In the following, responses to reviewer comments are shown in bold typeface.

Anonymous Referee #1

I agree that showing the full suite of maps and associated Taylor diagrams for individual fields would be overwhelming and relegating some of these to the Supplementary information is a good idea. However, I think Figure S5, or a similar one for annual mean data, could be incorporated into the main text. The paper has only 6 figures and I think an additional one summarizing the various models' skill in a Taylor diagram for each of the fields considered (except O2: see below point 4) is a good idea.

Figure S5 is now incorporated into the main text excluding O_2 as the reviewer suggests in point 4.

There are a few things missing from the model description:

(a) The grid resolution should be stated. This is highly relevant to issues discussed, such as computational cost and deficiencies in the modelled ocean circulation. NEMO at e.g. 1 or 2 degrees resolution gives a very different circulation.
(b) There should be a brief description of the algorithms used for carbon chemistry and gas exchange (e.g., which equations were used to calculate the equilibrium constants).These models are fairly mature and not the main source of error in ocean biogeochemistry models (and I assume they were standardized across the six models used here although this is not actually stated), but a brief description is nonetheless required.
(c) None of the ecosystem model descriptions say anything about calcification or calcite dissolution. This relates directly to interpretation of the modelled vertical profiles of DIC and alkalinity, and to the anomalous distribution of pCO2 in the equatorial zone in some of the models (see below points 1 and 5).

We are grateful to the referee for pointing out these omissions. A thorough description of the grid resolution is now given at the start of the manuscript section on experimental design:

"All participating models made use of a common version (v3.2) of the NEMO physical ocean general circulation model (Madec, 2008) coupled to the Los Alamos sea-ice model (CICE) (Hunke and Lipscomb, 2008). This physical framework is configured at approximately 1×1 degree horizontal resolution (ORCA100; 292×362 grid points), with a focusing of resolution around the equator to improve the representation of equatorial upwelling. Vertical space is divided into 75 fixed levels, which increase in thickness with depth, from approximately 1m at the surface to more than 200m at 6000m. Partial level thicknesses are used in the specification of seafloor topography to improve the representation of deep water circulation. Vertical mixing is parameterized using the turbulent kinetic energy scheme of Gaspar et al., (1990), with modifications made by Madec (2008). To ensure that the simulations were performed by the different modelling groups using an identical physical run, a Flexible Configuration Management (FCM) branch of this version of NEMO was created, and all biogeochemical models were implemented in parallel within this branch and run separately."

A brief description of the equations used for carbon chemistry and gas exchange in each of the models is now included at the end of the model description section:

"… including ocean carbonate chemistry and air-sea exchange (HadOCC, Diat-HadOCC – Dickson & Goyet 1994, Nightingale et al., 2000; MEDUSA - Blackford et al., 2007; PlankTOM-6, PlankTOM-10 - Orr et al., 1999; ERSEM - Artoli et al., 2012)."

We also now include brief descriptions of the calcification and $CaCO_3$ dissolution schemes models employ:

"In the case of calcium carbonate ($CaCO_3$) production, the models utilised a range of different parameterisations. HadOCC and Diat-HadOCC use a simple empirical relationship that ties $CaCO_3$ production to primary production. MEDUSA relates $CaCO_3$ production to export production, with a PIC:POC ratio (particulate inorganic carbon:particulate organic carbon ratio) dependent on calcite saturation state. In PlankTOM-6 and PlankTOM-10, coccolithophore algae are explicitly modelled, with a fixed PIC:POC ratio. ERSEM relates $CaCO_3$ production to export production driven by nanophytoplankton losses, with a variable PIC:POC ratio dependent on temperature, nutrient limitation and calcite saturation state. Meanwhile, $CaCO_3$ dissolution was a simple exponential function of depth in the HadOCC models, with the other models modifying similar vertical dissolution with reference to the ambient saturation state of $CaCO_3$."

Main conceptual points:

(1) When the errors are relatively uniform across models and are therefore attributed to errors in circulation there is little discussion of the underlying physical processes. Vertical gradients of DIC and alkalinity are weak in the Southern Ocean, which could conceivably be attributed to excessive vertical mixing. But I think there is a biological element that is not considered here. Modelled vertical gradients are much stronger for DIC than for alkalinity, which I would attribute to the ecosystem models exporting POC but negligible PIC. If it were purely due to circulation I doubt there would be such a difference between the two.

Regarding the Southern Ocean, the following text has been added in the results section where vertical profiles are discussed:

"As Figure S7 shows, this common problem of vertical homogeneity between the models is driven by systematic biases in vertical mixing in this region, as well as known errors in ocean circulation (e.g. Yool et al., 2013)."

Regarding the Equatorial Pacific, the following text has been added in the results section, together with a series of supplementary figures that illustrate model POC and PIC export:

"The source of this bias in surface alkalinity is, at least in part, due to disparity in modelled $CaCO_3$ production in this region. As Supplementary Figures S8-S10 show, PlankTOM6, PlankTOM10 and ERSEM export negligible particulate inorganic carbon (PIC; Figure S9) relative to particulate organic carbon (POC; Figure S8) in this region. This results in low rain ratios (Figure S10) and the divergence of DIC and alkalinity performance of these models in this region. The lack of PIC export in these models runs contrary to observations (e.g. Dunne et al., 2007), but reflects the current difficulty in modelling $CaCO_3$ production – which HadOCC, Diat-HadOCC and MEDUSA-2 circumvent by simplistic empirical parameterisations."

I also think that the x axes on Figures 5 and 6 (and S6 and S7, but see below note Re: 10550/12) should be rescaled to reduce white space. This is particularly true for the case of DIC in the equatorial Pacific. Some of these profiles don't show much vertical structure, so wasting half of the available space is a bad idea. The boxes themselves could also be made a bit wider. (Also the vertical axes are nonlinear and need some explanation. If it is a logarithmic scale, say so. If it is an arbitrary 'telescoping' this needs to be stated explicitly.)

The vertical depth profiles have now been revised, reducing white space and stating in legends that the vertical scaling is logarithmic (log_{10}).

(2) The Conclusion does an admirable job of spelling out the implications of different strategies for model formulation, and the arguments for continuing development of more complex models even if they do not have greater skill with respect to e.g. DIC and pCO2. But I have two caveats here:

(a) One issue that is not mentioned is model diversity. Given that no model is shown to be the most skillful by all metrics, and all are most or least skillful by at least one metric, a central conclusion that can be drawn from this work is that it is important that the international climate modelling community maintain a diverse suite of models and do not 'converge' on a few similar ones.

The following has been added to the manuscript conclusions:

"As no model is found to have the highest skill across all metrics and all are most or least skilful for at least one metric, our results suggest that it is in the interest of the international climate modelling community to maintain a diverse suite of ocean biogeochemical models."

(b) I don't care for the false dichotomy of improved climate simulations vs "scientific exploration" in the final paragraph. Adequately addressing some issues previously raised with respect to unresolved climate feedbacks (e.g., DMS) will certainly require more complex ocean biology models.

The text that the reviewer is referring to here has now been removed.

(3) I think the conclusion that no model is demonstrably better or worse than any other is not really consistent with the data. In Table 3 (see also Figure S5), not only does ERSEM show the weakest correlation for pCO2, chlorophyll and primary production, but these correlation coefficients are consistently the smallest by a wide margin and are in all cases not meaningfully different from zero. It does better for nitrate, DIC and alkalinity but these are weak diagnostics for the reasons discussed (e.g. 10547/18-19). I don't think the claim made on 10551/23-27 that in some cases "models of greater biological complexity tend to equate to improved model skill" is justified by ERSEM having (marginally) higher skill for surface nitrate.

The general conclusion that no model is demonstrable better or worse than any other has now been modified to: "no model is shown to consistently outperform all other models". The ERSEM based justification that in some cases models of greater biological complexity tend to equate to improved model skill has now been removed.

(4) I don't think surface O2 is a useful diagnostic, and the authors should consider removing it entirely (e.g., Table 3, Figure S5 and especially Figure 4). At the surface, biological processes play a negligible role in the distribution of O2, as is noted in the text (10548/21-23). Figure 4 summarizes the rank order of model skill on different metrics, with no consideration of how large the differences are. Do they really want this analysis to be biased by inclusion of an essentially meaningless diagnostic for which the differences among models are negligible?

We have considered the reviewer's advice and removed surface oxygen as an intercomparison variable.

(5) The pCO2 fields in the tropical upwelling zones in the more complex models (ERSEM, PlankTOM) look almost like a mirror image of the expected pattern, with lower pCO2 associated with recently upwelled waters (Figure 1). I agree that this probably results from excessive alkalinity in the upwelled water (10550/10-11, Figure 6). But these authors do not go into much depth about the underlying processes. Clearly these models are not removing alkalinity from the surface layer by biogenic sedimentation at anything like real-world rates. By failing to consider (or even describe) the calcification and calcite dissolution models and by too casually dismissing the Southern Ocean alkalinity errors as deriving from circulation, they miss an opportunity to delve into the source of errors that are on the surface quite pathological. No one is going to accept a model in which cold, DIC-rich water upwelled to the surface in the tropics has a pCO2 below atmospheric.

As described above, the text has been expanded in several locations regarding :

- **PIC and POC production in the models, with a particular reference to the Equatorial Pacific**
- **Evidence concerning physical deficiencies in, especially, the Southern Ocean**
- **A more complete description of the calcification and dissolution submodels by the different BGC models**

Some details:

10539/6 "Dynamic Green Ocean Models" Is this really a class of models? I thought it was just the name that a particular group gave to their own model (which may have since evolved into a suite of related models, but that still doesn't really justify calling it a class or type of model). Anyway the abbreviation is never used and is not necessary (see also 10544/1-2).

This abbreviation has now been removed.

10540/6 "direct human exploitation of the seas" I don't think there is any evidence for such top-down forcing of the kind of fields considered in this paper.

We agree with the referee, and this text has now been removed.

10540/23 "What controlled the variations in atmospheric trace gas over the geological past including those measured by isotopes?" What controlled variations in atmospheric trace gas concentrations and isotopic composition over the geological past?

This text has been changed as the reviewer recommends:

“What controlled variations in atmospheric trace gas concentrations and isotopic composition over the geological past?”

10540/28 I don’t think it’s accurate to say that IPCC ’produced’ the data archive.

This text has been changed:

“In addition, the ESM model archive is increasingly being used by activities within …”

10541/4 "how will climate change affect oceanic primary production" ocean

This text has been changed as the reviewer suggests.

10541/8 I would consider citing the more recent and more comprehensive paper by Harvey 2008 (10.1029/2007JC004373) in place of or in addition to Khesghi 1995. The older paper is in a somewhat obscure journal and is cited in the more recent one.

This additional reference has been added as the reviewer recommends.

10541/21 "following the same experiment protocol" experimental

This text has been changed as the reviewer recommends.

10543/14 "a dimethyl sulphide (DMS) sub-model for cloud feedbacks" I would delete "for cloud feedbacks" as it is not relevant to the present experiment.

This text has been changed as the reviewer recommends.

10544/2 add "level" after "trophic"

This text has been changed as the reviewer recommends.

10545/3 "the marine biology" biota

This text has been changed as the reviewer recommends.

10545/16-17 makes it sound like the pCO2 data came from SeaWiFS

This text has been changed.

10545/25 the GLODAP data product is not a climatology

This text has been changed.

10546/3 "the biogeochemical pathway through which the vast majority of marine ecosystems ultimately obtain energy" I would not word it like this. Phytoplankton photosynthesis represents the vast majority of the primary energy source to marine ecosystems. But I have trouble envisioning what is meant by a majority of ecosystems.

This text has been changed.

10546/10 delete "and in part related to preceding points"

This text has been deleted.

10546/25 "circumference axis" I have not heard this term before and Googling it turns up only a few marginally relevant examples. Taylor calls it the azimuthal position.

This text has been changed as the reviewer recommends.

10548/24 "Figure 4 summarises Table 3" Figure 4 summarizes the data in Table 3

This text has been changed as recommended by the reviewer.

10548/28-29 "field metric" Another jargony and probably unnecessary term. I would just delete "field". (see also 10552/1, 7)

This text has been changed as the reviewer recommends.

10549/22 "much shallower gradients with depth" Not clear what "shallower" means here. Weaker? I don't think it means there is a 'cline' at a shallower depth, although that is true in some cases. Please reword and clarify.

The reviewer is correct. We have changed the text as they recommend,

"… with much weaker gradients with depth …"

10549/27 "ocean physics deficiencies" errors in ocean circulation

This text has been changed as the reviewer recommends.

10550/6 delete "values"

This text has been deleted.

10550/6 "MONSooN" I don't think the name of the machine is relevant here and anyway the acronym is never used.

This text has been changed as the reviewer recommends.

10550/12 and 20 There are two supplemental figures numbered S7

The supplementary figure labels have been corrected.

10550/21-22 "This unsurprisingly reflects the significant cost of performing ocean physics operations on biogeochemical tracers." I'm not sure this sentence is necessary at all, but maybe it could be modified to

something like "reflecting the significant cost of applying advection and mixing terms to each tracer" and appended to the previous one.

This text has been changed as the reviewer recommends.

10550/26 It looks to me like "computational cost" means something other than total CPU time or wall-clock time here but I can't tell exactly what.

Computational cost does only mean CPU time. The text here has been changed to clarify this.

"Computational timing tests (CPU time) were carried out …"

10551/11,14 delete "of"

This text has been changed as the reviewer recommends.

10551/12 "shown to generally have higher" shown to have generally higher

This text has been changed as the reviewer recommends.

10551/20 delete "the oceanographic regions of"

This text has been deleted.

10551/21 "possibly because their biological export production can more easily be tuned to maintain the observed vertical gradients" Is there any reason to believe that these models were tuned to reproduce depth profiles in these specific regions?

For all models, some degree of tuning of production and export occurred prior to this study, albeit in physical frameworks different (to varying degrees) to that used here. In the case of the less complex models, tuning is typically more straightforward as they have less state variables and, as a result, simpler, more directly-amenable parameterisations. Tuning in the more complex models is more difficult where "community" properties, such as production, are a product of a greater number of (explicit and dynamic) ecological actors. Tuning during this study was limited or absent between models, but some models, such as HadOCC and MEDUSA, may have benefitted from being previously tuned within the NEMO framework (albeit a different version and grid configuration). However, as noted - and illustrated - in Yool et al. (2013) for MEDUSA, tuning remains difficult for 3D performance as improvements in short-duration simulations can easily turn into degraded performance when simulations are spun out longer. The text has been amended to draw the reader's attention to some of these aspects.

10552/7 add a comma after "(Table 4)"

This text has been corrected.

10552/10-11 "depths of 1000 m" less than?

This text has been corrected.

10552/13 "discrepancies within the physical ocean model" errors?

This text has been changed as the reviewer recommends.

10552/15 "For alternative fields such as DIN in the Southern Ocean and Equatorial Pacific (Supplement Fig. S7), however, models have both positive and negative biases" For other fields, such as DIN in the Southern Ocean and Equatorial Pacific (Supplement Fig. S7), models have both positive and negative biases

This text has been changed as the reviewer recommends.

10552/21-22 "also tend to represent additional factors" are also able to represent additional factors

This text has been changed as the reviewer recommends.

10553/5 "Specifically, the HadOCC and MEDUSA-2 models that were previously implemented within NEMO v3.2 were “familiar” with this ocean model’s configuration and flaws." Meaning, I assume, that the developers of these models were familiar with NEMO and had some opportunity to tune the ecosystem to a circulation similar to that used in this experiment. Please be more specific. Models of this sort do not learn on their own.

The text here has been amended. The following has also been added:

“Tuning during this study was limited or absent between models, but some models, such as HadOCC and MEDUSA, may have benefitted from being previously tuned within the NEMO framework (although in a different version and grid configuration).”

10553/7-8 "the ERSEM model ... had a distinct disadvantage" which is what?

The text here has been removed.

10553/9 delete "found"

This text has been deleted.

10553/10 change "settings" to "values"

This text has been changed.

10553/18-19 "a bottom-up approach to model skill assessment" I can’t tell what this means, and the term does not appear to have been used by Vetter et al.

This text has now been removed.

Table 2 I would change "Prokaryotes" to "Heterotrophic bacteria" (assuming that is what it means). Prokaryotes is a (mostly obsolete) taxonomic category rather than a functional/biogeochemical one, and some other groups in this table are mostly made up of prokaryotes.

The term Prokaryotes was originally used because this category also contains Archaea. We have now changed this to "Picoheterotrophs" focusing on size and functionality rather than phylogeny.

Anonymous Referee #2

The manuscript does not make clear how its findings are substantially different from previous studies of a similar nature, such as: Kriest et al., 2010, doi:10.1016/j.pocean.2010.05.002 Friedrichs et al., 2007, doi:10.1029/2006JC003852. I think that the authors need to present a strong case about how their work is new, compared to existing literature.

We thank the reviewer for pointing out this oversight. The following introductory paragraph has been added to the manuscript to better contextualize our work:

"Previous authors have performed biogeochemical model intercomparisons with parallels to this study (e.g. Friedrichs et al., 2007; Kriest et al., 2010; Steinacher et al., 2010; Popova et al., 2012). These have differed from this study, and each other, in a number of ways. For instance, this study is 3D rather than 1D (cf. Friedrichs et al., 2007); global rather than regional (cf. Popova et al., 2012); uses identical rather than diverse physics (cf. Steinacher et al., 2010); and spans a more functionally diverse range of biogeochemical models (cf. Kriest et al., 2010). The latter two factors, in particular, distinguish this study, permitting us to both formally separate the impact of physics from that of biogeochemical dynamics, and to do so across a broad range of model complexity from NPZD through to state-of-the-art PFT models with considerable ecological sophistication. This study is still constrained by the use of a single ocean circulation, and by a bespoke gradation of model complexity (PlankTOM6 and PlankTOM10 partially inform this). Nonetheless, this study represents an intercomparison along separate lines to those previously conducted."

Specific Comments

We know from previous work that the fidelity of the ocean physical model plays a large role in the behavior of ocean BGC models. Some studies that put the same OBGC model into different GCMs are: Doney et al., 2004, doi: 10.1029/2003GB002150 Najjar et al., 2007, doi:10.1029/2006GB002857 Dunne et al., 2013, doi:10.1175/JCLI-D-12-00150.1 Séférian et al., 2012, doi:10.1007/s00382-012-1362-8 With this in mind, it is important for the authors to describe how well their configuration of NEMO, and how well it performs. What is the spatial and vertical resolution of the model? What physical parameterizations are used? Describe the biases in the fields: SST, MLD, MOC. This is particularly relevant to the Southern Ocean comparisons, where it is suggested that ocean physics deficiencies are causing the OBGC biases. How much were the BGC model parameters tuned?

A description of the horizontal and vertical model resolution and some of the physical parameterisations used is now given at the start of the experimental design section of the manuscript:

"All participating models made use of a common version (v3.2) of the NEMO physical ocean general circulation model (Madec, 2008) coupled to the Los Alamos sea-ice model (CICE) (Hunke and Lipscomb, 2008). This physical framework is configured at approximately 1×1 degree horizontal resolution (ORCA100; 292×362 grid points), with a focusing of resolution around the equator to improve the representation of equatorial upwelling. Vertical space is divided into 75 fixed levels, which increase in thickness with depth, from approximately 1m at the surface to more than 200m at 6000m. Partial level thicknesses are used in the specification of seafloor topography to improve the representation of deep water circulation. Vertical mixing is parameterized using the turbulent kinetic energy scheme of Gaspar et al., (1990), with modifications made by Madec (2008). To ensure that the simulations were performed by the different modelling groups using an identical physical run, a

Flexible Configuration Management (FCM) branch of this version of NEMO was created, and all biogeochemical models were implemented in parallel within this branch and run separately."

We have also added a new Supplementary Figure (S7), and some text, to briefly outline performance issues with our NEMO simulation.

"Supplementary Figure S7 shows an intercomparison of the common NEMO physics with observations for several key physical fields. In terms of SST, NEMO represents observed patterns well, although simulates a warmer Gulf Stream and noticeably cooler temperatures in the vicinity of the Labrador Sea. In conjunction with fresher salinities in the North Atlantic (results not shown), these differences result in shallower depths of the mixed layer and pycnocline in this region. By contrast, in the Southern Ocean both mixed layer depths and the modelled pycnocline are markedly deeper than in observations. This latter regional bias has biogeochemical consequences across all of the models examined here (see later)."

The following description of model tuning has been added to the end of the experimental design section of the manuscript:

"For all models, some degree of tuning occurred prior to this study, albeit in physical frameworks different (to varying degrees) to that used here. Tuning during this study was limited or absent between models, but some models, such as HadOCC and MEDUSA, may have benefitted from being previously tuned within the NEMO framework (although in a different version and grid configuration). "

There is a comment in the discussion "model developers were afforded a limited opportunity to tune parameter settings". Please elaborate on this in the model descriptions. Previous work, like Kriest et al. (2010) and Friedrichs et al. (2007) demonstrate that models generally perform poorly if they are not tuned. If their 'limited opportunity' was not sufficient, then what's the point of this analysis? If these models were serious candidates for inclusion in a CMIP class ESM, they would be given more than a 'limited opportunity' to tune parameter settings.

As noted above, a short description of the extent of model tuning has been added to the end of the experimental design section of the manuscript. However, note that tuning in 3D models is typically performed continuously over a number of months or years as developers use their biogeochemical models to tackle research questions - and discover discrepancies in their performance. Here, only a few months were available, and it is not unlikely that the models could be improved by a more extended period of use within the framework used. This is the point of the remark in the discussion. However, the models were not fatally compromised by this limited period, and there is, anyway, no natural end to such ad hoc tuning. The advent of computationally efficient 3D tuning schemes, such as that used by Kriest et al., (2010), promise much in this regard, and similar future studies will doubtless utilise such approaches to ensure that model performance is optimal.

The model evaluation is too brief. Please relate biases in surface fields to processes, e.g. primary productivity and biological export.

Supplementary figures and the following manuscript text have been added, relating biases in pCO_2 to alkalinity and PIC production:

"The negative pCO_2 biases in the equatorial Pacific exhibited by the PlankTOM6, PlankTOM10 and ERSEM models may be explained, at least in part, by the positive biases that these models show for surface alkalinity in this region (Figure S3). The models with positive pCO_2 biases in the equatorial Pacific (HadOCC, Diat-HadOCC and MEDUSA-2), do not have negative surface alkalinity biases in this region but values are much closer to observations (Figure S3). The root of these alkalinity biases lies in variation in PIC production by the models in this region…"

"The source of this bias in surface alkalinity is, at least in part, due to disparity in modelled $CaCO_3$ production in this region. As Supplementary Figures S8-S10 show, PlankTOM6, PlankTOM10 and ERSEM export negligible particulate inorganic carbon (PIC; Figure S9) relative to particulate organic carbon (POC; Figure S8) in this region. This results in low rain ratios (Figure S10) and the divergence of DIC and alkalinity performance of these models in this region. The lack of PIC export in these models runs contrary to observations (e.g. Dunne et al., 2007), but reflects the current difficulty in modelling $CaCO_3$ production – which HadOCC, Diat-HadOCC and MEDUSA-2 circumvent by simplistic empirical parameterisations."

The following text has also been added:

"Surface DIN concentrations are influenced by both the efficiency of primary production and the efficiency of remineralisation both of which differ between models. Although we don't explore the differences in remineralisation, the models which show positive DIN biases in the equatorial Pacific (HadOCC, Diat-HadOCC and MEDUSA-2), are generally shown to also have positive integrated primary production biases in this region (Figure S1). To a lesser extent the reverse is true of the models with negative DIN biases in the equatorial Pacific (PlankTOM10 and ERSEM)."

The evaluation makes almost no mention of previous literature on OBGC model skill assessment that can guide the analysis. For instance, please see the special issue of Journal of Marine Systems on this topic http://www.sciencedirect.com/science/journal/09247963/76/1 A drawback of the Taylor diagrams is that it omits information on mean bias. For plots 1-3 and S1-S4, please add mean field values for models and observations to the plots. This could be done in the corner of the maps or in the legend.

The manuscript now includes reference to:

Doney, S. C., Lima, I., Moore, J. K., Lindsay, K., Behrenfeld, M. J., Westberry, T. K., Mahowald, N., Glover, D. M. and Takahashi, T.: Skill metrics for confronting global upper ocean ecosystem-biogeochemistry models against field and remote sensing data. J. Mar. Syst., Skill assessment for coupled biological/physical models of marine systems 76, 95–112, 2009.

Stow, C. A., Jolliff, J., McGillicuddy Jr., D. J., Doney, S. C., Allen, J. I., Friedrichs, M. A. M., Rose, K. A. and Wallhead, P.: Skill assessment for coupled biological/physical models of marine systems. J. Mar. Syst., Skill assessment for coupled biological/physical models of marine systems 76, 4–15, 2009.

As well as

Jolliff, J. K., Kindle, J. C., Shulman, I., Penta, B., Friedrichs, M. A. M., Helber, R., and Arnone, R.A.: Summary diagrams for coupled hydrodynamic-ecosystem model skill assessment, J. Marine Syst., 76, 64-82, 2009.

Mean field values for observations and models have been added to figure legends in both the main manuscript and supplementary material as the reviewer recommends.