit of the ERSEM model data to the three component model shows an overabundance of diatoms and large phytoplankton and an underestimate of picophytoplankton at almost all chlorophyll concentrations. Nevertheless,

it is important to stress that the three population community model is an appropriate emergent property for open ocean outside the polar regions.

Similarly, both Figs. 2 and 3 show that the model also has a modest surplus of diatoms and large phytoplankton at low chlorophyll concentrations in this simulation, which coincides with a low proportion of picophytoplankton. It is likely that this fault is caused by the same factors that cause the microphytoplankton PFT to dominate the community in a small number of cases. However, it is likely to be mitigated in most of the ocean by a lower silicate concentrations leading to slightly stronger silicate limitation for diatoms.

The model data was limited to the top 40 m of the surface ocean, and the relationship breaks down in Arctic waters in the model. Hints of the breakdown of the community structure appear in the in situ data (Brewin et al., 2015), but are seen clearly in the model. It became clear that the large phytoplankton and diatom functional types are in excess at low chlorophyll concentrations in the modelled community structure. In addition, the picophytoplankton chlorophyll as a fraction of the community behaves much more like the concave empirical fit to data seen in Hirata et al. (2011) than like the convex three-population model of Brewin et al. (2010) at low chlorophyll concentrations.

Despite these limitations, ERSEM was very successful at reproducing the overall shape of the community structure and natural balance of phytoplankton abundance between the four PFTs. This means that the combination of the nutrient affinity, growth rates, photosynthetic behaviour and predation rates ecosystem functions were modelled appropriately enough to bring out a natural emergent community structure. This is a robust and well known emergent property that can be reproduced successfully by ERSEM, despite the problems in the prescribed physical simulation.

P6109_L14-16: This sentence requires a reference

We removed this line.

P6109_17: Scientific usage? I think all usages are scientific in this context. A reference would seem appropriate here as well.

We changed this sentence to:

In remote sensing and modelling usage, this ratio also plays a~significant role in the calculation of phytoplankton biomass from ocean colour and in the modelling of primary production and \citet{Sathyendranath2009, Geider1997}.

P6109_L23: I would think that photoacclimation was not described first by Polimene et al. (2014). Previous literature should be considered, as for instance, MacIntyre, H., T. Kana, T. Anning, and R.

Geider (2002), Photoacclimation of photosynthesis irradiance response curves and photosynthetic pigments in microalgae and cyanobacteria, J. Phycol., 38, 17–38.

We referenced MacIntyre(2002) now.

P6110_L1: Please provide a reference for this sentence.

We referenced Geider1987

P6110_L8: What do you mean by mechanical?

We changed the text to:

> The direct measurement of phytoplankton carbon biomass is difficult because of the challenge of separating the phytoplankton from the other particulate organic matter \citet{Graff2015}.

P6110_L17: Data is plural

We changed was to were.

P6110_:27: Quantile. A quartile is 25%

We changed all instances of quartile to quantile.

P6111_L24: See comment above on density histogram

We changed density histogram to histogram distribution everywhere.

P6111_L245 Average of data or all the model levels within 40 m?

We changed the text to clarify it's all model levels in the top 40m.

P6112_L27: Define the extension of Arctic and Antarctic oceans (also for the previous section)

We added cut off coordinates:

> The model data is shown as a~logarithmically scaled two dimensional histogram distribution with in blue-scale, and the model distribution is taken as all model points in the top 40\,\unit{m} of the monthly climatology of the final ten years of the simulation, excluding

shallow seas, inland seas, the Southern Ocean from 45\degree South, and the Arctic from 55\degree North.

P6112_L23: This is a repetition of the previous concept

The concept is repeated for clarity and emphasis.

P6113_L1: I do not understand why this is a consequence of the previous analysis. See main comment 2.

We changed text to:

> This indicates that the combination of temperature, light and nutrients in the modelled global ocean result in a reasonable and natural response in the modelled phytoplankton. This natural response in the phytoplankton indicates that the combination of model parameters describing the phytoplankton community in this simulation are a valid parameter set for the modelled physical and biogeochemical environment they inhabit.

P6113_L3-4: You can only compare the fits. There may be points in the original distribution of Sathyendranath et al (2009) that also fall below the line. Actually, these points do affect the slope and you can comment on that. This should also be discussed further at the end.

I feel that we should discuss the group of points below the main body as well as a comparison of the fits. This grouping of points is what the eye is drawn toward in figure 4. These paragraphs have been moved into the discussions section and expanded. The discussion for section 4.2 now reads:

> Together, the particulate organic carbon to chlorophyll and phytoplankton carbon to chlorophyll ratios from Sect. 3.2 demonstrate that the phytoplankton biomass forms an appropriate fraction of the particulate organic matter. This means that the balance of producers to the rest of the organic matter, including zooplankton and detritus, is similar to nature over the range of observed total community chlorophyll.
>
> In both Figs. 4 and 5, there is a region where the phytoplankton carbon is much less than the Sathyendranath et al. (2009) fit, and this group of data appears to have a different slope than the bulk of the data. The phytoplankton carbon in the data points are almost entirely composed of diatoms and large phytoplankton, near the polar regions, and account for approximately 2.5 % of the dataset. Similarly to Sect. 3.1, this region highlights that the issues of the abundance of silicate caused by excessive mixing, the favouring of diatoms in low light regions and the relatively low grazing pressure on microphytoplankton from zooplankton at low phytoplankton biomass concentrations. When the polar, shallow and inland sea regions to a depth of 40 m are included in the analysis, the number of points included in these regions increases up to approximately 11 % of the dataset. This is another indication that either the model has not captured the behaviour of the high latitude regions, or the emergent property breaks down in these regions. Unfortunately, Sathyendranath et al.

(2009) did not include any data from Polar regions in the winter that could be used to test this hypothesis.

The carbon to chlorophyll relationship has many knock-on effects in the model: it influences the entire carbon cycle and has a huge impact on the calculation of total global primary production. When the model successfully reproduces the carbon to chlorophyll ratio in a global ocean simulation, this is an indicator that it has a good enough approximation of the roles of light, temperature and nutrient limitation in each of the plankton functional types. The fact that the model reproduces the natural range of behaviours of the carbon to chlorophyll ratio highlights that the ecosystem model functions appropriately over a range of environments. However, there may be other ecosystem functions that are currently absent in the model that could affect this ratio, such as the presence of higher trophic levels, anthropogenic detritus sources, or an improved model of photo-inhibition.

P6113_L14: I would say just “carbon cycle”. Carbonates in the ocean are one of the components of the carbon cycle, not a separate one.

We removed the word: carbonate.

P6114_L13: This is now an empirical probability density function and it is called histogram. Why the tick labels are hidden? The distributions should share the same scale if the areas are normalized to 1.

I would argue that it is not an PDF, on account of the data shown here is still discrete, not continuous, and it shows modelled and observed events, instead of a probability. As such, it is a frequency distribution. We changed the text accordingly and added y axis ticks and labels to the plot.

P6115_l16: This statement is also linked to the main comment 2 and should be put in the discussion. Why would you exclude the possibility to get an overall acceptable relationship for dysfunctional reasons? See for instance Flynn, K. J. (2010), Ecological modelling in a sea of variable stoichiometry: Dysfunctionality and the legacy of Redfield and Monod, Prog. Oceanogr., 84(1–2), 52–65.

This was moved to the discussion. The statement in question is:

> None of the detritus fields have any limitations on their stoichiometric variability. This means that the models POC : PON ratio can vary according to local conditions and predation, and the overall particulate organic matter stoichiometry in ERSEM is not susceptible to tuning via a small number of parameters. In order for all of these interlocking and competing components to reproduce the POC : PON ratio variability in the global ocean, it requires all the phytoplankton functional types, zooplankton functional types and

detrital fields to be balanced and healthy.

We added the following to this statement:

> One downside of the simultaneous analysis of all PFTs is the introduction of the risk of compensating errors. This occurs when errors in the ratio from one group is balanced by a similar but opposite error in another group. For this reason, we recommend that model validators also check each the C:N ratio for each PFT individually.

P6115_L23: I cannot understand the reason of this comment here? Seems like a fragment from another discussion.

This comment was removed.

P6116_L1-9: This sentence should also go in the discussion. The issue of grid resolution is pertinent to all the analyses done and not only to the POC:PON ratio.

This paragraph was moved to the discussion.

P6116_L10: It is not clear why the Gaussian distribution is mentioned here. Parameter estimation is clearly a function of the sample size, but 40k data are usually sufficient to capture the major shape of the distribution. I think the authors could do an analysis of skewness and kurtosis if they want to qualify the differences in the distributions. Or just limit the comment to the one at lines 17-18.

We removed this text about distribution shape vs number of data.

P6116_L19: Does this happen because of model parameter constraints? Also, I cannot clearly see that excess POC:PON is better captured. measuring the skewness would probably help to quantify this.

It's certainly not a direct constraint, as there are no limits on the C:N ratio in most of the fields that constitute the POC/PON ratio. Changed wording to reflect this.

We also removed the reference to the excess POC:PON, removing the need to discuss skewness.

The new discussion section is now:

> Figure 6 shows a comparison of the ratio of particulate organic carbon to particulate organic nitrogen in the model and in an distribution of in situ measurements. In ERSEM, none of the detritus fields have any limitations on their stoichiometric variability. This means that the models POC : PON ratio can vary according to local conditions and predation, and the overall particulate organic matter stoichiometry in ERSEM is not susceptible to tuning via a

small number of parameters. In order for all of these interlocking and competing components to reproduce the POC : PON ratio variability in the global ocean, it requires all the phytoplankton functional types, zooplankton functional types and detrital fields to be balanced and healthy.

While the model captures the central tendency of the in situ data, it does not capture the range of observed POC : PON ratios or the shape of the distribution seen in the data. The model underestimates the frequency of POC : PON ratios below 5 and above 6.5. It is possible that some of this difference is due to spatial bias and uneven sampling of the in situ data. In that case, it may be possible to capture more of the shape of the Martiny et al. (2013) data by sub-sampling the model data to match the distribution used to produce their data. However, the model data shown here is a monthly mean of a 1 ◦ by 1 ◦ square of ocean. The variability seen when taking an instantaneous measurement of the concentration in a bottle of seawater will always be more extreme that the mean value of a 1 ◦ by 1 ◦ square of ocean. Secondly, in this work, we attempt to validate the models ecosystem function over a large scale without the use of point-to-point matching.

The model's narrow POC : PON distribution is reflected in its the standard deviation (0.61), which is much lower than that seen in data (2.46). However, the range of stoichiometric variability seen in measured POC : PON data is underestimated in the model. Furthermore, there are precisely zero model data with a POC : PON ratio below 4.3 or above 16.5, whereas the Martiny et al. (2013) data ranges from 2. to 20. This behaviour in the model is linked to the fixed maximum luxury buffer of nitrogen relative to carbon in all the phytoplankton functional types. This maximum nitrogen buffer translates to a fixed minimum value of the POC : PON ratio, which is maintained as it cascades through the trophic levels. On the other end of the scale, there is a minimum requirement of nitrogen to carbon, below which the phytoplankton are nitrogen limited and do not grow.

One downside of the simultaneous analysis of all PFTs is the introduction of the risk of compensating errors. This occurs when an error in the C : N ratio from one group is balanced by a similar but opposite error in another group. For this reason, we recommend that model validators also check each the C : N ratio for each PFT individually.

While the model does not reproduce the standard deviation or the tails of the distribution seen in data, the ERSEM simulation was particularly successful at reproducing the mode of the Martiny et al. (2013) POC : PON ratio. This means that the most common values in the modelled POC : PON ratio are the same as the most common values in the in situ measurement of POC : PON. The reproduction of the mode of the dataset by the ERSEM model is a strong indication that the most common behaviour of the POC : PON relationship is appropriately simulated. The ability to reproduce this ratio from the combination of four phytoplankton functional types, three zooplankton functional types and three classes of particulate organic detritus, all of which have variable stoichiometry (except mesozooplankton), is a sign of the validity of models parameterisation of the balance of carbon to nitrogen.

P6116_L29: common, not comment

Good spot, we changed it.

P6117_L5: Why only in modelling? I think the authors should add some references here, as for instance taken from Sterner, R. W., and J. J. Elser (2002), Ecological stoichiometry: the biology of elements from molecules to the biosphere, Princeton University Press, Princeton, NJ.

We reworded this text and added citation to \citet{Sterner2002}:

> Stoichiometry is the balance of each element in organisms and in the ecosystem \citet{Sterner2002}.

P6117_L6: Redfield connected this ratio to the one found in particulate organic matter, thus linking its origin to living organisms.

We **c**hanged text to rephrase Redfield contribution.

> Redfield observed co-variability in the concentrations of dissolved nitrate and phosphate in seawater and in the composition of plankton, \citep{Redfield1934}.

P6117_L27: Michaelis

We changed the spelling**.**

P6118_L26: It is not completely clear to me the reason for this comment here. It would certainly be pertinent in an overall discussion of the validation method, that demonstrates how all the ecosystem functional properties analyzed here are actually neglecting (or better making implicit) the role of bacteria which is instead considered in models of the ERSEM type.

We removed this paragraph.

P6119_L4: better add “dissolved” detrital fields

We removed this paragraph.

P6119_L15: This is a very interesting original contribution that extends the work done by Moore at al. (2013). Did you use the spatial max and min of the climatological distribution? Please make sure that the figure caption report that this is not an estimate from Moore et al. paper but it’s your own original work.

Added to figure 6 caption:

> The typical in situ value and observed range from \citet{Moore2013} are shown as square markers with horizontal bars. The vertical bars associated with the square markers were calculated from World Ocean Atlas data\citep{Garcia2010} and GEOTRACES \citep{Henderson2007}.

Changed the “Moore et al” column to “Observations” in table 4, and added notes indication the data origin.

P6120_L4: I am a bit confused here with these ratios. Fig. 6 shows N:C and the original paper showed C:N. I think you mean that their lower cut off value (in their original figure) was 2.0, not in Fig. 6.

We changed the wording, hopefully this will be clearer:

> For instance, Moore 2013 cite an observed maximum phytoplankton N:C quota of 0.169, but the Martiny 2013 dataset has observational data all the way down to their cut off point, which was a C:N ratio of 2:1. Note that Moore 2013 used the N:C ratio, but Martiny 2013 used a C:N ratio.

P6120_L14: This is also a bit confused. Please make clear from the beginning (Pag. 6119, when presenting the Moore estimates and your original contribution for inorganic ranges) that the reported ranges are a combination of existing literature values and additional estimates .

I hope this is clearer now, given the explanation in the previous point. This section has been moved to the discussion section.

P6120_L18: You use the word “observed”, but according to your previous discussion there ratios are estimated because computed from ranges of numerators and denominators that are not necessarily correlated.

We replaced “observed in nature” with “estimated from the WOA dataset”, here and elsewhere:

> The model appears to have a~fixed lower limit of the dissolved inorganic nitrogen:DIC ratio that is higher than the minimum nitrogen:carbon ratio estimated from the WOA dataset.

P6120_L20: I think this should be expanded in the discussion

The original text was:

> This is a problem with the model parametrisations that has also been seen in Sect. 3.3 which will need to be addressed in future parametrisations.

We moved to section 4.4 of the discussion, which is much larger now:

The stoichiometric variability of particulate organic matter against the ratio of inorganic nutrients to DIC was shown in Fig. 7. This figure showed the typical organic nutrient : carbon ratio on the x axis against the typical inorganic nutrient : carbon ratio on the y axis. In addition to the typical values and the models distributions, this figure also showed a range of values that have been observed for each element. This figure demonstrated that the range of stoichiometric behaviours present in the model match those measured in situ. The ratios of the inorganic nutrients to carbon in the model do not typically get at low as the estimates of this ratio, except iron. This is a problem with the model parameterisations that has also been seen in Sect. 4.3 which will need to be addressed in future parameterisations. This is an example of the use to emergent property validation as a tool to direct future model development efforts.

The ratio of nitrogen to carbon is shown in blue in Fig. 7. The model captured the mean organic N : C ratio, but had a wider range of values than that quoted by Moore et al. (2013). However, the maximum value of the N : C ratio has been extended from 0.169 to 0.5 after the Martiny et al. (2013) results have been included. The model underestimated both the mean inorganic ratio and the range of variability in the inorganic N : C ratio. The model appears to have a fixed lower limit of the dissolved inorganic nitrogen:DIC ratio that is higher than the minimum nitrogen:carbon ratio estimated from the WOA dataset.

The mean organic phosphorus : carbon ratio is overestimated by the model, but the mean inorganic P : C ratio is underestimated. The range of the inorganic and organic P : C ratios were underestimated by the model relative to the Moore et al. (2013) data. However, both the organic and inorganic phosphorus in the model show a wide range of behaviours, reflecting those seen in nature. Similarly to the nitrogen case, the lowest values of the models dissolved inorganic phosphorus to DIC ratio is higher than the lowest values seen in the estimated DIP : DIC range. This means that the inorganic phosphorus in the model never gets as depleted as the minimum observed ratio estimate. However, the range of the inorganic P : C ratio from the WOA estimates has no indication of the frequency distributions of the P : C ratio range, so it is entirely possible that this extremely low inorganic P : C ratio is also relatively rare. In addition, the model data is the mean of a 1 ◦ by 1 ◦ patch of ocean, whereas the observational data originates from the mean of a one litre bottle, which would imply that we can expect fewer extreme values in the model. The modelled P : C ratio also illustrates the effect of external resource limitation: as the inorganic P : C ratio decreases, the organic P : C ratio simultaneously decreases. This figure can not indicate whether the drop in the organic P : C ratio occurs in the phytoplankton, zooplankton, or detritus, or in some combination of all three groups. Due to the trophic cascade of the POC : PON ratio described in Sect. 4.3, we postulate that this effect is caused by the modelled phytoplankton becoming nutrient stressed in low phosphorus environments.

The ERSEM model has four pelagic silicon fields: diatom silicon, inorganic silicate, and medium and large detritus silicon. The Si : C ratio in Moore et al. (2013) and in Fig. 7 are strictly limited to diatoms; there are no quotas associated with particulate organic silicon in other components of the ecosystem shown here. The modelled ratio of silicon to carbon,

shown in purple in Fig. 7, captured the range of variability in the inorganic version of the ratio. The range of the dissolved inorganic silicon to DIC ratio was estimated from World Ocean Atlas, Garcia et al. (2010), and is shown as a vertical purple dashed line in Fig. 7. As only diatom silicon are included in this figure and ERSEM has very little variability in the silicate stoichiometry for diatoms, there is no variability in the organic component for silicate. There is only a very slim range of Si : C ratios allowed in the modelled diatoms because the diatoms Si : C quota is set close to the optimal Si : C quota. External silicate limitation directly reduces carbon assimilation and respiration losses may cause small fluctuations in the diatoms Si : C quota (Butenschön et al., 2015). This means that the nutrient stress effect seen in the P : C ratio is not seen in the Si : C ratio. Instead of regulating the internal quota at low inorganic Si : C ratios, diatom growth is reduced and the community structure changes to disfavour diatoms.

The mean organic iron : carbon ratio in the model is lower than the same ratio in Moore et al. (2013): the model underestimated the mean organic ratio by an order of magnitude. However, the Fe : C ratio is the only inorganic nutrient:carbon ratio where the model captured the estimated range. While there is an atmospheric iron source from dust, the model does not include any atmospheric, riverine or hydrothermal sources of nitrogen, silicon or phosphorus. The nitrogen, silicon and phosphorus shown in this paper have been circulated, consumed and recycled for more than 100 simulated years and the relationship between organic and inorganic nutrients, and nutrients against carbon are still all representative of nature. On the other hand, the iron cycle is nudged towards what is observed in nature by an climatological surface deposition, and through hydroxide precipitation and saturation removal of excess iron. This means that the distribution of inorganic iron is not an emergent property of the model, but rather a tuned outcome. These nudges are needed because the iron cycle is ERSEM is much less complex than that seen in a nature. An example of a more natural iron model is Tagliabue et al. (2009), which has three bio-available forms of iron and two complexed forms of iron. Despite this, as the inorganic Fe : C ratio in our model decreases, the organic ratio also decreases, indicating that the organic community become increasingly nutrient stressed in low iron environments.

Anthropogenic nutrient loading is expected to increasingly influence nutrient cycles in the ocean and this may lead to shifts in the nutrient balance, (Paerl, 1997). Unfortunately, this model does not include any anthropogenic nutrient loading, or indeed any flux of riverine nutrients, and the iron dust deposition is forced with an annual climatology.

Overall, Fig. 7 informs about the relationship between the inorganic and organic component of the stoichiometric balance. Effectively, it illustrates whether nutrient limitation and nutrient stress are parameterized in a way that reflects nature. Much of the modelled organic matter appears to be iron poor and phosphorus rich relative to nature. The model never captures the lowest dissolved inorganic nitrate, phosphate, or silicate concentrations. It might be expected that the model will produce a wider range of quotas than the historic datasets as the ocean is vastly under-sampled relative to the model. On the other hand, the model is the mean of a 1 ◦ by 1 ◦ patch of ocean, and the data is typically the mean of a one

litre bottle, which would imply less variability in the model. Furthermore, some of the in situ data may originate in coastal datasets which have a higher spatial variability than would be seen in a coarse global model.

P6120_L21: Does this mean that the organic component retain more P? I think this deserves discussion because it may be linked to the “coastal” origin of the ERSEM model, where P is usually the limiting nutrient. I would link this to the discussion suggested above.

This text is:

> The dataset clearly illustrated the effect of external resource limitation: as the inorganic ratio decreases, the organic ratio also decreases, indicating that the phytoplankton become increasingly nutrient stressed in low phosphorus environments.

While re-working this text, we noticed that nothing can be said about the phytoplankton stoichiometry when looking at POC:PON ratio due to the risk of compensating errors. This section has been moved into the discussions and changed to:

> This figure does not indicate whether the drop in the organic P\,:\,C ratio occurs in the phytoplankton, zooplankton, or detritus, or in some combination of all three groups. [...]

see above for full text of the discussion section.

P6121_L1: This sentence is not completely clear to me. Are you referring to the model or observation data set? Also, I see this occurring only below a certain value of the inorganic range, and for Fe and P. This definitely deserve some more discussion in a later section.

This line is:

> However, the inorganic range shown in figure has no indication of the frequency or locations of such low concentrations.

Which was replaced with:

> However, the range of the inorganic P : C ratio from the WOA estimates has no indication of the frequency distributions of the P : C ratio range, so it is entirely possible that this extremely low inorganic P : C ratio is also relatively rare. In addition, the model data is the mean of a 1 ◦ by 1 ◦ patch of ocean, whereas the observational data originates from the mean of a one litre bottle, which would imply that we can expect fewer extreme values in the model. The modelled P : C ratio also illustrates the effect of external resource limitation: as the inorganic P : C ratio decreases, the organic P : C ratio simultaneously decreases. This figure can not indicate whether the drop in the organic P : C ratio occurs in the phytoplankton, zooplankton, or detritus, or in some combination of all three groups.

P6121_L10: What is the dashed line in Fig. 6? It seems purple, so is that related to Si? Are diatoms allowed to store Si internally in the cell? How is this variability regulated? Please refer to the ERSEM equations when possible.

The vertical dashed line is associated with silicon. It was necessary to visually separate the silicon and nitrogen as they have an identical mean organic ratio, but different inorganic ranges. The silicon discussion section has been moved into the discussions section. The discussion section now includes:

> The range of the dissolved inorganic silicon to DIC ratio was estimated from World Ocean Atlas, Garcia et al. (2010), and is shown as a vertical purple dashed line in Fig. 7. As only diatom silicon are included in this figure and ERSEM has very little variability in the silicate stoichiometry for diatoms, there is no variability in the organic component for silicate. There is only a very slim range of Si : C ratios allowed in the modelled diatoms because the diatoms Si : C quota is set close to the optimal Si : C quota. External silicate limitation directly reduces carbon assimilation and respiration losses may cause small fluctuations in the diatoms Si : C quota (Butenschön et al., 2015). This means that the nutrient stress effect seen in the P : C ratio is not seen in the Si : C ratio. Instead of regulating the internal quota at low inorganic Si : C ratios, diatom growth is reduced and the community structure changes to disfavour diatoms.

P6121_L16: Why do you use the term deficit? I think the model is just following the allowed rules, that is varying between the minimum structural ratio and the maximum luxury storage (if considered). This implies that there is a discrepancy between the parametrised values and the ones found in nature. I do completely agree that the final surface iron concentration is a tuned outcome of the combination between atmospheric deposition and scavenging rates, and this should be reported in the discussion and conclusions. The authors are not aware of this but there is an upcoming new paper by Tagliabue that does show that all the existing global ocean iron models have the same surface iron distribution but completely different combinations of input and scavenging rates. This means that their analysis have been capable to find this!

The original text was:

> The organic iron:carbon has a deficit in the model, relative to the same ratio in Moore et al. (2013): the model underestimated the mean organic ratio by an order of magnitude. However, the Fe : C ratio is the only inorganic nutrient:carbon ratio where the model captured the measured in situ range.

We have replaced the phrasing around the world deficit, but also moved this to the discussion section.

> The mean organic iron : carbon ratio in the model is lower than the same ratio in Moore et al. (2013): the model underestimated the mean organic ratio by an order of magnitude. However, the Fe : C ratio is the only inorganic nutrient:carbon ratio where the model

captured the estimated range.

We also look forward to reading the upcoming Tagliabue paper.

P6122_L5-19: I think this discussion on the role of external sources should be in the "Discussion"! It is general and not only linked to this section

This section has been expanded and moved to discussion:

> While there is an atmospheric iron source from dust, the model does not include any atmospheric, riverine or hydrothermal sources of nitrogen, silicon or phosphorus. The nitrogen, silicon and phosphorus shown in this paper have been circulated, consumed and recycled for more than 100 simulated years and the relationship between organic and inorganic nutrients, and nutrients against carbon are still all representative of nature. On the other hand, the iron cycle is nudged towards what is observed in nature by an climatological surface deposition, and through hydroxide precipitation and saturation removal of excess iron. This means that the distribution of inorganic iron is not an emergent property of the model, but rather a tuned outcome. These nudges are needed because the iron cycle is ERSEM is much less complex than that seen in a nature. An example of a more natural iron model is Tagliabue et al. (2009), which has three bio-available forms of iron and two complexed forms of iron. Despite this, as the inorganic Fe : C ratio in our model decreases, the organic ratio also decreases, indicating that the organic community become increasingly nutrient stressed in low iron environments.

Discussion: This is a summary, not a discussion. I think that it is an overstatement to say that many of the features seen here would not be visible in a flat comparison of model and data. This analysis is indeed powerful, but I see it as complementary to the other analyses (see main comment 2). There are other points that needs to be expanded, including some of the discussion that has been already done at the end of the result presentations and that are common to more than one emergent

property. Some more specific comments follow

Many of the suggestions in this review are about moving contents into the discussion, so this section now looks and reads very differently. It became so big that we even needed to add subheadings to the discussion section! We have toned down the statements about the power of this method, and characterised the sentence about the features being visible in point to point methods:

> Furthermore, many of the features seen here, such as the community structure, would not be explicitly apparent in a~point-to-point comparison of model to observations.

We also added a final section to the discussion:

> **4.5 The role of emergent properties**

Combined together, these relationships have informed about community structure and balance of C : Chl in phytoplankton, the ratio of POC : PON in particulate organic matter, the stoichiometric flexibility of POM and dissolved inorganic nutrients. While some selection cuts have been necessary to reduce the impact of non-physical behaviour, the combina tion of the relationships can be used to validate the ecosystem model without relying on the model to reproduce an historic measurement at exactly the right place and time. These methods should allow a validation of the behaviour of the BGC model irrespective of the quality of localised features of the physical model. In this way, these tools compliment current validation methods, such as the point to point, that may not function ideally in an inappropriately parameterized physical ocean model. Furthermore, many of the features seen here, such as the community structure, would not be explicitly apparent in a point-to-point comparison of model to observations.

While it has since expanded beyond its original remit, the European Regional Seas Ecosystem Model was originally built as a model for simulating temperate shelf seas. This work confirms the work of Vichi et al. (2007) by demonstrating using emergent properties that many of ERSEMs design choices and parameterisation are still appropriate in a global context. Furthermore, these emergent relationships were not explicitly parameterized in model development; all of them arise naturally out of a combination of many other ecosystem functions.

The combination of these well known phenomena have allowed a test of the majority of the modelled fields throughout the surface ocean. However, these relationships do not cover all aspects of the model. They do not inform about the food web such as the balance of zooplankton and detritus functional types to each other and to the phytoplankton functional types, or about the bacterial community or benthic environment in the model. Also these relationship do not cover important fluxes such as primary production, air-sea gas exchange, or export from the surface layers.

The nutrient cycles of carbon, nitrogen, phosphorus and iron are all influenced by the bacterial loop. The bacterial biomass does not contribute their biomass to the calculation of particulate organic matter used in sections 3.3 or 3.4, but the bacterial functional type competes with the phytoplankton for the inorganic nitrogen and phosphorus. The bacteria is an additional food source for zooplankton whilst also competing with the zooplankton by scavenging non-silicon particulate detritus. In addition, the bacterial group only excretes to the dissolved organic matter detrital fields. This means that the bacterial functional type wields influence over a significant portion of the pelagic model. Unfortunately, this work could not locate an emergent property to investigate the bacterial behaviour in the model. Some potential avenues where emergent properties may be found in the future include the ratio of primary production to bacterial production or the bacterial growth efficiency.

The models grazers biomass were implicitly included in the POC : PON, POC : Chl and the stoichiometric relationships. However, there are no metric included here to study the zooplankton by themselves. The authors are not aware of any stable global emergent

property for describing the grazer community.

These emergent relationships were selected to reduce the impact of spatial biases, but these relationships may still be influenced by uneven in situ data spatial and temporal coverage. This bias could potentially be resolved by using a point to point analysis for the emergent properties, however this may limit the scope and the power of the emergent property validation. These emergent properties also require the assumption that the property can be extended to cover the entire ocean. Some emergent properties have not been tested in shallow or high latitude seas, and may not hold across all marine environments.

P6123_L15-30: how is the behaviour of diatoms connected to the ratios shown in Fig. 6? You should give more context here because it is difficult to get back to the results and look for the discussion you are referring to.

We moved a large part of section 3.1's discussion to here, so this reads very differently now. It should be clearer now. See above.

P6123_L28: you said that the distribution is not Gaussian, so why mentioning this here

in the discussion?

We removed this phrase.

P6124_L6: I think the word nutrients is a bit misleading here. Do you refer to inorganic or organic nutrients? The power of the ERSEM-like parametrisations (originally from the ERSEM of the 90's) are that nutrients are indeed biogeochemical constituents and they can flow between various forms and components.

We changed the wording, added inorganic:

The ratios of the inorganic nutrients to carbon in the model do not typically get at low as the estimates of this ratio, except iron.

P6124_L12: I would say a combination of deposition and scavenging

We added:

the iron cycle is nudged towards what is observed in nature by an climatological surface deposition, and through hydroxide precipitation and saturation removal of excess iron.

P6124_L24-28: I agree, but I recall that it was already demonstrated in 2007 by Vichi, M., S.

Masina, and A. Navarra (2007), A generalized model of pelagic biogeochemistry for the global ocean ecosystem. Part II: numerical simulations, J. Mar. Sys., 64, 110–134.

Changed text to:

> This work confirms the work of Vichi et al. (2007) by demonstrating using emergent properties that many of ERSEMs design choices and parameterisation are still appropriate in a global context.

P6125_L2: I think this statement is too generic and requires some more discussion. Not all physiological parameterizations are well known, otherwise we would not have so many biogeochemical models with different combinations of terms.

I agree, many of the ecosystem functions are not well understood. We changed the text to:

> Furthermore, these emergent relationships were not explicitly parameterized in model development; all of them arise naturally out of a~combination of many other ecosystem functions.

P6125_L7: Was there a benthic parameterization?

There was no explicit benthic community, but the model contained a simple recycling benthic system, which can be seen in the new ERSEM diagram, above. We've rewording benthic community to benthic environment.

Conclusions: It would be good to have a future outlook on the applications of such a method: use it to compare with other models with different degrees of complexity or with fixed quota for certain nutrients. Also, the statement related to hydrodynamics needs some more discussion as suggested in my main comment 5.

We have added a future outlook paragraph:

> In future works, it should be possible to use emergent properties to test hypotheses during biogeochemical model development. As an example, if some process X is expected to influence an emergent property Y, a comparison of the emergent property Y in the presence and absence of X may yield insight as to the value of that process in models. Similarly, a comparison of the emergent property Y in two models with alternative parameterisations for X may facilitate the selection of a parameterisation. A second potential avenue for future research would be to test the independence of the biogeochemical model from the physical model by comparing the emergent properties from the same BGC model coupled against multiple physical models. Finally, it's important to remember that each emergent property that the model fails to reproduce is a new direction for future model development.