Dear Editor,

We would like to thank the reviewers for agreeing to review this paper again, and for their constructive criticisms. We would also like to thank Reviewer #1 for accepting the revised manuscript in its current form.

Reviewers 2 and 3 have raised their concerns about the function fitting. We have now added more details on the fitting in section 2.1, 2.2, and 2.3 (page 5-6) to ensure that the fitting process, which is crucial to this study, is clear. We have further done another fitting using narrower concentration ranges; between 0.001 to 20 mmol m-3, for dissolved inorganic nitrogen (DIN) and 0.001 to 10 mmol m-3, for phytoplankton and zooplankton concentrations, and confirmed that there are no significant changes in parameters when the range of DIN and phytoplankton concentrations have been changed (please see Table 1 and Reviewer #2's comment for p4 l29). We have also discussed (see annotated manuscript page 16 line 10-11 and 26-27) how the fitting still maintains the phenomenological similarity (such as when fitting nutrient uptake with DIN in the lower concentration range (up to 5 mmol m-3), we can still see that at lower concentrations the function deviates).

Below is our point by point response to all comments made by two reviewers. The reviewer comments are included in italic and responses are in bold. Please note that all the line numbers and pages in this response refer to the annotated version attached.

We hope that the response would be satisfactory, and we look forward to your decision.

Kind Regards,

Authors.

Response to the reviewers:

Reviewer #2

The updated manuscript shows some improvement in places and the authors have added some useful analyses. However, some of my main criticisms of the manuscript still apply.

General comments

1) *In my first review I wrote that "The results of the default run can have knock-on effects on the ensemble: in multiple parts of the manuscript the authors note that when there is a large bias between the model (ensemble) and the observation, that the ensemble spread is too low when really other model aspects may be to blame for the bias. In other words, problems with the parametrization, the physical model, or the 1D nature of the model cannot be explained by structural uncertainty in the biogeochemical model." To make it more explicit: if the ensemble does not envelop the observations, this does not necessarily imply that the ensemble spread is too low. A bad model (physical model, unresolved BGC*

processes) with a large bias may be to blame. This is now touched on in the discussion section but ignored before, and phrases like "the observation is outside the ensemble range ... making the ensemble spread too narrow" appear too often, especially in the phytoplankton phenology section 3.4 (blooms are often controlled by physical processes).

AR: We agree that the large bias, especially in the physical model, may be to blame for the ensemble not covering the in situ observations. We have updated the text to emphasise this (page 16 line 35 to page 17 line 2). Also, we have replaced 'ensemble spread too wide/narrow' with appropriate text to convey that the observations are outside the ensemble range.

2) *With regard to my earlier comment (reviewer 2, general comment (2)): The plot in the response shows little difference between the regular and the log-transformed fit. This is surprising to me. The linear fit produces a somewhat good match to the sigmoidal function only for high phytoplankton concentrations. This can be easily explained as the linear scale emphasizes high phytoplankton values. For the log-scale fit, what I suggested was to fit the linear function across the range of phytoplankton concentrations shown in the figure using the same log-scaling. I would expect a much lower mortality value in this case, much closer to 0.05 than the 0.085 that were obtained. This may yield significantly different results. The best idea would probably be to let the data determine the fit, i.e. minimize the misfit for the values occurring in the model output (or the observations if these are available). In any case, since many of the results hinge on which functional form is used, the authors need to include more detail on how exactly the fit was obtained. This should include the range of phytoplankton concentrations considered in the fit.*

AR: Perhaps the closer value to the linear fit occurred because we previously used a slightly higher range of phytoplankton between 0.01 – 50 mmol N m-3. We have now done the fitting on linear scale using phytoplankton concentration between 0.001 – 10 mmol N m-3, to capture a more realistic range based on the model outputs. The fitted parameters within this refined range remained very similar to our initial linear fit using 0.001- 100 mmol N m-3, ensuring robustness of the linear fit in the case. This have also been observed in nutrient uptake and grazing. However, if we log transform phytoplankton concentration before the fit (which we avoided earlier to ensure that we use the definition of functional forms, which are defined for actual concentrations, and not for log-transformed concentrations), the maximum mortality value for 'linear functional form' changes to 0.04. But, even in this case, the mean absolute errors (measures of goodness of fit) between the hyperbolic (default) function and other functions show lower misfits for all the linear fittings compared with the log-transformed fittings, as summarised on Table 1. Therefore, we have decided to stick to linear fitting in this study. But following the suggestion of the referee, we have added further details on fitting in the method section (section 2.1 page 5, line 1-4, section 2.2 page 5, line 16-19, section 2.3, page 6, line 1-2 and 5-7), and in the discussion section (page 16, line 17-23 on using log transform parameter (0.04) for the linear mortality functional form).

Table 1. Comparison between log transform and linear-scale function fitting parameter values, with phytoplankton concentration ranging from 0.001-10 mmol N m-3, and its mean absolute errors. The numbers in brackets are the original fit (concentration ranging from 0.001-50 mmol N m-3).

functional form	log transform parameter	mean abs error	non log transform parameter	mean abs error
sigmoidal	k = 1.20 (1.10)	0.007 (0.002)	k = 0.74 (0.74)	0.007 (0.002)
linear	μ=0.04 (0.085)	0.05 (0.13)	μ= 0.094 (0.097)	0.009 (0.008)
quadratic	μ= 0.04 (0.02)	0.01 (0.003)	μ= 0.05 (0.05)	0.01 (0.003)

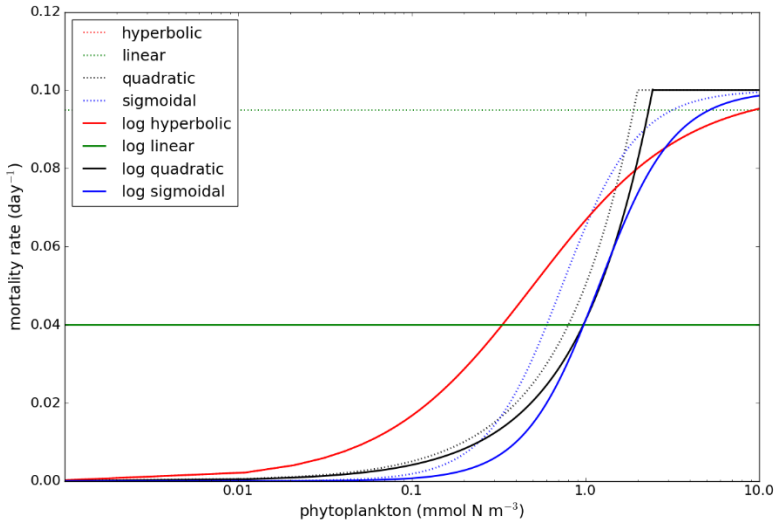

Figure 1. Mortality functional forms optimised against hyperbolic function. The range of phytoplankton used was 0.001 to 10 mmol N m-3

At the risk of sounding pedantic: there are still many sentences in the manuscript that are not clearly formulated and that can be interpreted in different ways. Readers familiar with the topic will likely know what is meant but others will not. I have pointed out some of these instances below but I would recommend that the authors go through the manuscript again carefully.

AR: Thank you for this suggestion. We have included all the suggestions, and have also gone through the manuscript for further textual improvements.

Specific comments:

p1

l3: "We assessed the impact of structural sensitivity ... by modelling the chlorophyll and nitrogen concentrations at five different oceanographic stations spanning three different regimes": This appears to imply (incorrectly) that structural sensitivity is linked to the model location.

AR: Thank you for noticing this, we did not mean that structural sensitivity is linked to the model location, but we are testing to observe the effect on structural sensitivity at different oceanographic regions. We have rephrased this in the revised manuscript to '….. by modelling the chlorophyll and DIN concentrations. The model is run at five different oceanographic stations

spanning three different regimes, over a 10-year timescale to observe the effect in different regions' (page 1, line 4-6).

l3: "nitrogen": Do you mean "nitrate"? This comment applies to later mentions of nitrogen as well.

AR: We use the combination of inorganic nitrate, nitrite, and ammonia as initial condition for dissolved inorganic nitrogen. We have rephrased nitrogen into dissolved inorganic nitrogen (DIN).

l6: Agreed, though there are other applications for BGC models than climate change assessments.

AR: We have added assessing the impact of anthropogenic input on biogeochemical cycles in the marine ecosystem, and producing decadal reanalyses (page 2, line 9-10)

p2

l24: Mention some causes of mortality here.
AR: These have been added to the revised manuscript. We added '…, due to diseases or implicit higher trophic levels…' (page 2, line 29).

p3

l14 "demonstrated ... that linear density-dependent mortality produces the biggest difference in diatoms". As opposed to other phytoplankton types? This is not clear.
AR: The most notable difference has been observed in diatoms. We have rephrased this sentence in the manuscript to make it clearer (page 3, line 21).

p4

l29: "The difference in shape of the optimised functional forms are more obvious at low nutrient concentrations." Likely due to the way the fitting is performed, see general comment (1).

AR: See our response to fitting above. We have changed the range and still observed that at lower DIN concentrations (below 1 mmol m-3) the functions still deviate the most. We have however changed the range of DIN concentration (between 0.001 – 20 mmol m-3) we used to fit the function. This reduction in range also has not changed the shape-defining parameters for other functional forms (U2= 0.74, U3= 1.12, U4= 0.60). We have added some text to mention these: ' The fit is done within the nutrient concentrations of 0.001 – 20 mmol N m-3 and are discretised into 1000 intervals. The difference in shape of the optimised functional forms are more obvious between 0.1 to 1 mmol m-3' (page 5, line 2-3).

p5

l3: "graze on small phytoplankton, non-diatoms, and detritus": This is confusing, as small phytoplankton and non-diatoms are the same variable.

AR: This has been rephrased in the manuscript. We have removed small phytoplankton and keeping non-diatoms (page 5, line 7).

l13: "hyperbolic": at this point it is not clear what the hyperbolic refers to, please use "hyperbolic" and "sigmoidal" when G_1 and G_2 are introduced initially, in line 6.
AR: This has been rephrased in the manuscript.

l28: Here it is very important to note if the area below the function is computed in log-space or not and over what range of concentrations. Figure 1 makes it look like a log transformation was used and concentrations between 0.001 and 10 mmol N m^-3 were considered in the fitting.

AR: The area below the function is not computed in log space, and the range of concentrations were between 0.001 and 100 mmol N m-3. In the new fit, (see the general response 2), we have used 0.001 and 10 mmol N m-3 instead. We have added text to detail the fitting (page 6, line 1-2 'calculated as the area below the function …. *In linear terms between 0.001 to 10 mmol m-3*' and line 5-7 for the detail about concentration range).

p6

l7: "as" -> "is"

AR: This has been changed in the revised manuscript (page 6, line 14)

l30: "mortalities" -> "mortality formulations"

AR: This has been changed in the revised manuscript (page 7, line 6)

p7

l18: There are several 1D models now, do they all use the same depth levels (at every station)?

AR: We only use one 1D biogeochemical model, but we use several NEMO outputs for the physical input. Our 1D biogeochemical model has the same level thickness resolution for all the stations, however shallower stations have fewer levels, such as Cariaco and L4.

l21: The station abbreviations should be mentioned earlier when the stations are first mentioned (section 2.5).

AR: This has been changed in the revised manuscript (please see page 6, line 30-33).

p8

l3: I don't quite understand the reasoning behind using the same concentration at each depth level. Are the initial conditions quickly forgotten? Mention this. A short model spinup could have been useful.

AR: Since station L4 is only 50m deep, we assume that the concentration is similar at all depths during the winter start of the run (1st January 1998). We have mentioned that L4 is a shallow station and therefore using similar biogeochemical input at all depth levels. We have now added 'Since the maximum depth in this station is only 50 m deep….' (page 8, line 10).

l17: "for to" something is missing here.

AR: We have corrected this mistake. What we meant was for the seasonal variations.

p10

l9: How does the ensemble range not cover the whole ensemble and what exactly is meant by "if we only allow one process function at a time to change"?

AR: Allowing one process function at a time means that we only change functional form in one process, whilst keeping other processes the default function, therefore the new smaller ensemble (11 members) range is different from that of the old larger ensemble (128 members). We have rephrased this to be clearer (page 10, line 18-19)

p13

l23: "phenology metrics" -> "observed phenology metrics" or "phenology metrics obtained from observations"

AR: Thank you, this has been changed in the revised manuscript (page 14, line 1).

l27: How can a range be late; rephrase?

AR: We meant that the dates within the inter-quartile range are later than for in situ observations. This has been revised in the revised manuscript (page 14, line 6).

l28 and again l30: "For initiation...": add "bloom"

AR: This has been added in the revised manuscript (page 14, line 7).

p17

l23: This long sentence mentions "active prey selection" twice.

AR: The first active prey selection has been removed from the sentence.

Fig 1: Should it be "nitrate" in the axis label in (a)? Same comment applies to other figures.

AR: We have changed this to DIN

Fig 3: missing units in (b) and (e)

AR: We have added the units in (b) and (e)

Fig 7: "(d)" is used twice

AR: We have changed (d) to (e)

Fig 9: The order in which panels are labeled is confusing and inconsistent. It lead to a problem in the description as well. White line is in panel (a) not (b).

AR: We have changed the labelling, please also see figure 3 and 14.

Fig 10: Use same y-axis across all panels.

AR: We have changed this

Fig 11: "(e)-(f)" should be "(e)-(h)"

AR: We have changed this

Reviewer #3

As already noted in my first review, I really appreciate the very thorough and exhaustive attempt to investigate the sensitivity of a biogeochemical model that will be, at a later stage, used for projections of climate change and other applied tasks. This paper could provide important information on global biogeochemical model sensitivity with respect to structural uncertainty, even for users not applying this particular model. However, so far, and for this purpose, I find the discussion of results somehow incomplete. I think discussing the outcome of these experiments before the background of earlier studies that also deal with global model uncertainty and performance would enhance the paper's impact, and put this work into a broader context.

1) *For example, I am not convinced that the results indicate a larger importance of structural sensitivity compared to uncertainties related to physical or parametric uncertainty. To my impression, so far there is little (with respect to the metrics applied) evidence that mean or median perform much better than the default model (see below, point B). I think this should*

be discussed more, e.g. before the background of the study by Kwiatkowski et al. (2014), who showed that "No model is shown to consistently outperform all other models across all metrics." when comparing six models (among them MEDUSA) against surface tracers such as DIN and Chl. Likewise, Galbraith et al. (2016) or Kriest (2017) found that models of varying complexity performed quite similar with respect to metrics such as RMSE etc. The importance of parametric uncertainty was addressed in several studies, even at a global scale. There are also many studies that deal with the uncertainty due to physics (starting with the "classic" study by Najjar et al., 2007), and show a large impact of circulation or phenomena at smaller (meso) scales.

AR: Thank you for acknowledging the importance of our study. We have revised in parts to highlight this importance. However, we do not think that structural sensitivity has larger importance compared to physical input or parametric uncertainty. We are trying to emphasise that using similarly shaped functional forms may yield different plankton dynamics, and such particular study have not been done as much as for parametric uncertainty, or physical uncertainty. We have also noted in the discussion section that 'a coupled biogeochemical model is only as good as its physical model', meaning that this input may have larger importance than structural sensitivity (page 16 line 33 to page 17 line 2) and how unresolved biogeochemical processes also affect the model output especially at the coastal and oligotrophic regions (page 17, line 26-29 and page 18, line 23-27). There are also no metrics to determine whether one sensitivity is more important than the others. We are trying to generate a range of model results from perturbing the model equations alone, with fixed parameters (apart from the shape defining parameters, such as half saturation coefficient) and quantify them.
Further, it should be noted from the results table (Table 3 in the manuscript), that only at station BATS where there is a large bias, the RMSE for default run is better than the ensemble mean, and we have stated this in the result section. But in all other stations (ALOHA (ensemble mean RMSE 0.095 < default 0.097 for chlorophyll profile), L4, PAP, Cariaco), the ensemble mean and median have lower RMSEs than the default run. We have now emphasised this important result within the discussion section (see page 17, line 11-21).

2) *Secondly, different equations, as applied in this study, necessarily result in different nutrient or food affinities of plankton, as briefly discussed by the authors; therefore, to some extent the structural sensitivity also includes some parametric sensitivity. Again, this might also be worthwhile discussing (see below, point A, and specific comments for p15).*
I would thus recommend an overhaul of the discussion section, as already suggested by Reviewer #1, to give this paper a wider impact and "to address the big picture".

    a. *Authors response to my comment 2 on fitting procedure of different structural forms. "However, we are trying to capture the whole range of nutrient and phytoplankton at all the different region, [...]". As shown in Figures 3, 9 and 14, the full range of observed and simulated nutrient concentration is between ~0-10 mmol/m3, and the range of Chl ~0-1 (PAP, Cariaco), 0-0.3 (ALOHA and BATS) and ~0-6 (L4), i.e. more than an order of magnitude lower than used for curve fitting. I would suggest to give the range over which the functions were fitted in the paper (method section).*

        AR: Thank you for the suggestion, we have updated the discussion in places to

include these points. Also, we have tried fitting the function between 0.001-10 mmol m^-3 and we found that there are no significant changes in the parameter values if we use linear fitting (please see the response for Reviewer #2 general comments 2, table 1). We have now included further details on fitting (method section 2.1 page 5, line 1-4, section 2.2 page 5, line 16-19, section 2.3, page 6, line 1-2 and 5-7). The information about the range has been included in the manuscript. We also agree that structural sensitivity is also dependent on parametric sensitivity, especially in linear mortality (please see the response to Reviewer #2 general comments 2), and we have included this in the manuscript (please see discussion section, page 16, line 17-22).

Further, in their response the authors state that

"Suppose we are optimising the nutrient uptake on the similar range of station BATS and ALOHA (with maximum nitrogen and phytoplankton concentration of 5 mmol N m- 3, shown on Figure 3, although at stations like Cariaco, PAP, and L4, we may see nitrogen larger than 5 mmol N m-3), the functions still deviate at low nitrogen and phytoplankton concentration."

This is very important information, and I would suggest to mention this in the paper, together with a full description on how the fitting was carried out. (Was the range between 0-100 discretized, and if so, into how many bins? Were the discrete values equally distributed over the range?)

I still think this issue (fitting different functional forms, the side effects on affinities and rates in different oceanic regimes) might merit more in depth discussion and detailed presentation.

AR: We have mentioned this in the method section in the manuscript, and we have now noted clearly that we are fitting in the range between 0.001-20 mmol m-3 for DIN (nutrient uptake) and 0.001-10 mmol m-3 (zooplankton grazing and plankton mortality) for phytoplankton and zooplankton, with 1000 intervals, similar to the observations. We found that there isn't much change in the shape defining parameter values (please also see the response for Reviewer #2 general comment no 2 and for p4 l29).

b. *I find the analysis of model performance for the different structural types (ensemble mean or median vs default) somewhat biased and incomplete: For example, Table 3 shows, that out of the 51 criteria listed in that table for r, RMSE and Bias for DIN and Chl (profile, surface, integrated), the default model is the best model w.r.t. to the bias and also r, compared to ensemble mean or median. It also outperforms mean and median at BATS, and, all three criteria combined, is as good as the ensemble mean at ALOHA. Also, for RMSE some of the values do not seem to provide a clear evidence that the ensemble mean or median is better: for example, at ALOHA the RMSE for DIN and Chl (profile or surface) is very similar among the different models.*

AR: We have stated in the results section that all of the statistical metrics at BATS are better in the default run. But in all other stations the results are promising. For example, in ALOHA, for the surface cholorophyll, the RMSE of ensemble mean (0.05) is better than the default run (0.07), and chlorophyll profile RMSE in ensemble mean is also lower (0.97) than the default run (0.10). In all other stations, chlorophyll profiles' RMSEs are lower than default run: Cariaco, ens mean= 0.83 < default= 0.87, PAP ens mean =0.06 < default= 0.18, and L4, ens mean = 0.42 < default= 0.83. In terms of correlation, in nitrogen, the ensemble mean in stations that are not BATS show either better or similar correlation to the default run: ALOHA, ens mean= 0.77 = default, Cariaco, ens mean= 0.78 > default= 0.76, PAP ens mean =0.23 > default= 0.21, and L4, ens mean = 0.70 > default= 0.52. Nonetheless, we agree that the model bias didn't improve, and we have explained the additional cause of model bias (see page 16, line 33- page 17, line 2 for physical input bias and page 17, line 24-29 and page 18, line 23-27 for unresolved biogeochemical process), and further discussed why the default run produces lower bias these in page 17, line 13-17).

Specific comments:

p1, line 6-7: "that describe the key biogeochemical processes" - I would suggest to rephrase this to "that describe some key biogeochemical processes", as, at a larger scale, phytoplankton light affinity, recycling and sinking organic matter may also be regarded as key processes.

AR: This has been changed accordingly in the revised manuscript (page 1, line7)

p1, line 13: "Changing mortality" - which mortality: both zooplankton and phytoplankton?

AR: Yes both zooplankton and phytoplankton, and this has been included in the abstract (page 1, line 14)

p1, line 14-15: "The RMSEs between in situ observations and the ensemble mean and median are mostly reduced compared to the default model output." - See above, point B. Why not address the bias or correlation coefficient here, and in the discussion?

AR: We have added the statement and further explanation in the discussion section (page 17, line 13-19).

p2, Introduction: This is probably just a minor point, but usually global models are NPZD models, with detritus (and/or DOP) being an important component in the recycling and vertical redistribution of nutrients.

AR: NPZD is now referred to (page 2, line 7).

p2, line 17ff: "Moreover, in order to investigate the effect of global climate change and anthropogenic activities in the ocean, marine biogeochemical models are now being embedded into earth system models." - Simple marine biogeochemical models were first embedded into GCMs in the

90ies, and into ESMs possibly around 2005 (e.g., UVic, Schmittner et al., 2005). Perhaps it would be better to rephrase this as "[...] activities in the ocean, MORE COMPLEX marine biogeochemical models are now being embedded into earth system models."

AR: This has been changed accordingly, thank you for the suggestion (page 2, line 21)

p3, line 14: "A few studies have investigated the effects of biogeochemical process formulations, e.g. Yool et al. (2011) has demonstrated [...]" – Perhaps better "A few studies have investigated the effects of biogeochemical process formulations. For example, Yool et al. (2011) have demonstrated [...]" ?

AR: This has been changed accordingly, thank you for the suggestion (page 3, line 19)

p4, section 2.1: I suggest to give range of nutrient concentrations over which the functions were fitted, and some details of the fitting procedure here (see A).

AR: We have added the range of nutrients in this section (0.001 – 20 mmol m-3).

p5, line 5 "which are high quality food sources" - In the model? In reality? Is this of importance here, and if so, how is it embedded in the model?

AR: meaning that it is more nutritious compared to detritus and non-diatoms. This has been rephrased (page 5, line 9-10)

p5, line 12: "These functions both become constant at a maximum grazing rate." - Perhaps better: "These functions approach a maximum grazing rate at high concentrations of prey."

AR: This has been rephrased accordingly, thank you for the suggestion (page 5, line 16-17).

p5, line 25: "as shown on Fig. 1(b)." - Fig 1(c)?

AR: This has been changed, thank you for noticing (page 5, line 20).

p5, line 27ff "integrated over the range of prey density" - Give range here (see p4, section 2.1)

AR: We have stated the range of phytoplankton concentration, and changed prey density into phytoplankton concentrations (page 6, line 1-2).

P6, line 5: "From a previous 3-D MEDUSA run, in the oligotrophic regions show" – skip "in".

AR: We have removed this from the sentence. Thank you for the suggestion (page 6, line 12).

p6, line 9: "MEDUSA also contains both slow and fast detritus sinking factors." - What does this mean? Sinking speeds? Remineralisation length scales? Perhaps better "MEDUSA also parameterises slow and fast sinking detritus"?

AR: Thank you for the suggestion, we have rephrased this sentence accordingly (page 6, line 16).

p6, line 14f: "We chose a lower sinking rate of 0.1 m day to prevent depletion of state variables particularly at the shallower stations." - Does this mean sinking organic matter is buried at the sea floor? If so, I would mention it here.

AR: Yes, the remineralisation rate is slower than the sinking rate, and therefore loss occurs to the seafloor. We have mentioned this in the revised manuscript (page 6, line 19).

p7, line 18f: "Our 1D model uses these same 63 depth levels vertical resolution." - The same as the global (MEDUSA) model? Then, on p8, line 9: "The model is simulated at 37 depth levels, [...]" - I am a bit confused - How many levels did the 1D setup have - 63 or 37?

AR: The model uses 37 depth levels, but the resolution is similar to the 63 depth levels vertical resolution, instead of the 75 mentioned earlier (please see page 7, line 22-26)

p7, line 18f: "we use the integrated nitrogen over 200 m (integrated nitrogen / depth)" - is this meant to be "dissolved inorganic nitrogen averaged over the upper 200 m"? I would suggest to refer to "averaged" when appropriate (also in some of the figure captions); Additionally, if nutrients are mean it should be "nitrate" or DIN or "dissolved inorganic nitrogen", because "nitrogen" alone can also include the organic forms.

AR: Thank you for the suggestion. What we meant was averaged over the upper 200m, and dissolved inorganic nitrogen and this has been changed accordingly.

p 9, line 25: "Chlorophyll and nitrogen profile 10 year means"? What does this mean? "Observed mean profiles of chlorophyll and DIN"?

AR: Yes. We have rephrased this sentence, thank you for the suggestion (page 10, line 1).

p15, lines 15ff: "In order to maintain phenomenological similarity, these functions are calibrated using non-linear least squares, while keeping the maximum process rates fixed." - To me it seems as if the phenomenological similarity is maintained with respect to the rates integrated over the nutrient and Chlorophyll range considered; i.e. between 0-100 mmol/m3 or mg/m3. This could downweigh the affinity at very low nutrient or chlorophyll concentrations. (See above, A.) Thus, phenomenological similarity is maintained with regard to a certain criterion. As stated by the authors in their reply, fitting the curves over a narrower range results [0-5] in almost the same curves, which would strengthen the above statement of phenomenological similarity. I would suggest to add a few sentences about this here.

AR: Thank you, we have included a few sentences about this in the discussion section and methods (methods page 5, line 2-3 and line 18-19, and in discussion page 16, line 10-11 and line 26-27).

p15, lines 19ff: "Applying structural sensitivity in the 1-D framework has also allowed a large parameter space of concurrent variations to be explored for several different oceanographic regions, and with minimal computational cost." - How is the parameter space explored, particularly when considering the above statement about phenomenological similarity?

AR: What we meant was large range of process variability, we have rephrased this in the revised manuscript (page 16, line 2-3).

p15, line 24: "have" -> has

AR: We have changed this in the revised manuscript

p15, line 25: "This is expected as at low concentrations, using the G2 function would graze more phytoplankton, as shown on Fig. 1(b)." - Again, here the effect of the fitting procedure for different functions shows up: perhaps adjusting the k's of the different function such that they become similar at very low concentrations might have resulted in similar results at the oligotrophic stations BATS and ALOHA? Thus, the effect of different functional forms seems to be mingled with the effects of parametric uncertainty. Of course, this cannot be avoided, but I think I think it could be discussed more.

AR: We have also tried fitting it using a lower range of phytoplankton concentrations, but it seems that the functions still deviate greatly in the lower concentration range (please see page 16 line 10-11). Also see our response to Referee #2, page 2.

p16, lines 19ff: "At most of the stations, the ensemble mean produced lower RMSE compared to the default run, suggesting that the structural ensemble with a wide range of predictions covering the in situ observations, is likely to produce a mean field closer to the observation, than a single-structure model." - I am still not fully convinced about this - see my above comment B.

AR: We have included the discussion on model bias and correlation in the discussion section (page 17, line 13-19). However, higher correlations in ensemble mean and median have been observed in nitrogen correlation in all of the stations but BATS (please see table 3 in the manuscript).

p18, line 8: "widely used in the community" - In which community?

AR: We have removed this sentence in the manuscript

References:

Galbraith et al. (2016), Complex functionality with minimal computation: Promise and pitfalls of reduced-tracer ocean biogeochemistry models. Journal of Advances in Modeling Earth Systems, doi:10.1002/2015MS000463

Kriest (2017), Calibration of a simple and a complex model of global marine biogeochemistry, Biogeosciences, doi:10.5194/bg-14-4965-2017

Najjar et al. (2007), Impact of circulation on export production, dissolved organic matter, and dissolved oxygen in the ocean: Results from Phase II of the Ocean Carbon-cycle Model Intercomparison Project (OCMIP-2), Global Biogeochemical Cycles, doi:10.1029/2006GB002857, 2007

Schmittner et al. (2005), A global model of the marine ecosystem for long-term simulations: Sensitivity to ocean mixing, buoyancy forcing, particle sinking, and dissolved organic matter cycling, Global Biogeochemical Cycles, doi:10.1029/2004GB002283

A perturbed biogeochemistry model ensemble evaluated against in situ and satellite observations

Prima Anugerahanti[1], Shovonlal Roy[1,2], and Keith Haines[3]

[1]Department of Geography and Environmental Science, University of Reading, Whiteknights, Reading, RG6 6AB, UK
[2]School of Agriculture, Policy, and Development, University of Reading, Whiteknights, Reading, RG6 6AR, UK
[3]Department of Meteorology, University of Reading, Whiteknights campus Early Gate, Reading, RG6 6BB, UK

Correspondence: Prima Anugerahanti (p.anugerahanti@pgr.reading.ac.uk), Shovonlal Roy (shovonlal.roy@reading.ac.uk)

Abstract. The dynamics of biogeochemical models are determined by the mathematical equations used to describe the main biological processes. Earlier studies have shown that small changes in the model formulation may lead to major changes in system dynamics, a property known as structural sensitivity. We assessed the impact of structural sensitivity in a biogeochemical model of intermediate complexity by modelling the chlorophyll and ~~nitrogen concentrations~~ dissolved inorganic nitrogen (DIN) concentrations. The model is run at five different oceanographic stations spanning three different regimes: oligotrophic, coastal, and the abyssal plain, over a 10-year timescale to observe the effect in different regions. A 1-D MEDUSA ensemble was used with each ensemble member having a combination of tuned function parameterizations that describe some of the key biogeochemical processes, namely nutrient uptake, zooplankton grazing, and plankton mortalities. The impact is quantified using phytoplankton phenology (initiation, bloom time, peak height, duration, and termination of phytoplankton blooms) and statistical measures such as RMSE, mean, and range for chlorophyll and nutrients. The spread of the ensemble as a measure of uncertainty is assessed against observations using the Normalised RMSE Ratio (NRR). We found that even small perturbations in model structure can produce large ensemble spreads. The range of 10-year mean surface chlorophyll concentration in the ensemble is between 0.14-3.69 mg m^{-3} at coastal stations, 0.43-1.11 mg m^{-3} on the abyssal plain, and 0.004-0.16 mg m^{-3} at the oligotrophic stations. Changing ~~mortality and~~ both phytoplankton and zooplankton mortalities and the grazing functions have the largest impact on chlorophyll concentrations. The in situ measurements of bloom timings, duration, and terminations lie mostly within the ensemble range. The RMSEs between in situ observations and the ensemble mean and median are mostly reduced compared to the default model output. The NRRs for monthly variability suggest that the ensemble spread is generally narrow (NRR 1.21-1.39 for ~~nitrogen~~ DIN and 1.19-1.39 for chlorophyll profiles, 1.07-1.40 for surface chlorophyll, and 1.01-1.40 for depth integrated chlorophyll). Among the five stations, the most reliable ensembles are obtained for the oligotrophic station ALOHA (for the surface and integrated chlorophyll and bloom peak height), for coastal station L4 (for inter-annual mean), and for the abyssal plain station PAP (for bloom peak height). Overall our study provides a novel way to generate a realistic ensemble of a biogeochemical model by perturbing the model equations/parameterizations, which will be helpful for the probabilistic predictions.

1 Introduction

Major changes in ocean biogeochemistry have been driven by anthropogenic activities, leading to ocean acidification, eutrophication, and increased levels of dissolved inorganic carbon (Gehlen et al., 2015; Bopp et al., 2013; Doney, 2010). To understand how the ocean ecosystem responds to these changes, marine biogeochemical models have been developed. The majority of these models focus on the lower trophic food-webs and explicitly represent dissolved nutrients, phytoplankton~~and zooplankton(NPZ~~, zooplankton, and detritus (NPZ-D). These models are then coupled with physical general circulation models to address and predict the impact of climate change in the ocean ecosystems (Doney et al., 2012; Yool et al., 2013; Butenschön et al., 2016), to assess the impact of anthropogenic input on biogeochemical cycles in the marine ecosystem (Bopp et al., 2005) , and to produce decadal reanalyses (Ford et al., 2012).

Marine biogeochemical model development began with simple NPZ models, and has become steadily more complex with increasing computing power and knowledge of ocean biogeochemistry (Anderson, 2005; Anderson et al., 2015). NPZ models consist of three compartments: nutrients as the primary resource, phytoplankton as the primary producers, and zooplankton as herbivores or grazers. Such models have been used to investigate the range of possible ecosystem behaviours before coupling them to a physical model (Franks, 2002) and seeking to represent observations at particular sites (Fasham et al., 1990; Robinson et al., 1993). More advanced biogeochemical models represent more processes and feedbacks compared to the NPZ models (Raick et al., 2006), covering much more of the lower-trophic food web (Anderson, 2005). Inclusion of cell size representations (Berelson, 2002; Le Quèrè et al., 2005), different phytoplankton functional types, such as calcifiers and dimethyl sulphide producers (Le Quèrè et al., 2005), and the addition of important micronutrients, such as iron to permit phytoplankton growth limitation (Yool et al., 2011, 2013), are now part of many biogeochemical models. Moreover, in order to investigate the effect of global climate change and anthropogenic activities in the ocean, more complex marine biogeochemical models are now being embedded into earth system models. For example, the Model of Ecosystem Dynamics, nutrient Utilisation, Sequestration, and Acidification (MEDUSA) (Yool et al., 2011, 2013) is the chosen biogeochemical component for the UK Earth System Model, as it has high spatial correlation with patterns of pCO_2, DIC, and alkalinity (Cox and Kwiatkowski, 2013; Kwiatkowski et al., 2014).

Despite becoming more complex (Anderson, 2005), the basic interactions among nutrients, phytoplankton, and zooplankton are still at the heart of all marine biogeochemical models. These interactions are governed by four primary processes, represented in the simplest NPZ models: nutrient uptake, grazing by zooplankton, phytoplankton and zooplankton mortality, due to diseases or implicit higher trophic levels (Yool et al., 2011). These processes are functions of the state concentrations and can be parameterized by different functional forms along with adjustable parameters. Biogeochemical models therefore have different sources of uncertainty, such as the physical input (Sinha et al., 2010; Doney, 1999; Hemmings and Challenor, 2012), parameters (Oschlies and Schartau, 2005; Friedrichs et al., 2006, 2007), and the model structure associated with how the ecosystem is represented, either by the number of model compartments and linkages (Friedrichs et al., 2007; Kriest et al., 2010;

Ward et al., 2013), or its mathematical formulations (Anderson et al., 2010; Flora et al., 2011; Adamson and Morozov, 2013; Aldebert et al., 2016). Sensitivity analyses show that small changes in the structural process formulation often produce larger changes in system dynamics, compared to varying parameter values alone (Wood and Thomas, 1999; Fussmann and Blasius, 2005; Levin and Lubchenco, 2008; Flora et al., 2011; Adamson and Morozov, 2013; Aldebert et al., 2016), a result known as structural sensitivity (Wood and Thomas, 1999; Flora et al., 2011; Adamson and Morozov, 2013). A study by Aldebert et al. (2016) shows that parameter values are weakly correlated to food-web dynamics compared to the model formulations, as equilibrium dynamics are determined by the choice of functional forms.

Structural sensitivity may be less significant in models built on well-tested mechanisms as in the physical sciences, however in biogeochemical models the process functional terms are all gross simplifications. This is even more problematic if the processes are poorly understood so that justification for any specific representation is weak (Adamson and Morozov, 2013). Often it is difficult to implement the functional relations that are observed in the laboratory into a large-scale ecosystem with heterogeneous populations (Englund and Leonardsson, 2008). It is known from studies of simple predator-prey models that similarly shaped equations often lead to completely different stability and oscillatory model dynamics (Fussmann and Blasius, 2005; Roy and Chattopadhyay, 2007). Moreover, a specific functional form may not capture all details of the biological processes, for example, the Michaelis-Menten type function for grazing, commonly known as the 'Holling Type II', fails to correctly describe what happens to grazers' movements when satiation has been reached (Flynn and Mitra, 2016). These discrepancies from simple interaction models suggest that complex biogeochemical models need to be tested by altering their default functional forms (Anderson and Mitra, 2010; Anderson et al., 2010).

A few studies have investigated the effects of biogeochemical process formulations~~, e. g.,~~. For example Yool et al. (2011) has demonstrated in an intermediately complex model, that linear density-dependent mortality produces the biggest difference in diatoms compared to non-diatoms and zooplankton, with concentrations at ~~high~~ mid latitudes being twice as high, compared with sigmoidal, quadratic, or hyperbolic forms. The choice of zooplankton grazing equations affect phytoplankton concentration dramatically in a model with five plankton types, PlankTOM5.2 (Le Quèrè et al., 2005). The Holling type II grazing function produces 30% less total surface phytoplankton concentration compared to the sigmoidal (Holling type III) function, in the North Atlantic and North Pacific (Anderson et al., 2010). However Anderson et al. (2015) shows that when two similarly-shaped photosynthesis-irradiance curves (Smith and the exponential) function, were used in an NPZ-detritus model, the concentration of chlorophyll during the spring bloom was only slightly higher (0.2 mg m^{-3}) for the exponential function, with little difference in phytoplankton dynamics (Anderson et al., 2015).

Since the individual compartments of models interact with one another, any biological perturbation is likely to affect the whole ecosystem dynamics. In climate modelling, perturbed physics ensembles have been developed to investigate multiple parameter uncertainty (Murphy et al., 2007; Tinker et al., 2016), and multiple parametrization (functional) uncertainties (Subramanian and Palmer, 2017). Inspired by these studies, here we attempt to generate a perturbed biogeochemical ensemble where model equations are varied by embedding different functional forms to describe the core processes, similar to the multi-parameterization ensembles in physical models. We implement this framework in the MEDUSA model (Yool et al., 2011, 2013), which is a lower trophic level model with two phytoplankton functional types, distinguished as large diatoms and small

non-diatoms, two zooplankton types represented by mesozooplankton and microzooplankton, and three nutrients: silicic acid, iron, and ~~inorganic nitrogen. Nitrogen~~ dissolved inorganic nitrogen (DIN). DIN is the primary currency of the model, similar to NPZ models, but MEDUSA allows phytoplankton to have different C:N ratios and Si:N ratios for diatoms. Diatoms utilise the silicic acid and can only be grazed by mesozooplankton. MEDUSA also includes an iron submodel developed by Parekh et al. (2005) based on Dutkiewicz et al. (2005), in which iron is separated into "free" iron and iron bound to organic ligands. Iron is removed by scavenging and added to the ocean by aeolian deposition.

We assess of the uncertainty arising from the MEDUSA model's equations from ensemble outputs generated using possible functional form combinations within the NPZ compartments. For simplicity we use a 1-D version of MEDUSA-1.0 model (Yool et al., 2011; Hemmings et al., 2015), and produce results for five oceanographic stations covering abyssal plain, oligotrophic, and coastal regimes. Apart from the model outputs on concentration of nutrients and chlorophyll, we also examine the emergent properties using phytoplankton phenology metrics. The performance of the ensemble mean, median, and the default MEDUSA run are compared with monthly and inter-annual values from in situ observations at those stations. We assessed the spread of the ensemble using the Normalised RMSE Ratio (NRR) which assesses the likelihood of the observations fitting the ensemble range. Section 2 describes the equations used and how the ensemble is run. The assessment of the uncertainty in terms of chlorophyll concentrations, phytoplankton phenology, and comparisons with the observations are described in section 3, and are further discussed in section 4.

2 Method

To explore structural uncertainty we first make the functional forms representing key processes more similar to each other by tuning the shape-defining parameters. For example, for Holling type II and Holling type III, we fix the maximum rates of each process, and implement a non-linear least squares method to optimize the half saturation coefficients so that the overall shapes are similar. This approach is used for nutrient uptake (4 functional forms), phytoplankton mortality (4 functional forms), and zooplankton mortality (4 functional forms), as in the subsections below. Table 1 shows the equations and parameter values.

2.1 Nutrient uptake

Alongside light, nutrient concentration limits the growth of phytoplankton. In MEDUSA the standard hyperbolic monod, hereafter U_1, function is the default. The growth of cells monotonically increases with ambient nutrient concentration, and halts when nutrients become scarce. If nutrient concentrations are high, the rate of uptake saturates. Other mathematical functions show similar properties including (i) Sigmoidal (Fennel and Neumann, 2014) U_2, (ii) the exponential (Ivlev, 1961), U_3, and (iv) trigonometric functions (Jassby and Platt, 1976), U_4. All these functions include a shape-defining parameter, k, which for monod and sigmoidal can be interpreted as a half saturation constant, and a maximum uptake rate, V_{p^T}, which is a function of temperature (Eppley, 1972): $V_{p^T} = V_p 1.066^T$, where V_p is the maximum growth rate when temperature, T, is at 0°Celsius. The uptake function of different phytoplankton types and nutrients use similar functions but different parameter values for k, summarised in Table 1, obtained by minimising the sum squared difference with U_1. The nutrient uptake functions after

optimization are shown in Fig. 1(a). The fit is done over the nutrient concentration ranging from 0.001 to 20 mmol N m^{-3} and are discretised into 1000 intervals. The difference in shape of the optimised functional forms are more obvious ~~at low nutrient concentrations.~~ between 0.1 to 1 mmol N m^{-3}. The fitting was done based on non-linear least square optimisation method using python's curve_fit function from scipy.optimise.

2.2 Zooplankton grazing

In MEDUSA, both phytoplankton and zooplankton are grouped into "small" and "large" categories. The small zooplankton, represented by the microzooplankton, graze on ~~small phytoplankton,~~ non-diatoms ~~,~~and detritus, with the more nutrient rich, higher quality, non-diatoms preferred over detritus. Larger zooplankton, represented by mesozooplankton have a broader range of prey, including both microzooplankton and diatoms, which are ~~high~~ higher quality food sources compared to non-diatoms and detritus. When describing multiple grazing functions, the zooplankton grazing rate is often defined using either the Holling type II hereafter, G_2 or Holling type III hereafter, G_1, function with maximum grazing rate g_m, and a weighted preference on the different food sources p_n (Fasham et al., 1990). The preference parameter changes through the year as a function of the food ratio. G_2 and G_1 grazing on prey Pa are described in Table1. In MEDUSA, the default multiple grazing parameterisation is based on the sigmoid Holling type III (Ryabchenko et al., 1997) function. Apart from the weighted preference, both of these functions include a half saturation constant k_x, where x is the zooplankton type.

These functions ~~both become constant at~~ approach a maximum grazing rate ~~.~~at high concentrations of prey. During the fitting process, the range of phytoplankton and microzooplankton concentration used was 0.001 to 10 mmol m^{-3}, and discretised in 1000 intervals. At low zooplankton concentrations the sigmoidal response has lower grazing rates than the hyperbolic, and therefore, the sigmoidal curve has a more rapid increase in predation rate before becoming saturated (Edwards and Yool, 2000), shown on Fig. 1(~~b~~c). Preferences for food types are kept the same as MEDUSA's default parameters, with terms summarized in Table 1.

2.3 Plankton mortality

MEDUSA has both density independent and density dependent mortality rates for all the phytoplankton and zooplankton types. Density-independent loss is modelled by a linear function representing plankton metabolic loss which was kept unchanged. Density-dependent loss includes processes such as higher-trophic grazing and disease. In MEDUSA these processes are modelled using the hyperbolic function of plankton concentration (Fasham et al., 1993). Alternative functions can describe the density-dependent mortality, and we use the combinations of hyperbolic (ρ_1, ξ_1), linear (ρ_2, ξ_2), quadratic (ρ_3, ξ_3), and sigmoidal (ρ_4, ξ_4) functions to describe the phytoplankton (ρ) and zooplankton (ξ) mortalities (equations and abbreviations are shown on Table 1). Similar to grazing and nutrient uptake, the functional forms have different maximum rates for each plankton type. These maximum rates are made the same for all the different functions.

Of the four different mortality functions, linear and quadratic functions are most different in shape, as shown on Fig. 1(~~b~~c). Using the linear term is similar to a change in the value of maximum mortality rate, μ. To make the linear function similar to the sigmoidal and hyperbolic functions, the maximum ~~grazing~~ mortality rate is set so that the total loss integrated over the range

of ~~prey density~~ phytoplankton concentrations (calculated as the area below the function representing the total loss in linear terms, between 0.001 to 10 mmol m^{-3}) is similar to that for the hyperbolic curve. The quadratic term, instead of asymptoting, continues to grow with plankton abundance. In order to keep this similar to other forms, after reaching a certain concentration the function is switched to linear, so that the rate plateaus at high abundance. For sigmoidal mortality, the default μ are not changed but the half-saturation constant, k_M is optimised. The optimised mortality functions are shown in Fig. 1(c). The range of phytoplankton and zooplankton concentrations used during the fitting process was between 0.001 - 10 mmol m^{-3}, and discretised within 1000 intervals. A distinctive feature of these functional forms after optimisation is that the quadratic mortality rate remains low until phytoplankton concentration reaches ~~10 mmol N~~ 1.0 mmol m^{-3}, and the linear function shows consistently high plankton mortality (Fig. 1(c)).

2.4 Model Parameters

Apart from sinking rate, maximum growth, and grazing rates, parameters not listed in Table 1 are kept at their default values (Yool et al. (2011) shown on table 1-4). From a previous 3-D MEDUSA run, ~~in~~ the oligotrophic regions show a low 'background' chlorophyll concentration (Yool et al., 2011) so to raise this concentration a higher maximum growth rate and lower grazing rate has been used. The maximum uptake rate, V_p , ~~as~~ is 0.8 day^{-1}, similar to that in the HadOCC model (Palmer and Totterdell, 2001). For zooplankton grazing, similar to NPZ models (Fasham et al., 1990; Fasham, 1995; Anderson et al., 2015) we use 1 day^{-1} as the maximum grazing rate, g_m. MEDUSA also ~~contains~~ parameterises both slow and fast detritus sinking factors. It is assumed that the latter sinks rapidly relative to the model time-step, and remineralisation of the detrital nitrogen and silicon is done implicitly. In the default model 3 m day^{-1} is used for the slow sinking detritus, however over long runs we found this leads to downward loss of nutrients from the euphotic zone, to the sea floor. Earlier studies have used lower detrital sinking rates (Steele and Henderson, 1981; Fasham et al., 1990; Lacroix and Gregoire, 2002; Raick et al., 2006), between 0 to 1.25 m day^{-1} and other study have suggested to use 0 m day^{-1} (Ward et al., 2013). We chose a lower sinking rate of 0.1 m day^{-1} to prevent depletion of state variables particularly at the shallower stations.

2.5 Running the Model and Generating the Ensemble

MEDUSA is run in the Marine Model Optimization Testbed (MarMOT-1.1) (Hemmings and Challenor, 2012; Hemmings et al., 2015), a site-based mechanistic emulator, where simulations are run in 1-D. MarMOT was developed to investigate the effect of sensitivity in plankton model simulations, especially in regard to parameter and environmental inputs (Hemmings and Challenor, 2012). Despite some uncertainties associated with the differences in physical forcing, fluxes, and initial values of biogeochemical properties, using 1-D simulations to approximate 3-D model behaviour for calibrating models based on specific sites has improved the 3-D models' predictive skill (Oschlies and Garçon, 1999; Oschlies and Schartau, 2005; Kane et al., 2011; McDonald et al., 2012). The 1-D MEDUSA is run at five oceanographic stations: ~~PAP~~Porculpine Abyssal Plain Sustained-Obervatory (PAP-SO, hereafter PAP), A long time Oligotrophic Habitat Assessment (ALOHA), Bermuda Atlantic Time Series (BATS), ~~ALOHA, BATS,~~ Cariaco, and L4 shown in Fig. 2. These are chosen as they represent different oceanographic regimes: abyssal plain (PAP), oligotrophic (ALOHA, BATS), and coastal (Cariaco, L4).

At each oceanographic station, all combinations of the optimized functional forms (as described in subsection 2.1, 2.2, and 2.3), are then embedded into the 1-D MEDUSA code. The same process function is always used for both diatoms and non-diatoms, or mesozooplankton and microzooplankton. The ensemble model at each station is initialized using in situ measurements of chlorophyll, inorganic nitrogen, silicic acid, and iron, and the ensemble is run over 10 years starting from January 1998. This provides a total number of 128 combinations, arising from 4 types of nutrient uptake, 4 phytoplankton ~~mortalities~~mortality formulations, 2 types of zooplankton grazing, and 4 zooplankton mortalities.

2.5.1 Physical input

Physical input files consist of gridded values of vertical velocity (m day^{-1}), vertical diffusion coefficient (m^2 day^{-1}), and temperature (°C), which are applied at each depth level. Additionally, time series of downwelling solar radiation (W m^{-2}) and mixed layer depth (m) are also used as input. These are obtained from the 5-day mean output of the Nucleus for European Modelling of the Ocean (NEMO) model, using the Met Office Forecast Ocean Assimilation Model (FOAM), which controls the physical parameters and therefore the biogeochemical tracers every 5 days. The FOAM-NEMO system assimilates *in situ* satellite SST, sea-level anomaly, sea-ice concentration, temperature, and salinity profile data, in order to make the physical system more realistic (Storkey et al., 2010). However, assimilating physical data directly into a coupled physical-biogeochemical model often does not improve the simulation of the ecosystem. For example when assimilation is used in the 3-D HadOCC model it overestimates the nutrient concentrations due to spurious vertical velocities (Ford et al., 2012; Ourmières et al., 2009).

To avoid overestimating surface nutrients the vertical velocities from the FOAM system were capped at the 90th and 10th quantiles, and the 10-year mean of the vertical velocity is also removed. This means that there is no time mean vertical velocity. These adjustments gave a better long-term vertical structure to the nutrient and other distributions. Since input data on the vertical diffusivity was not stored in FOAM, we used values from NEMO ORCA025-N102 output from January 1998-December 2001 and from ORCA0083-N01 from January 2002-December 2007, both obtained from the CEDA Group workspace web (http://gws-access.ceda.ac.uk/public/nemo/#_top). These physical inputs are 5-day averaged and are available at 75 depth levels (from 0.5 to 6000 m) for NEMO-FOAM and ORCA0083-N01, and 63 depth level (spanning from 6 to 5800 m) for NEMO ORCA024-N102, with less depth intervals than the 75 levels. The level thickness increases exponentially as the depth goes deeper. Our 1D model uses these same 63 depth ~~levels vertical resolution~~ level thickness of vertical resolution in order to minimise computational costs.

2.5.2 Biogeochemical input and validation data

The 1D MEDUSA ensemble is run at five oceanographic stations: ~~Porcupine Abyssal Plain Sustained Observatory (PAP-SO, hereafter, PAP), A long time Oligotrophic Habitat Assessment (ALOHA), Bermuda Atlantic Time Series (BATS)~~PAP, ALOHA, BATS, Cariaco, and L4. The input for the biogeochemical environment are the initial conditions for the 11 primary tracers (state variables) including; dissolved organic nitrogen (DIN), non-diatom, diatom, silicon in diatom, silica, detritus, microzooplankton, mesozooplankton, non-diatom chlorophyll, diatom chlorophyll, and iron (mmol m^{-3}), along with the model parameter values. Initial conditions for chlorophyll, silicate, iron, and ~~nitrogen~~ DIN concentrations and are taken from the in situ data

at the five oceanographic stations. We did not use spin up runs when initialising, as discussed in the supplementary material section S1. Location coordinate, data source, and maximum depth are summarised in Table 2 and the stations locations are shown in Fig. 2. After initialization, in situ data from these stations are used to validate the model results. For station PAP, we also use SeaWIFS-derived chlorophyll-a data with 9 km spatial resolution and 8-day averages provided by GlobColor (http://hermes.acri.fr/) for validating the surface chlorophyll.

At these stations, the DIN consists of ammonia, nitrate, and nitrite, however at oligotrophic stations like ALOHA the ammonium is below the detection limit (Hawaii Ocean Time Series), and therefore DIN only consists of nitrate and nitrite. At PAP we use the initial condition from one of MarMOT's test stations, located at 50°N, 20°W (Hemmings et al., 2015), since the nitrate data were only collected between 30-400 m. At station L4 ~~, chlorophyll and nitrogen~~ chlorophyll and DIN data were collected from the surface from 1999-2008. ~~Therefore the~~ Since the maximum depth in this station is only 50 m deep, the initial concentrations for chlorophyll and ~~nitrogen~~ DIN are the same at every depth (total chlorophyll = 0.27 mg m^{-3}, ~~nitrogen~~ DIN = 6 mmol m^{-3}). Other inputs that are not available at the websites mentioned above, such as microzooplankton, mesozooplankton, and detritus were taken from the nearest test stations. In the oligotrophic stations, 75 % of total chlorophyll was allocated initially to the non-diatom phytoplankton since these dominated the water column (Villareal et al., 2012). At the other stations half of the total chlorophyll goes into the diatoms.

For validation of the model, we consider the total chlorophyll-a concentration, instead of separating diatoms and non-diatoms. The model is simulated at 37 depth levels, spanning from 6-1200 m to minimise computational cost, apart from station L4, with maximum depth is 50 m, and Cariaco, where the maximum depth for the physical input is available down to 500 m. The level thickness is similar to that in 63 depth levels. At the lowest level, vertical velocity and diffusion are set to zero and this level is a sink for detritus. Additionally, apart from the physical input files a time series for soluble iron flux from dust deposition is applied, but this is constant using the average value from (Mahowald et al., 2009).

2.6 Model Metrics

We use statistical metrics including correlation coefficient, root-mean squared error (RMSE), bias, ensemble range, and 10-year mean depth profiles of ~~nitrogen~~ DIN and chlorophyll and integrated chlorophyll. For surface chlorophyll, apart from the metrics above we use the mean chlorophyll abundance each year in order to see inter-annual variability, and monthly abundance for ~~to~~ the seasonal variations. A similar approach is applied to ~~nitrogen~~DIN, however we use the ~~integrated nitrogen~~ averaged DIN over 200 m (integrated ~~nitrogen~~ DIN / depth) to calculate the inter-annual mean and monthly abundance. These statistical metrics are compared with in situ data. We also consider the phenological aspects of the phytoplankton spring bloom, which are useful ecological indicators for detecting natural and anthropogenic impacts on the pelagic ecosystem (Platt and Sathyendranath, 2008). We consider seven phenology indicators as metrics, including an initiation time where the chlorophyll concentration exceeds a certain threshold, at half the concentration of the bloom peak. When the bloom concentration starts to diminish, we derived a termination time, where bloom concentration falls below the same threshold. The number of days when chlorophyll concentration is higher than the threshold is the bloom duration. The concentration at the bloom peak and the date it takes place, are also included as indicators. We also note the amplitude of the bloom, which is half of the peak height minus

the minimum chlorophyll concentration. These indicators are derived using the method described in appendix A, and applied to all ensemble outputs for each year.

In an ensemble forecast system, an ensemble with good reliability is the one that is statistically consistent with the observations, such that the observation is statistically indistinguishable from the ensemble members. In order to assess the value of the ensemble probability distribution we must assess the consistency of the ensemble spread as well as the ensemble mean error (Moradkhani and Meskele, 2010). A simple method is discussed by Anderson (2001) which takes the ratio R_a of RMSE of the ensemble mean and the mean RMSE of all the ensemble members which has the expectation value $E[R_a] = \sqrt{\frac{(n+1)}{2n}}$, where n is the number of ensemble members. This is called the Normalised RMSE Ratio (NRR= $R_a/E[R_a]$) where the desirable ensemble spread is expected to have NRR=1. If the NRR >1 then the spread is too small, and NRR <1 indicates that the ensemble spread is too large. We may expect different NRR values for different metrics and also for variability on different timescales, such as monthly or inter-annual data. This method has previously been used to set the number of ensemble members in data assimilation (Moradkhani et al., 2006; Roy et al., 2012).

3 Results

First the ensemble range and mean are compared with the observational fields, followed by the error statistics for the ensemble mean/median, the default run, and the ensemble range in order to assess whether the ensemble spans the observational data. Then variability are compared with the in situ data, followed by the NRR to assess the ensemble spread, and phytoplankton bloom phenology.

3.1 Abyssal Plain

In station PAP, in situ nitrate was only measured from mid 2002 to mid 2004 with a maximum depth of 400 m and chlorophyll from mid 2003 to mid 2005 with maximum depth of 200 m, as in Table 2. Surface chlorophyll is derived from SeaWIFs (8-day averaged) and is available for the full 10-year time series (see supplementary Fig. S5).

Distinct seasonality is simulated by the ensemble mean. High nitrate concentrations at the surface occur during winter (December-April) and decline in summer. From the seasonal mean ~~nitrogen~~ DIN profile in Fig. 7(a), the ensemble inter-quartile range shows later bloom peaks compared to in situ, and mean ~~nitrogen~~ DIN concentrations are high during months when both the ensemble and in situ ~~nitrogen~~ DIN decline (between May to June), and show an earlier spike of ~~nitrogen~~ DIN in July instead of September, and therefore underestimating the increase of ~~nitrogen~~ DIN between October to December. These shared errors make the ensemble spread (NRR=1.25) still too narrow for the phenological metrics.

Chlorophyll concentration starts to decline at a depth of $\sim$ 50 m, which also corresponds to the decline in the chlorophyll inter-quartile (between 25$^{\text{th}}$ and 75$^{\text{th}}$ percentile) range shown on Fig. 3(b). Chlorophyll also shows seasonality, similar to that of ~~nitrogen~~DIN. The in situ data show high concentrations of chlorophyll in the top 70 m during May-June, coinciding with the shallowing mixed layer depth. In the model this occurs earlier in spring (between end of April to May), and slightly deeper, to 100 m, as summarised in Fig. 3(a) and (c).

~~Chlorophyll and nitrogen profile 10-year means~~ Observed mean profiles of chlorophyll and DIN are within the ensemble range, with ensemble spread on the narrow side, with NRR= 1.20 and 1.25 for chlorophyll and ~~nitrogen~~ DIN respectively. For chlorophyll and ~~nitrogen~~ DIN profiles, the ensemble median shows the highest correlation and lowest RMSE and bias, compared to the default or the ensemble mean. High RMSEs in ~~nitrogen~~ DIN occur from ensemble members that contain the U_4G_2 combinations, as shown in Fig. 13(~~t~~j), which also correspond to high ~~nitrogen~~ DIN mean (< 9 mmol m^{-3}). High chlorophyll profile RMSEs (>0.62) are produced from members that combine G_1 with $\rho_1\xi_2$, $\rho_3\xi_3$, and $\rho_1\xi_4$, and this coincides with high chlorophyll mean (> 0.7 mg m^{-3}). Surface chlorophyll 10-year mean and RMSEs (>0.8) are notably high when combining U_1 with $\rho_1\xi_2$, $\rho_3\xi_3$, and $\rho_1\xi_4$, as summarised in Fig. 12(e) and (j).

When compared to satellite-derived chlorophyll-a, the surface chlorophyll at this station shows higher correlations and lower RMSEs than in other regions, especially with the ensemble median, compared to the default run or the ensemble mean. In years when satellite-derived chlorophyll is not within the ensemble range this is due to the ensemble overestimating the satellite-derived chlorophyll (supplementary material Fig. S5), giving a slightly narrow ensemble spread (NRR=1.29). There is also inter-annual decline in satellite-derived chlorophyll, ($r =$ -0.21, $p < 0.05$). Six ensemble members capture this decline in surface chlorophyll, although with weaker correlations ($r =$ -0.14 ($\pm$0.06), $p < 0.05$). In certain years (1998, 1999, and 2001) the observed chlorophyll are not within the ensemble range, Fig. 4(a), making the ensemble spread too narrow, with NRR of 1.26.

The range of surface chlorophyll annual mean is 0.7 mg m^{-3}. If we only allow one process function at a time to change ~~the~~ (by changing functional form in one process at a time, whilst keeping the other processes at their respective default functions), the new ensemble range reduces to 0.58 mg m^{-3}, covering 84% of the all ensemble members. If the original MEDUSA parameters are used, the interannual chlorophyll fits the ensemble slightly better, but the ~~nitrogen~~ DIN fit gets worse. The results from using MEDUSA parameters and in situ ~~nitrogen~~ DIN concentration as initial condition can be found in the supplementary material, section S2 and S3. For monthly data (Fig. 5(a)), the ensemble shows only slightly earlier peak chlorophyll concentrations in May compared to in situ, which occur in either May or June. However, since the ensemble mean and median overestimate the satellite-derived chlorophyll during months of high chlorophyll (> 0.5 mg m^{-3}), during peak chlorophyll in May the satellite-derived chlorophyll is outside the ensemble range.

3.2 Oligotrophic Ocean

In oligotrophic regions nutrients are scarce at the surface but may be plentiful at deeper depths (Dave and Lozier, 2010; Lipschultz, 2001). Figure 9(d) and (e), show that the ensemble range decreases as the depth increases, with high ensemble range found at depths between 3-50 m. At ALOHA all ensemble members have similar distributions for in situ ~~nitrogen~~DIN, Fig. 9(e), with ~~nitrogen~~ DIN levels > 1.0 mmol m^{-3} only found $\sim$ 150 m depth. However, for inter-annual means, Fig. 4(b), the ensemble inter-quartile range mostly overestimates the in situ observations. There is also an increasing trend of in situ ~~nitrogen~~ DIN ($r = 0.69$, $p < 0.03$), which is not captured by the ensemble, Fig. 6(b), leading to an NRR of 1.38. This overestimation is also observed in the seasonal mean, Fig. 7(b), and in situ data is rarely within the ensemble range.

At BATS, the ~~nitrogen~~ DIN concentration in the top 200 m is clearly overestimated, Fig. 9(k), with ~~nitrogen~~ DIN levels $>$ 1.0 mmol m^{-3} at $\sim$ 10 m (with some members occasionally showing such concentrations at 3 m). Consequently, the overall mean ~~nitrogen~~ DIN concentration is overestimated, as indicated by the positive bias in the ensemble mean, in Table 3. Similar to ALOHA, ~~nitrogen~~ DIN inter-annual and seasonal means are overestimated, summarised in Fig.6(c) and 7(c), respectively. This results in narrow ensemble spread (NRR= 1.40). As at ALOHA, an increasing trend in ~~nitrogen~~ DIN is observed ($r = 0.67$, $p < 0.03$), but only 28.9% of the ensemble results, which uses G_2 as its grazing function shows similar trend.

Another feature of the oligotrophic ocean is a deep chlorophyll maximum (DCM) that occurs below the mixed layer (Fennel and Boss, 2003). In Fig. 9(b) and 9(h), high chlorophyll concentrations are simulated by the ensemble mean between 70-90 m in BATS and up to 150 m in ALOHA. A DCM occurs when lower chlorophyll is detected at the surface, which roughly matches with the in situ profiles at ALOHA (see Fig. 9(c) and Letelier et al. (2004)) and BATS (Fig. 9(i)) although the depth of the ensemble DCM at both stations is 10-20 m shallower than in situ. The high subsurface chlorophyll coincides with a higher ensemble range, with the range decreasing with depth. However neither BATS nor ALOHA show the continuous DCM seen in the in situ profiles, Fig. 9(c) and 9(i).

The majority of ensemble members underestimate in situ 10-year mean chlorophyll profile concentrations, especially at BATS where all ensemble members show positive bias for both surface and integrated chlorophyll profiles. Figure 11(a) and (b) show that ensemble members with G_2, $\rho_2\xi_2$, $\rho_2\xi_3$, $\rho_2\xi_4$, $\rho_3\xi_1$, and U_3 produce lowest chlorophyll concentrations (< 0.08 mg m^{-3}), which then coincide with high RMSE, shown in Fig. 11(e) and (f). The magnitude and range of RMSEs at BATS are highest, and the default run has lower RMSEs than ensemble mean and median, summarised in Table 3. Ensemble members that use U_1G_1 and U_4G_1 show highest profile 10-year mean concentrations of chlorophyll at both stations, although when paired with $\rho_3\xi_3$ and $\rho_1\xi_4$, the RMSEs increase. High ~~nitrogen~~ DIN concentrations are almost always observed when U_3 and U_2 were used in these oligotrophic regions, summarised in Fig. 13(a) and (b).

As for profiles, surface chlorophyll at ALOHA (supplementary Fig. S2) has lower RMSEs and higher 10-year mean concentrations compared to BATS, summarised in Table 3. Ensemble members with low surface chlorophyll concentrations are the same as those with low chlorophyll profiles, although high surface chlorophyll RMSEs also coincide with high surface concentrations, summarised in Fig. 12(a) and (f). The low RMSEs for surface chlorophyll at ALOHA are also reflected in the NRR (1.07), and the ensemble almost always encompasses the in situ observations, (see supplementary material Fig. S2). During summer months (June-September), most ensemble members still underestimate the in situ monthly means, summarised in Fig. 5(b), and peak values are not always covered by the ensemble.

Figure 4(b) and (c), show no distinct inter-annual variability at either ALOHA or BATS, and in most years the default run is closer to in situ. Model output at BATS have a lower 10-year mean surface chlorophyll than in situ data and most ensemble members underestimate the surface inter-annual means, making the ensemble spread appear narrow both in the 10-year and inter-annual means, shown on Table 3 and 4. Changing the functional forms one at a time produces an ensemble range of 0.11 and NRR =1.39, slightly wider than the whole ensemble, summarised in Fig. 8. At BATS in 2004, a high in situ chlorophyll mean was recorded (0.65 mg m^{-3}) and none of the ensemble members captured this high mean, see the supplementary material Fig. S1 and 4(c), which therefore narrows the NRR value for annual mean at BATS. At ALOHA the range for inter-annual

means is too wide, with NRR= 0.84. This is broader than the overall time series mean, as the in situ inter-annual means are mostly within the 75th quartile, making the mean RMSE of the ensemble higher than the RMSE of the ensemble mean (0.043 and 0.025 respectively). However, when only one process is perturbed at a time, the NRR becomes narrow (1.17), and the in situ is no longer within the inter-quartile range, as shown on Fig. 8 and summarised in Table 4. In 1999, the ensemble mean and median is higher than the in situ, which is even clearer in the inter-annual mean of the primary production.

Unlike at PAP, there is no distinct seasonality in chlorophyll. At ALOHA during months of low chlorophyll ($<$ 0.1 mg m^{-3}, occurring in July-October), most ensemble members underestimate surface chlorophyll. At BATS, in situ concentrations are clearly underestimated during the same months, summarised in Fig. 5. At ALOHA, between December to May, when in situ chlorophyll seasonal means are $>$ 0.1 mg m^{-3}, the in situ data are within the inter-quartile range, but at BATS this only happens when in situ chlorophyll means decrease from 0.20 to 0.08 mg m^{-3} in May.

At ALOHA the ensemble mean and median produce smaller RMSEs for both chlorophyll and ~~nitrogen although~~ DIN in the depth profiles compared to BATS. Bias in the default run is still smaller than for the ensemble mean and median. However for surface and integrated chlorophyll the ensemble mean and median have lower bias than the default. At BATS low RMSEs and bias with high correlation coefficient are produced by the default run compared to ensemble mean and median. At both stations, integrated chlorophyll ensemble mean and median have smaller RMSEs and a better correlation coefficient, compared to the default run. At ALOHA, NRR for the integrated chlorophyll is closer to 1 than for the surface and chlorophyll profiles. However the default run in oligotrophic regions generally produces higher chlorophyll and lower ~~nitrogen~~ DIN concentrations compared to the ensemble mean and median. This also matches better with in situ data as the correlation coefficient, *r* is higher. Using U_1G_1 is seen to give rise to higher chlorophyll concentrations.

3.3 Coastal

In the coastal stations, in situ observations show strong seasonality, shown on Fig. 14(c), (f), (g), and (h). This has not been captured by the ensemble mean in station Cariaco, shown in Fig. 14(a), with chlorophyll concentration almost constant above 0.7 mg m^{-3} in the upper 30 m and the surface (supplementary Fig. S3), shown in Fig. 5(d). Since the range of seasonal chlorophyll is wide, apart from in August and November, in situ concentrations are within the ensemble range. A decline of chlorophyll has been recorded at Cariaco from 2004 (Taylor et al., 2012), and this is captured by the ensemble mean, median, and default (r = -0.72, $p < 0.05$, r = -0.66, $p < 0.05$, and r =-0.35, $p < 0.05$ respectively). Similar to chlorophyll, ~~nitrogen~~ DIN from the ensemble also shows no seasonality, see Fig. 7(d). Nonetheless, in situ concentrations are still mostly within the ensemble range, apart from November, where in situ ~~nitrogen~~ DIN decreases to 5.32 mg m^{-3}. However in 2001, and between 2005-2006, inter-annual downwelling of nutrients is well reproduced, Fig. 14(d). Despite the lack of seasonality, annual means of chlorophyll and ~~nitrogen~~DIN, are mostly within the ensemble range, Fig. 6(f) and 4(f), with the NRR 0.78 and 1.15 for chlorophyll and ~~nitrogen~~ DIN respectively.

At station L4 the in situ and ensemble means both show seasonality of ~~nitrogen~~ DIN with high concentration ($>$ 8 mmol m^{-3}) during November to February, and close to zero ($>$ 0.1 mmol m^{-3}), during summer months, consistent with the observation from Smyth et al. (2010). The interquartile range often overestimates the surface ~~nitrogen~~ DIN concentrations, especially

during the sharp decline in April, May, September, and October, shown on Fig. 7(e). For inter-annual means, in situ ~~nitrogen~~ DIN concentrations are mostly within the ensemble range, apart from 2001 and 2002, but this results in NRR of 1.29. For chlorophyll, Fig. 14(g), sharp peaks are observed during spring (March-April) and fall (September), coinciding with the sharp decline of ~~nitrogen~~ DIN in spring. However this has not been represented in the model, where only one peak is simulated between May-June, summarised in Fig. 5(e). If only diatom chlorophyll concentration is shown, the two bloom events are clearer, especially in the default run (see supplementary material Fig. S4). The ensemble mostly overestimates the in situ during non-bloom periods (in situ range = 0.09-2 mg m^{-3}, ensemble range= 0.28-3.13 mg m^{-3}), so that the surface chlorophyll is not fully captured by the ensemble, and therefore NRR value is 1.21, indicating a too narrow spread.

At both stations, for the inter-annual mean, the ensemble range always includes the observations (Fig. 4(c) and (d)), the NRR values are given in Table 3. Weak positive correlations of surface chlorophyll for the ensemble mean is seen at both stations, and these are better, and with smaller RMSEs than the default run, apart from ~~nitrogen~~ DIN at L4. Similar to the oligotrophic stations, the integrated chlorophyll shows better correlation with in situ, compared to both surface and chlorophyll profiles. At L4, the ensemble mean shows high RMSE for surface ~~nitrogen~~DIN, but low RMSE for surface chlorophyll, summarised in Table 3.

Although from Table 3, in situ surface chlorophyll concentrations are slightly overestimated by the ensemble mean, most other ensemble outputs are underestimated at Cariaco, except for ensemble members that use the combinations $\rho_2\xi_3$, $\rho_1\xi_2$, $\rho_3\xi_3$, and $\rho_1\xi_4$. This in turn makes the ensemble spread appear too narrow in the NRR. Unlike the oligotrophic regions, these high chlorophyll concentrations coincide with higher RMSE (> 1.7). Higher ~~nitrogen~~ DIN concentrations (>1.2 mmol m^{-3}) with high RMSEs (> 1.5) are also associated with the same ensemble members, summarised in Fig 13(c). These same ensemble members however show relatively low ~~nitrogen~~ DIN concentration (>5 mmol m^{-3}) at station L4. The chlorophyll mean at L4 shows that high concentrations (> 0.2 mg m^{-3}) are produced when the model uses $\rho_2\xi_3$,$\rho_1\xi_2$, $\rho_3\xi_3$, and $\rho_1\xi_4$ combinations. These coincide with high RMSEs, especially in members which pair U_2 and $\rho_2\xi_3$, $\rho_1\xi_2$, $\rho_3\xi_3$, and $\rho_1\xi_4$. High ~~nitrogen~~ DIN concentrations (> 9 mmol m^{-3}) are produced by U_4G_2, with correspondingly high RMSE.

Surface chlorophyll at these coastal stations also has a higher relative range than other stations, with L4 higher than Cariaco, see Table 3. The wider spread for annual means compared to the monthly data over 10-years have been observed at all the stations, including the coastal stations. Despite having lower range than L4 in terms of surface 10-year means, for the inter-annual means (Fig 4(d) and (g)), the NRR value for Cariaco is too small (0.78), indicating the ensemble spread is wider than necessary, which is also observed for inter-annual primary production, shown in Fig. 10(b) and (c). However, if processes are perturbed one at a time, the NRR is closer to the ideal ensemble range (NRR= 0.90). On the other hand, at L4 the in situ annual mean is reliable (NRR=1.001), since the in situ chlorophyll is close to the ensemble median (see Fig. 4(e)). However if the ensemble is reduced by only perturbing one process at a time the NRR increases to 1.36, and the in situ data is no longer within the ensemble mean, shown on Fig. 8, despite the range still covering 86% of the full ensemble (2.14).

3.4 Phytoplankton Phenology

At most stations, the observed phenology metrics are covered by the ensemble range. There are differences in the timing of phenological events between the ensemble mean, median, and default run, ranging from a couple of days to a couple of weeks, as shown in Table 4. The timing of initiation, bloom peak, and termination show wide interquartile ranges for all stations and can lie between ∼20 and 100 days earlier than the in situ timing, apart from stations PAP and ALOHA, see Fig. 15(b). At stations PAP and ALOHA the bloom initiation and timing dates from the inter-quartile range is at least ∼40 days too early. However, the ensemble mean and median at station L4 and Cariaco are later than in situ timings. For bloom initiation both stations are two months late but are stil within the ensemble range. In terms of the timing of the bloom peak and termination, these are up to 3 months and 120 day late respectively.

BATS has the largest range of phenological timings, especially in termination time. For bloom initiation, the in situ timing is within the interquartile range and only three days earlier than the ensemble median. However, at ALOHA the initiation time shows more inter-annual variability (supplementary Fig. S6) and may occur in June, August and October, as well as in December and January. This causes the mean observed initiation time to end up in May. From Fig. 15(a), the ensemble run shows a mean initiation time between late January and April instead and so the observations fall outside the ensemble range. Due to this variable initiation, although peak bloom time is within the full ensemble range at ALOHA, the timing is outside the 75^{th} and 25^{th} percentile range, making the ensemble spread too narrow (NRR=1.35). The peak chlorophyll at ALOHA shown in Fig. 5(b), where high (> 0.1 mg m^{-3}) chlorophyll monthly means are recorded in June, August, and September as well as December and January, yet the ensemble mean and median show highest concentrations only in January and February, also placing the bloom timing outside the inter-quartile range, see Fig. 15(b). At BATS the earliest initiation in the ensemble is mid January, but the earliest in situ initiation occurs in February. Therefore, peak bloom time from the ensemble at BATS are usually later than in situ. However, ensemble estimates of bloom peaks for 30°N, where BATS is located, agree with a study by Racault et al. (2012), who identify early April as the peak time. Although the range of peak bloom time in BATS is very high (174 days), the NRR suggests that this range is still narrow (1.17), ~~because most ensemble members produce blooms between April and May, and~~ as the ensemble does not cover the in situ timing~~occurs on 29 March and so is still outside the interquartile range~~. The large full range is caused by some ensemble members blooming much later. However, since the in situ timing is earlier, it is not within most of the ensemble range, so the overall ensemble spread appears narrow.

Both coastal stations show in situ initiation typically happens in mid-March, which is usually within the ensemble range, which spans 100 days (between the end of February and late June). The ensemble means show later initiation, with the 75^{th} and 25^{th} spanning mid April to end of May for Cariaco, and between early and mid May for L4. This later timing is also clear in peak bloom times, shown on Fig. 15(b). Figure 5(e) shows the in situ bloom at L4 is one to two months overestimated by the ensemble. Cariaco is the only station with peak bloom time, duration, and termination outside the ensemble range, due to the lack of chlorophyll seasonality, as noted in section 3.3, also resulting in higher NRR values.

Initiation timing is captured best at station PAP, with the ensemble median's initiation averaging only eight days earlier than for the satellite-derived chlorophyll, resulting in NRR=1.14 for initiation, closer to one compared to other stations. A typical

North Atlantic bloom happens during spring (Raymont, 1980), however most blooms at PAP occur in late May-early June, as shown in Table 4 and Fig. 5. Later blooms are recorded from satellite-derived chlorophyll-a in 2005, three months later than the average and much later than the ensemble mean and median, although the bloom timing is still within the ensemble range, although the NRR value is 1.31. At L4, also in the North Atlantic, the spring bloom is in April, but most ensemble members show later initiation and peak bloom time, mostly in June. Due to this delay the NRR values at L4 indicate that the ensemble range is too narrow, although still within the full ensemble range. Ensemble mean and median at PAP show good agreement with in situ termination date, and although other station termination times are also within the ensemble range, most are later than the inter-quartile range. However, at ALOHA, located at 22°N, the ensemble median for termination at the end of August agrees with the observations from Racault et al. (2012).

Compared to running only the default MEDUSA, where only a single bloom peak is produced, the ensemble range mostly encompasses the in situ peak amplitudes, shown on Fig. 15(c). Only at BATS are in situ peak height and amplitude outside the ensemble range. This is expected since most of the ensemble members underestimate in situ chlorophyll. At Cariaco, in situ peak heights are within the ensemble range, but observed peaks are higher (mean= 3.5 mg m^{-3}, maximum peak= 7.7 mg m^{-3}), and the ensemble reaches less than half of the in situ peak (mean= 1.2 mg m^{-3}, maximum height= 5 mg m^{-3}). This underestimate of the peak and bloom amplitude, results in NRR of 1.40 and 1.39 respectively. Ensemble members with higher peak and amplitudes are also those with higher chlorophyll biases. Despite the narrow ensemble range, at L4 chlorophyll peaks are within the 75^{th} and 25^{th} range box, and amplitudes are within the full spread. ALOHA and PAP have reliable ensemble spreads according to their NRR values for peak height (see Table 4).

Similar to peak heights, the bloom durations at most stations are within the ensemble range, apart from station Cariaco~~,~~ ~~which shows the narrowest ensemble spread according to its NRR~~. The duration at Cariaco is overestimated because the peak is very wide (up to 143 bloom days). This, along with the late initiation of the bloom, results in a three month late termination. At ALOHA, duration is outside the 75^{th} and 25^{th} quartile box, since the peak is also much broader compared to in situ blooms. This results in too narrow ensemble mean according to the NRR. The opposite is true at BATS where in situ peaks are generally broader, and the ensemble members with lower chlorophyll concentration showing narrower peaks, and a greater range in bloom durations, which consequently lowers the NRR value.

4 Summary and Discussion

In this paper we have investigated structural sensitivity, associated with the mathematical formulation of the processes in an intermediately complex biogeochemical model by generating its ensemble outputs of chlorophyll and ~~nitrogen~~ DIN and comparing them with a single default run, and with in situ observations at five oceanographic stations. The ensemble consists of 128 ensemble members, each with different process function combinations. In order to maintain phenomenological similarity, these functions are calibrated using non-linear least squares, while keeping the maximum process rates fixed, and using the range of concentrations that have been observed in all of the stations. We have chosen nutrient uptake, zooplankton grazing, and plankton mortalities to vary, as these are the core processes of every marine biogeochemical model, from the simplest to the most

complex. Through this approach, we provide a perturbed biology ensemble conditioned upon process structural uncertainties. Applying structural sensitivity in the 1-D framework has also allowed a large ~~parameter space of concurrent variations~~ range of process variability to be explored for several different oceanographic regions, and with minimal computational cost. From these assessments, we find that small perturbations in model structure can produce a wide range of results regarding chlorophyll and nutrient concentration as well as phytoplankton phenology. Compared to parametric sensitivity studies in biogeochemical model, studies about structural sensitivity are rather limited.

Our findings reveal that in all regions, the Holling Type II (G_2) grazing function lowers the chlorophyll concentrations, which ~~have~~ has also been observed by Anderson et al. (2010) at low concentrations. The nutrients respond in the opposite direction with enhanced ~~nitrogen~~ DIN concentrations. This is expected as at low concentrations, using the G_2 function would graze more phytoplankton, as shown on Fig. 1(b). Even though fitting the curve in lower concentration range similar to the oligotrophic regions (0.001-0.5 mmol m^{-3}), higher grazing rate in G_2 is still apparent in lower concentrations ($>$ 0.2 mmol m^{-3}). Pairing G_2 with the linear (ρ_2) mortality of phytoplankton, which constantly removed the phytoplankton regardless of the phytoplankton concentrations, will reduce the chlorophyll concentration even further; but the opposite will happen when G_2 is paired with linear zooplankton mortality. Yool et al. (2011) has similarly shown that using a linear mortality causes the biggest changes in phytoplankton concentrations compared to quadratic and sigmoidal. In contrast, the default phytoplankton (ρ_1) and sigmoidal zooplankton mortality (ξ_4) produce the highest chlorophyll concentrations in all regions, similar to the experiment from Yool et al. (2011). Suppose we use less than half of the current maximum mortality for both ρ_2 and ζ_2, then the deviation in phytoplankton concentrations from the default run would not be as apparent (for example, mean surface chlorophyll obtained from running an ensemble member with $U_1 G_1 \rho_2 \zeta_2$ in station ALOHA using $\mu_{nd}, \mu_d, \mu_{mi}$= 0.04 day^{-1} and μ_{me}= 0.08 day^{-1} is 0.12 mg m^{-3} (default function is 0.11 mg m^{-3}), increases from 0.070 mg m^{-3}). This shows that structural sensitivity to some extent captures the parametric sensitivity as well. However compared to the lower maximum mortality, our current parameter set shows lower error during the fitting process, and in order to be consistent with other functional forms, we decided to use the current parameter set. In terms of nutrient uptake, the exponential (U_3) and sigmoidal (U_2) show inefficient uptake, as they produce low chlorophyll and especially high nitrogen concentrations, as shown on Fig. 12(a), (b), 13(a), and (b), especially in the oligotrophic region. Even though the functional forms have been optimised, the most deviations are observed when nitrogen is $<$ 1 mmol N m^{-3} shown in Fig. 1(a). This deviation still occurred when the concentration range is reduced to 0.001 - 5 mmol m^{-3}. This makes the U_3 and U_2, which uptake less nitrogen in low nitrogen concentrations, produce high ~~nitrogen~~ DIN and low chlorophyll. However, the effect is not as noticeable compared to using G_1 or G_2.

These disparities between the ensemble members that are caused by the structural differences therefore garner some range. Stations that have produced high chlorophyll concentrations also have high ensemble range. For example, in station Cariaco where chlorophyll concentration is high, despite the discrepancy between the in situ seasonal pattern and the ensemble, the range still covers the in situ concentrations. The mismatch between the observation and ensemble is mostly caused by the physical dynamics, in which the upwelling of nutrients that feeds the phytoplankton is not well simulated by the assimilated vertical velocity. This emphasises that despite using the ensemble approach, a ~~biogeochemical~~ coupled-biogeochemical model is only

as good as its physical model (Doney, 1999), as the physical component such as mixing and upwelling dictate the seasonal pattern, phytoplankton community structure, and primary production (Sinha et al., 2010). Even though there are discrepancies, the chlorophyll profile at Cariaco has an NRR value closest to 1, due to the large range in the ensemble. However, in the annual mean of chlorophyll and primary production (Fig. 4(d) and 10(b)) the large range makes the ensemble spread too large. The ensemble range is considerable even if the equation of only one process is changed at a time, which accounts for at least 80% of the full ensemble range. This has been observed at all of the stations, summarised in Table 4 and Fig. 8, emphasising that perturbing functional forms will produce a large range of model results. In some cases, this reduced range may be statistically more meaningful than the full range. For example, compared with the full ensemble, the reduced ensemble range for Cariaco's annual-mean chlorophyll gives an NRR closer to unity. Therefore, it may be possible through a further study to systematically reduce the number of ensemble members, whilst retaining a realistic ensemble range, which will reduce computational costs.

At most of the stations, the ensemble mean produced lower RMSE and higher nitrogen correlation compared to the default run, as shown on Table 3, suggesting that the structural ensemble with a wide range of predictions covering the in situ observations, is likely to produce a mean field closer to the observation, than a single-structure model. However in terms of correlation coefficient and bias in chlorophyll concentration, the default run outperforms the ensemble mean and median, especially in the oligotrophic regions, as shown on Table 3. This may be because using default function, the model produces higher chlorophyll and lower DIN than the ensemble mean and median, and in the oligotrophic regions, the ensemble tend to overestimate DIN and underestimate the chlorophyll. Reducing the number of ensemble members, in a further study, may improve the bias and correlation in ensemble mean and median, as some of the ensemble members contributes to this high bias, especially those which uses ρ_2 and G_2 . ~~Even~~

Nonetheless, in stations such as BATS, where the in situ chlorophyll is underestimated by most ensemble members and the RMSE is higher in ensemble median and mean, the monthly means (Fig. 5(c)) of the in situ values (during months of high chlorophyll) are within the ensemble range. Similar monthly-mean chlorophyll patterns have also been observed in PAP, ALOHA, Cariaco and L4 (with some exceptions in summer month) (Fig. 5(b)), whereby the in situ chlorophyll in generally within the ensemble range. We further note that, considerable model bias such as lower modelled concentrations of chlorophyll, compared to the in situ data, has been observed for the default 3-D MEDUSA model itself, e.g., in the subtropical gyre (Yool et al., 2011). This may be due to the absence of nitrogen fixers and picoplankton in MEDUSA, which cause the increase of plankton concentration in the summer (White et al., 2015), or due to the fact that phytoplankton uptake equation in MEDUSA does not allow phytoplankton to acclimatise in the oligotrophic region through optimum uptake kinetics (Smith et al., 2009; Yool et al., 2011).

Apart from the model's state variables such as chlorophyll and nutrient concentrations, we have looked into the model-derived phytoplankton phenology because of its importance to marine ecosystems e.g. importance of the timing of phytoplankton blooms for the survival of zooplankton and fish larvae (Cushing, 1990). The timing of the blooms has also been shown to control the variability of pCO_2 in the sub-polar region (Bennington et al., 2009). Despite having a reliable spread in the annual mean, in terms of phytoplankton phenology, stations such as L4 show some mismatch with the observation. In situ initiation, bloom timing, and duration in L4 are earlier than in most of the ensemble members, although still lying within the

ensemble range and some ensemble mean timings (termination and peak bloom time) are similar to the satellite observations at this latitude (Racault et al., 2012). When in situ chlorophyll is fitted with a smooth curve, the highest peak mostly occurs during spring (March-April). But model metrics, including ensemble mean and median, are noisy, and peaks mostly fall in the summer (May-July). Moreover, at L4, distinct phytoplankton blooms occur twice a year: first in spring and the second in autumn (Smyth et al., 2010). These blooms are sometimes well simulated, e.g. in Fig. 14(g) and 5(d), but are not as distinct as in situ measurements due to the variability of the model. Some of these discrepancies may also be caused by the way zooplankton select their prey in MEDUSA. In a study by Sailley et al. (2014) grazing selection based on total prey concentration would result in rapid nutrient turn-over, which results in a single peak event, but if the selection is based on the stoichiometry of C:N, the nutrients would regenerate slower, and therefore result in two chlorophyll peaks. However, the difference in peak timing does not affect the duration of the blooms, and the in situ duration is well within the ensemble inter-quartile range. More generally, discrepancies in predicting bloom timing by large-scale biogeochemical models are also reported in many studies, e.g., Henson et al. (2017) and Kostadinov et al. (2017). Henson et al. (2017) shows that compared with the satellite data, the 3-D MEDUSA 2.0 (Yool et al., 2013) model estimates spring bloom start date ∼50 days late, and in southern hemisphere, model estimates subtropical bloom start date ∼50 days earlier. By generating an ensemble of 7 CMIP5 models, Kostadinov et al. (2017) highlighted that the difference in bloom timing between the model ensemble and satellite-derived chlorophyll can be more than one month over most of the ocean. This agrees with our study (see, Table 4), as most of our ensemble members have earlier bloom initiation dates, and the difference between the ensemble mean and in situ timing of bloom, e.g. PAP and L4, are more than one month. Additionally, the whole ensemble range produced by this study can help to provide an uncertainty range for the timing of phytoplankton blooms. The ensemble range almost always encompasses the observed annual mean, peak height, and amplitude. Therefore it may be suitable to use the ensemble model in order to forecast these phenological aspects. Further, it may also be possible to improve the accuracy of the ensemble range, by systematically removing certain ensemble members in a future study.

Finally, the unresolved discrepancy between in situ observations and large number of process ensemble results, such as in the oligotrophic stations for ~~nitrogen~~ DIN and L4 for phytoplankton peak timings, emphasise that the inclusion of some missing processes, such as active prey selection, and species would improve the performance of the model (Friedrichs et al., 2007; Kriest et al., 2010; Sailley et al., 2014) and functional forms which describe chemostat experiments, such as the droop function or the active prey selection (Sailley et al., 2014) are not as structurally sensitive as the logistic equations (Aldebert et al., 2018). Additionally, MEDUSA uses logistic functions such as Monod and Holling type III equations to describe its processes and are well known to be structurally sensitive (Aldebert et al., 2018). We did not include equations that allow such selection or species, as in this paper we tried to ensure that all the equations have similar properties to the default MEDUSA, in order to show that perturbing the structure of the model equations would result in different plankton and nutrient dynamics. Therefore, comparing the performance of model complexity and the ensemble method was beyond the scope of this study.

5 Conclusions

Our study highlights that it is important to conduct structural sensitivity analyses in addition to parameter sensitivity analyses and it is crucial to include mathematical functions that can capture sufficient information of the key biogeochemical processes known from experimental studies. However, none of the deterministic functions can capture all details of these processes (Anderson et al., 2010), therefore we have introduced a method whereby instead of having only one default model output, we have an ensemble generating a range of possible outcomes arising from alternative model structures. We have explored the structural sensitivity of the 1-D version of MEDUSA, the ocean biogeochemistry component of UK-ESM1~~, which is becoming widely used in the community~~. This study emphasises that small perturbations in MEDUSA process equations can produce very different model results, and the ensemble of perturbations generally encompasses the in situ observations. Therefore, our study shows promise that the ensemble of a single biogeochemical model resulting from perturbing the model structure, may produce meaningful prediction ranges of its state variables. However, our study is based on 1-D simulation, and further study with a 3-D biogeochemical model would help extend results to the global ocean. It may also be possible to further minimise the computational costs by systematically reducing the number of ensemble members whilst retaining a realistic ensemble range. Further studies could include varying the weighting of ensemble members, or reducing the number of model combinations to improve the ensemble range and to assess properly different plankton functional types and dissolved inorganic carbon. Such a perturbed biology ensemble may also be useful for data assimilation e.g. with satellite-derived chlorophyll.

Data availability. The raw model outputs will be available at Pangaea after the manuscript has been published and upon request from the authors (p.anugerahanti@pgr.reading.ac.uk, shovonlal.roy@reading.ac.uk)

Appendix A: Determining phytoplankton phenology

Before determining the initiation time, bloom timing must be identified. This is done by taking the ten years of surface chlorophyll output and breaking it down into individual years. These are then rearranged into two datasets: January-December and June-May, and the date of maximum chlorophyll concentration in each year is determined. If the peak timing occurs mostly towards the end or the beginning of the year, June to May datasets are used instead of the former. The timing is then adjusted if the calendar year has changed.

The initiation is determined by the day that chlorophyll concentration exceeds a given threshold. However, since in situ chlorophyll has some data gaps and modelled chlorophyll is not smooth, some studies have fitted a function or model to the datasets to make the chlorophyll data smoother (Platt et al., 2009; Sapiano et al., 2012; Brody et al., 2013). Here we use a 5th order polynomial curve to get a smooth fit of the bloom peaks in the data (Fig. A1), from which phenology metrics are calculated. After being fitted, a threshold of half the bloom peak concentration is chosen. To find the **peak time**, the date at which maximum chlorophyll concentration is achieved in the fitted curve is determined, and this date is used as a reference to calculate other metrics. **Amplitude** is then calculated as half of the highest peak minus the minimum concentration. **Initiation** is

the day when chlorophyll concentration goes just above the threshold towards the maximum (Brody et al., 2013). **Termination** of the bloom is defined when concentration falls below the threshold (Racault et al., 2012). If two peaks are detected the termination of the spring bloom is determined when the first bloom reduces to its minimum, just before the second bloom starts (in the first valley). **Duration** of the bloom is simply the total number of days on which chlorophyll concentration is above the threshold or termination minus initiation.

This phenology is useful to see how the bloom develops and terminates, whether the concentration increases rapidly and decreases slowly or vice versa. The phenology is summarised in Fig. A1. The curve fitting method is only applied if the data shows potential outliers especially in higher concentrations. If there is only one prominent bloom each year, as at stations ALOHA and BATS, and the data is smooth, the regular threshold method (when the concentration is above 50% of the maximum bloom, and the associated initiation and termination times), without fitting the data with a curve is applied. To avoid results being affected by how bloom phenology is determined, the same method is used for determining the metrics from both in situ and model output.

Competing interests. The authors declare that they have no conflict of interest

Acknowledgements. This study was funded by the Bakrie Center Foundation (grant no. 1307/BCF-SK/RSCH/VII/2015), Indonesia. The authors would like to thank Kevin White, for his advise on this study and John Hemmings, for providing the most recent version of the MarMOT code. We would like to thank Ruth Airs and Denise Cummings for providing the L4 chlorophyll and nutrient measurements. This study uses various oceanographic station data and we would like to thank all crews and scientists involved in collecting, processing, and making the data publicly available. Three anonymous reviewers are acknowledged for their valuable comments and suggestions.

References

Adamson, M. W. and Morozov, A. Y.: When can we trust our model predictions? Unearthing structural sensitivity in biological systems, Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences, 469, 20120500, https://doi.org/10.1098/rspa.2012.0500, 2013.

Aldebert, C., Nerini, D., Gauduchon, M., and Poggiale, J. C.: Does structural sensitivity alter complexity–stability relationships?, Ecological Complexity, 28, 104–112, https://doi.org/10.1016/j.ecocom.2016.07.004,2016.

Aldebert, C., Kooi, B. W., Nerini, D., and Poggiale, J. C.: Is structural sensitivity a problem of oversimplified biological models? Insights from nested Dynamic Energy Budget models, Journal of Theoretical Biology, 448, 1–8, https://doi.org/10.1016/j.jtbi.2018.03.019,2018.

Anderson, J. L.: An Ensemble Adjustment Kalman Filter for Data Assimilation, Monthly Weather Review, 129, 2884–2903, https://doi.org/10.1175/1520-0493(2001)129<2884:AEAKFF>2.0.CO;2, 2001.

Anderson, T. R.: Plankton functional type modelling: Running before we can walk?, Journal of Plankton Research, 27, 1073–1081, https://doi.org/10.1093/plankt/fbi076, 2005.

Anderson, T. R. and Mitra, A.: Dysfunctionality in ecosystem models: An underrated pitfall?, Progress in Oceanography, 84, 66–68, 2010.

Anderson, T. R., Gentleman, W. C., and Sinha, B.: Influence of grazing formulations on the emergent properties of a complex ecosystem model in a global ocean general circulation model, Progress in Oceanography, 87, 201–213, https://doi.org/10.1016/j.pocean.2010.06.003, 2010.

Anderson, T. R., Gentleman, W. C., and Yool, A.: EMPOWER-1.0: An Efficient Model of Planktonic ecOsystems WrittEn in R, Geoscientific Model Development, 8, 2231–2262, https://doi.org/10.5194/gmd-8-2231-2015, 2015.

Bennington, V., Mckinley, G. A., Dutkiewicz, S., and Ullman, D.: What does chlorophyll variability tell us about export and air-sea CO_2 flux variability in the North Atlantic?, 23, 1–11, 2009.

Berelson, W.: Particle settling rates increase with depth in the ocean, Deep-Sea Research II, 49, 237–251, 2002.

Bopp, L., Aumont, O., Cadule, P., Alvain, S., and Gehlen, M.: Response of diatoms distribution to global warming and potential implications: A global model study, Geophysical Research Letters, 32, 1–4, https://doi.org/10.1029/2005GL023653, 2005.

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, 6225–6245, https://doi.org/10.5194/bg-10-6225-2013, 2013.

Brody, S. R., Lozier, M. S., and Dunne, J. P.: A comparison of methods to determine phytoplankton bloom initiation, Journal of Geophysical Research: Oceans, 118, 2345–2357, https://doi.org/10.1002/jgrc.20167, 2013.

Butenschön, M., Clark, J., Aldridge, J. N., Icarus Allen, J., Artioli, Y., Blackford, J., Bruggeman, J., Cazenave, P., Ciavatta, S., Kay, S., Lessin, G., Van Leeuwen, S., Van Der Molen, J., De Mora, L., Polimene, L., Sailley, S., Stephens, N., and Torres, R.: ERSEM 15.06: A generic model for marine biogeochemistry and the ecosystem dynamics of the lower trophic levels, Geoscientific Model Development, 9, 1293–1339, https://doi.org/10.5194/gmd-9-1293-2016, 2016.

Cox, P. M. and Kwiatkowski, L.: Assessment of the iMarNet Ocean Biogeochemical Models, Tech. rep., University of Exeter, Exeter, http://imarnet.org/pdf/iMarNet_Model_Evaluation_-_CoxKwiatkowski, 2013.

Cushing, D.: Plankton production and year-class strength in fish populations - an update of the match mismatch hypothesis., Advances in Marine Biology, 26, 249–293, http://www.sciencedirect.com/science/article/pii/S0065288108602023, 1990.

Dave, A. C. and Lozier, M. S.: Local stratification control of marine productivity in the subtropical North Pacific, Journal of Geophysical Research: Oceans, 115, 1–16, https://doi.org/10.1029/2010JC006507, 2010.

Doney, C.: Major challenges confronting marine biogeochemical modeling, Global Biogeochemical Cycles, 13, 705–714, 1999.

Doney, S. C.: The growing human footprint on the planet, Science, 328, 1512–1516, https://doi.org/10.1126/science.1185198, 2010.

Doney, S. C., Ruckelshaus, M., Emmett Duffy, J., Barry, J. P., Chan, F., English, C. A., Galindo, H. M., Grebmeier, J. M., Hollowed, A. B., Knowlton, N., Polovina, J., Rabalais, N. N., Sydeman, W. J., and Talley, L. D.: Climate Change Impacts on Marine Ecosystems, Annual Review of Marine Science, 4, 11–37, https://doi.org/10.1146/annurev-marine-041911-111611, 2012.

Dore, J. E., Letelier, R. M., Church, M. J., Lukas, R., and Karl, D. M.: Summer phytoplankton blooms in the oligotrophic North Pacific Subtropical Gyre: Historical perspective and recent observations, Progress in Oceanography, 76, 2–38, https://doi.org/10.1016/j.pocean.2007.10.002, 2008.

Dutkiewicz, S., Follows, M. J., and Parekh, P.: Interactions of the iron and phosphorus cycles: A three-dimensional model study, Global Biogeochemical Cycles, 19, 1–22, https://doi.org/10.1029/2004GB002342, 2005.

Edwards, A. and Yool, A.: The role of higher predation in plankton population models, Journal of Plankton Research, 22, 1085–1112, https://doi.org/10.1093/plankt/22.6.1085, 2000.

Englund, G. and Leonardsson, K.: Scaling up the functional response for spatially heterogeneous systems, Ecology Letters, 11, 440–449, https://doi.org/10.1111/j.1461-0248.2008.01159.x, 2008.

Eppley, R. W.: Temperature and phytoplankton growth in the sea, Fishery bulletin, 70, 1063–1085, https://doi.org/163346, 1972.

Fasham, M. J. R.: Variations in the seasonal cycle of biological production in subarctic oceans: A model sensitivity analysis, Deep-Sea Research Part I, 42, 1111–1149, https://doi.org/10.1016/0967-0637(95)00054-A, 1995.

Fasham, M. J. R., Ducklow, H. W., and McKelvie, S. M.: A nitrogen-based model of plankton dynamics in the ocean mixed layer, Journal of Marine Research, 48, 591–639, https://doi.org/10.1357/002224090784984678, 1990.

Fasham, M. J. R., Sarmiento, J. L., Slater, R. D., Ducklow, H. W., and Williams, R.: Ecosystem behaviour at Bermuda station"S" and ocean weather station "India": A general circulation moel and observational anlysis, Global Biogeochemical Cycles, 7, 379–415, 1993.

Fennel, K. and Boss, E.: Subsurface maxima of phytoplankton and chlorophyll: Steady-state solutions from a simple model, Limnology and Oceanography, 48, 1521–1534, https://doi.org/10.4319/lo.2003.48.4.1521, 2003.

Fennel, W. and Neumann, T.: Introduction to the Modelling of Marine Ecosystem, Elsevier Science, second edn., 2014.

Flora, C., David, N., Mathias, G., Andrew, M., and Jean-Christophe Poggiale, P.: Structural sensitivity of biological models revisited, Journal of Theoretical Biology, 283, 82–91, https://doi.org/10.1016/j.jtbi.2011.05.021, 2011.

Flynn, K. J. and Mitra, A.: Why Plankton Modelers Should Reconsider Using Rectangular Hyperbolic (Michaelis-Menten, Monod) Descriptions of Predator-Prey Interactions, Frontiers in Marine Science, 3, 165, 1–17 https://doi.org/10.3389/fmars.2016.00165, 2016.

Ford, D. A., Edwards, K. P., Lea, D., Barciela, R. M., Martin, M. J., and Demaria, J.: Assimilating GlobColour ocean colour data into a pre-operational physical-biogeochemical model, Ocean Science, 8, 751–771, https://doi.org/10.5194/os-8-751-2012, 2012.

Franks, P. J. S.: NPZ models of plankton dynamics: Their construction, coupling to physics, and application, Journal of Oceanography, 58, 379–387, https://doi.org/10.1023/A:1015874028196, 2002.

Friedrichs, M. A., 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., and Wiggert, J. D.: Assessment of skill and portability in regional marine biogeochemical models: Role of multiple planktonic groups, Journal of Geophysical Research: Oceans, 112, 1–22, https://doi.org/10.1029/2006JC003852, 2007.

Friedrichs, M. A. M., Hood, R. R., and Wiggert, J. D.: Ecosystem model complexity versus physical forcing: Quantification of their relative impact with assimilated Arabian Sea data, Deep-Sea Research Part II: Topical Studies in Oceanography, 53, 576–600, https://doi.org/10.1016/j.dsr2.2006.01.026, 2006.

Fussmann, G. F. and Blasius, B.: Community response to enrichment is highly sensitive to model structure., Biology letters, 1, 9–12, https://doi.org/10.1098/rsbl.2004.0246, 2005.

Gehlen, M., Barciela, R., Bertino, L., Brasseur, P., Butenschön, M., Chai, F., Crise, A., Drillet, Y., Ford, D., Lavoie, D., Lehodey, P., Perruche, C., Samuelsen, A., and Simon, E.: Building the capacity for forecasting marine biogeochemistry and ecosystems: recent advances and future developments, Journal of Operational Oceanography, 8, s168–s187, https://doi.org/10.1080/1755876X.2015.1022350, 2015.

Hawaii Ocean Time Series: Analytical methods and results, http://hahana.soest.hawaii.edu/hot/methods/inuts.html.

Hemmings, J. C. P. and Challenor, P. G.: Addressing the impact of environmental uncertainty in plankton model calibration with a dedicated software system: The marine model optimization testbed (MarMOT 1.1 alpha), Geoscientific Model Development, 5, 471–498, https://doi.org/10.5194/gmd-5-471-2012, 2012.

Hemmings, J. C. P., Challenor, P. G., and Yool, A.: Mechanistic site-based emulation of a global ocean biogeochemical model (MEDUSA 1.0) for parametric analysis and calibration: An application of the Marine Model Optimization Testbed (MarMOT 1.1), Geoscientific Model Development, 8, 697–731, https://doi.org/10.5194/gmd-8-697-2015, 2015.

Henson, S. A., Yool, A., Cole, H. S., Hopkins, J., and Martin, A. P.: Detection of climate change-driven trends in phytoplankton phenology, https://doi.org/10.1111/gcb.13886, Global Change Biology, 00, 1–11, 2017.

Ivlev, V.: Experimental Ecology of The Feeding of The Fishes, Yale University Press, New Haven, CT, 1961.

Jassby, A. D. and Platt, T.: Mathematical Formulation of the Relationship Between Photosynthesis and Light for Phytoplankton, Limnology and Oceanography, 21, 540–547, https://doi.org/10.4319/lo.1976.21.4.0540, 1976.

Kane, A., Moulin, C., Thiria, S., Bopp, L., Berrada, M., Tagliabue, A., Crépon, M., Aumont, O., and Badran, F.: Improving the parameters of a global ocean biogeochemical model via variational assimilation of in situ data at five time series stations, Journal of Geophysical Research: Oceans, 116, 1–14, https://doi.org/10.1029/2009JC006005, 2011.

Kostadinov, T. S., Cabré, A., Vedantham, H., Marinov, I., Bracher, A., Brewin, R. J., Bricaud, A., Hirata, T., Hirawake, T., Hardman-Mountford, N. J., Mouw, C., Roy, S., and Uitz, J.: Inter-comparison of phytoplankton functional type phenology metrics derived from ocean color algorithms and Earth System Models, Remote Sensing of Environment, 190, 162–177, https://doi.org/10.1016/j.rse.2016.11.014, 2017.

Kriest, I., Khatiwala, S., and Oschlies, A.: Towards an assessment of simple global marine biogeochemical models of different complexity, Progress in Oceanography, 86, 337–360, https://doi.org/10.1016/j.pocean.2010.05.002, 2010.

Kwiatkowski, L., Yool, A., Allen, J. I., Anderson, T. R., Barciela, R., Buitenhuis, E. T., Butenschön, M., Enright, C., Halloran, P. R., Le Quèrè, C., De Mora, L., Racault, M. F., Sinha, B., Totterdell, I. J., and Cox, P. M.: IMarNet: An ocean biogeochemistry model intercomparison project within a common physical ocean modelling framework, Biogeosciences, 11, 7291–7304, https://doi.org/10.5194/bg-11-7291-2014, 2014.

Lacroix, G. and Gregoire, M.: Revisited ecosystem model (MODECOGeL) of the Ligurian Sea: seasonal and interannual variability due to atmospheric forcing, Journal of Marine Systems, 37, 229–258, https://doi.org/http://dx.doi.org/10.1016/S0924-7963(02)00190-2, 2002.

Le Quèrè, C., Harrison, S. P., Prentice, I. C., Buitenhuis, E. T., Aumont, O., Bopp, L., Claustre, H., Leticia Cotrim Da Cunha, R. G., 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, Global Change in Biology, 11, 2016–2040, https://doi.org/10.1111/j.1365-2486.2005.01004.x, 2005.

Letelier, R. M., Karl, D. M., Abbott, M. R., and Bidigare, R. R.: Light driven seasonal patterns of chlorophyll and nitrate in the lower euphotic zone of the North Pacific Subtropical Gyre, Limnol. Oceanogr, 49, 508–519, https://doi.org/10.4319/lo.2004.49.2.0508, 2004.

Levin, S. and Lubchenco, J.: Resilience, robustness, and marine ecosystem-based management, BioScience, 58, 27–32, https://doi.org/10.1641/B580107, 2008.

Lipschultz, F.: A time-series assessment of the nitrogen cycle at BATS, 48, 1897–1924, 2001.

Mahowald, N. M., Engelstaedter, S., Luo, C., Sealy, A., Artaxo, P., Benitez-Nelson, C., Bonnet, S., Chen, Y., Chuang, P. Y., Cohen, D. D., Dulac, F., Herut, B., Johansen, A. M., Kubilay, N., Losno, R., Maenhaut, W., Paytan, A., Prospero, J. M., Shank, L. M., and Siefert, R. L.: Atmospheric iron deposition: global distribution, variability, and human perturbations., Annual review of marine science, 245–278, https://doi.org/10.1146/annurev.marine.010908.163727, 2009.

McDonald, C., Bennington, V., Urban, N., and McKinley, G.: 1-D test-bed calibration of a 3-D Lake Superior biogeochemical model, Ecological Modelling, 225, 115–126, https://doi.org/10.1016/j.ecolmodel.2011.11.021, 2012.

Moradkhani, H. and Meskele, T. T.: Probabilistic Assessment of the Satellite Retrieval Error Translation to Hydrologic Response, in: Satellite Rainfall Applications for Surface Hydrology, edited by Gebremichael, M. and Hossain, F., chap. II, pp. 235–241, Springer, Netherlands, 2010.

Moradkhani, H., Hsu, K., Hong, Y., and Sorooshian, S.: Investigating the impact of remotely sensed precipitation and hydrologic model uncertainties on the ensemble streamflow forecasting, Geophysical Research Letters, 33, 1–5, https://doi.org/10.1029/2006GL026855, 2006.

Murphy, J. M., Booth, B. B. B., Collins, M., Harris, G. R., Sexton, D. M. H., and Webb, M. J.: A methodology for probabilistic predictions of regional climate change from perturbed physics ensembles., Philosophical transactions. Series A, Mathematical, physical, and engineering sciences, 365, 1993–2028, https://doi.org/10.1098/rsta.2007.2077, 2007.

Oschlies, A. and Garçon, V.: An eddy-permitting coupled physical-biological model of the North Atlantic. 1. Sensitivity to advection numerics and mixed layer physics, Global Biogeochemical Cycles, 13, 135–160, https://doi.org/10.1029/98GB02811, 1999.

Oschlies, A. and Schartau, M.: Basin-scale performance of a locally optimized marine ecosystem model, Journal of Marine Systems, 63, 335–358, 2005.

Ourmières, Y., Brasseur, P., Lévy, M., Brankart, J.-m., and Verron, J.: On the key role of nutrient data to constrain a coupled physical– biogeochemical assimilative model of the North Atlantic Ocean, Journal of Marine Systems, 75, 100–115, https://doi.org/10.1016/j.jmarsys.2008.08.003, 2009.

Palmer, J. and Totterdell, I.: Production and export in a global ecosystem model, Deep-Sea Research I, 48, 1169–1198, https://doi.org/10.1016/S0967-0637(00)00080-7, 2001.

Parekh, P., Follows, M. J., and Boyle, E. A.: Decoupling of iron and phosphate in the global ocean, Global Biogeochemical Cycles, 19, 1–16, https://doi.org/10.1029/2004GB002280, 2005.

Platt, T. and Sathyendranath, S.: Ecological indicators for the pelagic zone of the ocean from remote sensing, Remote Sensing of Environment, 112, 3426–3436, https://doi.org/10.1016/j.rse.2007.10.016, 2008.

Platt, T., White, G. N., Zhai, L., Sathyendranath, S., and Roy, S.: The phenology of phytoplankton blooms: Ecosystem indicators from remote sensing, Ecological Modelling, 220, 3057–3069, https://doi.org/10.1016/j.ecolmodel.2008.11.022, 2009.

Racault, M.-F., Le Quèrè, C., Buitenhuis, E., Sathyendranath, S., and Platt, T.: Phytoplankton phenology in the global ocean, Ecological Indicators, 14, 152–163, https://doi.org/10.1016/j.ecolind.2011.07.010, 2012.

Raick, C., Soetaert, K., and Grégoire, M.: Model complexity and performance: How far can we simplify?, Progress in Oceanography, 70, 27–57, https://doi.org/10.1016/j.pocean.2006.03.001, 2006.

Raymont, J.: Plankton and Productivity in the Oceans, Pergamon Press, London, 1980.

Robinson, C. L. K., Ware, D. M., and Parsons, T. R.: Simulated annual plankton production in the northeastern Pacific Coastal upwelling Domain, Journal of Plankton Research, 15, 161–183, https://doi.org/10.1093/plankt/15.2.161, 1993.

Roy, S. and Chattopadhyay, J.: Enrichment and stability: A phenomenological coupling of energy value and carrying capacity, BioSystems, 90, 371–378, https://doi.org/10.1016/j.biosystems.2006.10.001, 2007.

Roy, S., Broomhead, D. S., Platt, T., Sathyendranath, S., and Ciavatta, S.: Sequential variations of phytoplankton growth and mortality in an NPZ model: A remote-sensing-based assessment, Journal of Marine Systems, 92, 16–29, https://doi.org/10.1016/j.jmarsys.2011.10.001, 2012.

Ryabchenko, V. A., Fasham, M. J. R., Kagan, B. A., and Popova, E. E.: What causes short-term oscillations in ecosystem models of the ocean mixed layer?, Journal of Marine Systems, 13, 33–50, https://doi.org/10.1016/S0924-7963(96)00110-8, 1997.

Sailley, S. F., Polimene, L., Mitra, A., Atkinson, A., and Allen, J. I.: Impact of zooplankton food selectivity on plankton dynamics and nutrient cycling, Journal of Plankton Research, 37, 519–529, https://doi.org/10.1093/plankt/fbv020, 2014.

Sapiano, M. R. P., Brown, C. W., Schollaert Uz, S., and Vargas, M.: Establishing a global climatology of marine phytoplankton phenological characteristics, Journal of Geophysical Research: Oceans, 117, 1–16, https://doi.org/10.1029/2012JC007958, 2012.

Sinha, B., Buitenhuis, E. T., Le Quere, C., and Anderson, T. R.: Progress in Oceanography Comparison of the emergent behavior of a complex ecosystem model in two ocean general circulation models, Progress in Oceanography, 84, 204–224, https://doi.org/10.1016/j.pocean.2009.10.003, 2010.

Smith, S., Yamanaka, Y., Pahlow, M., and Oschlies, A.: Optimal uptake kinetics: physiological acclimation explains the pattern of nitrate uptake by phytoplankton in the ocean, Marine Ecology Progress Series, 384, 1–12, https://doi.org/10.3354/meps08022, 2009.

Smyth, T. J., Fishwick, J. R., Al-moosawi, L., Cummings, D. G., and Harris, C.: A broad spatio-temporal view of the Western English Channel observatory, 32, 585–601, https://doi.org/10.1093/plankt/fbp128, 2010.

Steele, J. H. and Henderson, E. W.: A Simple Plankton Model, The American Naturalist, 117, 676–691, 1981.

Storkey, D., Blockley, E., Furner, R., Guiavarc'h, C., Lea, D., Martin, M., Barciela, R. M., Hines, A., Hyder, P., and Siddorn, J.: Forecasting the ocean state using NEMO:The new FOAM system, https://doi.org/10.1080/1755876X.2010.11020109, 2010.

Subramanian, A. C. and Palmer, T. N.: Ensemble superparameterization versus stochastic parameterization: A comparison of model uncertainty representation in tropical weather prediction, Journal of Advances in Modeling Earth Systems, 9, 1231–1250, https://doi.org/10.1002/2016MS000857, 2017.

Taylor, G. T., Muller-karger, F. E., Thunell, R. C., Scranton, M. I., Astor, Y., and Varela, R.: Ecosystem responses in the southern Caribbean Sea to global climate change, 109, 19 315–19 320, https://doi.org/10.1073/pnas.1207514109, 2012.

Tinker, J., Lowe, J., Pardaens, A., Holt, J., and Barciela, R.: Uncertainty in climate projections for the 21st century northwest European shelf seas, Progress in Oceanography, 148, 56–73, https://doi.org/10.1016/j.pocean.2016.09.003, 2016.

Villareal, T. A., Brown, C. G., Brzezinski, M. A., Krause, J. W., and Wilson, C.: Summer Diatom Blooms in the North Pacific Subtropical Gyre : 2008 – 2009, 7, 4, e33109, https://doi.org/10.1371/journal.pone.0033109, 2012.

Ward, B. A., Schartau, M., Oschlies, A., Martin, A. P., Follows, M. J., and Anderson, T. R.: When is a biogeochemical model too complex? Objective model reduction and selection for North Atlantic time-series sites, Progress in Oceanography, 116, 49–65, https://doi.org/10.1016/j.pocean.2013.06.002, 2013.

White, A., Letelier, R. M., Whitmire, A. L., Barone, B., Bidigare, R. R., Church, M. J., and Karl, D. M.: Phenology of particle size distributions and primary productivity in the North Pacific subtropical gyre (Station ALOHA), Journal of Geophysical Research Oceans, 120, 7381–7399, https://doi.org/10.1002/2015JC010897, 2015.

Wood, S. and Thomas, M.: Super-sensitivity to structure in biological models, The Royal Society, 266, 565–570, https://doi.org/10.1098/rspb.1999.0673, 1999.

Yool, A., Popova, E. E., and Anderson, T. R.: MEDUSA-1.0: A new intermediate complexity plankton ecosystem model for the global domain, Geoscientific Model Development, 4, 381–417, https://doi.org/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, Geoscientific Model Development, 6, 1767–1811, https://doi.org/10.5194/gmd-6-1767-2013, 2013.

Tables

Table 1. Parameter values for resource uptake (U), zooplankton grazing (G), and plankton mortalities (ρ and ξ for phytoplankton and zooplankton respectively), described using similar functional forms (shown in Fig. 1). In grazing equation, g_m represents maximum grazing rate, P_a is the prey, and p_a denotes the grazing preference. Starred equations are the default functional responses in MEDUSA.

Process/ Plankton type	Symbol	Meaning	Parameter value (mmol m^{-3})			
Nutrient Uptake (U)			Monod* (U_1) $\frac{n}{n+k}$	Sigmoidal (U_2) $\frac{n^2}{n^2+k^2}$	Exponential (U_3) $1-\exp(-\frac{n}{k})$	Trigonometric (U_4) $\frac{2}{\pi}\arctan\left(\frac{n}{k}\right)$
Non-diatom	kN_{nd}	shape defining constant for nitrogen	0.5	0.74	1.12	0.60
	kFe_{nd}	shape defining constant for iron	0.33 $\times 10^{-3}$	0.49 $\times 10^{-3}$	0.74 $\times 10^{-3}$	0.40 $\times 10^{-3}$
Diatom	kN_d	shape defining constant for nitrogen	0.75	1.12	1.68	0.91
	kSi_d	shape defining constant for silicon	0.75	1.12	1.68	0.91
	kFe_d	shape defining constant for iron	0.67 $\times 10^{-3}$	0.99 $\times 10^{-3}$	1.50 $\times 10^{-3}$	0.81 $\times 10^{-3}$

Grazing (G)			Holling type III* (G_1) $g_m \frac{p_a Pa^2}{k_g^2+p_a Pa^2+p_b Pb^2}$	Holling type II (G_2) $g_m \frac{p_a Pa^2}{k_g(p_a Pa+p_b Pb)+p_a Pa^2+p_b Pb^2}$
Microzooplankton	k_{mi}	half saturation constant	0.80	0.46
	pmi_{nd}	grazing preference for non-diatom	0.75	0.75
	pmi_{det}	grazing preference for detritus	0.25	0.25
Mesozooplankton	k_{me}	half saturation constant	0.30	0.17
	pme_{nd}	grazing preference for non-diatom	0.15	0.15
	pme_{det}	grazing preference for detritus	0.15	0.15
	pme_d	grazing preference for diatoms	0.35	0.35
	pme_{mi}	grazing preference for microzooplankton	0.35	0.35

Mortality (ρ,ξ)			Hyperbolic* (ρ_1,ξ_1) $\mu \frac{P}{P+k_M} P$	Linear (ρ_2,ξ_2) μP	Quadratic (ρ_3,ξ_3) μP^2	Sigmoidal (ρ_4,ξ_4) $\mu \frac{P^2}{P^2+k_M^2} P$
Non-diatom	μ_{nd}	maximum rate (day^{-1})	0.10	0.10	0.05	0.10
	k_{Mnd}	half saturation constant	0.50	-	-	0.74
Diatom	μ_d	maximum rate (day^{-1})	0.10	0.10	0.05	0.1
	k_{Md}	half saturation constant	0.50	-	-	0.74
Microzooplankton	μ_{mi}	maximum rate (day^{-1})	0.10	0.10	0.05	0.10
	k_{Mmi}	half saturation constant	0.50	-	-	0.74
Mesozooplankton	μ_{me}	maximum rate (day^{-1})	0.20	0.20	0.07	0.20
	k_{Mme}	half saturation constant	0.75	-	-	1.12

Table 2. Location, data source, and available depth range for the five oceanographic stations

Station	Location	Source	depth range
ALOHA	22°45'N, 158°00'W	http://hahana.soest.hawaii.edu/hot/hot-dogs/interface.html	5-5000 m
BATS	32°50'N, 64°10'W	http://bats.bios.edu/	4-4000 m
Cariaco	10°30'N, 64°40'W	http://imars.marine.usf.edu/cariaco	1-1310 m
L4	50°15'N, 4°12.3'W	http://www.westernchannelobservatory.org.uk/data.php (available upon request)	surface
PAP	49°N, 16.5°W	http://projects.noc.ac.uk/pap/data	7-400m

Table 3. Error statistics, 10-year mean, and NRR of chlorophyll (mg m^{-3}) and ~~nitrogen~~ DIN (mmol m^{-3}) concentration at five stations for the default run, ensemble mean, ensemble median, and the ensemble range (ensemble maximum - ensemble minimum). These are calculated from surface to 200 m depth, starting from January 1998 to December 2007. Bias is (model output) – (in situ observation). Bold text indicate the smallest RMSE. At Station L4 error statistics and mean are taken from the surface and starts from January 1999 for chlorophyll and June 2000 for ~~nitrogen~~DIN. For station PAP, error statistics are taken from 2002-2004 since in situ data is only available during that time.

		DIN profile				Chlorophyll profile				Surface chlorophyll				Integrated chlorophyll			
Stations		r	RMSE	Bias	Mean	r	RMSE	Bias	Mean	r	RMSE	Bias	Mean	r	RMSE	Bias	Mean
PAP	Ens mean	0.23	3.26	0.61	6.59	0.42	0.32	0.06	0.48	0.45	0.51	0.22	0.66				
		(±0.07)	(±2.57)	(±5.13)	(±5.24)	(±0.37)	(±0.73)	(±0.68)	(±0.75)	(±0.38)	(±0.73)	(±0.68)	(±0.76)				
	Ens median	0.23	**3.16**	0.54	6.38	0.49	**0.29**	0.003	0.42	0.54	**0.46**	0.15	0.60				
	Default run	0.21	3.32	-0.20	5.64	0.28	0.40	0.18	0.59	0.36	0.57	0.30	0.74				
	In situ				5.83				0.42				0.44				
	NRR		1.25				1.20				1.29						
ALOHA	Ens mean	0.77	**1.06**	0.67	1.20	0.22	**0.10**	-0.06	0.06	0.22	**0.05**	-0.01	0.10	0.69	**2.73**	-0.72	3.80
		(±0.03)	(±0.19)	(±0.39)	(±0.39)	(±0.49)	(±0.04)	(±0.11)	(±0.11)	(±0.47)	(±0.09)	(±0.13)	(±0.14)	(±0.60)	(±5.49)	(±7.09)	(±10)
	Ens median	0.77	**1.06**	0.68	1.18	0.14	0.11	-0.07	0.05	0.13	**0.05**	-0.01	0.07	0.56	3.3	-1.17	3.34
	Default run	0.77	1.09	0.61	1.10	0.28	**0.10**	-0.03	0.09	0.27	0.07	0.03	0.11	0.70	4.71	1.25	5.77
	In situ				0.50				0.12				0.08				4.52
	NRR		1.39				1.29				1.07				1.01		
BATS	Ens mean	0.56	1.39	1.16	1.77	0.19	0.33	-0.12	0.05	0.22	0.33	-0.12	0.05	0.39	52.13	-19.39	6.18
		(±0.38)	(±0.84)	(±1.00)	(±1.01)	(±0.37)	(±0.05)	(±0.16)	(±0.16)	(±0.58)	(±0.15)	(±0.05)	(±0.15)	(±0.54)	(±9.40)	(±21)	(±14)
	Ens median	0.55	1.39	1.16	1.77	0.11	0.33	-0.12	0.05	0.06	0.34	-0.12	0.05	0.27	**23.30**	-17.71	4.51
	Default run	0.58	**0.73**	0.62	1.35	0.23	**0.31**	-0.07	0.10	0.28	**0.31**	-0.07	0.09	0.43	48.58	-10.77	13.14
	In situ				0.98				0.17				0.15				23.90
	NRR		1.38				1.39				1.40				1.40		
Cariaco	Ens mean	0.78	**2.97**	0.61	5.39	0.29	**0.83**	-0.02	0.49	0.13	**1.23**	0.02	0.77	0.41	**17.73**	-1.05	11.47
		(±0.08)	(±0.49)	(±2.54)	(±2.54)	(±0.34)	(±0.42)	(±0.93)	(±0.93)	(±0.22)	(±0.33)	(±1.90)	(±0.57)	(±0.40)	(±7.90)	(±17)	(±17)
	Ens median	0.76	3.24	0.51	5.29	0.20	0.88	-0.18	0.32	0.072	1.29	-0.29	0.46	0.29	19.46	-5.51	7.00
	Default run	0.76	3.29	0.59	5.37	0.22	0.87	-0.09	0.42	0.11	1.27	-0.18	0.57	0.34	18.71	-3.86	8.65
	In situ				4.78				0.51				0.76				12.52
	NRR		1.25				1.19				1.21				1.17		
L4	Ens mean	0.70	2.94	1.56	4.52					0.25	1.05	0.42	1.76				
		(±0.14)	(±2.13)	(±4.06)	(±4.06)					(±0.33)	(±1.67)	(±2.61)	(±2.61)				
	Ens median	0.68	3.10	1.73	4.69					0.21	**1.02**	0.27	1.61				
	Default run	0.52	**2.67**	1.12	4.08					0.31	1.13	0.83	2.17				
	In situ				2.96								1.34				
	NRR		1.31								1.21						

Table 4. Surface annual mean and phytoplankton phenology from in situ, ensemble mean, median, and default run. The range and NRR in the bracket are the values for changing the functional form one process at a time (shown on Fig. 8).

Stations		Annual Mean (mg m^{-3})	Initiation Time	Bloom (mg m^{-3})	Peak Height (mg m^{-3})	Amplitude (mg m^{-3})	Duration	Termination
PAP	Ens mean	0.61	01 Apr	07 May	2.07	0.96	95	26 Jul
	Range	$\pm 0.70(0.58)$	± 51	± 45	± 2.98	± 1.63	± 99	± 124
	NRR	1.26 (1.37)	1.14	1.31	1.08	1.09	1.42	1.60
	Ens med	0.55	12 Apr	15 May	2.03	0.95	87	22 Jul
	Default run	0.71	03 Apr	05 May	2.1	0.96	99	21 Aug
	In situ	0.44	20 Apr	03 Jun	1.52	0.44	95	24 Jul
ALOHA	Ens mean	0.07	21 Mar	21 Apr	0.14	0.047	62	15 Aug
	Range	$\pm 0.13(0.11)$	± 89	± 119	± 0.28	± 0.11	± 95	± 119
	NRR	0.84 (1.17)	1.35	1.29	0.97	1.19	1.56	1.28
	Ens med	0.063	26 Mar	02 May	0.14	0.05	85	24 Aug
	Default run	0.10	14 Mar	18 Apr	0.25	0.096	66	10 Aug
	In situ	0.084	08 May	26 May	0.14	0.048	47	23 Jun
BATS	Ens mean	0.047	02 Mar	12 Apr	0.1	0.043	89	06 Jul
	Range	$\pm 0.14(0.11)$	± 187	± 174	± 0.42	± 0.19	± 116	± 198
	NRR	1.40 (1.39)	1.18	1.17	1.42	1.42	1.08	1.20
	Ens med	0.038	28 Feb	06 Apr	0.08	0.033	95	02 Aug
	Default run	0.091	06 Mar	25 Apr	0.29	0.13	65	19 Jun
	In situ	0.17	25 Feb	29 Mar	0.58	0.27	93	28 May
Cariaco	Ens mean	0.61	20 May	22 Jul	1.09	0.38	133	30 Sep
	Range	$\pm 1.53(1.29)$	± 101	± 66	± 2.61	± 0.86	± 63	± 61
	NRR	0.78 (0.90)	1.48	1.40	1.39	1.42	1.88	1.55
	Ens med	0.37	22 May	14 Jul	0.83	0.34	110	06 Sep
	Default run	0.46	21 May	22 Jul	0.98	0.39	122	19 Sep
	In situ	0.61	16 Mar	21 Apr	2.39	1.15	76	01 Jun
L4	Ens mean	1.65	13 May	06 Jun	3.25	1.13	64	17 Aug
	Range	$\pm 2.48(2.14)$	± 100	± 82	± 3.12	± 1.50	± 78	± 167
	NRR	1.00 (1.36)	1.49	1.42	1.32	1.48	1.22	1.19
	Ens med	1.49	18 May	07 Jun	3.09	1.13	70	18 Sep
	Default run	2.03	19 Apr	08 Jun	3.73	1.3	94	11 Aug
	In situ	1.20	09 Mar	11 Apr	3.58	1.64	80	28 May

: Figure Captions

Figure 1. Nearly identical curves which describes resource uptake (a), zooplankton grazing (b), and phytoplankton mortality (c). Figure (a) shows four uptake functions, which have been optimised to the default uptake function, monod (U_1). Figure (b) shows two grazing functional forms, the holling type III (G_1) and type II (G_2) functions. Four phytoplankton mortality functions are shown on figure (c), whereby hyperbolic is the default function. The optimisation method is describe in section 2.1, 2.2, and 2.3. The range for DIN in (a) is between 0.001 and 20 mmol m^{-3}, and phytoplankton in (b) and (c) are 0.001 and 10 mmol m^{-3}. Table 1 describes the function's equations and parameters.

Figure 2. SeaWIFs-derived mean 1998 chlorophyll-a (mg m^{-3}) overlain with the 5 oceanographic stations time series site (Red dots). These stations are located in different oceanic regions: oligotrophic (ALOHA and BATS), coastal (L4 and Cariaco), and abyssal plain (PAP).

Figure 3. Chlorophyll and ~~nitrogen~~ DIN profiles from ensemble mean ((a) and (d) respectively), in situ observations ((c) and (f) for chlorophyll and ~~nitrogen~~ DIN respectively), and 75th and 25th quartile range of concentrations at each depth ((b) for chlorophyll and (e) for ~~nitrogen~~DIN) at station PAP. The range are obtained by averaging the concentrations from all ensemble members for 10 years at each depths. Black dots in the second column show the mean concentration of the ensemble mean over the time series (from January 1998-December 2007). White solid line in (a) shows mixed layer depth.

Figure 4. Inter-annual mean of surface chlorophyll from all the study sites ((a)-(e)) and the 10-year annual mean (g), all measured in mg m^{-3}. The boxplots show the ensemble annual means. Blue cross is the in situ observation, red open circle, black dot, and blue stars are the ensemble mean, median, and the default run respectively. The blue box is the 75th (top) and 25th (bottom) quartiles. Red line is the median. The whiskers are the ensemble minimum and maximum mean of surface chlorophyll. Annual mean values and NRR are described in Table 4.

Figure 5. 10-year monthly mean surface chlorophyll from all the study sites ((a)-(e)), showing the seasonal dynamics of surface chlorophyll (mg m^{-3}). The boxplots show the ensemble seasonal means. Blue cross is the in situ observation, red open circle, black dot, and blue stars are the ensemble mean, median, and the default run respectively. The blue box is the 75th (top) and 25th (bottom) quartiles. The red line is the median. The whiskers are the ensemble minimum and maximum mean of surface chlorophyll. In station PAP, in situ data for December is not available due to low light and high cloud cover.

Figure 6. Inter-annual variability of DIN averaged over 200 m~~integrated nitrogen~~, from all the study sites ((a)-(e)), and the annual mean (f). Since the in situ data for PAP does not always cover the first 200m, the overall mean ~~nitrogen~~ DIN concentration from all depth is used instead. For station L4, in situ ~~nitrogen~~ DIN is only collected on the surface. Blue cross is the in situ observation, red open circle, black dot, and blue stars are the ensemble mean, median, and default run respectively. The blue box is the 75th(top) and 25th(bottom) quartiles. Red line is the median, and the whiskers are the ensmeble minimum and maximum of the ~~integrated nitrogen~~averaged DIN. In station L4 and PAP data for ~~nitrogen~~ DIN is only available from 2000-2007 and 2002-2004 respectively.

Figure 7. 10-year monthly mean of DIN averaged over 200 m ~~integrated nitrogen~~ from all the study sites ((a)-(e)), showing the seasonal dynamics of ~~nitrogen~~ DIN (mmol m^{-3}). For station PAP, the ~~nitrogen~~ DIN shown is the overall profile, and in L4, the in situ ~~nitrogen~~ DIN concentration is only available at the surface. The boxplot shows the ensemble monthly means. Blue cross is the in situ observation, red open circle, black dot, and blue stars are the ensemble mean, median, and the default run respectively. The blue box is the 75th (top) and 25th (bottom) quartiles. The red line is the median. The whiskers are the ensemble minimum and maximum mean of ~~integrated nitrogen~~averaged DIN. In station PAP, the in situ data is only collected from 2002-2004 and L4 from 2000-2007.

Figure 8. Annual mean of surface chlorophyll when changing only one process at a time (blue box), overlain with annual mean of all ensemble members (green box) at five oceanographic stations. Ensemble mean and median plotted in the figure (shown in red open circle and black closed circle), are the from the 128 ensemble members.

Figure 9. Time series (from January 1998-December 2007) of ensemble mean and in situ, and range of chlorophyll and ~~nitrogen~~ DIN concentrations at oligotrophic stations. Station ALOHA is shown on (a)-(f) and BATS is shown on (g)-(l). White solid line in (~~b~~a) and (g) represents mixed layer depth. (~~b~~e), (~~d~~f), (~~h~~k), and (~~j~~l) are the 75th and 25th percentile range of chlorophyll ((~~b~~e) for ALOHA and (~~h~~k) for BATS) and ~~nitrogen~~ DIN ((~~d~~f) for ALOHA and (~~j~~l) BATS) over the depth. The range is obtained by averaging the chlorophyll and ~~nitrogen~~ DIN concentrations of each ensemble members over the time series at each depth. Black dots in (~~b~~e), (~~d~~f), (~~h~~k), and (~~j~~l) are the mean of the ensemble. Ensemble mean chlorophyll profiles (shown on (a) and (g)) and ~~nitrogen~~ DIN ((~~e~~c) and (~~k~~l)) are obtained from all of the ensemble members. *In situ* chlorophyll are shown in (~~e~~b) and (~~i~~h), and ~~nitrogen~~ DIN are shown in (~~g~~d) and (~~l~~j), for ALOHA and BATS respectively.

Figure 10. Mean integrated primary production averaged over 200m that are available in (a) ALOHA and (b) Cariaco, and (c) the annual mean. The NRR for ALOHA and Cariaco are 1.12 and 0.80 respectively.

Figure 11. Chlorophyll profile 10-year means ((a)-(d)) and its RMSEs ((e)-(~~f~~h)) at four oceanographic station from all of the ensemble members. Station L4 is not included as chlorophyll data is only taken at the surface. These are arranged by the lowest chlorophyll (top left) mean to the highest (bottom right), depending on the oceanographic regions.

Figure 12. 10-year mean and RMSE of surface chlorophyll (mg m^{-3}) at five stations from all ensemble members. The first panel ((a)-(e)) shows surface chlorophyll mean and RMSEs are shown on the second panel ((f)-(j)). Concentrations and RMSEs are arranged by the lowest chlorophyll (top left) mean to the highest (bottom right), depending on the oceanographic regions. For station PAP, the sequence is sorted based on coastal station. The y-axis shows combination of uptake (U_1, U_2, U_3, and U_4) and grazing (G_1 and G_2), and x-axis shows combinations of phytoplankton (ρ) and zooplankton (ξ) mortalities.

Figure 13. 10-year mean and RMSE of ~~nitrogen~~ DIN (mmol m^{-3}), at five stations from all ensemble members. The first panel ((a)-(e)) shows ~~nitrogen~~ DIN mean and RMSEs are shown on the second panel ((f)-(j)). Concentrations and RMSEs are arranged by the lowest chlorophyll (top left) mean to the highest (bottom right), depending on the oceanographic regions. For station PAP, the sequence is sorted based on coastal station. The y-axis shows combination of uptake (U_1, U_2, U_3, and U_4) and grazing (G_1 and G_2), and x-axis shows combinations of phytoplankton (ρ) and zooplankton (ξ) mortalities.

Figure 14. Time series of chlorophyll and ~~nitrogen~~ DIN profile of ensemble mean, their range, and in situ concentrations at the coastal stations Cariaco (a-f) and L4 (g-h) from January 1998-December 2007. (a) and (~~d~~c) show chlorophyll and ~~nitrogen~~ DIN ensemble mean at Cariaco respectively. White solid line in (a) is the mixed layer depth. (~~b~~e) and (~~e~~f) shows the 75^{th} and 25^{th} percentile of chlorophyll and ~~nitrogen~~ DIN concentrations at each depth. The black dots are the mean of the ensemble. These range are obtained form the 10-year mean concentrations at each depth. Since in situ chlorophyll and ~~nitrogen~~ DIN were taken at the surface in station L4, only surface time series were shown in (g-h). The grey shades on chlorophyll, shown in (g), and ~~nitrogen~~DIN, shown in (h) time series show 75^{th} and 25^{th} percentile of the range. Blue and red dots are in situ concentrations for chlorophyll and ~~nitrogen~~ DIN respectively.

Figure 15. Phytoplankton phenology metrics at the five stations. Blue cross is the in situ, red, black, and blue dots are the ensemble mean, median, and the default run respectively. The timings and concentrations are averaged annually from January 1998 to December 2007.

Figure A1. Determining phenology using a combination of threshold method and curve fit at station L4, here the initiation is when the fitted curve is above 50% of the maximum peak, however the termination is on the first valley.

: Figures

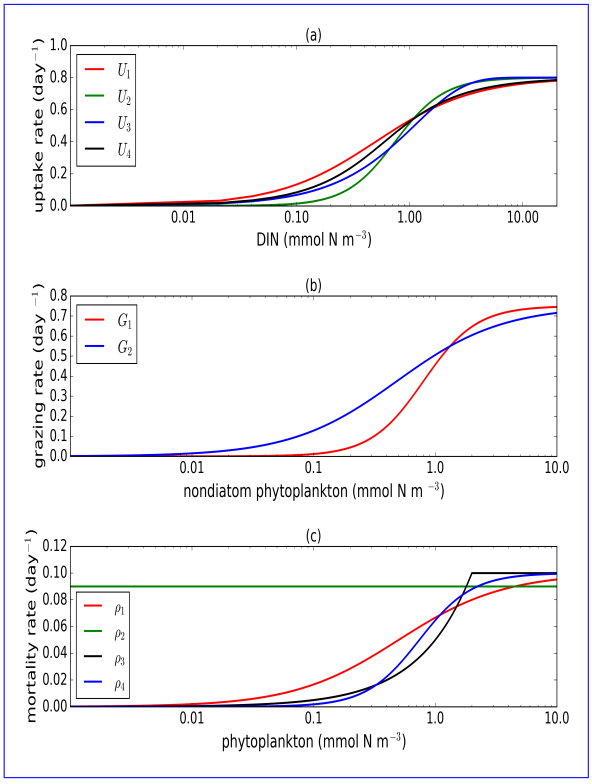

Figure 1. Nearly identical curves which describes resource uptake (a), zooplankton grazing (b), and phytoplankton mortality (c). Figure (a) shows four uptake functions, which have been optimised to the default uptake function, monod (U_1). Figure (b) shows two grazing functional forms, the holling type III (G_1) and type II (G_2) functions. Four phytoplankton mortality functions are shown on figure (c), whereby hyperbolic is the default function. The optimisation method is describe in section 2.1, 2.2, and 2.3. The range for DIN in (a) is between 0.001 and 20 mmol m^{-3}, and phytoplankton in (b) and (c) are 0.001 and 10 mmol m^{-3}. Table 1 describes the function's equations and parameters.

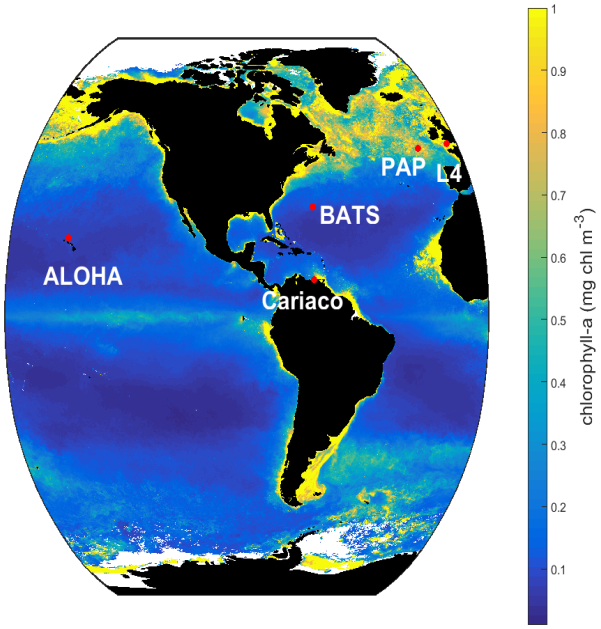

Figure 2. SeaWIFs-derived mean 1998 chlorophyll-a (mg m^{-3}) overlain with the 5 oceanographic stations time series site (Red dots). These stations are located in different oceanic regions: oligotrophic (ALOHA and BATS), coastal (L4 and Cariaco), and abyssal plain (PAP).

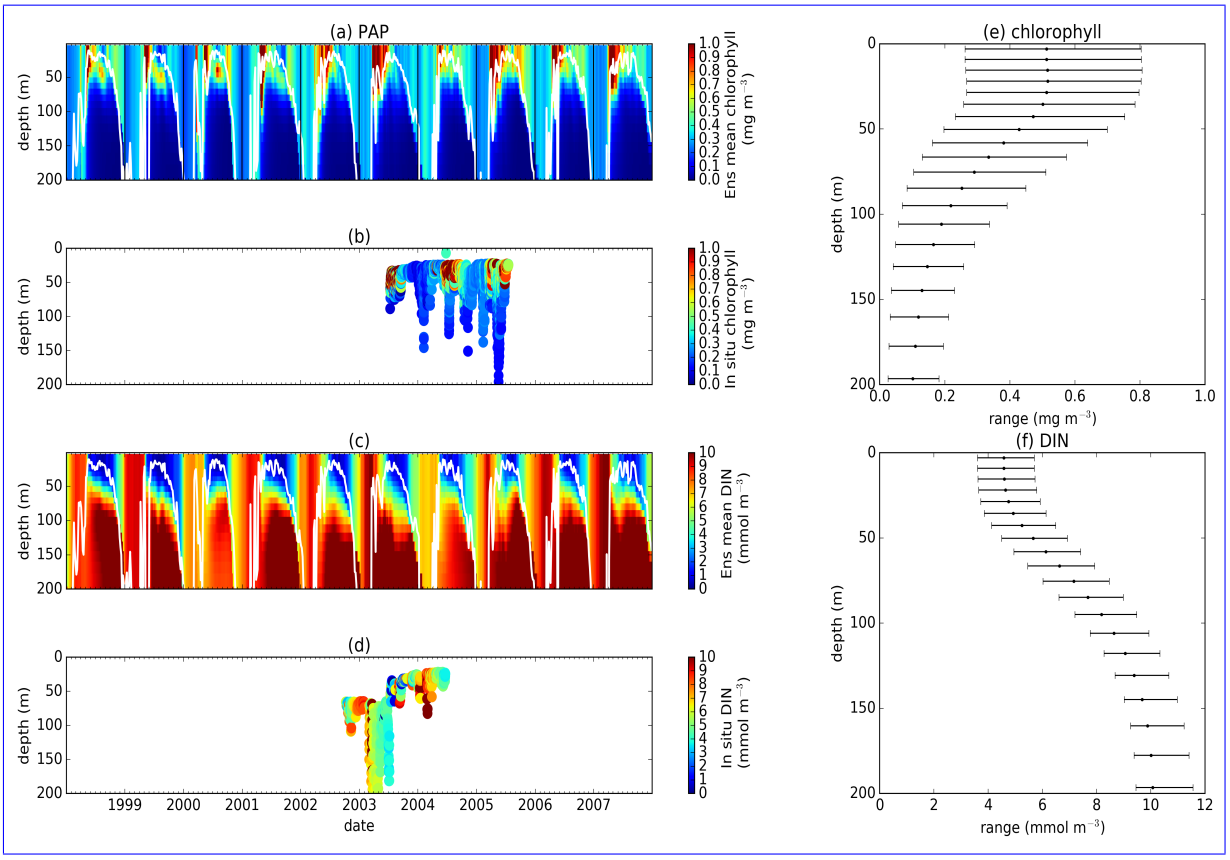

Figure 3. Chlorophyll and ~~nitrogen~~ DIN profiles from ensemble mean ((a) and (d) respectively), in situ observations ((c) and (f) for chlorophyll and ~~nitrogen~~ DIN respectively), and 75th and 25th quartile range of concentrations at each depth ((b) for chlorophyll and (e) for ~~nitrogen~~DIN) at station PAP. The range are obtained by averaging the concentrations from all ensemble members for 10 years at each depths. Black dots in the second column show the mean concentration of the ensemble mean over the time series (from January 1998-December 2007). White solid line in (a) shows mixed layer depth.

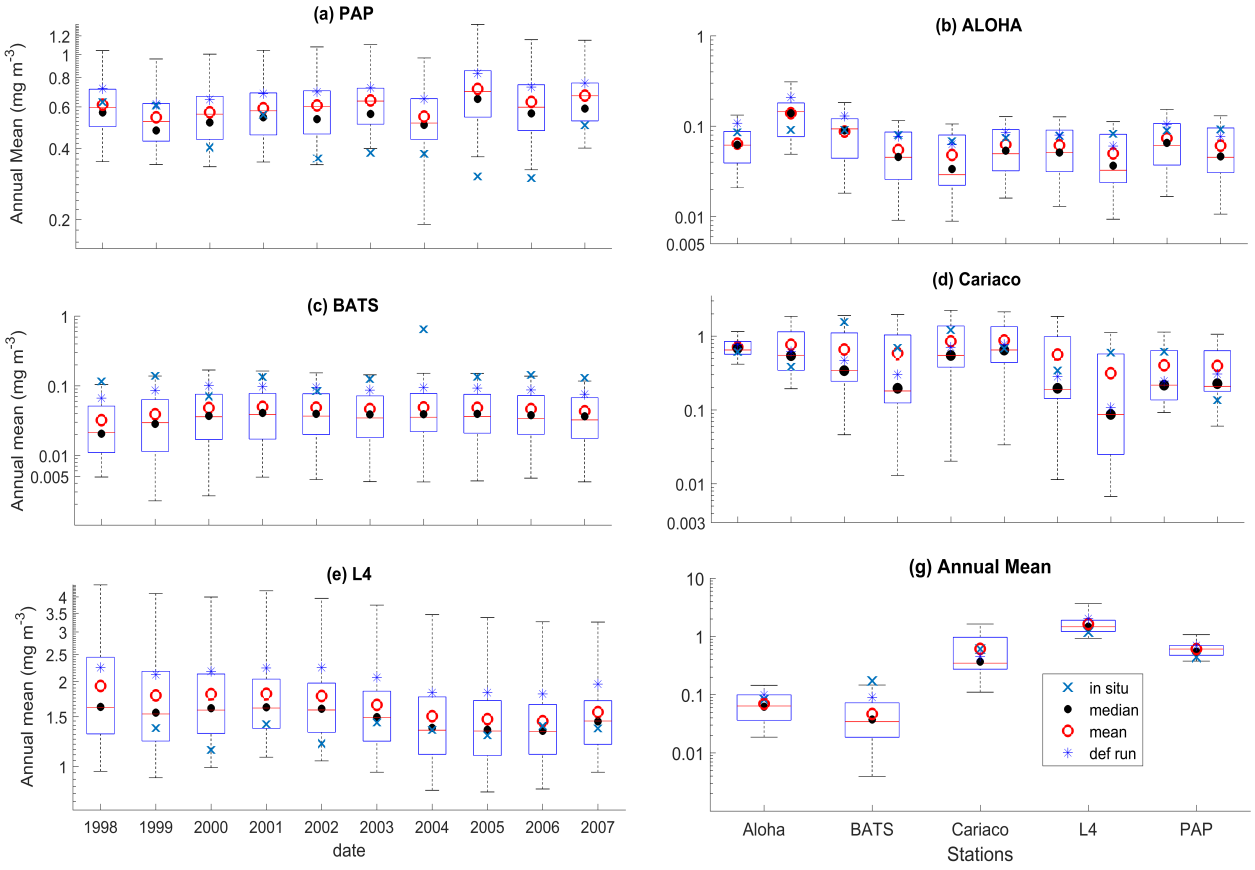

Figure 4. Inter-annual mean of surface chlorophyll from all the study sites ((a)-(e)) and the 10-year annual mean (g), all measured in mg m^{-3}. The boxplots show the ensemble annual means. Blue cross is the in situ observation, red open circle, black dot, and blue stars are the ensemble mean, median, and the default run respectively. The blue box is the 75th (top) and 25th (bottom) quartiles. Red line is the median. The whiskers are the ensemble minimum and maximum mean of surface chlorophyll. Annual mean values and NRR are described in Table 4.

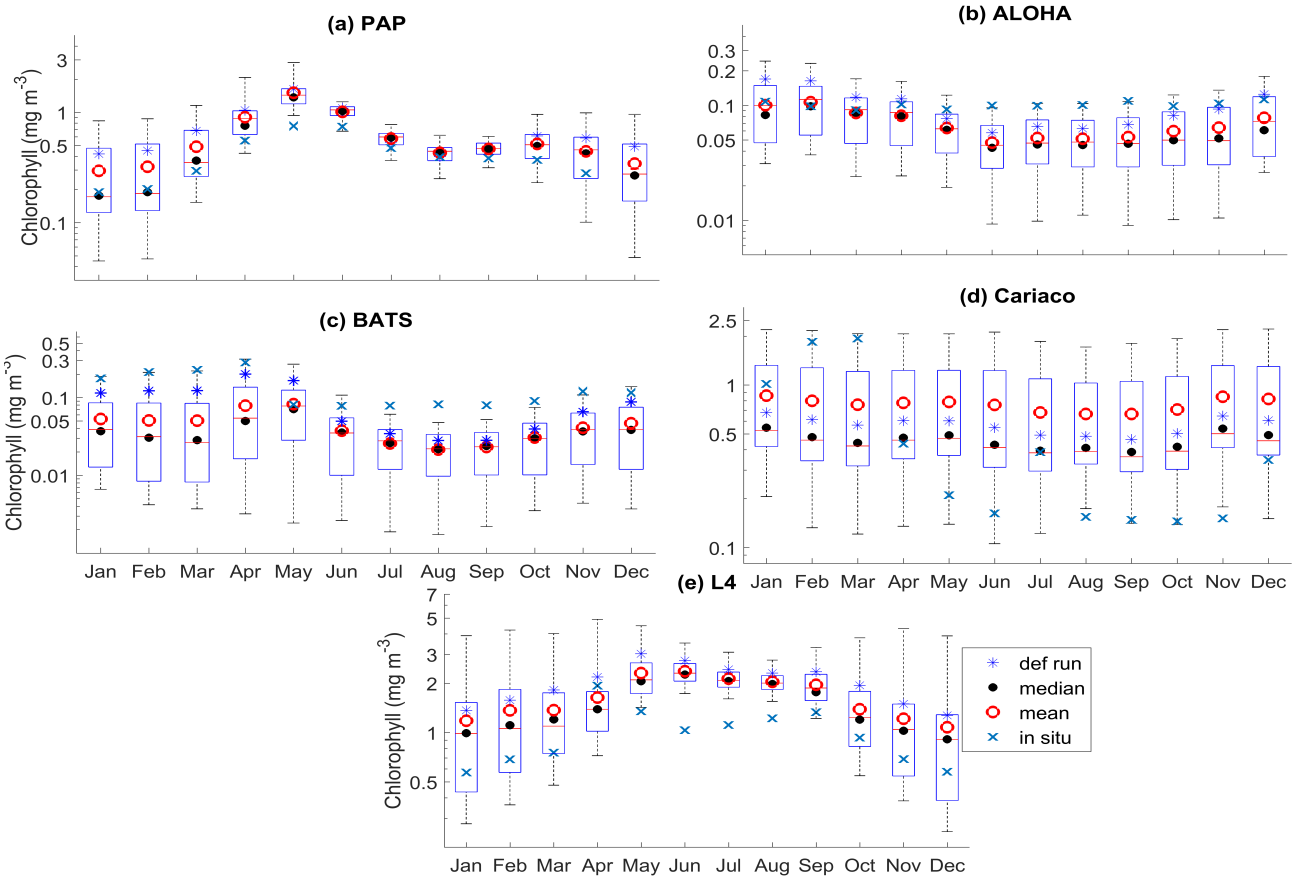

Figure 5. 10-year monthly mean surface chlorophyll from all the study sites ((a)-(e)), showing the seasonal dynamics of surface chlorophyll (mg m^{-3}). The boxplots show the ensemble seasonal means. Blue cross is the in situ observation, red open circle, black dot, and blue stars are the ensemble mean, median, and the default run respectively. The blue box is the 75th (top) and 25th (bottom) quartiles. The red line is the median. The whiskers are the ensemble minimum and maximum mean of surface chlorophyll. In station PAP, in situ data for December is not available due to low light and high cloud cover.

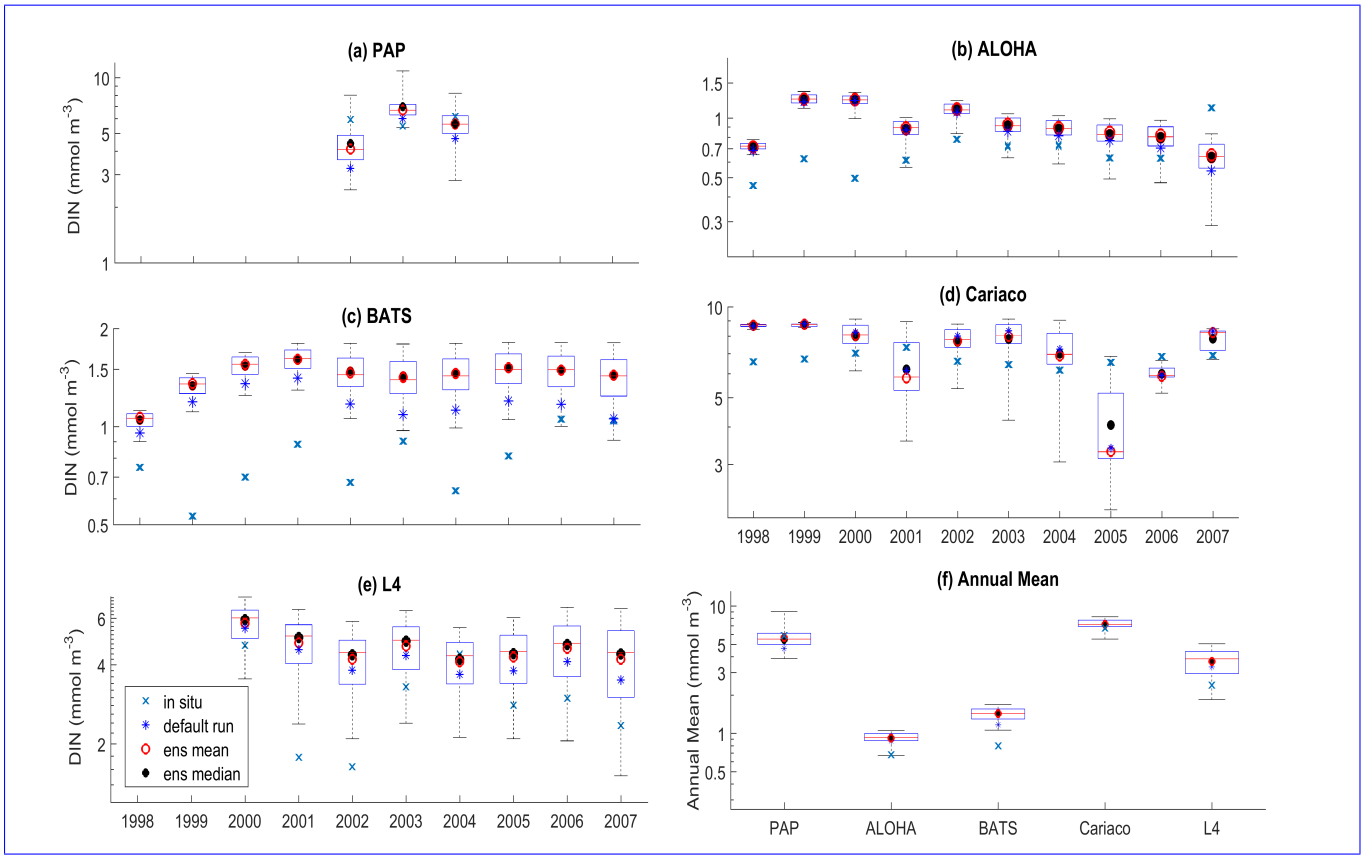

Figure 6. Inter-annual variability of DIN averaged over 200 m~~integrated nitrogen~~, from all the study sites ((a)-(e)), and the annual mean (f). Since the in situ data for PAP does not always cover the first 200m, the overall mean ~~nitrogen~~ DIN concentration from all depth is used instead. For station L4, in situ ~~nitrogen~~ DIN is only collected on the surface. Blue cross is the in situ observation, red open circle, black dot, and blue stars are the ensemble mean, median, and default run respectively. The blue box is the 75th(top) and 25th(bottom) quartiles. Red line is the median, and the whiskers are the ensmeble minimum and maximum of the ~~integrated nitrogen~~averaged DIN. In station L4 and PAP data for ~~nitrogen~~ DIN is only available from 2000-2007 and 2002-2004 respectively.

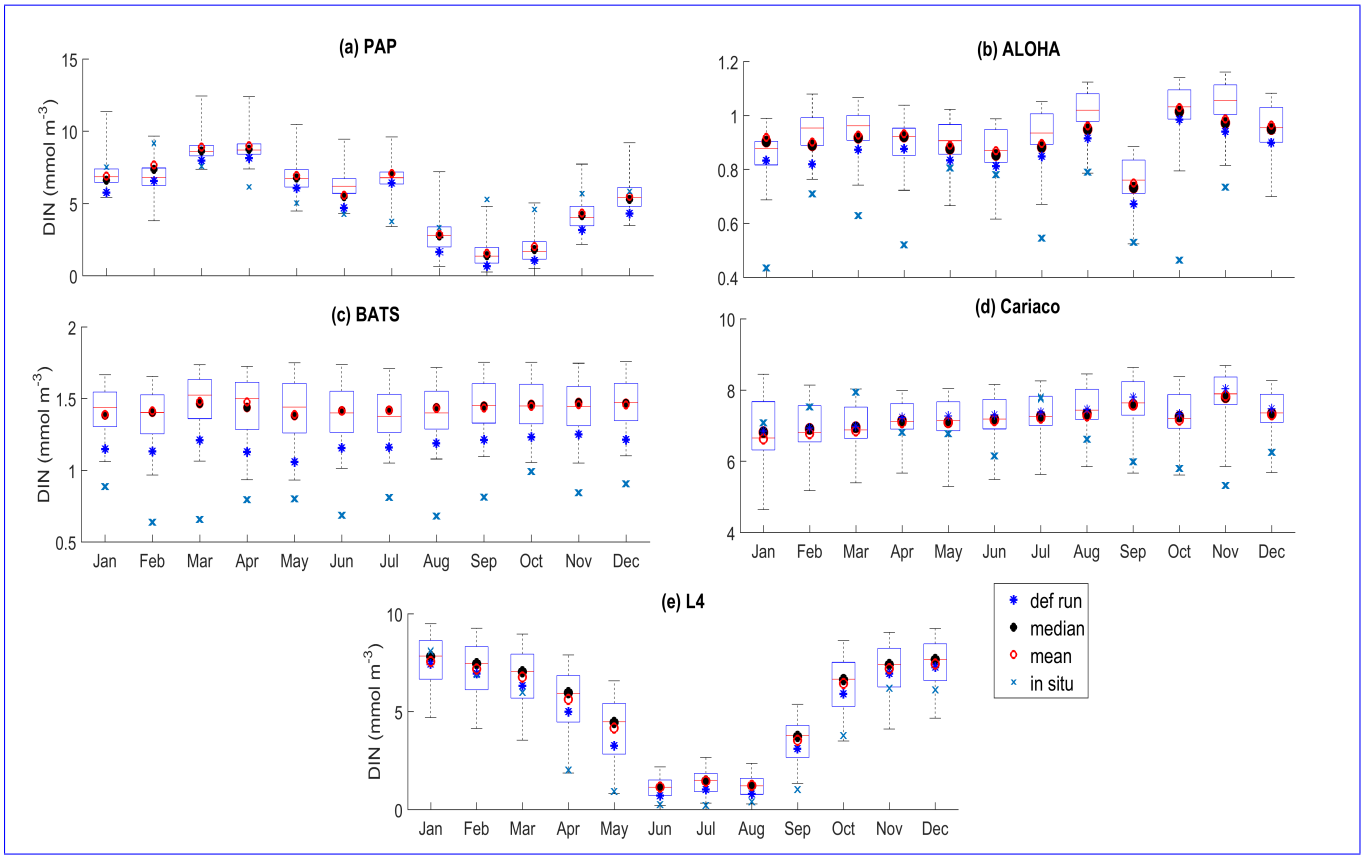

Figure 7. 10-year monthly mean of DIN averaged over 200 m ~~integrated nitrogen~~ from all the study sites ((a)-(e)), showing the seasonal dynamics of ~~nitrogen~~ DIN (mmol m^{-3}). For station PAP, the ~~nitrogen~~ DIN shown is the overall profile, and in L4, the in situ ~~nitrogen~~ DIN concentration is only available at the surface. The boxplot shows the ensemble monthly means. Blue cross is the in situ observation, red open circle, black dot, and blue stars are the ensemble mean, median, and the default run respectively. The blue box is the 75th (top) and 25th (bottom) quartiles. The red line is the median. The whiskers are the ensemble minimum and maximum mean of ~~integrated nitrogen~~averaged DIN. In station PAP, the in situ data is only collected from 2002-2004 and L4 from 2000-2007.

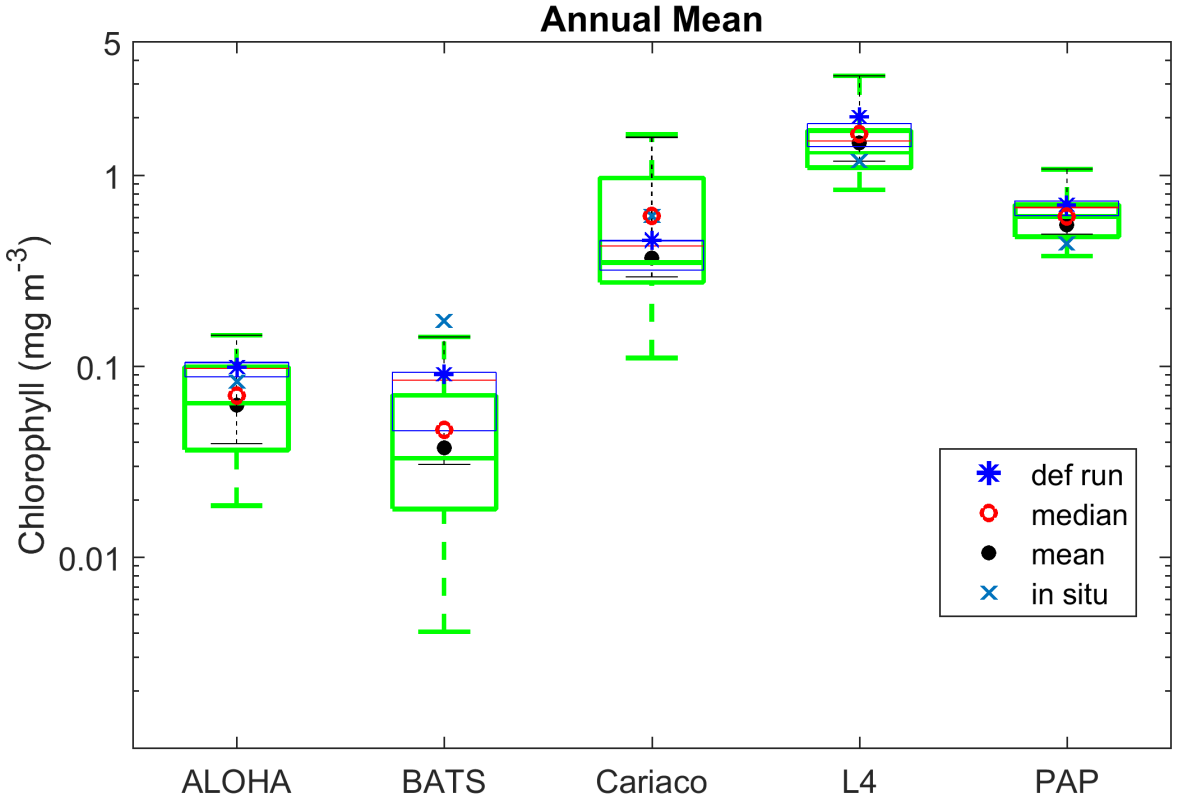

Figure 8. Annual mean of surface chlorophyll when changing only one process at a time (blue box), overlain with annual mean of all ensemble members (green box) at five oceanographic stations. Ensemble mean and median plotted in the figure (shown in red open circle and black closed circle), are the from the 128 ensemble members.

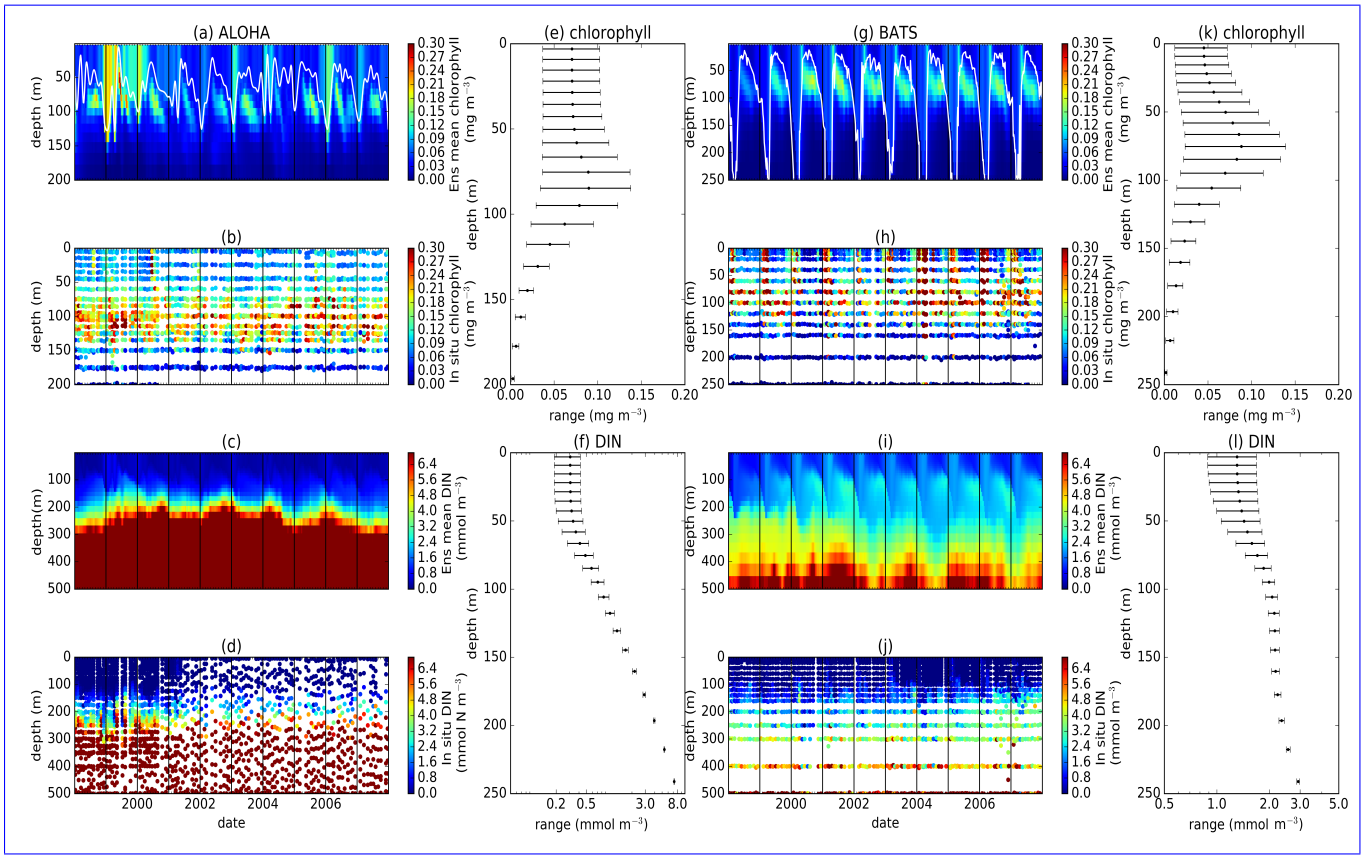

Figure 9. Time series (from January 1998-December 2007) of ensemble mean and in situ, and range of chlorophyll and ~~nitrogen~~ DIN concentrations at oligotrophic stations. Station ALOHA is shown on (a)-(f) and BATS is shown on (g)-(l). White solid line in (~~b~~a) and (g) represents mixed layer depth. (~~b~~e), (~~d~~f), (~~h~~k), and (~~j~~l) are the 75th and 25th percentile range of chlorophyll ((~~b~~e) for ALOHA and (~~h~~k) for BATS) and ~~nitrogen~~ DIN ((~~d~~f) for ALOHA and (~~j~~l) BATS) over the depth. The range is obtained by averaging the chlorophyll and ~~nitrogen~~ DIN concentrations of each ensemble members over the time series at each depth. Black dots in (~~b~~e), (~~d~~f), (~~h~~k), and (~~j~~l) are the mean of the ensemble. Ensemble mean chlorophyll profiles (shown on (a) and (g)) and ~~nitrogen~~ DIN ((~~e~~c) and (~~k~~l)) are obtained from all of the ensemble members. *In situ* chlorophyll are shown in (~~e~~b) and (~~j~~h), and ~~nitrogen~~ DIN are shown in (~~g~~d) and (~~j~~j), for ALOHA and BATS respectively.

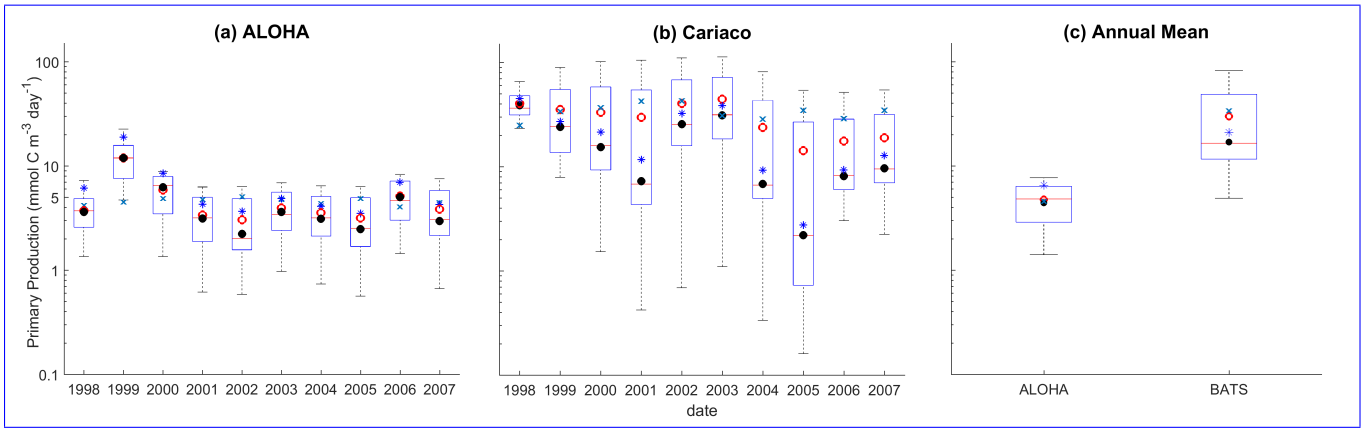

Figure 10. Mean integrated primary production averaged over 200m that are available in (a) ALOHA and (b) Cariaco, and (c) the annual mean. The NRR for ALOHA and Cariaco are 1.12 and 0.80 respectively.

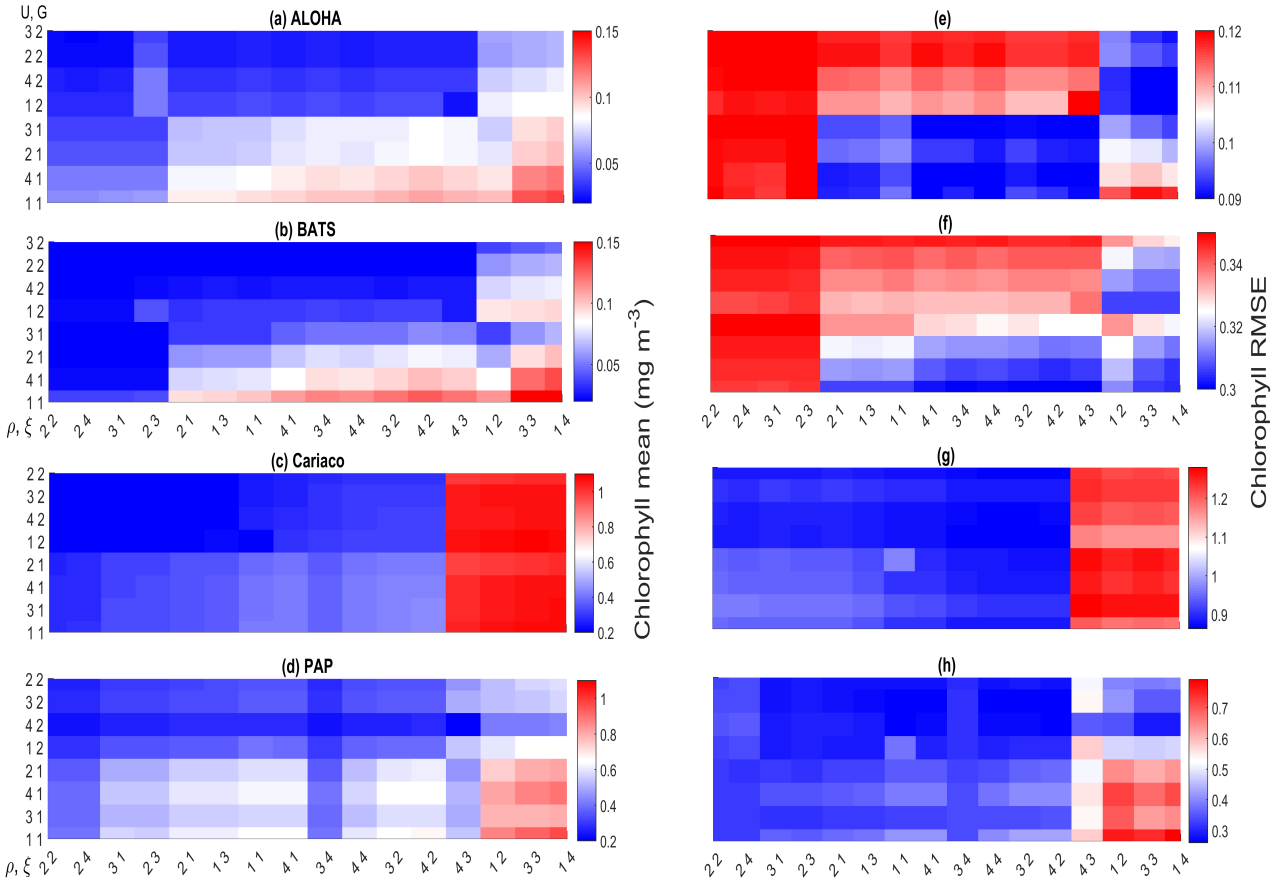

Figure 11. Chlorophyll profile 10-year means ((a)-(d)) and its RMSEs ((e)-(h)) at four oceanographic station from all of the ensemble members. Station L4 is not included as chlorophyll data is only taken at the surface. These are arranged by the lowest chlorophyll (top left) mean to the highest (bottom right), depending on the oceanographic regions.

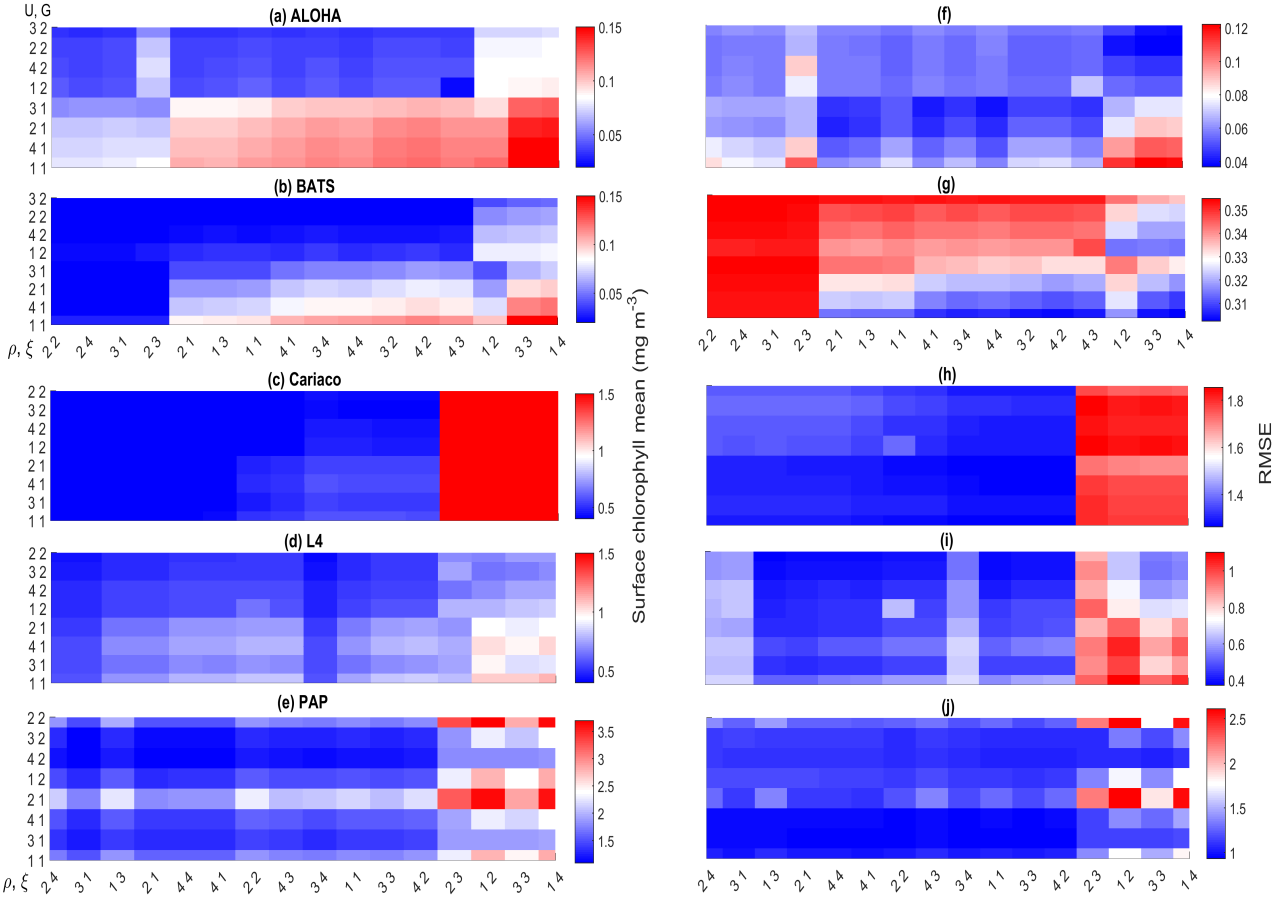

Figure 12. 10-year mean and RMSE of surface chlorophyll (mg m^{-3}) at five stations from all ensemble members. The first panel ((a)-(e)) shows surface chlorophyll mean and RMSEs are shown on the second panel ((f)-(j)). Concentrations and RMSEs are arranged by the lowest chlorophyll (top left) mean to the highest (bottom right), depending on the oceanographic regions. For station PAP, the sequence is sorted based on coastal station. The y-axis shows combination of uptake (U_1, U_2, U_3, and U_4) and grazing (G_1 and G_2), and x-axis shows combinations of phytoplankton (ρ) and zooplankton (ξ) mortalities.

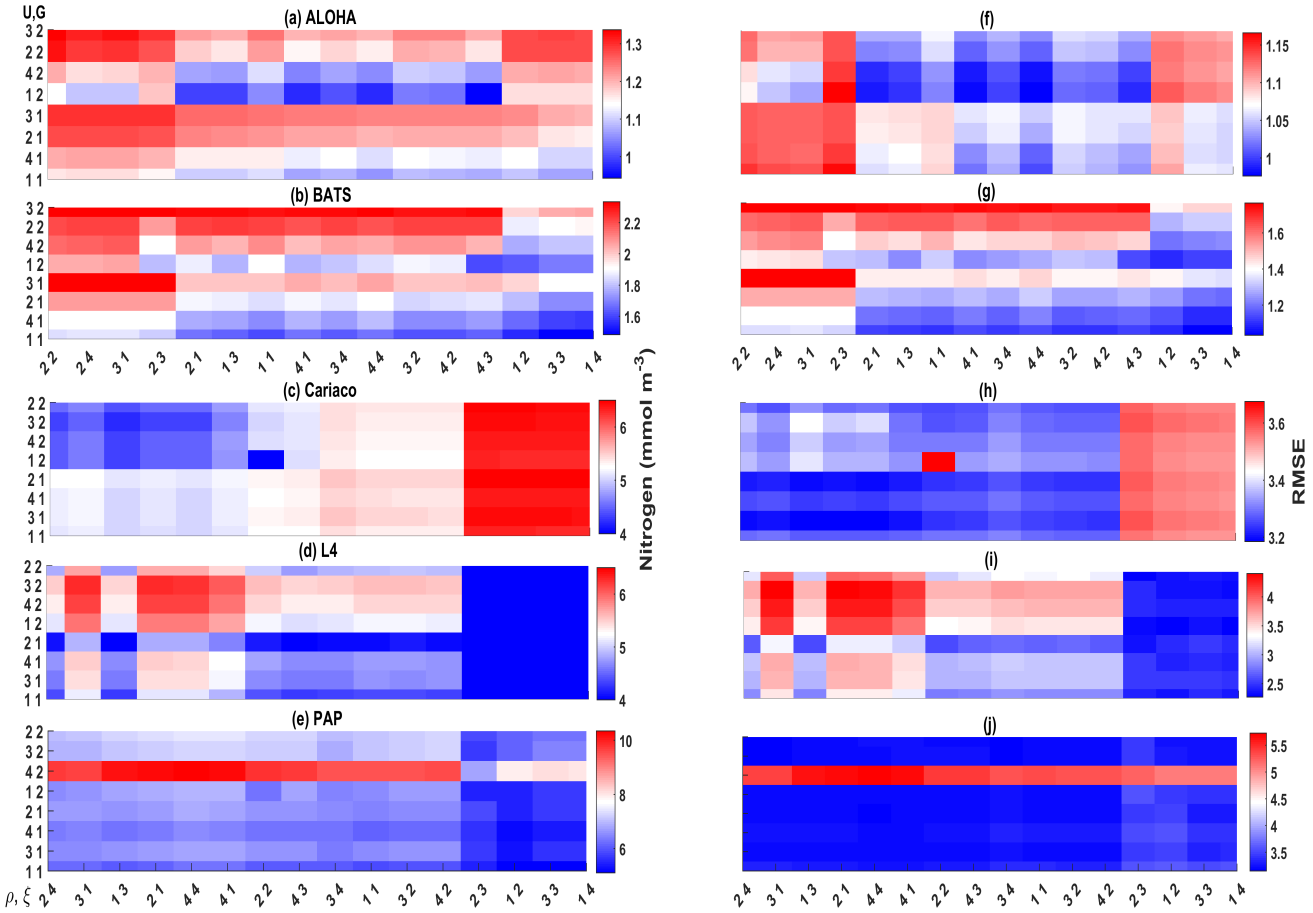

Figure 13. 10-year mean and RMSE of ~~nitrogen~~ DIN (mmol m^{-3}), at five stations from all ensemble members. The first panel ((a)-(e)) shows ~~nitrogen~~ DIN mean and RMSEs are shown on the second panel ((f)-(j)). Concentrations and RMSEs are arranged by the lowest chlorophyll (top left) mean to the highest (bottom right), depending on the oceanographic regions. For station PAP, the sequence is sorted based on coastal station. The y-axis shows combination of uptake (U_1, U_2, U_3, and U_4) and grazing (G_1 and G_2), and x-axis shows combinations of phytoplankton (ρ) and zooplankton (ξ) mortalities.

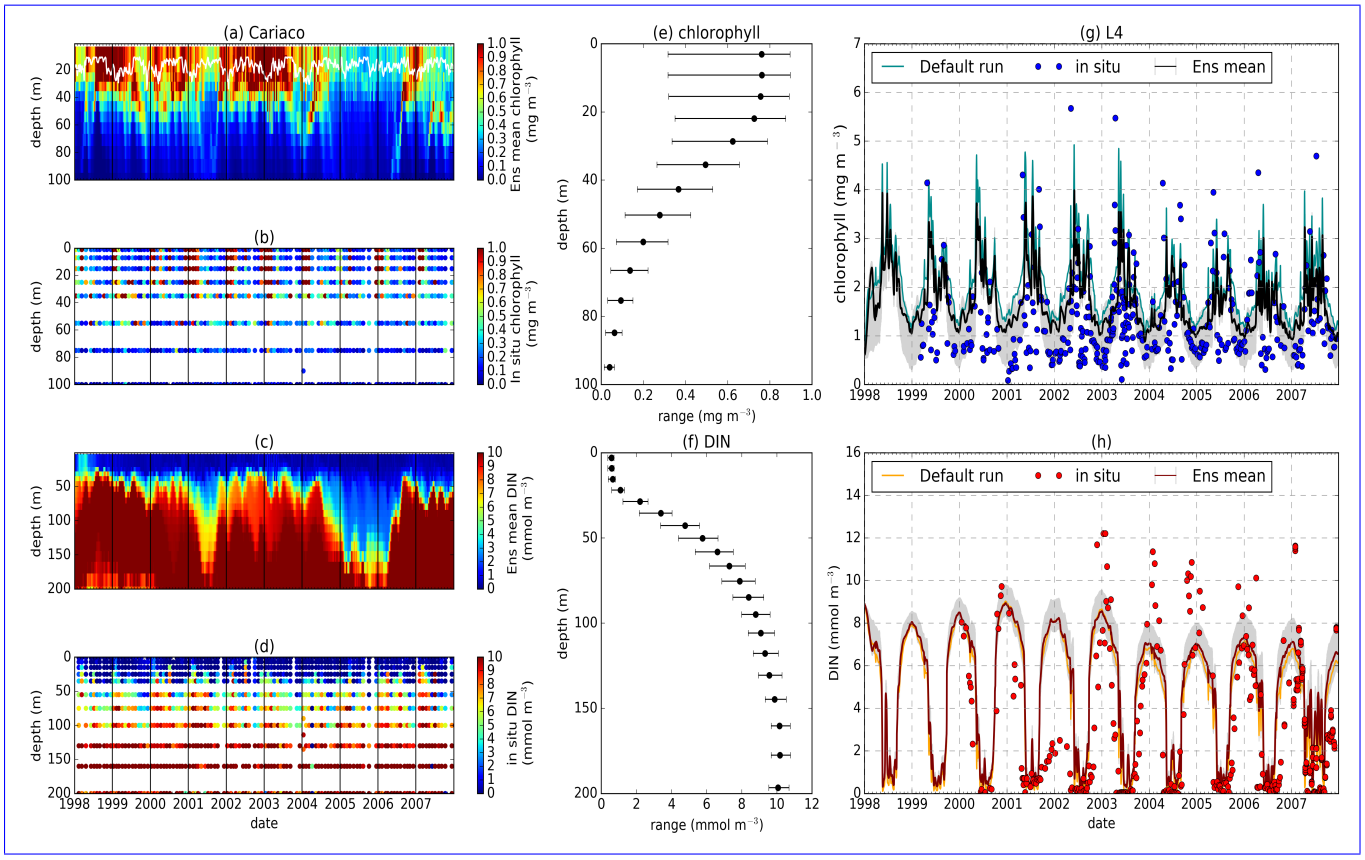

Figure 14. Time series of chlorophyll and ~~nitrogen~~ DIN profile of ensemble mean, their range, and in situ concentrations at the coastal stations Cariaco (a-f) and L4 (g-h) from January 1998-December 2007. (a) and (~~d~~c) show chlorophyll and ~~nitrogen~~ DIN ensemble mean at Cariaco respectively. White solid line in (a) is the mixed layer depth. (~~b~~c) and (~~e~~f) shows the 75th and 25th percentile of chlorophyll and ~~nitrogen~~ DIN concentrations at each depth. The black dots are the mean of the ensemble. These range are obtained form the 10-year mean concentrations at each depth. Since in situ chlorophyll and ~~nitrogen~~ DIN were taken at the surface in station L4, only surface time series were shown in (g-h). The grey shades on chlorophyll, shown in (g), and ~~nitrogen~~DIN, shown in (h) time series show 75th and 25th percentile of the range. Blue and red dots are in situ concentrations for chlorophyll and ~~nitrogen~~ DIN respectively.

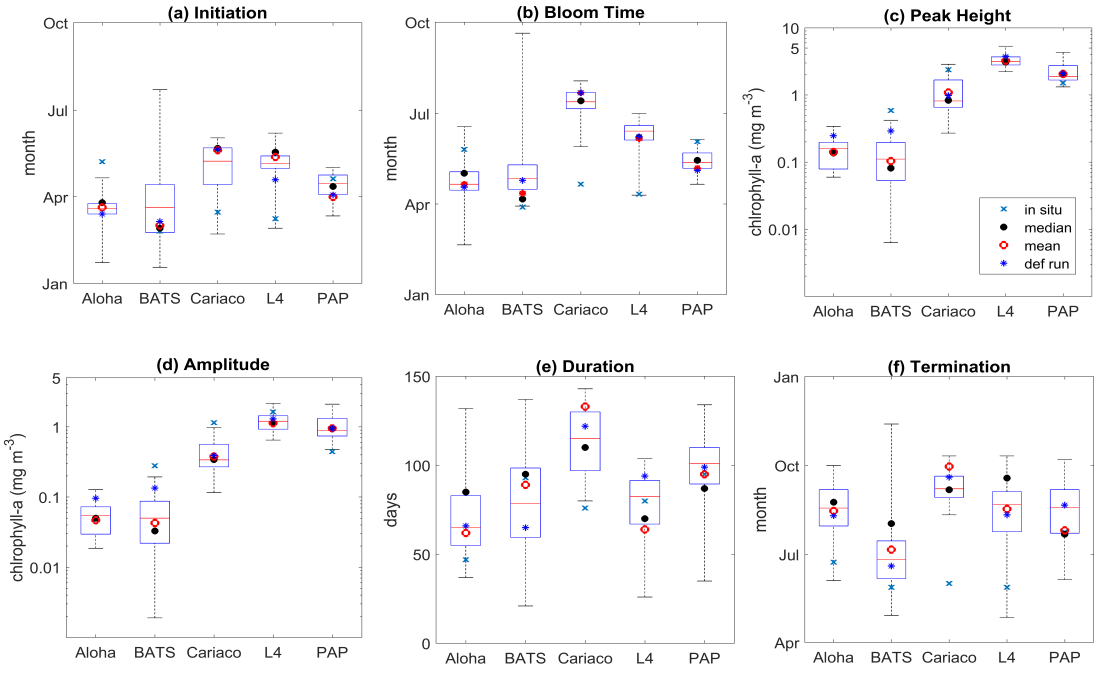

Figure 15. Phytoplankton phenology metrics at the five stations. Blue cross is the in situ, red, black, and blue dots are the ensemble mean, median, and the default run respectively. The timings and concentrations are averaged annually from January 1998 to December 2007.

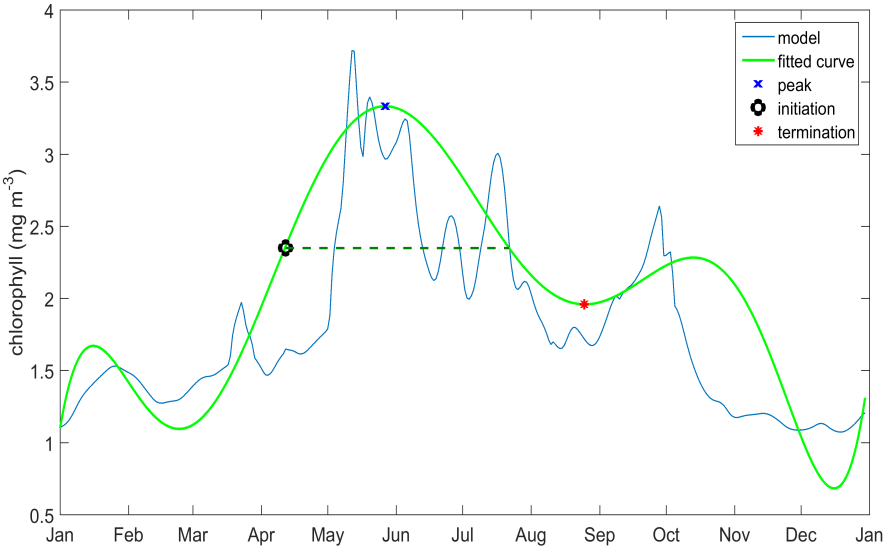

Figure A1. Determining phenology using a combination of threshold method and curve fit at station L4, here the initiation is when the fitted curve is above 50% of the maximum peak, however the termination is on the first valley.