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.

iMarNet: An ocean biogeochemistry model inter-comparison project within a common physical ocean modelling framework.

L. Kwiatkowski[1,2]*, A. Yool[3], J. I. Allen[4], T. R. Anderson[3], R. Barciela[5], E. T. Buitenhuis[6], M. Butenschön[4], C. Enright[6], P. R. Halloran[7], C. Le Quéré[6], L. de Mora[4], M.-F. Racault[4], B. Sinha[3], I. J. Totterdell[5] and P. M. Cox[1]

[1] College of Engineering, Mathematics and Physical Sciences, University of Exeter, Exeter EX4 4QF, UK;
[2] Department of Global Ecology, Carnegie Institution for Science, 260 Panama Street; Stanford, California, 94305, USA;
[3] National Oceanography Centre, University of Southampton Waterfront Campus, European Way, Southampton SO14 3ZH, UK;
[4] Plymouth Marine Laboratory, Prospect Place, West Hoe, Plymouth PL1 3DH, UK;
[5] Hadley Centre, Met Office, Exeter, EX1 3PB, UK;
[6] Tyndall Centre for Climate Change Research, School of Environmental Sciences, University of East Anglia, Norwich, UK;
[7] College of Life and Environmental Sciences, University of Exeter, Exeter EX4 4RJ, UK.

Correspondence to: L. Kwiatkowski (lkwiatkowski@carnegiescience.edu)

Deleted: Butenschön[3]

Formatted: Font: Bold

Abstract

Ocean biogeochemistry (OBGC) models span a wide range of complexities from highly simplified, nutrient-restoring schemes, through nutrient-phytoplankton-zooplankton-detritus (NPZD) models that crudely represent the marine biota, through to models that represent a broader trophic structure by grouping organisms as plankton functional types (PFT) based on their biogeochemical role (Dynamic Green Ocean Models) and ecosystem models which group organisms by ecological function and trait. OBGC models are now integral components of Earth System Models (ESMs), but they compete for computing resources with higher resolution dynamical setups and with other components such as atmospheric chemistry and terrestrial vegetation schemes. As such, the choice of OBGC in ESMs needs to balance model complexity and realism alongside relative computing cost. Here, we present an inter-comparison of six OBGC models that were candidates for implementation within the next UK Earth System Model (UKESM1). The models cover a large range of biological complexity (from 7 to 57 tracers) but all include representations of at least the nitrogen, carbon, alkalinity and oxygen cycles. Each OBGC model was coupled to the Nucleus for the European Modelling of the Ocean (NEMO) ocean general circulation model (GCM), and results from physically identical hindcast simulations were compared. Model skill was evaluated for biogeochemical metrics of global-scale bulk properties using conventional statistical techniques. The computing cost of each model was also measured in standardised tests run at two resource levels. No model is shown to consistently outperform all other models across all metrics. Nonetheless, the simpler models are broadly closer to observations across a number of fields, and thus offer a high-efficiency option for ESMs that prioritise high resolution climate dynamics. However, simpler models provide limited insight into more complex marine biogeochemical processes and ecosystem pathways, and a parallel approach of low resolution climate dynamics and high complexity biogeochemistry is desirable in order to provide additional insights into biogeochemistry – climate interactions.

Deleted: ¶ ¶ ¶ ¶ ¶ ¶ ¶ ¶ ¶ ¶ ¶ ¶ ¶ ¶ ¶ ¶ ¶ ¶

Deleted: ; DGOM

Deleted: or underperform

Deleted: that are easier to tune

Formatted: Font: Bold

1 Introduction

Ocean biogeochemistry is a key part of the Earth System: it regulates the cycles of major biogeochemical elements and controls the associated feedback processes between the land, ocean and atmosphere. As a result, changes in ocean biogeochemistry can have important implications for climate (Reid et al., 2009). Marine ecosystems are indirectly affected by anthropogenic environmental change (Jackson et al., 2001), particularly through climate-induced changes in physical properties and CO_2-induced ocean acidification. Understanding and quantifying the response of ocean biogeochemistry to global changes and their feedbacks with the Earth System is essential to improve our capacity to maintain ecosystem services this century and beyond.

With the recent publication of the Intergovernmental Panel on Climate Change (IPCC) 5th Assessment Report (AR5), global efforts are already underway to develop the next generation of Earth System Models (ESMs) to support climate policy development and any further IPCC Assessment Report. OBGC coupled to ESMs can help address a series of overarching scientific questions: How will the ocean contribute to atmospheric trace gas composition (e.g. CO_2, CH_4, N_2O, DMS) in a changing climate? Are there tipping points in marine biogeochemistry (e.g. oceanic anoxic events, methane hydrate release) that could be triggered by a changing climate? Are there interactions between ESM processes and society's management of resources (e.g. fisheries, land use, agriculture) in the marine environment? Furthermore, as ESMs are increasingly being evaluated based on their capacity to understand past variability (Braconnot et al., 2012), further questions might include: What controlled variations in atmospheric trace gas concentrations and isotopic composition over the geological past?

For an anticipated 6th IPCC assessment report it is generally considered that these global-scale questions, with direct implications for climate policies, will again be the main focus of ocean biogeochemical models within ESMs. In addition, the ESM model archive is increasingly being used by activities within the Inter-Sectoral Impact Model Intercomparison Project (http://www.pik-potsdam.de/research/climate-impacts-and-vulnerabilities/research/rd2-cross-cutting-activities/isi-mip/scientific-publications) to address socioeconomically-directed questions such as: How will climate change affect ocean primary production (e.g. Bopp et al., 2013), fisheries (Barange et al., 2014; Cheung et al., 2012), and harmful algal and jellyfish blooms (e.g. Codon et al., 2013, Gilbert et al., 2014)? What is the potential for geoengineering schemes such as ocean fertilisation (Buesseler & Boyd, 2003) and alkalinity addition (Kheshgi, 1995; Harvey, 2008) to affect the climate system, and how do they affect the rest of the Earth System?

Deleted: ¶ ¶ ¶ ¶ ¶

Deleted: both the direct human exploitation of the seas and the indirect

Deleted: the

Formatted: Font: +Body (Calibri), English (U.S.)

Formatted: Font: +Body (Calibri), English (U.S.)

Formatted: Font: +Body (Calibri), English (U.S.)

Deleted: including those measured by isotopes

Deleted: produced by the IPCC

Deleted: for example

Deleted: some

Deleted: oceanic

Deleted: .

Within the UK, the Integrated Global Biogeochemical Modelling Network (iMarNet) project aims to advance the development of ocean biogeochemical models through collaboration between existing modelling groups at Plymouth Marine Laboratory (PML), National Oceanography Centre (NOC), University of East Anglia (UEA) and the Met Office-Hadley Centre (UKMO). As part of iMarNet we conducted an intercomparison of 6 current UK models, to help inform the selection of a baseline OBGC model for the next UK Earth System Model (UKESM1). This intercomparison focused on model skill at reproducing global-scale bulk properties - such as nutrient and carbon distributions - that broadly characterise the activity of marine biota (and, thus, the carbon cycle) in the ocean. To limit the role of errors originating with modelled physics, all of the examined model simulations were performed within the same physical ocean GCM, under the same external forcing and following the same experimental protocol. As all of the models examined have been previously published, our analysis does not include an assessment of their underlying biological fidelity (i.e. the extent to which structures, parameterisations and parameter sets of candidate models are *a priori* realistic). However, while primarily focused on model skill, the intercomparison also considers the computational cost of the models in relation to the realism that they offer.

Deleted: experiment

Formatted: Font: Calibri, 12 pt, Not Bold

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.

2 Method

2.1 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

Deleted: A Flexible Configuration Managemen (FCM) branch of this version 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.

Formatted: Font: 12 pt

Simulations were initialised at year 1890 from an extant physics-only spin-up (ocean and sea-ice), to minimise undesirable transient behaviour in ocean circulation. In terms of ocean biogeochemistry, all model runs made use of a common dataset of three-dimensional fields for the initialisation of major tracers. Nutrients (nitrogen, silicon and phosphorus) and dissolved oxygen in this dataset were drawn from the World Ocean Atlas 2009 (Garcia et al., 2010a; Garcia et al., 2010b), while dissolved inorganic carbon (DIC) and alkalinity were drawn from the Global Ocean Data Analysis Project (GLODAP) (Key et al., 2004). GLODAP does not include a DIC field that is directly valid for 1890, so a temporally-interpolated field was produced based on GLODAP's "pre-industrial" (i.e. ~1800) and "1990s" fields of DIC. As there is currently no comprehensive spatial dataset of the micronutrient iron, participating models were permitted to make use of different initial distributions of iron (typically those routinely used by the models in other settings). All other biogeochemical fields (e.g. plankton, particulate or dissolved organic material) were initialised to arbitrary small initial conditions.

After initialisation at 1890, the models were run for 60 years (1890-1949 inclusive) under the so-called “normal year” of version 2 forcing for common ocean-ice reference experiments (CORE2-NYF; Large and Yeager, 2009). Subsequently, the models were run under transient, interannual forcing from the same dataset (CORE2-IAF) for a further 58 years (1950-2007 inclusive). CORE2 provides observationally derived geographical fields of downwelling radiation (separate long- and short-wave), precipitation (separate rain and snow), and surface atmospheric properties (temperature, specific humidity and winds), and is used in conjunction with bulk formulae to calculate net heat, freshwater and momentum exchange between the atmosphere and the ocean.

Formatted: Font: 12 pt

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).

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).

2.2 Candidate model structures

The models evaluated within this study vary significantly in biological complexity. The key features of the participating models are summarized below:

HadOCC (Palmer & Totterdell, 2001)**:** the *Hadley Centre Ocean Carbon Cycle model (HadOCC)* model is a simple NPZD (Nutrient, Phytoplankton, Zooplankton, Detritus) representation that uses N nutrient as its base currency but with coupled flows of C, alkalinity and O_2. The model was the ocean biogeochemistry component of the UK Met Office's HadCM3 climate model, and was used for the first ever fully coupled carbon-climate study (Cox et al., 2000).

Diat-HadOCC (Halloran et al., 2010): is a development of the HadOCC model which includes two phytoplankton classes (diatoms and "other phytoplankton") and representations of the Si and Fe cycles, as well as a dimethyl sulphide (DMS) sub-model. The model is the ocean biogeochemistry component of HadGEM2-ES (Collins et al., 2011), the UK Met Office's Earth System model used to run simulations for CMIP5 and the Intergovernmental Panel on Climate Change (IPCC) 5th Assessment Report (AR5).

Deleted: for cloud feedbacks.

MEDUSA-2 (Yool et al., 2011; Yool et al., 2013)**:** *Model of Ecosystem Dynamics, nutrient Utilisation, Sequestration and Acidification (MEDUSA)* is an "intermediate complexity" plankton ecosystem model designed to incorporate sufficient complexity to address key feedbacks between anthropogenically-driven changes (climate, acidification) and oceanic biogeochemistry. MEDUSA-2 resolves a size-structured ecosystem of small (nanophytoplankton and microzooplankton) and large (microphytoplankton and mesozooplankton) components that explicitly includes the biogeochemical cycles of N, Si and Fe nutrients as well as the cycles of C, alkalinity and O_2. As such, MEDUSA-2 is broadly similar in structure to Diat-HadOCC, but includes several more recent parameterisations.

PlankTOM6 & PlankTOM10 (Le Quéré et al., 2005)**:** PlankTOM is a Dynamic Green Ocean Model that represents lower-trophic level marine ecosystems based on Plankton Functional Types (PFTs). A hierarchy of PlankTOM models exists that vary in the number of PFTs resolved. Two members drawn from this stable were used in iMarNet. PlankTOM6 includes six PFTs - diatoms, coccolithophores, mixed-phytoplankton, bacteria, protozooplankton and mesozooplankton - while PlankTOM10 includes an additional four PFTs - Nitrogen-fixers, *Phaeocystis*, picophytoplankton and macrozooplankton (Le Quéré et al. 2005; Buitenhuis et al., 2013). The models include the marine cycles of C, N, O_2, P, Si, a simplified Fe cycle, and three types of detrital organic pools including their ballasting properties and estimates the air-sea fluxes of CO_2, O_2, DMS, and N_2O. PlankTOM6 and PlankTOM10 were developed by an international community of ecologists and modellers to quantify the interactions between climate and marine biogeochemistry, particularly those mediated through CO_2. They make use of extensive synthesis of data for the parameterisation of growth rates of PFTs (e.g. Buitenhuis et al., 2006; 2010) and for the model evaluation (Buitenhuis et al., 2013).

Deleted: (DGOM)

ERSEM (Baretta et al., 1995; Blackford et al., 2004): *European Regional Seas Ecosystem Model (ERSEM)* is a generic lower-trophic level/ model designed to represent the biogeochemical cycling of C and nutrients as an emergent property of ecosystem interaction. The ecosystem is subdivided into three functional types: producers (phytoplankton), decomposers (bacteria) and consumers (zooplankton), and then further subdivided by trait (size, silica uptake) to create a foodweb. Physiological (ingestion, respiration, excretion and egestion) and population (growth, migration and mortality) processes are included in the descriptions of functional group dynamics. Four phytoplankton (picophytoplankton, nanophytoplankton, diatoms and non-siliceous macrophytoplankton), three zooplankton (microzooplankton, heterotrophic nanoflagellates and mesozooplankton) and one bacteria are represented, along with the cycling of C, N, P, Si and O_2 through pelagic (Blackford et al., 2004) and benthic (Blackford, 1997) ecosystems. ERSEM is used for shelf seas water quality monitoring and climate impact assessment, has been coupled to fisheries models (e.g. Barange et al., 2014), and is run operationally by the UK Met Office (e.g. Siddorn et al., 2007).

The intercomparison process required limited changes to model organisation and code, and models retained disparate parameterisations for several overlapping processes, 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). 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 $CaCO_3$.

The representation of biogeochemical cycles and biota in each model are summarized in Tables 1 and 2 respectively.

Deleted: the marine biology

Deleted: tables

2.3 Model evaluation

Assessment against observational datasets was made for a set of bulk ocean biogeochemical properties that were common across all models: pCO_2, alkalinity, dissolved inorganic carbon (DIC), dissolved inorganic nitrogen (DIN), chlorophyll and primary production. In all cases, model results were regridded to the same geographical grid (World Ocean Atlas) and guided by literature on appropriate skill metrics (e.g. Doney et al., 2009; Stow et al., 2009) model skill was assessed through statistical techniques such as global surface field standard deviation and spatial pattern correlation coefficients. In the biogeochemical regions of the North Atlantic, Equatorial Pacific and Southern Ocean, depth profiles of model outputs were also assessed against observations within the top 1000m of the water column.

Deleted: ,

Deleted: and dissolved oxygen (O_2).

Observational fields used within the model intercomparison are comprised of World Ocean Atlas 2009 DIN (Garcia et al., 2010a), chlorophyll (O'Reilly et al., 1998) and pCO_2 (Takahashi et al., 2009). Because of its biogeochemical importance, and the diversity in published estimates, observational primary production is an average of three empirical models: (Behrenfeld and Falkowski, 1997); (Carr et al., 2006) and (Westberry et al., 2008)- which are all estimates derived from satellite ocean colour and SST. The observational fields of chlorophyll and primary production used here represent averages over the 2000-2004 time period. This same period is used throughout the following analysis as a standard interval except in the case of DIC and alkalinity, which are analysed over the mean 1990-1999 period corresponding to the GLODAP data product.

Deleted: and O_2

Deleted: ; Garcia et al., 2010b), SeaWiFS

Deleted: climatology

These fields were selected for several reasons. Firstly, they are ocean or biogeochemical bulk properties for which there are global-scale observations. Secondly, these fields broadly represent foundational aspects of marine biogeochemical cycles. For instance, nutrients play a critical role in regulating the distribution and occurrence of marine plankton, while phytoplankton photosynthesis represents the vast majority of the primary energy source to marine ecosystems. Thirdly, the measurement of these fields is relatively well-defined with long-established standard methodologies. Properties that are directly related to biological entities, for instance biomass abundances, can be less precisely defined, difficult to match up with modelled quantities, or even absent from some models examined here. That said, the observational field of global scale primary production used here has a relatively high uncertainty because it is drawn from three methodologies which exhibit a large range (cf. Yool et al., 2013). Finally the examined properties are those which, if modelled poorly, legitimately cast doubt over the wider utility of a biogeochemical model in an earth systems context. Model results always depart from observations, but systematic disagreement with these basic observations is strongly suggestive of problems with process representation within a model. The model comparison focuses on the mean and seasonal cycle. It does not include evaluation of variability over interannual or longer timescales, in part because of limited data availability.

Deleted: primary production by

Deleted: is the biogeochemical pathway through which

Deleted: ultimately obtain energy

Deleted: and in part related to preceding point

3 Results

3.1 Model skill assessment

3.1.1 Surface fields

Figures 1-3 (and Supplementary Figures S1-S3) show annual average fields from each of the models for a series of ocean properties, together with comparable observational fields. The figures also include a panel that shows the corresponding model-observation Taylor diagram (Taylor, 2001). These illustrate both the correlation between (azimuthal position) and relative variability (radial axis) of model and observations, such that models more congruent with observations generally appear closer to the reference marker on the x-axis of the diagram. As Taylor diagrams do not account for mean field biases (Joliff et al., 2009) these are provided separately in figure legends.

Deleted: S4

Deleted: circumference axis

Deleted: (Jolliff et al., 2009).

Figure 1 shows annual average surface pCO_2 fields for both models and observations, with correlation coefficients ranging from r=0.01 to r=0.68 (Takahashi et al., 2009). In general, the simpler models (HadOCC, Diat-HadOCC and MEDUSA-2) better capture the global spatial pattern of pCO_2 (r=0.54 to r=0.68), but they overestimate the standard deviation in global surface pCO_2 by up to a factor of 2. This overestimation of the variance in global surface pCO_2 is a result of high modelled pCO_2 values in the equatorial Pacific and in particular the

Eastern equatorial Pacific. In contrast, the more complex models (PlankTOM6, PlankTOM10 and ERSEM) perform considerably worse in terms of capturing global spatial patterns of surface ocean pCO_2. In particular, all three models underestimate the observed high pCO_2 values along the equatorial Pacific ocean as well as the high coastal pCO_2 values in that region, opposite to the bias found in simpler models. However, the PlankTOM models overall show comparable standard deviations in mean global surface pCO_2 to that seen in observations.

Formatted: Font: 12 pt

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 as discussed in greater detail below.

Figure 2 illustrates model performance for annual average surface dissolved inorganic nitrogen (DIN) concentrations. Here, all models capture global patterns relatively well, with correlation coefficients >0.8, in part because of the initialisation from observations in 1890. The model with the highest spatial pattern correlation coefficient is ERSEM, although it slightly underestimates the global variability of DIN. The other models have lower spatial pattern correlation coefficients and generally overestimate the global variability of DIN. PlankTOM6 performs below other models, while PlankTOM10 has similar performance as the simpler models. In general, aside from ERSEM and PlankTOM10, most models show elevated Pacific DIN, with the simpler models, MEDUSA-2 in particular, exhibiting high equatorial anomalies. Finally, while ERSEM shows good agreement throughout most of the world ocean, both the North Atlantic and North Pacific show anomalously low annual average DIN concentrations.

Formatted: Font: 12 pt

“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).”

Figure 3 shows low correlation (r<0.5) for annual surface chlorophyll concentrations for all models. The models with the highest correlation coefficients are PlankTOM10 (0.49) followed by MEDUSA-2 (0.36). All other models have correlation coefficients <0.2. Anomalously high chlorophyll values in the equatorial Pacific and, especially, the Southern Ocean significantly elevate the spatial variability of Diat-HadOCC above that of observations (and all other models). More generally, with the exception of PlankTOM10, all of the models show some degree of excess chlorophyll in the Southern Ocean, with Diat-HadOCC exhibiting very high concentrations in this relatively unproductive region.

In addition to the ocean properties shown in Figures 1-3, complementary figures for alkalinity, DIC and primary production can be found in the supplementary material (Figures S1-S3). In each case, global annual average fields are shown together with the corresponding Taylor diagram.

Deleted: ,
Deleted: and oxygen
Deleted: S4

Table 3 shows the correlation coefficients and standard deviations normalised relative to observations of the models for all six of the ocean properties (five surface fields plus depth-integrated primary production). These are additionally colour-coordinated according to the rank order of model performance, and the range of correlation coefficients over all of the models is shown for each field. As already suggested above, model performance varies both between fields and between models. All models perform consistently and relatively well for DIN and DIC in part because of the “memory” of initial distributions. Model performance varies more widely for pCO_2 and primary production and varies most widely for chlorophyll, although it is consistently poor across all models.

Deleted: seven
Deleted: six
Deleted: ,
Deleted: and, especially, oxygen,
Deleted: The excellent performance of all mode for surface oxygen reflects the dominance of temperature-based solubility and air-sea gas exchange over biological activity for this ocean property.

Figure 4 summarises the data in Table 3 by showing the distribution of performance rankings (both correlation coefficients and normalised standard deviations) across the selected fields for each model, i.e. the number of first, second, etc., rankings for each model. No model is shown to consistently outperform all other models across all metrics. Indeed all models perform best in at least one metric, and similarly all models perform worst in at least one metric. There is little discernable relationship between model complexity and model performance. Indeed Table 3 shows that for 4 out of 6 fields the best performing model in terms of correlation coefficients is a simpler model (i.e. HadOCC, Diat-HadOCC or MEDUSA-2) and for 5 out of 6 fields the best performing model in terms of normalised standard deviations is a more complex model (i.e. PlankTOM6, PlankTOM10 or ERSEM).

Deleted: or underperform
Deleted: field
Deleted: field
Deleted: 5
Deleted: 7
Deleted: 7

These findings in annual average model performance are found to be consistent when examined at monthly timescales (Figure 5).

Deleted: S5

3.1.2 Depth profiles

While the majority of biological activity in the ocean is concentrated in its surface layers, biogeochemical fields in the deep ocean have a complex structure created through the interaction of ocean physics with biologically-mediated processes such as export and remineralisation. As such, model performance cannot be solely assessed from surface fields of ocean BGC properties. To examine this, Figures 6 and 7 show the annual average depth profiles of DIC and alkalinity for three important regions: the North Atlantic (Atlantic 0-60°N), Southern Ocean (≥60°S) and Equatorial Pacific (Pacific Ocean 15°S-15°N).

In Figure 6, all models are shown to capture the DIC profile in the Equatorial Pacific though HadOCC, Diat-HadOCC and MEDUSA-2 are somewhat closer to observations than ERSEM and the PlankTOM models. A similar situation is seen in the North Atlantic where the depth profiles of MEDUSA-2, HadOCC and Diat-HadOCC are closest to observations, although surface agreement is greater than that at depth. All models are shown to perform relatively poorly in the Southern Ocean, with much weaker gradients with depth than observations. HadOCC, Diat-HadOCC and ERSEM show gradients that are marginally closer to that observed, but all of the models consistently fail to reproduce the observed >100 mmol m^{-3} surface-1000m increase. 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).

The annual average depth profiles of alkalinity are shown in Figure 7. In the North Atlantic, HadOCC and Diat-HadOCC are closer to observations while ERSEM and, particularly, MEDUSA-2 are further away from observations (but in opposite directions). Again, and for the same reasons as outlined above, no model performs well at capturing the depth profile observed in the Southern Ocean. In the Equatorial Pacific all of the models have similar alkalinity at depth but diverge from observations towards the surface. The near-surface depth profiles in HadOCC, Diat-HadOCC and MEDUSA-2 are closest to observations in that region. Alkalinity shows very little variability with depth in the PlankTOM6, PlankTOM10 and ERSEM models and is higher than observations in near-surface waters (>100 meq m^{-3}). This excess alkalinity may explain the broadly lower pCO_2 values visible in this region in Figure 1. 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 depth profiles of DIN and O_2 are given in the supplementary material (Figures S4-5).

Deleted: 5
Deleted: 6
Deleted: °
Deleted: °
Deleted: °
Deleted: °
Deleted: 5
Deleted: shallower
Deleted: This
Deleted: suggests that the cause lies with ocea physics deficiencies
Deleted: .
Deleted: 6. Again, and for the same reasons, nc model performs well at capturing the depth profi observed in the Southern Ocean
Deleted: values
Deleted: S6-7

3.2 Computational benchmarking

Computational timing tests (CPU time) were carried out relative to the ocean component of the HadGEM3 (Hewitt et al., 2011) model (ORCA1.0L75), on standard configurations of 128 and 256 processors on an IBM Power7 machine. As would be intuitively expected, the cost of candidate ocean biogeochemical models is found to be higher for models with more tracers regardless of the number of processors used. While there are deviations in both directions between the models, broadly there is a linear relationship between number of model tracers and compute cost (Figure S6) reflecting the significant cost of applying advection and mixing terms to each tracer.

Deleted: -

Deleted: (MONSooN).

Deleted: S7). This unsurprisingly reflects

Deleted: performing ocean physics operations biogeochemical tracers

Using ERSEM (the computationally most expensive model) increases computational cost approximately 6-fold relative to HadOCC when 128 processors are used. This relative increase in computational cost is reduced to approximately 4.5-fold when 256 processors are used. PlankTOM10 has the greatest relative reduction (36.6%) in computational cost when run on 256 processors as opposed to 128, although this model would still increase the total cost of the ocean component by a factor of 5 relative to a physics-only ocean, compared to a factor of 1.5 for HadOCC (Table 4).

4 Discussion

Our model comparison suggests that for global annual average surface fields, global monthly average surface fields and annual average depth profiles in three oceanographic regions there is little evidence that increasing the complexity of OBGC models leads to improvements in the representation of large scale ocean patterns of bulk properties. In some cases, the comparison suggests that simpler OBGC are closer to observations than intermediate or complex models for the standard assessment metrics used here.

The biologically simpler models HadOCC, Diat-HadOCC and MEDUSA-2 are shown to have generally higher global spatial pattern correlation coefficients of pCO_2, DIC and alkalinity at both annual and monthly temporal resolution (Figures 1, 5 and Table 3). The more complex models PlankTOM6, PlankTOM10 and, in the case of DIC, ERSEM, have annual and monthly standard deviations that are generally closer to observations than the simplest two models (HadOCC and Diat-HadOCC). As such, we find no robust relationship between model complexity and model skill at capturing global scale distributions of surface pCO_2, DIC and alkalinity. The biologically simpler models are shown to generally best capture the depth

Deleted: of

Deleted: have

Deleted: S5

Deleted: of

profiles of DIC and alkalinity in the North Atlantic and Equatorial Pacific (Figures 6-7), possibly because their biological export production can more easily be tuned to maintain the observed vertical gradients.

There are however ocean biogeochemical fields where models of greater biological complexity tend to equate to improved model skill. The annual and monthly global correlation coefficients of the PlankTOM models are shown to be closest to observations for chlorophyll and primary production fields (Figures 3 and Table 4). These PlankTOM models do not consistently produce the annual chlorophyll and primary production field standard deviations closest to observations (Table 4), however at monthly resolution their field standard deviations are the most consistent across models (Figure 5).

The comparison of depth profiles shows that despite all models being initialised from the same observational fields, there is quite a lot of divergence even at depths of less than 1000m. In some cases, such as alkalinity in the Southern Ocean (Figure 7), all models have a similar systematic bias compared to observations. This is suggestive of the influence of errors within the physical ocean model. That is, the ocean biogeochemistry may be influenced to a greater extent by the physical ocean model and hence there is a common response across models. For other fields such as DIN in the Southern Ocean and Equatorial Pacific (Figure S5), models have both positive and negative biases compared to observations suggestive of a greater relative role of the OBGC model than the physical model.

It is clear that more biologically complex models are required to more completely assess the impacts of environmental change on marine ecosystems. By representing processes that are not present in simpler models, the more complex models are also able to represent additional factors such as climatically-active gases (e.g. DMS, N_2O). Assessment of such representations however fell outside the scope of this paper. Models of intermediate complexity (e.g. Diat-HadOCC and MEDUSA-2) are shown in this inter-comparison to reproduce large scale ocean biogeochemistry features relatively well, yet minimise computational cost and have sufficient biological complexity to allow important ESM questions to be explored, including those that require an explicit iron cycle (e.g. ocean iron fertilisation).

It should be noted that models implemented within the NEMO physical ocean framework prior to this inter-comparison project had an advantage over those new to this framework. This is a somewhat unavoidable consequence of what is also one of this inter-comparison study's main strengths, namely that the models were adapted to use the same ocean physics framework. Specifically, the HadOCC and MEDUSA-2 model developers were familiar with NEMO v3.2 and had some previous opportunity to tune models. Linked to this is the

Deleted: oceanographic regions of the

Deleted: 5-

Formatted: Font: 12 pt

Deleted: ERSEM, the most complex model assessed, best captures the observed global annu spatial patterns in DIN both in terms of spatial correlation coefficient and standard deviation (Figure 2). At monthly resolution ERSEM global D fields consistently have the highest correlation coefficients however field standard deviations ar not always the closest to observations (Figure S5) should however be noted that all models have relatively high correlation coefficients for DIN. In addition the

Deleted:)

Deleted: S5

Deleted: 6

Deleted: discrepancies

Deleted: alternative

Deleted: S7), however,

Deleted: tend

Deleted: models that

Deleted: previously implemented within NEMC v3.2 were "

Deleted: "

Deleted: this ocean model's configuration

Deleted: flaws. Meanwhile, the ERSEM model, which previously

Deleted: limited use within the context of the global open ocean, had a distinct disadvantage.

question of how dependent the results were on parameter values. Although model developers were afforded a limited opportunity to tune parameters, given further time to tune one would expect improved performance, especially for those models that had not been previously implemented within NEMO v3.2.

Deleted: found

Deleted: settings

The rationale for the chosen fields of intercomparison was, as stated previously, that they are common across all models and are key facets of global marine biogeochemistry. It could however be argued that these bulk fields were insufficient to adequately assess all models and in particular the most complex models. Further analysis, beyond the scope of this paper will undertake as thorough an analysis of the biological components as each model will support.

Deleted: Specifically this will aim to establish if bottom-up approach to model skill assessment focused on relationships between properties (e.g Vetter et al., 2008) can identify differences in performance related to the OBGC model complexity.

Finally although computational cost is discussed as a pragmatic driver of OBGC model selection, it should be noted that computer power is continuously increasing and the intercomparison results presented here may differ for an alternative spatial resolution ocean grid requiring greater computational resources. In addition, ongoing efforts to transport passive ocean tracers on degraded spatial scales (e.g. Levy et al., 2012) have the potential to result in computational savings that would realistically permit the implementation of higher complexity OBGC models within ESMs.

5 Conclusions

The 6 ocean biogeochemical models analysed within this inter-comparison cover a large range of ecosystem complexity (from 7 tracers in HadOCC to 57 in ERSEM), and therefore result in a range of approximately 5 in computational costs (from increasing the cost of the physical ocean model by a factor of 2 to a factor of 10). Results suggest little evidence that higher biological complexity implies better model performance in reproducing observed global-scale bulk properties of ocean biogeochemistry.

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.

One priority for the next generation of Earth System Models (CMIP6) is to enhance model resolution in the hope that it will resolve some of the existing biases in climate models. This puts pressure on the computing time available for representing biological complexity. Our results suggest that intermediate complexity models (such as MEDUSA-2 and Diat-HadOCC) offer a good compromise between the representation of biological complexity (through their inclusion of an iron cycle) and computer time, given their relatively good performance in reproducing bulk properties. However, intermediate complexity models are limited in the detail to which they can address climate feedbacks and it may be that more complex models can in future provide additional insight, based on ongoing measurements and data syntheses.

The quest for increasing resolution in ESMs is unlikely to end soon, as the resolution needed to resolve eddies in the ocean (1/8 degree or less) needs to be achieved before important improvements in representing climate dynamics are achieved. Most ESMs being developed for the next CMIP phase will have a grid of 1/2 to 1/4 degree. Even with increasing computational power and schemes for accelerating transport of passive tracers (Levy et al., 2012) available, other priorities (e.g. ensemble simulations for risk assessments) may still make it difficult to prioritise the representation of biogeochemical complexity in ESMs. In order to achieve scientific progress on important questions of the interactions between marine biogeochemistry and climate, it is thus important that lower resolution ESMs that prioritise biogeochemical complexity are maintained and used in CMIP exercises in parallel to higher resolution models.

Deleted: Only with a dual low and high resoluti strategy can we ensure that the priorities of improving climate dynamics and those of scientific exploration can be achieved. Such a strategy would also help support a closer integration between the assessment of climate change science and that of climate chang impacts, and help ensure more integration betwe IPCC's working groups.

Acknowledgements

This work was funded by the UK Natural Environmental Research Council **Integrated Marine Biogeochemical Modelling Network to Support UK Earth System Research (i-MarNet) project** (NE/K001345/1) and the UK Met Office. MFR was partially funded by the EC FP7 GreenSeas project.

References

Antonov, J. I., Seidov, D., Boyer, T. P., Locarnini, R. A., Mishonov, A. V., Garcia, H. E., Baranova, O. K., Zweng, M. M., and Johnson, D. R.: World ocean atlas 2009, volume 2: Salinity, in: NOAA Atlas NESDIS 69, edited by: Levitus, S., US Government Printing Office, Washington, DC, USA, 184 pp., 2010.

Artioli, Y., Blackford, J. C., Butenschön, M., Holt, J. T., Wakelin, S. L., Thomas, H., Borges, A. V., and Allen, J. I.: The carbonate system in the north sea: Sensitivity and model validation, J. Marine Syst., 102-104, 1 - 13, 2012

Barange, M., Merino, G., Blanchard, J. L., Scholtens, J., Harle, J., Allison, E. H., Allen, J. I., Holt, J., and Jennings, S.: Impacts of climate change on marine ecosystem production in societies dependent on fisheries. Nat. Clim. Change, doi:10.1038/NCLIMATE2119, 2014.

Baretta, J. W., Ebenhoh, W., and Ruardij, P.: The European Regional Seas Ecosystem Model, a complex marine ecosystem model. Neth. J. Sea Res. 33, 233–246,1995.

Behrenfeld, M. J. and Falkowski, P. G.: Photosynthetic rates derived from satellite-based chlorophyll concentration, Limnol. Oceanogr., 42, 1–20, 1997.

Bopp, L., Resplandy, L., Orr, J. C., Doney, S. C., Dunne, J. P., Gehlen, M., Halloran, P., Heinze, C., Ilyina, T., Séférian, R., Tjiputra, J., and Vichi, M.: Multiple stressors of ocean ecosystems in the 21st century: projections with CMIP5 models, Biogeosciences, 10, 3627-3676, 2013.

Blackford, J. C.: An analysis of benthic biological dynamics in a North Sea ecosystem model, J. Sea Res., 38, 213-230, 1997.

Blackford, J. C., Allen, J. I., and Gilbert, F. J.: Ecosystem dynamics at six contrasting sites: a generic modelling study, J. Marine Syst., 52, 191-215, 2004.

Blackford, J. C. and Gilbert, F. J.: pH variability and CO2 induced acidification in the North Sea, J. Marine Syst., 64, 229-241, 2007.

Braconnot, P., Harrison, S. P., Kageyama, M., Bartlein, P. J., Masson-Delmotte, V., Abe-Ouchi, A., Otto-Bliesner, B., and Zhao, Y.: Evaluation of climate models using palaeoclimatic data. Nat. Clim. Change, 2, 417-424, 2012.

Buesseler, K. O., and Boyd, P. W.: Will ocean fertilisation work? Science, 300, 67-68, 2003.

Buitenhuis, E., Le Quéré, C., Aumont, O., Beaugrand, G., Bunker, A., Hirst, A., Ikeda, T., O'Brien, T., Piontkovski, S., and Straile, D.: Biogeochemical fluxes through mesozooplankton. Global Biogeochem. Cy., 20, 2006, GB2003, doi:10.1029/2005GB002511.

Buitenhuis, E., Rivkin, R., Sailley, S., Le Quere, C.: Biogeochemical fluxes through microzooplankton, Global Biogeochem. Cy., 24, 2010, GB4015, doi:10.1029/2009GB003601.

Moved (insertion) [1]

Formatted: Font: +Body (Calibri)

Moved (insertion) [2]

Formatted: Font: +Body (Calibri)

Moved (insertion) [3]

Formatted: Font: +Body (Calibri)

Moved (insertion) [4]

Formatted: Font: +Body (Calibri)

Moved (insertion) [5]

Formatted: Font: +Body (Calibri), 12 pt

Formatted: Space After: 0 pt, Line spacing single, Widow/Orphan control, Don't adjust space between Latin and Asian text, Don't adjust space between Asian text and numbe

Formatted: Font: 12 pt

Deleted: *, 24*

Formatted: Default Paragraph Font, Font: Calibri, 11 pt, Font color: Auto, English (U.S.

Buitenhuis, E. Vogt, M., Moriarty, R., Bednaršk, N., Doney, S., Leblanc, K., Le Quéré, C., Luo, Y. -W., O'Brien, C., O'Brien, T., Peloquin, J. Schiebel, R., and Swan, C.: MAREDAT: towards a world atlas of MARine Ecosystem DATa. Earth System Science Data, 5, 2013.

Carr, M. -E., Friedrichs, M. A. M., Schmeltz, M., Aita, M. N., Antoine, D., Arrigo, K. R., Asanuma, I., Aumont, O., Barber, R., Behrenfeld, M., Bidigare, R., Buitenhuis, E. T., Campbell, J., Ciotti, A., Dierssen, H., Dowell, M., Dunne, J., Esaias, W., Gentili, B., Gregg, W., Groom, S., Hoepffner, N., Ishizaka, J., Kameda, T., Le Quéré, C., Lohrenz, S., Marra, J., Mélin, F., Moore, K., Morel, A., Reddy, T. E., Ryan, J., Scardi, M., Smyth, T., Turpie, K., Tilstone, G., Waters, K., and Yamanaka, Y.: A comparison of global estimates of marine primary production from ocean color, Deep-Sea Res. Pt. II, 53, 741–770, 2006.

Cheung, W. W. L., Sarmiento, J. L., Dunne, J. P., Frölicher, T. L., Lam, V. W. Y., Palomares, M. L. D., Watson, R. and Pauly, D.: Shrinking of fishes exacerbates impacts of global ocean changes on marine ecosystems, Nat. Clim. Change, 3, 254-258, 2012.

Condon, R., Duarte, C. M., Pitt, K. A., Robinson, K., Lucas, C. H., Sutherland, K. R., Mianzan, H. W., Bogeberg, M., Purcell, J. E., Decker, M. B., Uye, S., Madin, L. P., Brodeur, R. D., Haddock, S. H. D., Malej, A., Parry, G. D., Eriksen, E., Quiñones, J. Acha, M., Harvey, M., Arthur, J. M., and Graham, W. M. :Recurrent jellyfish blooms are a consequence of global oscillations, P. Natl. Acad. Sci. USA, 110, 1000-1005, 2013.

Collins, W. J., Bellouin, N., Doutriaux-Boucher, M., Gedney, N., Halloran, P., Hinton, T., Hughes, J., Jones, C. D., Joshi, M., Liddicoat, S., Martin, G., O'Connor, F., Rae, J., Senior, C., Stich, S., Totterdell, I. J., Wiltshire, A., and Woodward, S.: Development and evaluation of an Earth-system model–HadGEM2, Geosci. Model Dev., 4, 1051–1075, doi:10.5194/gmd-4-1051-2011, 2011.

Cox, P. M., Betts, R. A., Jones, C. D., Spall, S. A., and Totterdell, I. J.: Acceleration of global warming due to carbon-cycle feedbacks in a coupled climate model, Nature, 408, 184-187, 2000.

Dickson, A. G., and Goyet, C.: Handbook of methods for the analysis of the various parameters of the carbon dioxide system in sea water. Version 2 (No. ORNL/CDIAC--74). Oak Ridge National Lab., TN (United States), 1994.

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.

Dunne, J. P., Sarmiento, J. L., and Gnanadesikan A.: A synthesis of global particle export from the surface ocean and cycling through the ocean interior and on the seafloor, Global Biogeochem. Cy., 21, 2007, GB4006, doi:10.1029/2006GB002907.

Formatted: Font: 12 pt

Moved (insertion) [6]

Formatted: Font: +Body (Calibri)

Friedrichs, M.A.M., Dusenberry, J.A., Anderson, L.A., Armstrong, R.A., Chai, F., Christian, J.R., Doney, S.C., Dunne, J., Fujii, M., Hood, R., McGillicuddy, D.J., Moore, J.K., Schartau, M., Spitz, Y.H., Wiggert, J.D.,: Assessment of skill and portability in regional marine biogeochemical models: Role of multiple planktonic groups. J. Geophys. Res. Oceans 112, C08001. 2007.

Friedlingstein, P., Cox, P., Betts, R., Bopp, L., Von Bloh, W., Brovkin, V., Cadule, P., Doney, S., Eby, M., Fung, I., Bala, G., John, J., Jones, C., Joos, F., Kato, T., Kawamiya, M., Knorr, W., Lindsay, K., Matthews, H. D., and Raddatz, T.: Climate–Carbon Cycle Feedback Analysis: Results from the C4MIP Model Intercomparison. J. Climate, 19, 337-3353, 2006.

Formatted: Font: +Body (Calibri)

Garcia, H. E., Locarnini, R. A., Boyer, T. P., Antonov, J. I., Baranova, O. K., Zweng, M. M., and Johnson, D. R.: World Ocean Atlas 2009, Volume 3: Dissolved Oxygen, Apparent Oxygen Utilization, and Oxygen Saturation, edited by: Levitus, S., NOAA Atlas NESDIS 70, US Government Printing Office, Washington DC, 344 pp., 2010a.

Formatted: Font: 12 pt

Kriest, I., Khatiwala, S., and Oschlies, A.: Towards an assessment of simple global marine biogeochemical models of different complexity. Prog. Oceanogr., 86(3), 337-360, 2010.

Garcia, H. E., Locarnini, R. A., Boyer, T. P., Antonov, J. I., Zweng, M. M., Baranova, O. K., and Johnson, D. R.: World Ocean Atlas 2009, Volume 4: Nutrients (phosphate, nitrate, silicate), edited by: Levitus, S., NOAA Atlas NESDIS 71, US Government Printing Office, Washington DC, 398 pp., 2010b.

Formatted: Font: 12 pt

Gaspar, P., Grégoris, Y., and Lefevre, J.-M.: A simple eddy kinetic energy model for simulations of the oceanic vertical mixing: Tests at station Papa and long-term upper ocean study site. J. Geophys. Res. Oceans 95, 16179–16193, 2010.

Glibert, P.M., Allen, J.I., Artioli, Y., Beusen, A., Bouwman, L., Harle, J., Holmes, R., Holt, J.: Vulnerability of coastal ecosystems to changes in harmful algal bloom distribution in response to climate change: projections based on model analysis Global Change Biol., 20(12), 3845–3858, 2014, doi: 10.1111/gcb.12662

Gnanadesikan, A., Slater, R. D., Gruber, N., and Sarmiento, J. L.: Oceanic vertical exchange and new production: a comparison between models and observations, Deep-Sea Res. II, 49, 363–401, 2002.

Gruber, N.: Warming up, turning sour, losing breath: ocean biogeochemistry under global change. Phil Trans R-Soc-A, 369, 1980-1996, 2011.

Halloran, P. R., Bell, T. G., and Totterdell, I. J.: Can we trust empirical marine DMS parameterisations within projections of future climate? Biogeosciences, 7, 1645-1656, 2010.

Harvey, L. D. D.: Mitigating the atmospheric CO2 increase and ocean acidification by adding limestone powder to upwelling regions. Journal of Geophysical Research: Oceans (1978–2012), 113(C4), 2008.

Hewitt, H. T., Copsey, D., Culverwell, I. D., Harris, C. M., Hill, R .S. R., Keen, A. B., McLaren, A. J., and Hunke, E. C.: Design and implementation of the infrastructure of HadGEM3: the next-generation Met Office climate modelling system, Geosci. Model Dev., 4, 223-253, 2011.

Hunke, E. C., and Lipscomb, W. H.: CICE: The los alamos sea ice model, documentation and software user's manual, version 4.0, Los Alamos National Laboratory Tech. Rep. LA-CC-06, 2008.

Jackson , D. A., Peres-Neto, P. R., and Olden. J. D.: What controls who is where in freshwater fish communities – the roles of biotic, abiotic, and spatial factors. Can. J. Fish. Aquat. Sci. 58, 157-170, 2001.

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.

Key, R. M., Kozyr, A., Sabine, C. L., Lee, K., Wanninkhof, R., Bullister, J. L., Feely, R. A., Millero, F. J., Mordy, C., and Peng, T. -H.: A global ocean carbon climatology: results from Global Data Analysis Project (GLODAP), Global Biogeochem. Cy., 18, GB4031, doi:10.1029/2004GB002247, 2004.

Kheshgi, H. S.: Sequestering atmospheric carbon dioxide by increasing ocean alkalinity, Energy, 20, 915-922, 1995.

Large, W., and Yeager, S.: The global climatology of an interannually varying air–sea flux data set, Clim. Dynam., 33, 341–364, 2009.

Le Quéré, C., Harrison, S. P., Prentice, I. C., Buitenhuis, E. T., Aumont, O., Bopp, L., Claustre, H., Da Cunha, L. C., Geider, R., Giraud, X., Klaas, C., Kohfeld, K. E., Legendre, L., Manizza, M., Platt, T., Rivkin, R. B., Sathyendranath, S., Uitz, J., Watson, A. J., and Wolf-Gladrow, D.: Ecosystem dynamics based on plankton functional types for global ocean biogeochemistry models, Glob. Change Biol., 11, 2016–2040, doi:10.1111/j.1365-2468.2005.01004.x, 2005.

Lévy, M., Resplandy, L., Klein, P., Capet, X., and Ethé, C.: Grid degradation of submesoscale resolving ocean models: benefits for offline passive tracer transport, Ocean Model., 48, 1-9, 2012.

Locarnini, R. A., Mishonov, A. V., Antonov, J. I., Boyer, T. P., Garcia, H. E., Baranova, O. K., Zweng, M. M., and Johnson, D. R.: World ocean atlas 2009, volume 1: Temperature, in: NOAA Atlas NESDIS 68, edited by: Levitus, S., US Government Printing Office, Washington, DC, USA, 184 pp., 2010.

Madec, G.: NEMO reference manual, ocean dynamic component: NEMO–OPA, Note du Pôle de modélisation, Institut Pierre Simon Laplace, Technical Report 27, Note du pôle de modélisation, Institut Pierre Simon Laplace, France, No. 27, ISSN No. 1288–1619, 2008.

Formatted: Font: 12 pt

Monterey, G. and Levitus, S.: Seasonal Variability of Mixed Layer Depth for theWorld Ocean, 96 pp. NOAA Atlas NESDIS 14, US Gov. Printing Office, Washington DC, 1997.

Nightingale, P., Malin, G., Law, C., Watson, A., Liss, P., Liddicoat, M., Boutin, J., and Upstill-Goddard, R.: In situ evaluation of air-sea gas exchange parameterizations using novel conservative and volatile tracers, Global Biogeochem. Cy., 14, 373–387, 2000.

O'Reilly, J. E., Maritorena, S., Mitchell, B. G., Siegal, D. A., Carder, K. L., Garver, S. A., Kahru, M., and McClain, C.: Ocean color chlorophyll algorithms for SeaWiFS, J. Geophys. Res., 103, 24937–24953, 1998.

Orr, J., Najjar, J., Sabine, C., and Joos, F.: Abiotic-HOWTO, OCMIP-2 Project, 29 pp., online available at: www.ipsl.jussieu.fr/OCMIP/, 1999.

Palmer, J. R., and Totterdell, I. J.: Production and export in a global ocean ecosystem model, Deep-Sea Res. Pt I, 48, 1169-1198, 2001.

Popova, E.E., Yool, A., Coward, A.C., Dupont, F., Deal, C., Elliott, S., Hunke, E., Jin, M., Steele, M., and Zhang, J.: What controls primary production in the Arctic Ocean? Results from an intercomparison of five general circulation models with biogeochemistry. J. Geophys. Res. Oceans 117, C00D12, 2012.

Reid, P. C., Fischer, A. C., Lewis-Brown, E., Meredith, M. P., Sparrow, M., Andersson, A. J., Antia, A., Bates, N. R., Bathmann, U., Beaugrand, G., Brix, H., Dye, S., Edwards, M., Furevik, T., Gangstø, R., Hátún, H., Hopcroft, R. R., Kendall, M., Kasten, S., Keeling, R., Le Quéré, C., Mackenzie, F. T., Malin, G., Mauritzen, C., Ólafsson, J., Paull, C., Rignot, E., Shimada, K., Vogt, M., Wallace, C., Wang, Z., and Washington, R.: Chapter 1 Impacts of the Oceans on Climate Change, Adv. Mar. Biol., 56, 1-150, 2009.

Siddorn, J. R., Allen, J. I.,Blackford, J. C., Gilbert, F. J., Holt, J. T., Holt, M. W., Osborne, J. P., Proctor, R., and Mills, D. K.: Modelling the hydrodynamics and ecosystem of the North-West European continental shelf for operational oceanography, J. Mar. Syst., 65, 417–429, doi:10.1016/j.jmarsys.2006.01.018, 2007.

Steinacher, M., Joos, F., Frölicher, T.L., Bopp, L., Cadule, P., Cocco, V., Doney, S.C., Gehlen, M., Lindsay, K., Moore, J.K., Schneider, B., and Segschneider, J.: Projected 21st century decrease in marine productivity: a multi-model analysis. Biogeosciences 7, 979–1005, 2010.

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.

Takahashi, T., Sutherland, S. C., Wanninkhof, R. Sweeney, C., Feely, R. A., Chipman, D. W., Hales, B., Friederich, G., Chavez, F., Sabine, C., Watson, A., Bakker, D. C. E., Schuster, U., Metzl, N., Yoshikawa-Inoue, H., Ishii, M. Midorikawa, T., Nojiri, Y., Kortzinger, A., Steinhoff, T., Hoppema, M., Olafsson, J., Arnarson, T. S., Tillbrook, B., Johannessen, T., Olsen, A., Bellerby, R., Wong, C. S., Delille, B., Bates, N. R., and de Baar, H. J. W.: Climatological mean and decade change in surface ocean pCO_2, and net sea–air CO_2 flux over the global oceans, Deep-Sea Res. Pt. II, 56, 554–577, doi:10.1016/j.dsr2.2008.12.009, 2009.

Taylor, K. E.: Summarizing multiple aspects of model performance in a single diagram, J. Geophys. Res., 106, 7183–7192, 2001.

Westberry, T., Behrenfeld, M. J., Siegel, D. A., and Boss, E.: Carbon-based primary productivity modeling with vertically resolved photoacclimation, Global Biogeochem. Cy., 22, GB2024, doi:10.1029/2007GB003078, 2008.

Yool, A., Popova, E. E., and Anderson, T. R.: Medusa-1.0: a new intermediate complexity plankton ecosystem model for the global domain, Geosci. Model Dev., 4, 381–417, doi:10.5194/gmd-4-381-2011, 2011.

Yool, A., Popova, E. E., and Anderson, T. R.: MEDUSA-2.0: an intermediate complexity biogeochemical model of the marine carbon cycle for climate change and ocean acidification studies, Geosci. Model Dev. , 6, 1767–1811, doi:10.5194/gmd-6-1767-2013, 2013.

Moved up [6]: G.,

Moved up [4]: M.,

Moved up [3]: K.,

Deleted: Jung,

Formatted: Font: +Body (Calibri)

Moved up [1]: P.,

Deleted: Viovy, N., Heimann M.: Analyzing the causes and spatial pattern of the European 2003 carbon flux anomaly in Europe using seven mode Biogeosciences, 5, 561-583, 2008.¶

Formatted: Font: +Body (Calibri)

Deleted: Feser, F., Freibauer, A., Geyer, R., Jone

Moved up [5]: C.,

Deleted: Papale, D., Tenhunen, J., Tomelleri,

Moved up [2]: E.,

Deleted: Vetter, M., Churkina,

Formatted: Font: +Body (Calibri)

Deleted: Reichstein, M., Zaehle, S., Bondeau, A Chen, Y., Ciais,

Formatted: Font: +Body (Calibri)

Deleted: Trusilova,

Formatted: Font: +Body (Calibri)

Formatted: Font: 12 pt

Table 1. Biogeochemical cycles represented in each candidate model.

	HadOCC	Diat-HadOCC	MEDUSA-2	PlankTOM6	PlankTOM10	ERSEM
N	✓	✓	✓	✓	✓	✓
P					✓	✓
Si		✓	✓	✓	✓	✓
Fe		✓	✓	✓	✓	✓
C	✓	✓	✓	✓	✓	✓
Alkalinity	✓	✓	✓	✓	✓	✓
O_2	✓	✓	✓	✓	✓	✓

Deleted: ¶
¶
¶

Formatted: Font: 12 pt

Deleted: ¶
¶
¶
¶
¶
¶
¶
¶
¶
¶
¶
¶
¶
¶
¶
¶
¶
¶
¶
¶
¶
¶

Formatted: Space Before: 0 pt, After: 10

Table 2. Composition of the marine ecosystems represented in each candidate model, along with the total number of biogeochemical tracers (including those detailed in Table 1).

	HadOCC	Diat-HadOCC	MEDUSA-2	PlankTOM6	PlankTOM10	ERSEM
Generic Phytoplankton	✓	✓		✓	✓	
Diatoms		✓	✓	✓	✓	✓
Large Phytoplankton						✓
Picophytoplankton			✓		✓	✓
Coccolithophores				✓	✓	✓
N_2 fixers					✓	
Flagellates						✓
Phaeocystis					✓	
Generic Zooplankton	✓	✓				
Microzooplankton			✓	✓	✓	✓
Mesozooplankton			✓	✓	✓	✓
Macrozooplankton					✓	
Heterotrophic Nanoflagellates						✓
Picoheterotrophs				✓	✓	✓
Tracers	**7**	**13**	**15**	**25**	**39**	**57**

Formatted: Font: 12 pt

Deleted: Prokaryotes

Deleted: ¶ ¶ ¶ ¶ ¶ ¶ ¶ ¶ ¶ ¶ ¶

Table 3. Model-observation correlation coefficients (*R*) and standard deviations normalised by the standard deviation of observations (σ) for all examined annual surface fields and depth integrated primary productivity. Colours indicate model ranking and are organised through the worst performing model in red to the best performing model in dark blue (through orange, yellow, green and light and dark blue).

Model	**pCO_2**		**DIN**		**Chl.**		**Alkalinity**		**DIC**		**Primary Production**	
	R	**σ**	***R***	**σ**	***R***	**σ**	***R***	**σ**	***R***	**σ**	***R***	**σ**
HadOCC	0.68	1.92	0.88	1.20	0.30	0.68	0.91	1.19	0.93	1.18	0.19	0.92
Diat-HadOCC	0.54	1.77	0.90	1.20	0.15	2.65	0.91	1.19	0.93	1.13	0.13	1.51
MEDUSA-2	0.64	1.56	0.85	1.21	0.36	0.40	0.88	1.14	0.92	1.17	0.64	1.10
PlankTOM6	0.34	1.03	0.79	1.20	0.32	1.08	0.70	0.88	0.75	0.96	0.47	0.61
PlankTOM10	0.29	0.94	0.88	1.19	0.50	0.43	0.58	1.16	0.65	1.08	0.53	0.74
ERSEM	0.01	2.04	0.94	0.95	0.04	0.91	0.84	1.18	0.86	1.07	-0.08	1.12
Range	0.64	1.09	0.15	0.26	0.46	2.64	0.33	0.31	0.28	0.23	0.72	0.90

Deleted: Oxygen
Deleted Cells
Formatted Table
Deleted: *R*
Deleted Cells
Deleted: σ
Deleted Cells
Deleted: 0.99
Deleted Cells
Deleted: 1.07
Deleted Cells
Deleted: 0.99
Deleted: 1.08
Deleted: 0.99
Deleted: 1.02
Deleted: 0.99
Deleted: 1.05
Deleted: 0.99
Deleted: 1.04
Deleted: 0.98
Deleted: 0.97
Formatted Table
Deleted: 0.01
Deleted: 0.11
Deleted: 67
Deleted: 25
Deleted: ¶

Table 4. Computational cost of each candidate model when coupled to the ocean component of HadGEM3, relative to a physics-only simulation with the same ocean model (ORCA1.0L75). A cost of 2.0 indicates that adding the biogeochemistry model doubles total simulation cost. Timings are shown for simulations carried-out on 128 and 256 processors of an IBM Power7 machine.

Model	Cost (128 processors)	Cost (256 processors)
HadOCC	1.75	1.48
Diat-HadOCC	2.36	1.88
MEDUSA-2	2.73	2.10
PlankTOM6	5.11	3.52
PlankTOM10	7.74	4.90
ERSEM	10.36	6.87

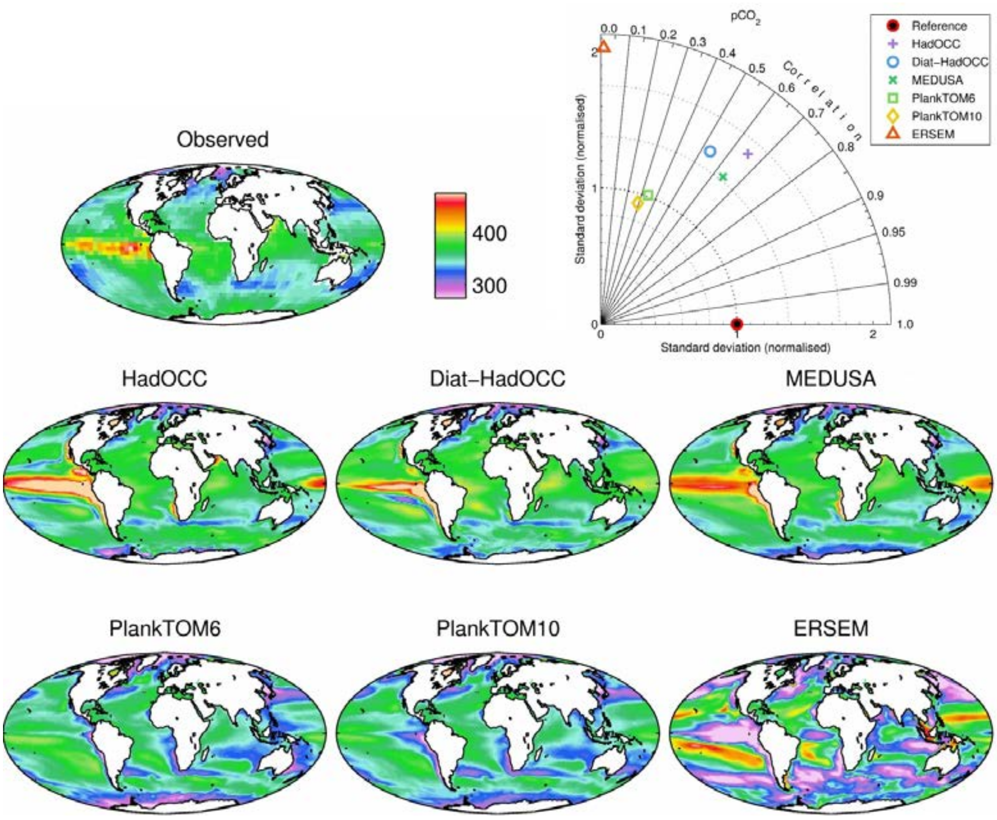


Figure 1. Observational (Takahashi et al., 2009; top left) and modelled annual average surface ocean pCO_2 (µatm) for year 2000. Mean field values: observations 357.7; HadOCC 368.8; Diat-HadOCC 369.2; MEDUSA 368.5; PlankTOM6 349.8; PlankTOM10 349.5; ERSEM 343.0.

Formatted: Space After: 10 pt

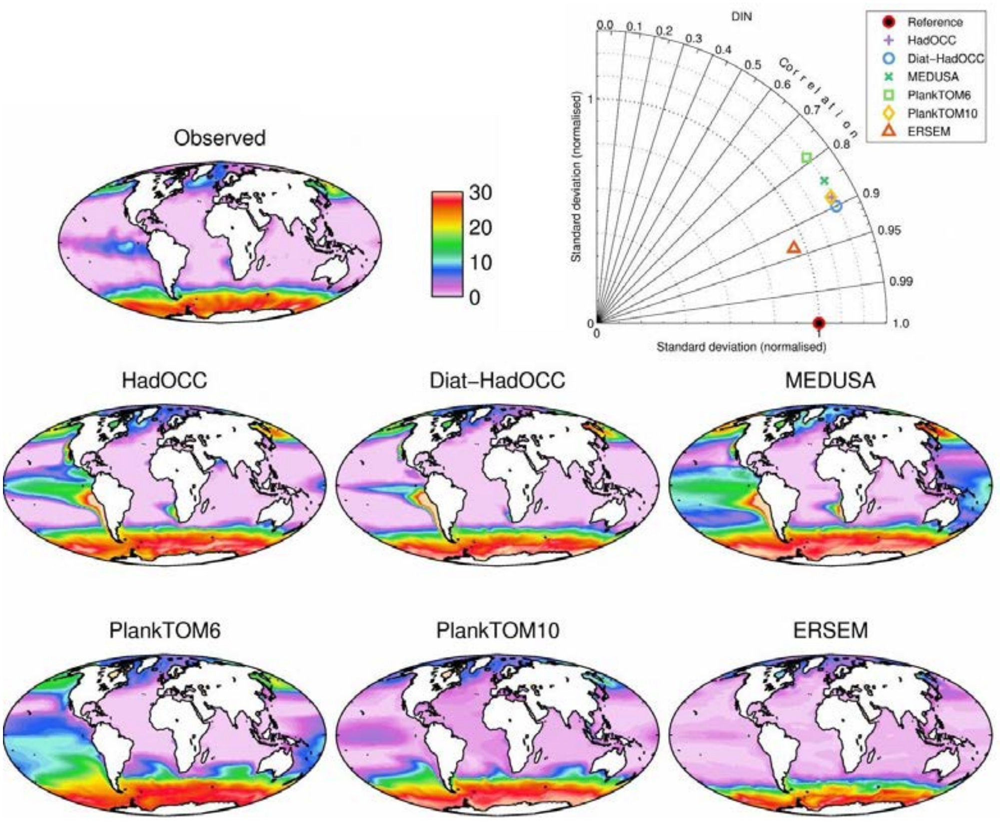


Figure 2. Observational (World Ocean Atlas, 2009; top left) and modelled annual average surface ocean Dissolved Inorganic Nitrogen (mmol m^{-3}) for the period 2000-2004. Mean field values: observations 5.24; HadOCC 7.88; Diat-HadOCC 6.33; MEDUSA 10.18; PlankTOM6 9.45; PlankTOM10 7.25; ERSEM 4.58.

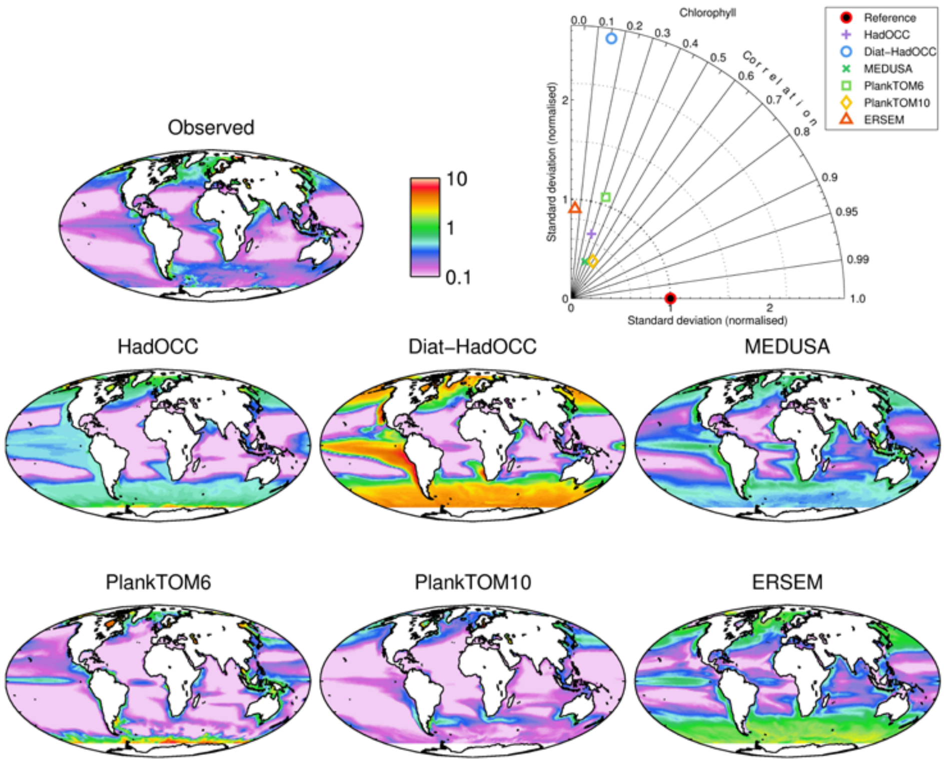
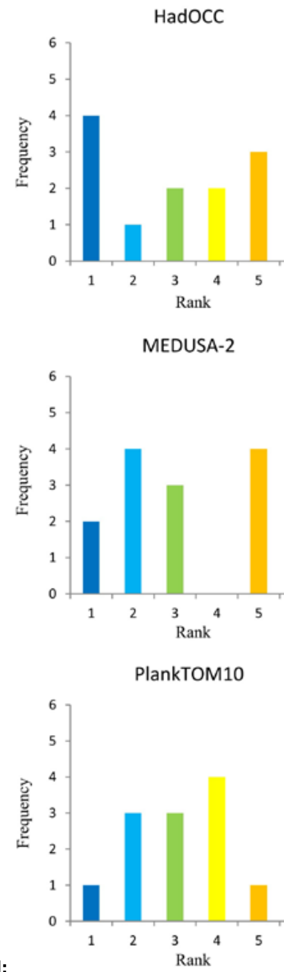


Figure 3. Observational (SeaWiFS; top left) and modelled annual average surface ocean chlorophyll (mg m^{-3}) for the period 2000-2004. To avoid biasing the plots, observational data and model output are only shown for regions in which all months were represented at least once across all of the sampled years. Mean field values: observations 0.215; HadOCC 0.347; Diat-HadOCC 1.170; MEDUSA 0.346; PlankTOM6 0.312; PlankTOM10 0.160; ERSEM 0.501.

Formatted: Space After: 10 pt

Deleted:



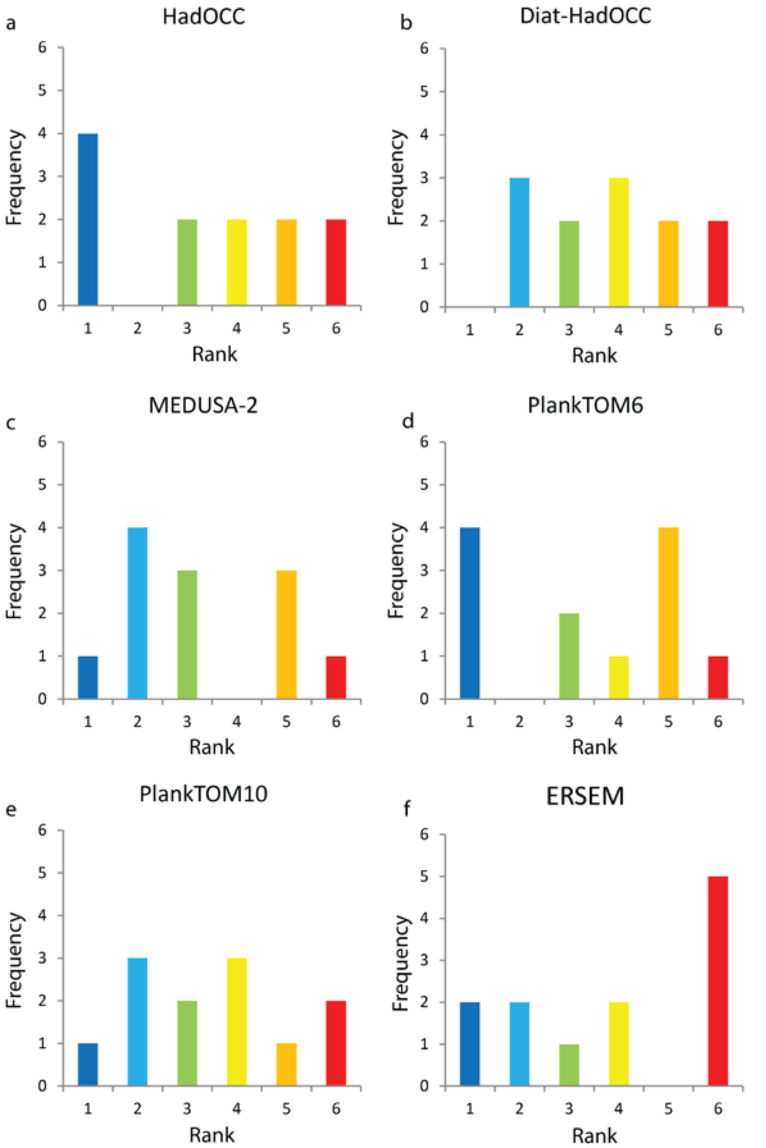


Figure 4. Frequency distributions of best- to worst-performances for each model, in terms of correlation coefficients and normalised standard deviations of annual surface fields and depth integrated primary productivity. Colours follow those of Table 1.

Formatted: Space After: 10 pt

Deleted: or

Deleted: table

Formatted: Font: 12 pt, Bold

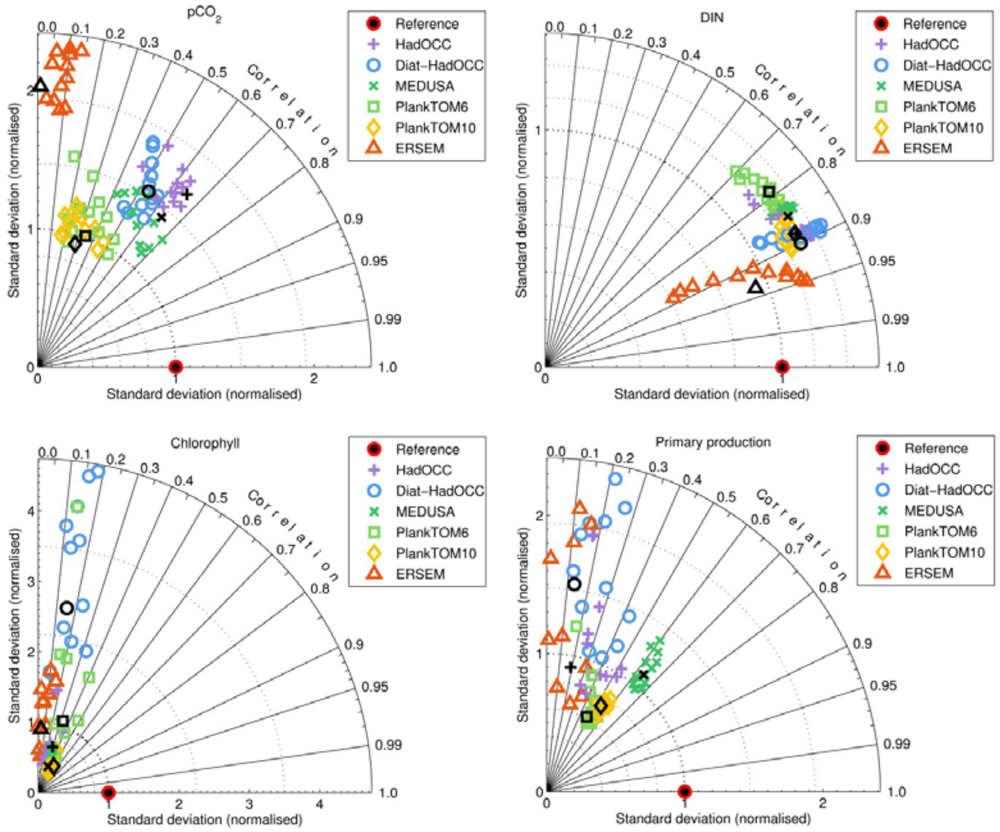


Figure 5. Monthly Taylor plots for pCO_2, Dissolved Inorganic Nitrogen (DIN), chlorophyll and primary production for all models relative to observations. Annual averages are shown in black. Note that negative correlation coefficients are not shown in the Taylor plot.

Deleted: ¶
¶
¶
¶

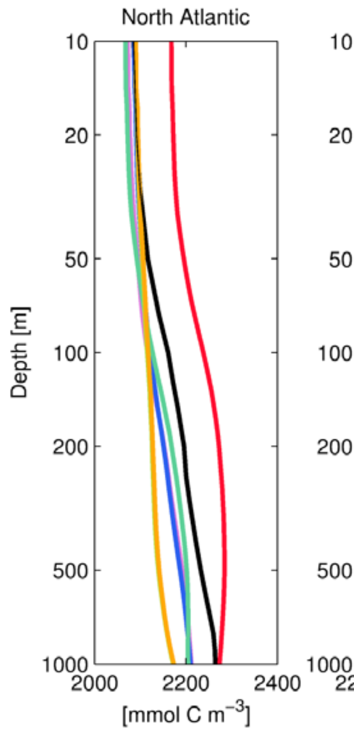
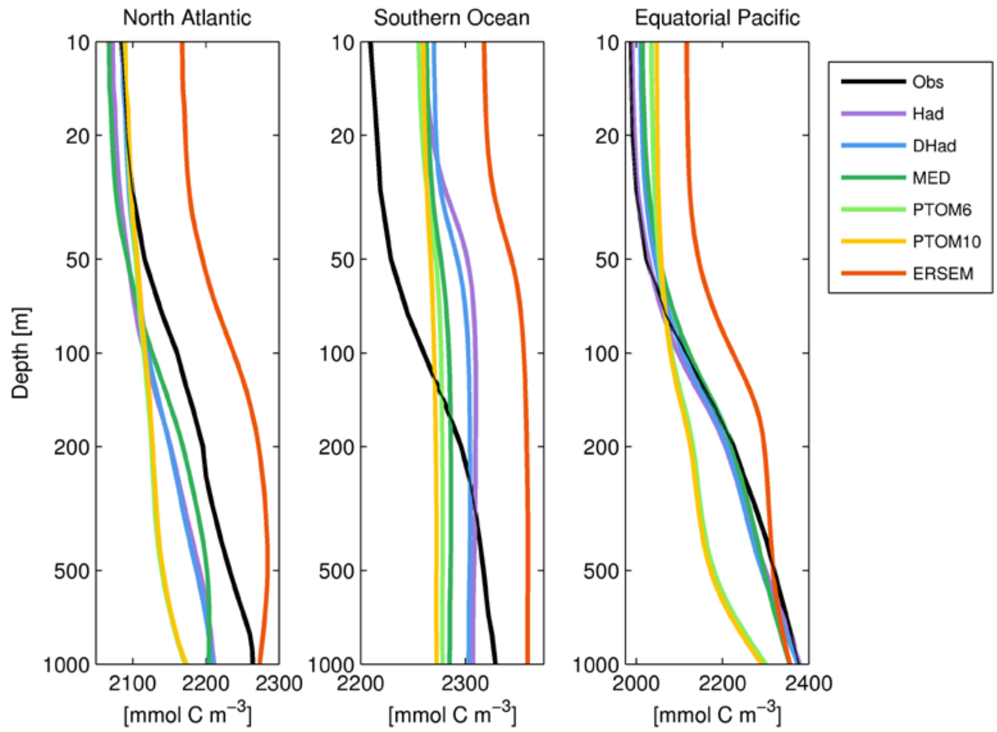


Figure 5.



1590

1591 Figure 6. Observed (black; GLODAP) and modelled profiles of Dissolved Inorganic Carbon
1592 (mmol C m^{-3}) in the North Atlantic (0°N to 60°N), Southern Ocean (90°S to 60°S) and
1593 Equatorial Pacific (15°S to 15°N). Vertical scaling is logarithmic ($\log_{10}$).

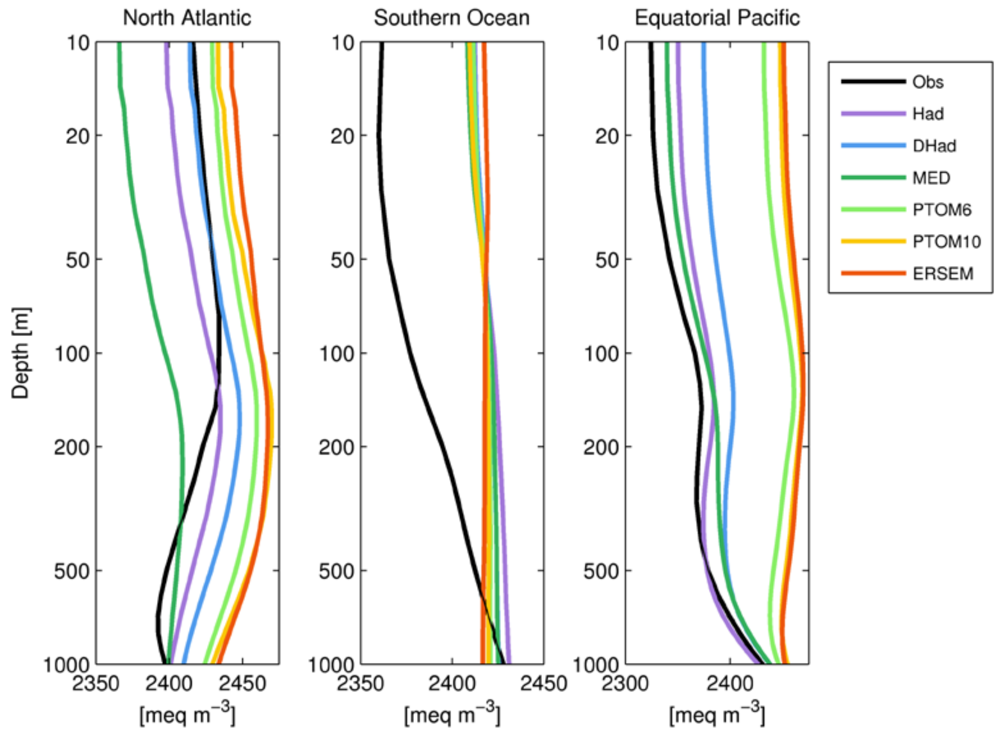
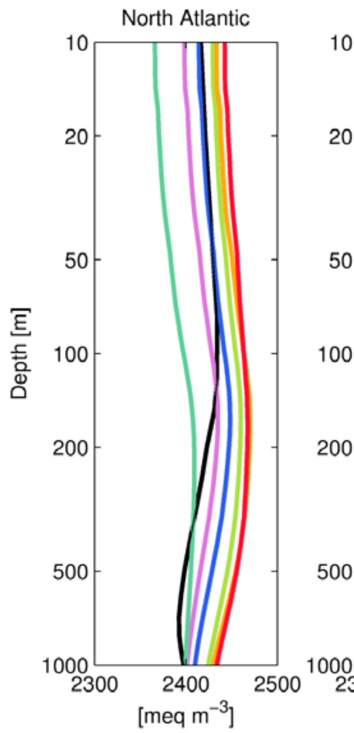


Figure 7. Observed (black; GLODAP) and modelled profiles of alkalinity (meq m^{-3}) in the North Atlantic (0°N to 60°N), Southern Ocean (90°S to 60°S) and Equatorial Pacific (15°S to 15°N). Vertical scaling is logarithmic (log$_{10}$).

Deleted: ¶
¶
¶



Formatted: Space After: 24 pt

Deleted: 6

Deleted: ¶

Supplementary online material

iMarNet: An ocean biogeochemistry model inter-comparison project within a common physical ocean modelling framework.

L. Kwiatkowski[1,2]*, A. Yool[3], J. I. Allen[4], T. R. Anderson[3], R. Barciela[5], E. T. Buitenhuis[6], M. Butenschön[4], C. Enright[6], P. R. Halloran[7], C. Le Quéré[6], L. de Mora[4], M.-F. Racault[4], B. Sinha[3], I. J. Totterdell[5] and P. M. Cox[1]

1 Surface fields

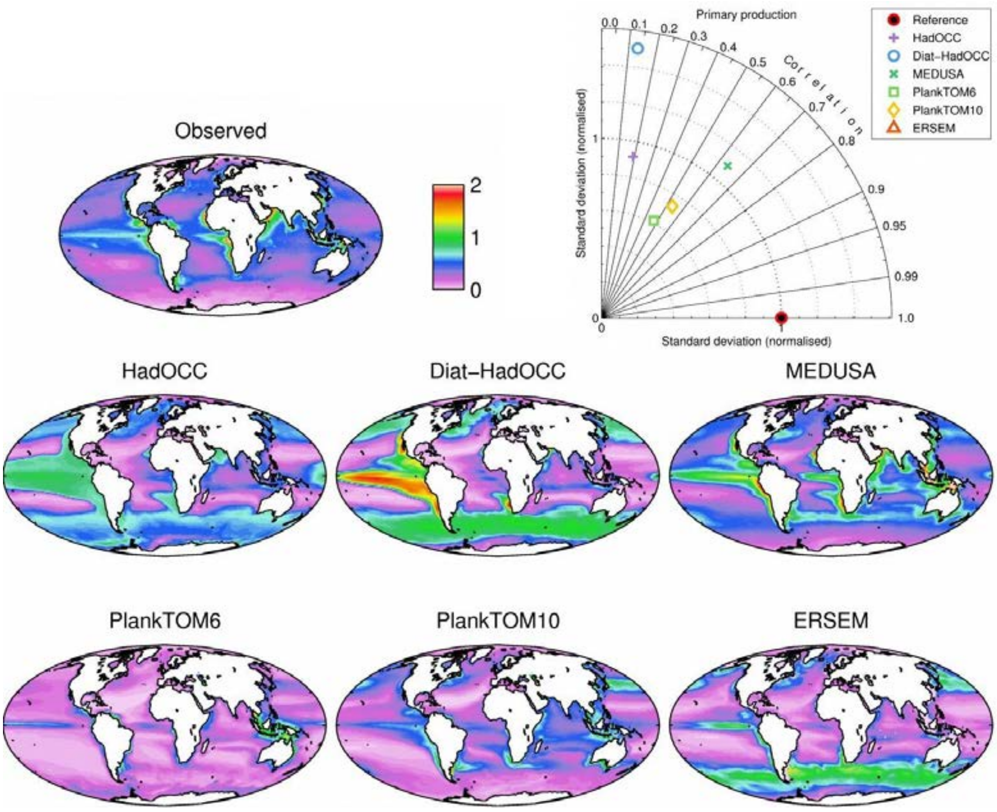


Figure S1. Annual average integrated primary production ($g\ C\ m^{-2}\ d^{-1}$) for all models and observations and a Taylor plot (top right) of all models relative to observations. Note that the models with negative correlation coefficients are not shown in the Taylor plot. Mean field values: observations 0.385; HadOCC 0.455; Diat-HadOCC 0.515; MEDUSA 0.473; PlankTOM6 0.188; PlankTOM10 0.289; ERSEM 0.365.

Deleted: gCm

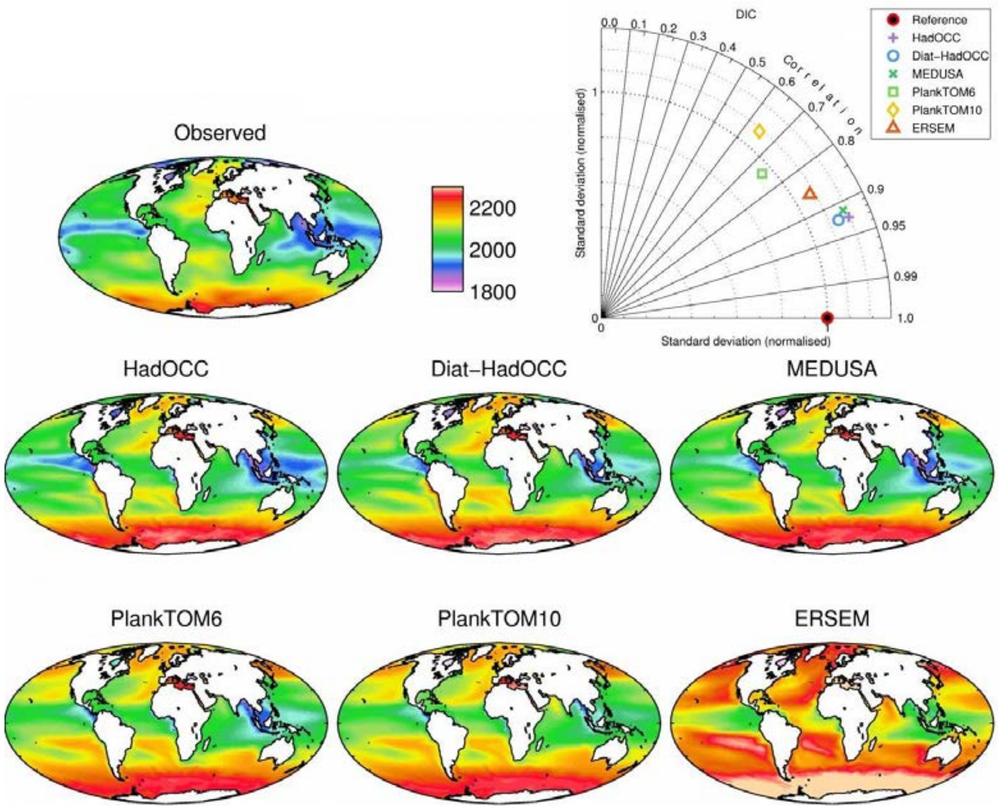


Figure S2. Annual average Dissolved Inorganic Carbon (DIC) (mmol C m^{-3}) for all models and observations and a Taylor plot (top right) of all models relative to observations. Mean field values: observations 2059.9; HadOCC 2089.6; Diat-HadOCC 2107.7; MEDUSA 2096.5; PlankTOM6 2125.4; PlankTOM10 2134.5; ERSEM 2199.7.

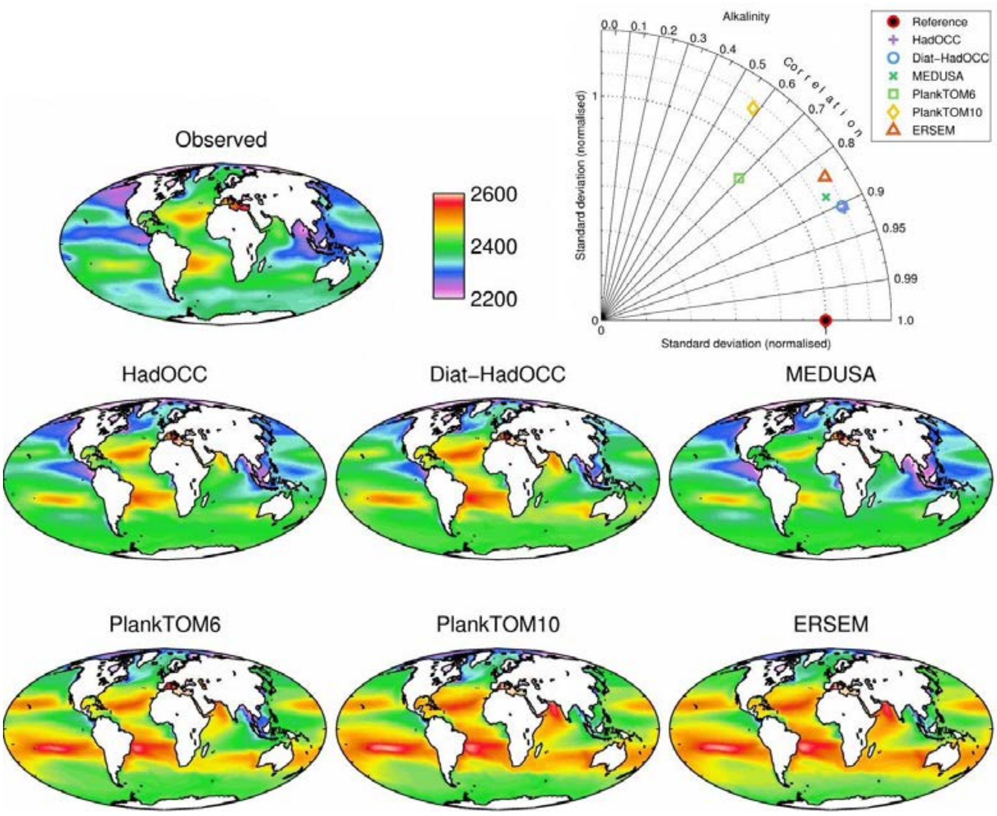
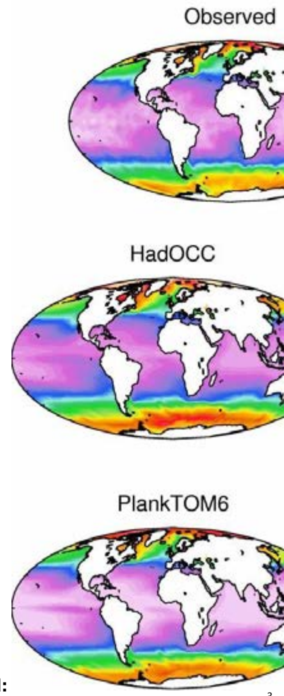


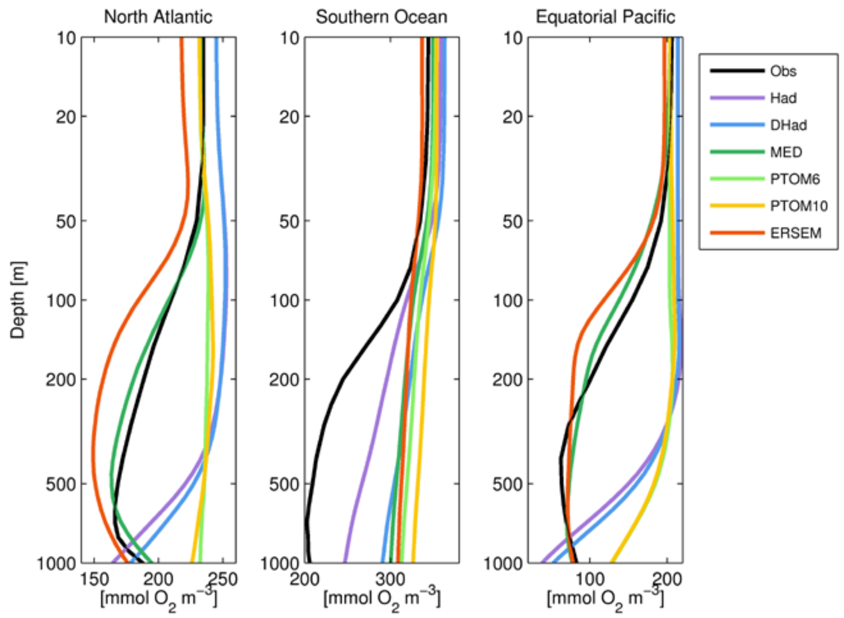
Figure S3. Annual average alkalinity (meq m^{-3}) for all models and observations and a Taylor plot (top right) of all models relative to observations. Mean field values: observations 2353.1; HadOCC 2377.8; Diat-HadOCC 2398.7; MEDUSA 2362.1; PlankTOM6 2434.4; PlankTOM10 2446.7; ERSEM 2450.9.



Deleted: Figure S4. Annual average oxygen (mmol O_2 m^{-3}) all models and observations and a Taylor plot (top right) of all models relative to observations. Note that the models with negative correlation coefficients are not shown in the Taylor plot.¶

Formatted: Font: 12 pt

2 Depth profiles



Deleted:

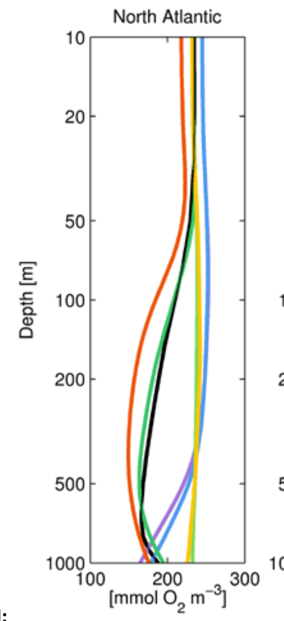
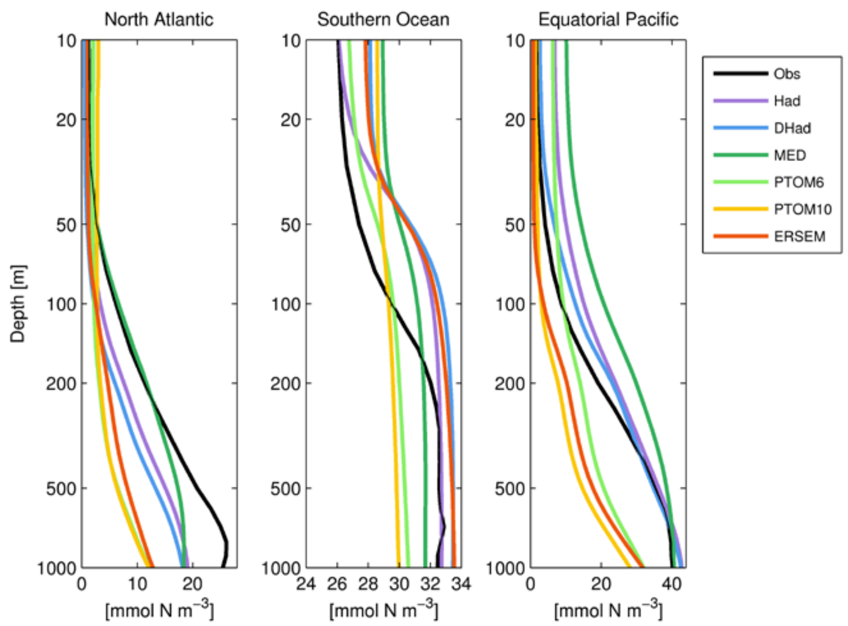


Figure S4. Oxygen depth profiles for the Equatorial Pacific, Southern Ocean and North Atlantic. Vertical scaling is logarithmic (log_{10}).

Deleted: S6



Deleted:

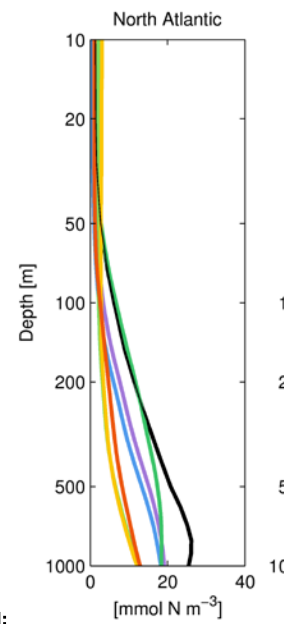


Figure S5. Dissolved Inorganic Nitrogen (DIN) depth profiles for the Equatorial Pacific, Southern Ocean and North Atlantic. Vertical scaling is logarithmic (log_{10}).

Deleted: S7

3 Computational cost

Deleted: ¶
¶

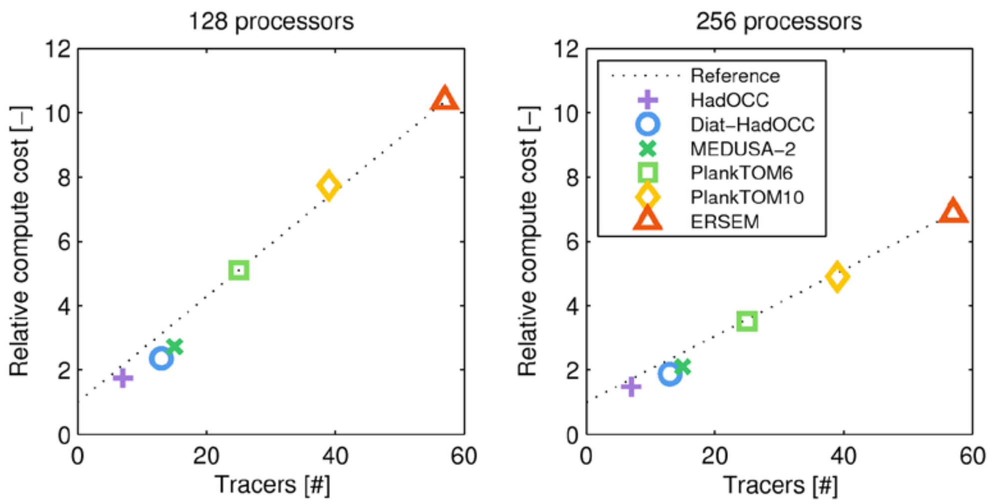


Figure S6. The compute cost of ocean biogeochemical models relative to a physics-only simulation against the number of model tracers.

Deleted: S8

4 Modelled physics

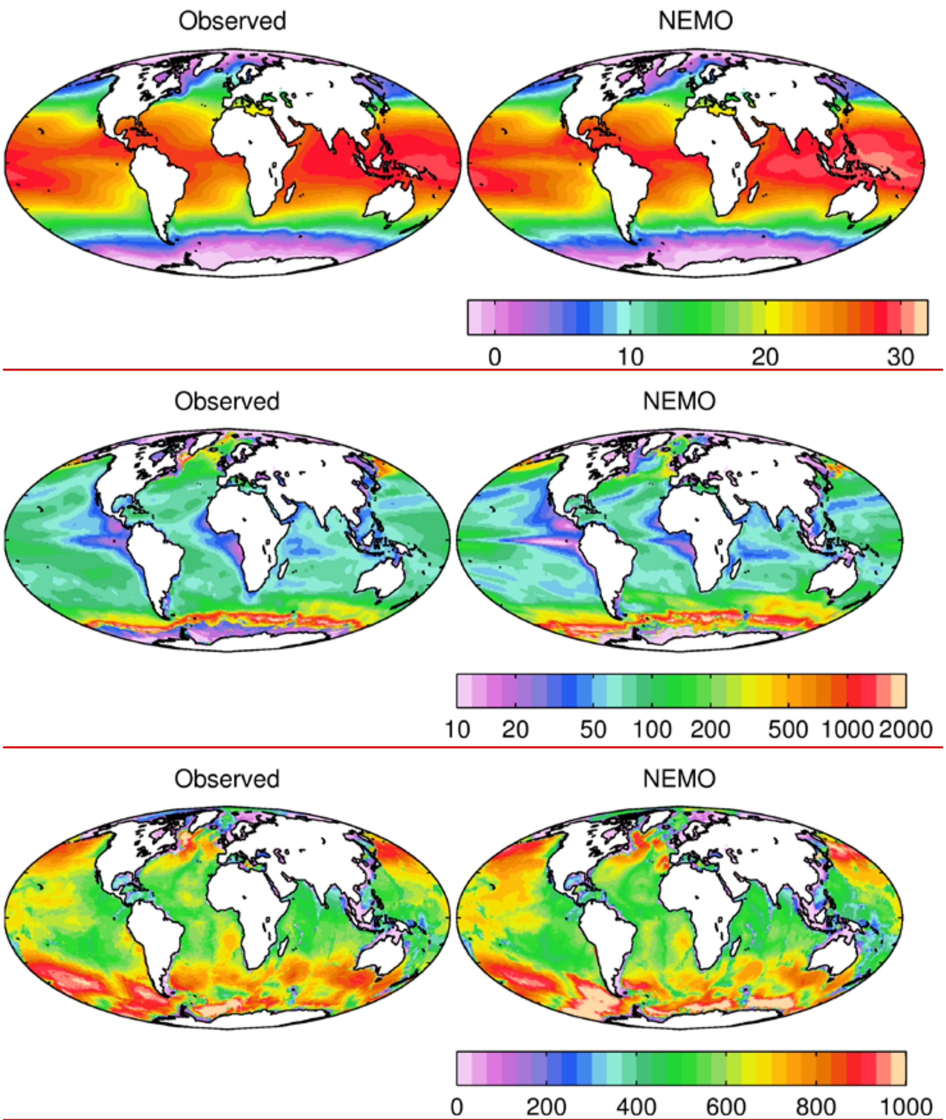


Figure S7. Comparison of observed (left) and modelled (right) mean annual sea surface temperature (top; °C), mixed layer depth (middle; m) and pycnocline depth (bottom; m). Observed fields are derived from the World Ocean Atlas 2009 fields of temperature and salinity (Locarnini et al., 2010; Antonov et al., 2010). Mixed layer depth is calculated using a 0.5°C criterion (Monterey & Levitus, 1997) and shown on a logarithmic scale, while pycnocline depth is calculated using the method of Gnanadesikan et al. (2002).

5 Export production

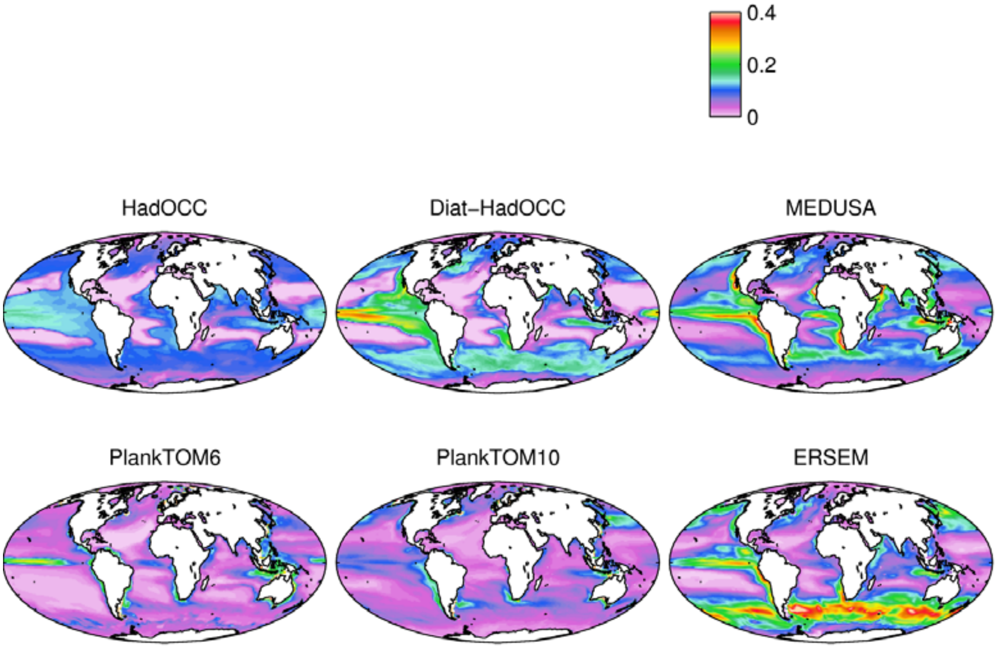


Figure S8. Annual average export flux of particulate organic carbon (POC) at 100m (g C m^{-2} d^{-1}) for all models. The flux is derived in each case from the final simulated year, 2007.

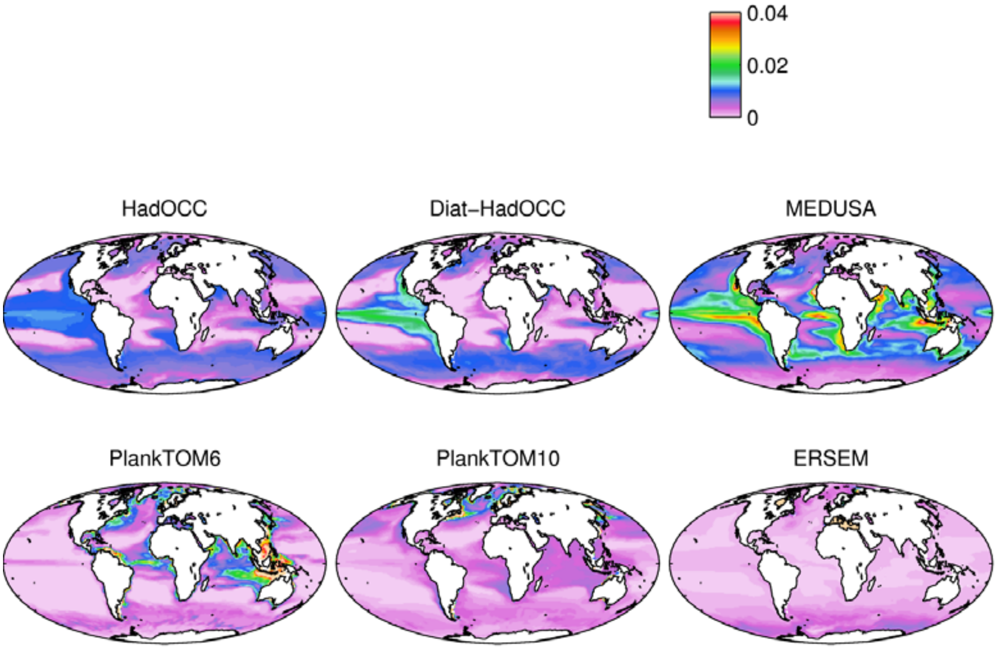


Figure S9. Annual average export flux of particulate inorganic carbon (PIC) at 100m (g C m^{-2} d^{-1}) for all models. The flux is derived in each case from the final simulated year, 2007.

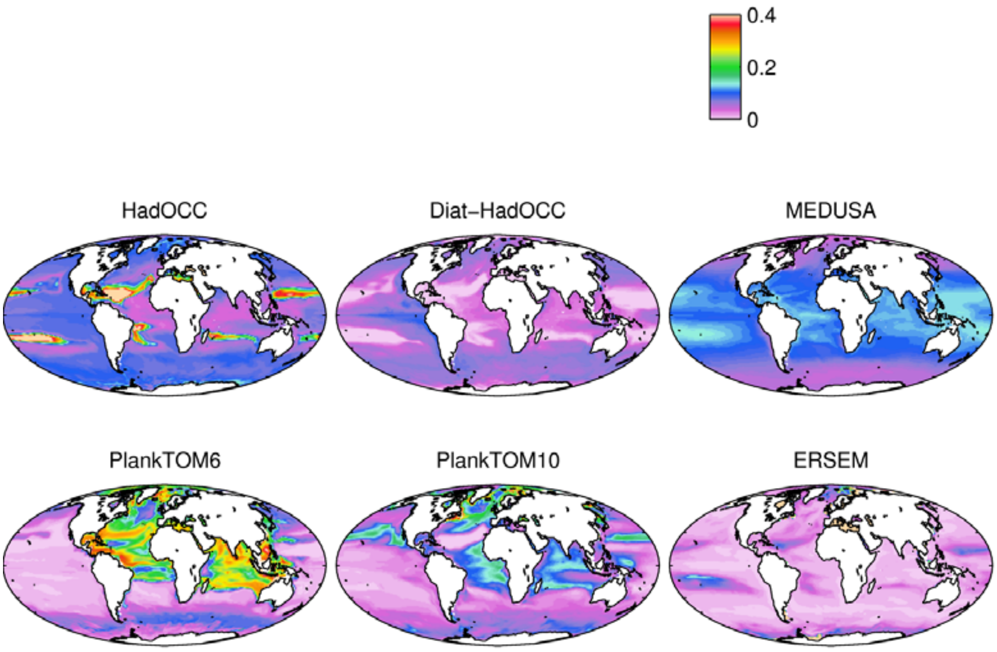


Figure S10. Annual average rain ratio (PIC:POC) of particulate material at 100m (-) for all models. The ratio is derived in each case from the final simulated year, 2007.