Dear Editor,

We would like to thank Reviewer #2 for agreeing to review this paper again, and for the constructive criticism.

As suggested by Reviewer #2, we have now changed the functions from U_1, U_2, U_3, and U_4 into U_h, U_s, U_e, and U_t for uptake, and similarly to the mortality functions. We have also re-structured section 3, whereby the observations are described first, followed by the ensemble simulations, their statistics, and NRR. The paragraphs and sentences in the Summary and Discussion section have been made shorter.

Below is our point by point response to all comments made by Reviewer #2. Please note that the line and page numbers corresponds to the annotated version attached. Reviewer comments (RC) are included in normal text and responses (AR) are in bold.

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

Kind Regards,

Authors.

Response to Reviewer #2:

RC: The updated manuscript has improved in many aspects. I have no more issues with methods and results that are presented but the way they are presented needs improvement before the manuscript can be published.

AR: Thank you reviewing the paper again and acknowledging the improvement of the current manuscript.

General comments:

RC: I found the results sections were difficult to read and not well structured. For example, section 3.1 starts out by describing the data. The second paragraph covers model DIN seasonality but in the last sentence also starts the model-data comparison. In the third paragraph, it is initially not clear if the chlorophyll data is described or the ensemble. The fourth paragraph then launches the model data comparison and mentions the same result from the second paragraph again (NRR=1.25 for DIN). This back and forth is confusing to the reader. I suggest to reorganize each results section: start out by describing the ensemble characteristics (which are the focus here) and then compare them to the data.

AR: Thank you for the suggestion. The data description at the beginning of section 3.1 is included because the abyssal station has a shorter data timespan compared to other stations. We find that it is easier to start with in situ observations then briefly explain the match or miss-match between the ensemble and the in situ. We therefore have re-structured the paragraphs in the results section by first describing the observations, followed by the ensemble simulation. Then we describe ensemble characteristics, along with the RMSEs, NRR, and other error statistics. These are

followed by the ensemble results when we only change one functional term in each member, keeping the others in the default forms. (Please see the revised section 3.1 (page 9-11), 3.2 (page 11-14), 3.3 (page 14-16), and the Phytoplankton Phenology in section 3.4 (page 17-19).

RC: Section 4 is very dense and difficult to read. Again there is a lack of structure and the first few paragraphs each cover a range of different aspects without much consistency. Some run-on sentences are almost impossible to follow (for example, p18, l15). The manuscript would greatly benefit from tidying up this summary section and a clear focus on the key results.

AR: Thank you for the suggestion, we have now made the sentences and paragraphs shorter, for example, for p18, l15, we have split it into two sentences, and make the sentences a bit shorter (page 24, line 17-22).

RC: In order to improve readability, I'd suggest to rename the functions that are used throughout the manuscript, so it is easier for the reader to keep track. For example, new names for the uptake functions could be:

U_1: U_hyp

U_2: U_sig

U_3: U_exp

U_4: U_tri

The same naming scheme can easily be applied to the other functions.

AR: We have explained which functional forms corresponds to the abbreviations (U_1, U_2, G_1, etc) in Table 1. We have now changed the function abbreviations to U_h, U_s, U_e etc. For grazing function, we keep Holling type II function as G_2, and the default Holling type III as G_1.

RC: Specific comments:

p2

l6: "NPZ-D": "NPZD" is much more commonly used.

AR: We have changed this to NPZD (page 2 line 6).

l31: Change "parameters" to "biological parameters".

AR: We have changed this to biological parameters (page 2 line 31).

p3

l23: Anderson et al. (2015) is cited twice in the same sentence.

AR: **The second Anderson et al. (2015) has been removed (page 3 line 26).**

p4

l10: Here the manuscript shifts briefly from present into past tense. Use on consistently.

AR: We have changed ‘assessed’ to assess (page 4 line 11).

p6

l11: "has" -> "have"

AR: This has been changed (page 6 line 11).

l18: "study" -> "studies"

AR: This has been changed (page 6 line 18).

l18: 0.1 is not lower than zero, rephrase to something like "towards the lower end of the range of literature values"

AR: This sentence has been rephrased (page 6 line 19).

l26: "the 3-D models' predictive skill" -> "the predictive skill of 3-D models"

AR: This has been changed (page 6 like 27).

l33: "ensemble model" -> "model ensemble"

AR: This has been changed (page 7 line 5).

p7

l2: "This provides a total number of 128 ..." This sentence should probably precede the previous one. Altogether, the ensemble generation is needs a bit more explanation.

AR: This has been changed, we have added more explanation before this sentence (page 7 line 4-6).

l9: What is "in situ satellite SST"?

AR: We have rephrased this to satellite derived sea surface temperature (page 7 line 13).

l15: "there is no" rephrase to "is zero"

AR: This has been changed (page 7 line 20).

p8

l12: This paragraph is very difficult to understand. How does a simulation from 6-1200m minimize the computational cost? Is the maximum depth of the model at Cariaco also at 500m? What does "The level thickness is similar to that in 63 depth levels." mean? Do the 2 stations with reduced depth have fewer layers? You do not need to answer these questions in your reply, please rephrase the paragraph so that the reader knows.

AR: Running the model from 6-1200 instead of 6-5800 minimise the computational cost, and this have been added in the paragraph. We have also added the maximum sampling depth for Cariaco changed this sentence to make it clearer and shows all the boundaries for the depth levels (please see page 8 line 16 – 22).

l19: "We use ..." The sentence does not say what the metrics are used for.

AR: We have added what the metrics are used for (page 8 line 26-27).

p9

l9: I don't this paragraph is very helpful. Instead of explaining in what order the metrics will be used (the order is not followed anyway: in 3.1, NRR is mentioned before RMSE), it could explain the order the following subsections will be in.

AR: We have changed the content of this paragraph and included the order the following subsection will be in (page 9 line 15-22).

p10

l11: "range": Is this the ensemble range? Make this explicit.

AR: Yes this means ensemble range, and we have changed this to ensemble range (page 11 line 9)

l11: "If we only allow one process function at a time to change": I still don't think the new explanation helps much to understand what is going on. The "at a time" seems to imply that something is changing over time. I would suggest: "by generating an ensemble in which the functional form of only one of the processes is modified"

AR: We have changed the explanation to: 'However if we only allow one process function to change in each ensemble member, keeping the other processes with their default functions' (page 11 line 9-12).

l13: "covering 84% of the all ensemble members": what does this mean, please explain better.

AR: 'all ensemble members' has been changed to 'the full ensemble (128 members)' (page 11 line 12).

p11

l5: "A DCM occurs when lower chlorophyll is detected at the surface": Not sure if a description of DCM is needed but this one is more confusing than helpful.

AR: We have rephrase this to 'Another feature of the oligotrophic ocean is a deep chlorophyll maximum (DCM) that occurs below the mixed layer, when the chlorophyll concentration in the surface is low {Fennel2003, Letelier2004}' (page 11 line 33 – page 12 line 1).

p13

l3: "the in situ" add "data"

AR: This has been removed from the manuscript.

l30: "At most stations, the observed phenology metrics are covered by the ensemble range.": "covered by" does not make it clear what you are trying to say, I would suggest to rephrase using "fall within".

AR: This sentence has been removed from the manuscript.

p15

l26: "Through this approach, we provide a perturbed biology ensemble conditioned upon process structural uncertainties." This is a very dense sentence

AR: We have changed this sentence to 'Through this approach, we provide a perturbed biology ensemble conditioned upon structural uncertainties in model formulation' to the sentence (page 20 line 9-10).

p16

l2: "Even though fitting..." This sentence is difficult to understand, please rephrase.

AR: We have rephrased this sentence to 'Even when fitting for a phytoplankton concentration range similar to oligotrophic regions… is applied' (page 20 line 20-21).

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 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 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 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, zooplankton, and detritus (~~NPZ-D~~NPZD). 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), biological 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. 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 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 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~~ assess 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~~U_h, 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~~U_s, (ii) the exponential (Ivlev, 1961), ~~U_3~~U_e, and (~~iv~~iii) trigonometric functions (Jassby and Platt, 1976), ~~U_4~~U_t. 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~~U_h. 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 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 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 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 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(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 ($\rho_1,\xi_1\rho_h,\xi_h$), linear ($\rho_2,\xi_2\rho_l,\xi_l$), quadratic ($\rho_3,\xi_3\rho_q,\xi_q$), and sigmoidal ($\rho_4,\xi_4\rho_s,\xi_s$) 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(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 mortality rate is set so that the total loss integrated over the range of 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 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, 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~~ have been used. The maximum uptake rate, V_p , 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 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~~ studies have suggested to use 0 m day^{-1} (Ward et al., 2013). We chose a ~~lower~~ sinking rate of 0.1 m day^{-1}, towards the lower end of the range of literature values 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 predictive skill of 3-D ~~models' predictive skill~~ models (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: Porculpine Abyssal Plain Sustained-Obervatory (PAP-SO, hereafter PAP), A long time Oligotrophic Habitat Assessment (ALOHA), Bermuda Atlantic Time Series (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.~~ Each ensemble member has at least one functional form changed from the default functions. This provides a total number of 128 combinations, arising from 4 types of nutrient uptake, 4 phytoplankton mortality formulations, 2 types of zooplankton grazing, and 4 zooplankton mortalities. The model ensemble 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.

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~~satellite-derived sea surface temperature, 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~~the time mean of vertical velocity is zero. 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 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: 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 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 DIN data were collected from the surface from 1999-2008. Since the maximum depth in this station is only 50 m deep, the initial concentrations for chlorophyll and DIN are the same at every depth (total chlorophyll = 0.27 mg m^{-3}, 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, instead of 6-5800 m to minimise computational cost, apart from station L4, with maximum depth ~~is~~ of 50 m, and Cariaco, where the maximum depth for the physical input is available down to 500 m, although the depth at which nutrients are sampled are down to 1310 m. The boundaries for the depth levels are as follows: 6, 12, 19, 25, 32, 39, 46, 54, 62, 71, 80, 90, 100, 112, 124, 137, 152, 168, 187, 207, 229, 254, 281, 312, 347, 386, 429, 477, 531, 591, 656, 729, 809, 896, 991, 1093, and 1200 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. Stations that have shallower maximum depths are run with fewer depth levels. 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 DIN and chlorophyll and integrated chlorophyll in order to compare the ensemble model with default model and how well it represent the observations. 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 the seasonal variations. A similar approach is applied to DIN, however we use the averaged DIN over 200 m (integrated 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~~ In this section, the ensemble mean and median, spread (NRR), and range are compared with the ~~in situ data~~observations. Inter-annual and monthly variabilities are considered, and both biological concentrations and phytoplankton bloom phenology are assessed. The Abyssal Plain station comparisons are discussed in section (3.1), followed by the ~~NRR to assess the ensemble spread, and phytoplankton bloom phenology~~two oligotrophic stations (3.2), and the two coastal stations (3.3).

3.1 Abyssal Plain

~~In station PAP , in situ nitrate was~~ The Abyssal Plain is represented by station PAP located in the North Atlantic. However in situ sampling is limited, with nitrate only measured from mid 2002 to mid 2004 ~~with~~ and to 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~~averages) and is available for the full 10-year time series (see supplementary Fig. S5).

~~Distinct seasonality is simulated by the ensemble mean. High nitrate concentrationsat the surface occur~~ The PAP data show seasonality in both chlorophyll and DIN concentrations, with high DIN during winter (December-April)~~and~~ , and a decline in summer~~. From the seasonal mean DIN profile in~~ , Fig. 3(b). The averaged DIN profile peaks in February, with a spike of high DIN in September, as shown in Fig. 7(a)~~, the ensemble inter-quartile range shows later bloom peaks compared to~~ . At around

70 m the highest concentrations of chlorophyll occur in May-June, summarised in Fig. 3(b), similar to that in the surface. An inter-annual decline has been observed in the satellite-derived chlorophyll, (r = -0.21, $p < 0.05$), shown on figure 4(a).

The ensemble mean reproduces the seasonality in averaged DIN in Fig. 7(a), but with later peak concentrations in March, April, compared to in situ, and with a secondary peak in ~~situ, and mean DIN concentrations are high during months when both the ensemble and in situ DIN decline (between May to June), and show an earlier spike of DIN in~~ July instead of September~~, and therefore underestimating the increase of DIN between October to December. These shared errors make the ensemble spread (NRR=1.25) still too narrow for the phenological metrics.~~

~~Chlorophyll concentration~~ . The ensemble mean chlorophyll also has a seasonal cycle but with chlorophyll confined to shallower depths than in situ, summarised in Fig. 3(a) and (e). The ensemble mean chlorophyll starts to decline ~~at a depth of ~~ below 50 m, which also corresponds to the decline in the chlorophyll inter-quartile (between 25th and 75th percentile) range with depth, shown on Fig. 3(e). ~~Chlorophyll also shows seasonality, similar to that of DIN. The in situdata 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~~ At the surface, the ensemble peak chlorophyll occurs in May, similar to the in situ, although peak concentrations are higher than in situ. The decline in surface chlorophyll in the observation has been captured by six ensemble members ($U_t\rho_b\xi_bG_2$, $U_t\rho_b\xi_sG_2$, $U_t\rho_l\xi_qG_2$, $U_t\rho_q\xi_lG_2$, $U_t\rho_q\xi_qG_2$, and $U_t\rho_s\xi_lG_2$), although with weaker correlations (r = -0.14 ~~(e).~~ $\pm$0.06), $p < 0.05$).

~~Observed mean profiles of~~ In situ chlorophyll and DIN ~~are~~ vertical profile means fall within the ensemble range, with ~~ensemble spread on the narrow side, with~~ NRR= 1.20 and 1.25 for chlorophyll and DIN respectively. For ~~chlorophyll and DIN profiles,~~ both the surface and upper layer chlorophyll the ensemble median ~~shows the highest correlationand lowest RMSEand bias, compared to the default or the ensemble mean. High RMSEs in DIN occur from ensemble members that contain the U_4G_2 combinations, as shown in Fig. 13(j), which also correspond to high DIN mean (< 9 mmol m). High~~ is better than the default model, showing a higher correlation, a lower RMSE, and lower bias, against the in situ and satellite-derived chlorophyll values, Table 3. The highest chlorophyll profile RMSEs (> 0.62 mg m^{-3}) are produced from ensemble members that combine G_1 with ~~$\rho_1\xi_2$, $\rho_3\xi_3$, and $\rho_1\xi_4$~~$\rho_b\xi_l$, $\rho_q\xi_q$, and $\rho_b\xi_s$, and this also coincides with high chlorophyll ~~mean~~ profile concentrations (> 0.7 mg m^{-3}). ~~Surface chlorophyll 10-year meanand RMSEs~~ However for DIN the ensemble mean/median RMSEs are higher than in other regions, which is traced to be due to ensemble members that contain the U_tG_2 combination which has a particularly high DIN bias ($>$ ~~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~~ 9 mmol m^{-3}), as shown in Fig. ~~12(e) and (~~13(j).

~~When compared to~~ At the surface, in years 1998, 1999, and 2001, the satellite-derived ~~chlorophyll-a, the surfacechlorophyll 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 rangethis is due to the ensemble overestimating the satellite-derived chlorophyll (supplementary material Fig. S5), giving aslightly narrow ensemble spread (NRR~~ within the inter-quartile range, however in other years, it is well below the ensemble inter-quartile box limits, Fig. 4(a). The ensemble spread for inter-annual means has NRR of 1.26, and there is also an effect on the overall 10-year ensemble spread (see Fig. S5 in the supplementary material) with an 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, . High surface chlorophyll concentrations with high RMSE (>0.8 mg m^{-3}) are seen when combining U_h with $\rho_b\xi_l$, $\rho_a\xi_a$, 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. $\rho_b\xi_s$, see Fig. 12(e) and (j), which is consistent with the largest errors in the profile average values. However the low chlorophyll concentrations (< 0.4 mg m^{-3}) that are produced when combining U_t and G_2 in the profile averages, is not reproduced in the surface chlorophyll.

The

The ensemble range of surface chlorophyll annual mean is 0.7 mg m^{-3}. If However if we only allow one process function at a time to change (by changing functional form in one process at a time, whilst to change in each ensemble member, keeping the other processes at their respective default functions)with their default functions, the new 11 member ensemble range reduces to 0.58 mg m^{-3}, covering 84still covering 83% of the all ensemble members. If full ensemble (128 members). When the original MEDUSA parameters are used, the interannual chlorophyll fits the ensemble ensemble fits the inter-annual surface chlorophyll observations slightly better, but the DIN fit gets worse. The results from using MEDUSA parameters and the original MEDUSA parameters (maximum growth and grazing rates) and including in situ DIN concentration as initial condition concentrations as initial conditions can be found in the supplementary material, section sections 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), during peak chlorophyll in May the satellite-derived chlorophyll is outside the ensemble range.

3.2 Oligotrophic Ocean

In oligotrophic regions nutrients The oligotrohpic region is represented at stations ALOHA and BATS. The nutrients in this region are scarce at the surface but may be plentiful at deeper depths (Dave and Lozier, 2010; Lipschultz, 2001). Figure 9(f) and (l), 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 DIN , Fig. 9(d), with DIN levels High DIN levels ($>$ 1.0 mmol m^{-3}only found) are only found below $\sim$ 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 , shown in Fig. 9(d) and (j) for ALOHA and BATS respectively. The annual means of the averaged DIN profile in the top 200 m are 0.68 mmol m^{-3} and 0.80 mmol m^{-3} for ALOHA and BATS respectively. In station ALOHA, an increasing trend of in situ DIN inter-annual in situ averaged DIN profile ($r = 0.69$, $p < 0.03$) , which is not captured by the ensemble, has been observed, shown on Fig. 6(b), leading to an NRR of 1.38. This overestimation is also observed in the seasonal mean, Fig. 7. In the oligotrophic region, seasonality has not been observed in both chlorophyll and DIN. However, there are months of low chlorophyll (< 0.1 mg m^{-3}) months that have been observed in July-October, as shown in Fig. 5(b) , and in situ data is rarely within the ensemble range. and (c). Another feature of the oligotrophic ocean is a deep chlorophyll maximum (DCM) that occurs below the mixed layer, when

the chlorophyll concentration in the surface is low (Fennel and Boss, 2003; Letelier et al., 2004). At both stations, the DCM is observed between 70-150 m depth, and continuously occurs throughout the year.

At The ensemble mean has reproduced DIN concentration distribution in station ALOHA as seen in Fig. 9(c) and (d). The ensemble range decreases as the depth increases, with a high ensemble range found at depths between 3-50 m, Fig. 9(f) and (l). However, at BATS, the DIN concentration in the top 200 m is clearly overestimated, Fig. 9(i), with DIN levels has higher DIN concentration (> 1.0 mmol m^{-3}) at ~ 10 m (with some members occasionally showing such concentrations at 3 m). Consequently, the overall mean DIN concentration is overestimated, as indicated by the positive bias in the ensemble mean, in Table 3. Similar to ALOHA, DIN inter-annual and seasonal means are overestimated, summarised in . This consequently leads to a higher annual mean of (> 1 mmol m^{-3}), and overestimation in monthly, Fig. 7(c) and inter-annual variability, Fig. 6(c) and 7(c), respectively. This results in narrow ensemble spread (NRR= 1.40). As of averaged DIN profile, for all the ensemble members. This higher averaged DIN profile concentration has also been observed at ALOHA, an whereby both the ensemble mean and median have annual means of > 0.9 mmol m^{-3}. The increasing trend in DIN is observed ($r = 0.67$, $p < 0.03$) , but only has also been observed in 28.9% of the ensemble resultsmembers, which uses G_2 as its grazing functionshows similar trend. .

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

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.

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(e) and 9(k), high chlorophyll concentrations are simulated by the ensemble mean DCM has also been observed at both stations in the ensemble mean. However, the depths at which the DCM is simulated are slightly shallower, between 70-90 m in BATS and up to 150 70-150 m in ALOHA. A DCMoccurs when lower chlorophyll is detected at the surface, which roughly matches with the in situ profiles at ALOHA (see None of these stations show continuous DCM, shown on Fig. 9 (b) and Letelier et al. (2004)) and BATS (Fig. 9(h)) although the a) and (g) for ALOHA and BATS respectively. The depth of the ensemble DCM at both stations is 10-20 m shallower than in situ. The high subsurface chlorophyll coincides with a DCM coincides with the 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. 9after the DCM depth. At the surface during months with high chlorophyll (> 1 mg m^{-3}), the in situ concentrations are within the inter-quartile range box shown on Fig. 5(b) and 9(h).

(c). Although during months of low concentration, most of the ensemble mean shows even lower chlorophyll (as low as 0.045 and 0.022 mg m^{-3} for ALOHA and BATS respectively).

~~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 surfaceand integrated chlorophyll profiles. Figure~~ The DIN concentrations at both oligotrophic stations are mostly overestimated by the ensemble mean/median, and the opposite has been observed in chlorophyll. At BATS all of the ensemble members overestimate chlorophyll profile, surface, and integrated. From Fig. 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~~ a low chlorophyll profile means (< 0.08 mg m^{-3}) are produced from ensemble members that combine G_2 with $\rho_l\xi_l$, $\rho_l\xi_q$, ~~which then coincide~~ $\rho_l\xi_s$, $\rho_q\xi_h$, and U_e, which coincides 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~~ the ensemble mean/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~~ High chlorophyll vertical profile means at both stations ~~, although when paired with $\rho_3\xi_3$ and $\rho_1\xi_4$, the RMSEs increase~~are produced from ensemble members that combine U_hG_1 and U_tG_1. High DIN ~~concentrations are almost always observed when U_3 and U_2 were used in these oligotrophic regions~~profile means are produced when U_e and U_s are combined with any mortality functions, summarised in Fig. 13(a) and (b). Combining these uptake functions with G_2 will also increases the mean DIN concentration even further, and therefore increase the RMSE in station BATS. Nevertheless, since overestimation of mean DIN concentration occurs in most of the ensemble members, and therefore the ensemble mean, the NRRs are high for both ALOHA and BATS (NRR=1.38 and 1.40 respectively).

~~As for profiles~~At station ALOHA, surface chlorophyll ~~at ALOHA~~ (supplementary Fig. S2) ~~has lower RMSEs and higher 10-year mean concentrations compared to BATS~~RMSEs from the ensemble mean and median are lower compared to the chlorophyll profile. The surface mean concentrations from the ensemble mean/median are also closer to the in situ concentration, summarised in Table 3. Ensemble members with low surface chlorophyll ~~concentrations are the same as those with low chlorophyll profiles~~means are consistent with the profile averaged values, although high surface chlorophyll ~~RMSEs~~ errors also coincide with high surface ~~concentrations~~mean, 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~~ The default run in oligotrophic regions generally produces higher chlorophyll and lower DIN concentrations compared to the ensemble mean and median. Default run also matches better with in situ data as the correlation coefficient, r is higher. Nevertheless, the ensemble mean and median at ALOHA also have lower bias and RMSE for surface chlorophyll compared to the default. However, at BATS the default run shows better RMSE and bias. For integrated chlorophyll in the oligotrophic region, the ensemble mean and median have smaller RMSEs and a better correlation coefficient, compared to the default run. The NRR for the integrated chlorophyll is closer to 1 compared to the surface and chlorophyll profiles.

~~Figure 4(b) and (c), show no distinct inter-annual variability at either ALOHA or BATS, and in most years the defaultrun 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) 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 = has low NRR (0.84. This is broader), lower 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 if we only allow one process function to change in each ensemble member, whilst keeping the other processes with their default function, the new ensemble produces higher NRR (1.17), and the in situ annual mean 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, 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, 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 min May.

At ALOHA the ensemble mean and median produce smaller RMSEs for both chlorophyll and 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 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 , Coastal zones are represented by stations Cariaco and L4, with in situ observations show showing strong seasonality, shown on in Fig. 14(b), (d), (g), and (h). This has not been captured by the ensemble mean At Cariaco, highest mean DIN profile concentration (> 7.5 mmol m^{-3}) at the top 200 m is observed in March and July, and lowest (< 5.5 mmol m^{-3}) in November, see Fig. 7(d). At L4 high concentrations (> 8 mmol m^{-3}) are observed between November-February, with very low values ($\sim$0.1 mmol m^{-3}), during summer months (Smyth et al., 2010). The in situ profiles at Cariaco show high chlorophyll concentrations ($>$1 mg m^{-3}) within the upper 30 m between December-February (see Fig. 14(b)) coinciding with the rise of nitrogen from depth to $\sim$30 m, seen in station Cariaco , shown in Fig. 14(a), with chlorophyll concentration almost

~~constant~~ d), increasing the nitrogen concentration to $\sim$ 5 mmol m^{-3} from $<$ 1 mmol m^{-3}. Similarly at L4, Fig. 14(g), sharp peaks of chlorophyll are observed during spring (March-April) and fall (September), coinciding with the sharp decline of DIN between April and July (from ~6 in March to 0.22 mmol m^{-3} in July), shown on Fig. 7(e), resulting in an annual mean of 2.40 mmol m^{-3} and 1.20 mg m^{-3} for DIN and chlorophyll respectively. During non-bloom periods, chlorophyll is observed from 0.09-2 mg m^{-3}, with peak concentrations up to 6.4 mg m^{-3}.

At Cariaco seasonal variability is absent from the ensemble for both chlorophyll and DIN. There are only two years simulating downwelling of DIN, in 2001 and between 2005-2006, shown in Fig. 14(c). The chlorophyll concentration is almost constant (above 0.7 mg m^{-3}) in the upper 30 m and at the surface (supplementary Fig. S3), apart from a decline in concentration to $\sim$ 0.5 mg m^{-3}, followed by a sharp chlorophyll peak in the winter (December-January) in 2007, 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~~ was noted at Cariaco from 2004 (Taylor et al., 2012), and this is captured by the ensemble mean, median, and the default (r = -0.72, $p < 0.05$, r = -0.66, $p < 0.05$, and r =-0.35, $p <$ 0.05 respectively). ~~Similar to chlorophyll, 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 DIN decreases to 5.32 mg m. However in 2001, and between 2005-2006, inter-annual downwelling of nutrients is well reproduced, Fig. 14(e). Despite the lack of seasonality, annual means of chlorophyll and DIN, are mostly within the ensemble range, Fig. 6(f) and 4(f), with the NRR 0.78 and 1.15 for chlorophyll and DIN respectively.~~

~~At station~~ At L4 the ~~in situ and ensemble means both show seasonality of DIN with high concentration (> 8 mmol m) during November to February, and close to zero (> 0.1 mmol m), during summer months, consistent with the observation from Smyth et al. (2010). The interquartile~~ ensemble shows seasonality although the inter-quartile range often overestimates the surface DIN concentrations, especially during the sharp decline in ~~April, May, September, and October~~spring and summer, shown on Fig. 7(e). ~~For inter-annual means, in situ 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 DIN in spring. However this has not been represented in the model, where only one peak is simulated~~ The two bloom events per year in the observations are not represented and the ensemble only simulates one peak between May-June, summarised in Fig. ~~5~~7(e). ~~If~~ However if only diatom chlorophyll concentration is shown, the two bloom events are ~~clearer~~captured, especially in the default run (see supplementary material Fig. S4). The ensemble ~~mostly overestimates the in situ~~ also shows higher concentration range during non-bloom periods (~~in situ range = 0.09-2 mg m, ensemble range =~~ ensemble range from 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 Cariaco although surface chlorophyll seasonality is not well reproduced, the ensemble range is wide so that in situ concentrations mostly fall within it, apart from August and November, summarised in Fig. 5(d). The annual mean of surface chlorophyll and averaged DIN in the top 200 m are also within the ensemble range, Fig. 4(f) and 6(f), with the NRR 0.78 and 1.15 for chlorophyll and the averaged DIN profile respectively. At L4, despite reproducing seasonality, ensemble concentrations for DIN and chlorophyll are overestimated, and the NRR value for DIN and chlorophyll are 1.31 and 1.21 respectively,

summarised on Table 3. At both stations, for the inter-annual mean, the ensemble range always ~~includes~~ encompasses 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~~ and small RMSEs are found for ensemble mean surface chlorophyll at both stations, ~~and these are better, and with smaller RMSEs than~~ which are improvements over the default run~~, apart from DIN at L4~~. Similar to the oligotrophic stations, the integrated chlorophyll shows better correlation with in situ measurements, compared to both surface and chlorophyll profiles~~. At L4, the ensemble mean shows high RMSE for surface DIN, but low RMSE for surface chlorophyll,~~ , with results 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~~ The overestimation of DIN and chlorophyll are produced when the model uses $\rho_l\xi_q$, $\rho_h\xi_l$, $\rho_q\xi_q$, and $\rho_h\xi_s$ combinations as these functional forms produced high chlorophyll means for both stations ($>$ ~~1.7~~0.8 mg m^{-3} in Cariaco profile, and $>$ 1.5 mg m^{-3} in L4 surface), with higher RMSEs, especially when U_s is also used. Higher DIN concentrations in Cariaco ($>$~~1.2~~ 5 mmol m^{-3}) with high RMSEs ($>$ ~~1.5~~3.4 mmol m^{-3}) are also associated with the same ensemble members, summarised in Fig 13(c). ~~These~~ However at L4 these same ensemble members ~~however~~ show relatively low DIN concentration ($>$5 mmol m^{-3}) ~~at station L4. The chlorophyll mean at L4 shows that high concentrations~~ with lower RMSEs. From Table 3, at Cariaco in situ surface chlorophyll concentrations are slightly overestimated by the ensemble mean but most other ensemble outputs are underestimated, except for ensemble members that use the combinations above. Unlike the oligotrophic regions, these high chlorophyll concentrations in the coastal stations coincide with higher RMSE ($>$ ~~0.2 mg m) are produced when the model uses $\rho_2\xi_3$, $\rho_1\xi_2$, $\rho_3\xi_3$, and $\rho_1\xi_4$ combinations~~ 1.7 mmol m^{-3}). ~~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 DIN concentrations ($>$ 9 mmol m)are produced by U_4G_2, with correspondingly high RMSE.~~

Surface chlorophyll at these coastal stations also has a higher ~~relative range than~~ ensemble range than at other stations, ~~with L4 higher than Cariaco, see~~ summarised on 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~~ Cariaco inter-annual ~~means~~ mean chlorophyll (Fig 4(d) and (g)), ~~the NRRvalue for Cariaco is too small (~~has NRR=0.78~~), indicating the ensemble spreadis wider than necessary, which is also observed for inter-annual primary production~~ , ~~shown~~ indicating this wide ensemble spread. Inter-annual primary production also shows wider spread compared to ALOHA, in Fig. 10(b) and (c). However, if ~~processes are~~ process functions are only perturbed one at a time ~~, the NRR is~~ the reduced ensemble has NRR=0.90, closer to the ideal ensemble range~~(NRR= 0.90). On the other hand, at~~ . At L4 the in situ annual mean ~~is reliable (~~of the full ensemble has NRR=~~1.001), since~~ 1.00, with the in situ chlorophyll ~~is~~ close to the ensemble median (see Fig. 4(e)). ~~However if~~ If the ensemble is reduced ~~by only perturbing one process at a time~~ to single process perturbations the NRR increases to 1.36, and the in situ data is no longer within the ensemble ~~mean~~range, 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~~ There are some relatively small differences in the timing of phenological events between the ensemble mean~~,~~ /median, and the default run, ranging from a couple of days to a couple of weeks, as shown in Table 4. ~~The~~ However the timing of initiation, bloom peak, and termination show wide ~~interquartile~~ ensemble inter-quartile 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~~ For this reason at most stations the observed phenology metrics fall within the full ensemble range. ~~In terms of the timing of the bloom peak and termination, these are up to 3 months and 120 day late respectively.~~ The ensemble range also mostly encompasses the in situ peak and amplitudes, shown on Fig. 15(c).

~~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~~ At station ALOHA, the observed initiation times show 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~~ average 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 75th and 25th percentile range, making the ensemble spread too narrow (NRR=1.35). The peak~~ The chlorophyll at ALOHA~~shown in~~ , Fig. 5(b), ~~where high~~ shows peak highs ($>$ 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~~ . At BATS, the initiation occurs mostly between January and February, ~~also placing the bloom timing outside the inter-quartile range, see~~ although in 2002, the initiation occurred in October. Bloom peaks generally occur a month later, and the terminations vary between April and May, apart from 2002 when it was in December. The height of the peak and amplitude at ALOHA are 0.14 and 0.05 mg m^{-3} respectively. At BATS these metrics have slightly higher chlorophyll concentrations, of 0.60 and 0.28 mg m^{-3} for peak and amplitude respectively. The duration of the bloom at ALOHA is rather short compared to other stations, $\sim$ 50 days, whereas BATS is $\sim$ 90 days.

At ALOHA Fig. 15(a) shows the ensemble run having initiation times between late January and April instead and so the observations fall outside the ensemble range, Fig. 15(b)~~. At BATS~~ , and the ensemble does not show as strong inter-annual variability as the observations (supplementary Fig. S6). The ensemble at BATS has the largest range of phenological timings, especially in termination time, and this matches the observations better. For bloom initiation, the in situ timing is within the inter-quartile range and only three days earlier than the ensemble median. However, since the earliest peak occurs in mid-January in the ensemble, the ~~earliest initiationin 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,~~ Nonetheless, the 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 bloom time. Although the ensemble range of peak bloom time ~~in BATS is very~~ at BATS is high (174 days), the NRR ~~suggests that this range is still narrow (~~=1.17~~)~~, as the ensemble does not cover all the in situ ~~timing. The large full range is caused by some ensemble members blooming much later~~timings.

At ALOHA ensemble bloom peak and amplitude inter-quartile ranges encompass the observations, with ensemble mean and median being very close to the observation. However, ~~since~~ at BATS, the in situ ~~timing is earlier, it is not within most of~~ observations for peak and amplitude are outside the ensemble range, consistent with section 3.2. The observed bloom duration at ALOHA and BATS are within the ensemble range, ~~so the overall ensemble spread appears narrow.~~ although the inter-quartile range at ALOHA shows longer bloom durations. For bloom termination, both stations show later termination, with the ensemble mean being almost two months late and a month late for ALOHA and BATS respectively. However, at ALOHA, located at 22°N, the ensemble median for termination at the end of August agrees well with the observations from Racault et al. (2012).

~~Both coastal stations show~~ For coastal stations L4 and Cariaco, the 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).~~ with peak bloom timing in April for both stations. At Cariaco the mean peak height is 3.5 mg m^{-3}, with mean amplitude 1.15 mg m^{-3}, shown in Fig. 15(c) and (d). At L4, the mean peak height is slightly higher (3.6 mg m^{-3}), with a higher amplitude (1.64 mg m^{-3}). Both stations have nearly similar bloom duration, of 76 and 80 days for Cariaco and L4 respectively. This makes the termination times for both stations very similar, which happen in June.

The ensemble means show later initiation, with the 75th and 25th spanning mid April to end of May for Cariaco, and between early and mid May for L4. However, the overall ensemble range covers the observed initiation, in Fig 15(a). This later timing is also clear in peak bloom times for both stations, shown on Fig. 15(b)~~.~~, whereby in L4, the inter-quartile range of the bloom occurs mostly in June, and the ensemble range for Cariaco, between the end of May and early August. Consequently, the in situ observations for Cariaco and L4 both falls outside the ensemble range. Figure 5(e) shows ~~the in situ~~ that the bloom at L4 is simulated by the ensemble one to two months ~~overestimated by the ensemble . Cariaco is the~~ later. At Cariaco the ensemble peak height and amplitude reach less than half of the in situ values (mean= 1. 10 mg m^{-3}) and amplitude (mean = 0.38 mg m^{-3}), which makes the in situ concentration fall outside the inter-quartile range for peak height and amplitude. This underestimate of the peak and bloom amplitude, results in NRR of 1.40 and 1.39 respectively. At L4 chlorophyll peaks are within the inter-quartile range, and amplitudes are within the full ensemble range. The bloom duration at Cariaco is also overestimated (up to 143 bloom days) and this, along with the late initiation of the bloom, results in a three month late termination. 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. At L4 the duration of the bloom is within the ensemble range, however, since the initiation and bloom timing of the inter-quartile range is later than the observation, the inter-quartile range also shows later termination time.

~~Initiation timing is captured best at~~ At station PAP, ~~with the ensemble median's initiation averaging only eight days earlier than for~~ the initiation from the satellite-derived chlorophyll ~~, resulting in NRR=1.14 for initiation, closer to one compared to other stations. A~~ occurs in April, see Fig. 15(a). Although typical North Atlantic ~~bloom happens during~~ blooms happen in

spring (Raymont, 1980), ~~however most~~ most peak blooms at PAP occur in late May-early June, as shown in Table 4 and Fig. 5. ~~Later blooms are recorded~~ Additionally, a late bloom in September from satellite-derived ~~chlorophyll-a~~ chlorophyll occurred in 2005, ~~three months later than the average and much later than the ensemble mean and median, although the bloom timing~~ making the mean bloom timing fall in June. The peak height is observed to be 1.52 mg m^{-3} with an amplitude of 0.45 mg m^{-3}. The duration of the bloom 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~~around three months (95 days), putting the mean termination in July.

The observed initiation time at PAP is within the ensemble inter-quartile range, and the ensemble median is only one week earlier than the observations. However, ~~at ALOHA, located at 22N, the~~ due to the inter-annual variability, the observed bloom peaks occur about a month later than the ensemble mean and median, although the bloom timing is still within the ensemble ~~median for termination at the end of August agrees with the observations from Racault et al. (2012).~~ range, with NRR value of 1.31. The ensemble mean produced higher peak chlorophyll (2.03 mg m^{-3}) and therefore higher amplitude. This puts the satellite-derived chlorophyll at the lower end of the ensemble range. The termination for ensemble mean and median is two days later and earlier respectively than the observed termination. This puts the satellite observed duration time within the ensemble inter-quartile range and very close to the ensemble mean duration.

~~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, maximum peak= 7.7 mg m), and the ensemble reaches less than half of the in situ peak (mean= 1.2 mg m, maximum height= 5 mg m). 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 75th and 25th 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. 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 75th and 25th 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. This is done by generating its ensemble outputs of chlorophyll and 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~~ across 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~~ structural uncertainties in model formulation. Applying structural sensitivity in the 1-D framework has also allowed a large range of process variability to be explored for several different oceanographic regions, and with minimal computational cost. The results are compared with a single default run, and in situ observations at five oceanographic stations. 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~~ models, studies of structural sensitivity much more limited.

Our findings reveal that in all regions, the Holling Type II (G_2) grazing function lowers the chlorophyll concentrations especially at low concentrations, which has also been observed by Anderson et al. (2010) ~~at low concentrations~~. The nutrients respond in the opposite direction with enhanced DIN concentrations. This is expected as at low concentrations, using the G_2 function ~~would~~ will graze more phytoplankton, as shown on Fig. 1(b). Even ~~though fitting the curve in lower~~ when fitting for a phytoplankton concentration range similar to ~~the~~ oligotrophic regions (0.001-0.5 mmol m^{-3}) is applied, higher grazing rate ~~in~~ with G_2 is still apparent in lower concentrations (~~>~~ < 0.2 mmol m^{-3}).

Pairing G_2 with the linear (~~ρ_2~~ρ_l) mortality of phytoplankton, which constantly ~~removed the~~ removes phytoplankton regardless of ~~the phytoplankton concentrations~~concentration, will reduce the chlorophyll concentration even further; but the opposite ~~will happen~~ happens 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 forms. In contrast, the default phytoplankton (~~ρ_1~~ρ_b) and sigmoidal zooplankton mortality (~~ξ_4~~ξ_s) produce the highest chlorophyll concentrations in all regions, similar to the experiment from Yool et al. (2011). ~~Suppose~~ If we use less than half of the current maximum mortality for both ~~ρ_2 and ζ_2~~ρ_l and ξ_2, then the deviation in phytoplankton concentrations from the default run ~~would not be as apparent(for~~ is less apparent. For example, mean surface chlorophyll obtained from running an ensemble member with ~~$U_1 G_1 \rho_2 \zeta_2$ in station ALOHA~~$U_b G_1 \rho_l \xi_l$ at 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~~ up from 0.07 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~~

For nutrient uptake, the exponential ($U_3 U_e$) and sigmoidal (~~U_2) show inefficient uptake, as they~~ U_s) forms are inefficient as ensemble members which contain these functions produce low chlorophyll and especially high nitrogen (DIN) concentrations, as shown on Fig. 12(a), (b), 13(a), and (b)~~, especially in the oligotrophic region~~. This is very apparent in oligotrophic regions. Even though the functional forms have been optimised, the ~~most deviations are observed~~ largest deviations occur when nitrogen is < 1 mmol N m^{-3}, shown in Fig. 1(a). This deviation still ~~occurred~~ occurs 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 DIN and low chlorophyll.~~ However, the effect is not as noticeable ~~compared to~~ as when 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~~ Stations that produce high chlorophyll concentrations also have high ensemble range~~. For example, in station~~ , for example, at Cariaco where chlorophyll concentration is high, despite the discrepancy ~~between the in situ seasonal pattern and the ensemble, the~~ with in situ seasonality, the ensemble 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 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~~ and 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~~ close to 1. However for annual mean chlorophyll and primary production (Fig. 4(d) and 10(b)) the ~~large range makes the ensemble spread~~ ensemble spread appears too large. ~~The ensemble range is considerable even if the equation of~~ Even in the reduced 11 member ensemble where only one process is changed~~at a time, which accounts for at least~~ , the range still covers 80% of the full ensemble range ~~. This has been observed at all of the stations, summarised in~~ of the surface chlorophyll annual mean, see Table 4 and Fig. 8~~, emphasising~~ . This emphasises 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~~ The mismatch between seasonal patterns in the observations and the ensemble, eg. at Cariaco, is mostly caused by the physical dynamics. At Cariaco the upwelling of nutrients that feeds the phytoplankton is not well simulated by the vertical velocity. This emphasises that despite using the ensemble approach, a coupled-biogeochemical model is only as good as its physical model (Doney, 1999), as the physical components such as mixing and upwelling dictate the seasonal pattern, phytoplankton community structure, and primary production (Sinha et al., 2010).

At most stations, the ensemble mean produced lower RMSE and higher nitrogen ~~correlation~~ correlations with in situ compared to the default run, as shown ~~on Table 3, suggesting~~ in Table 3. This suggests that the structural ensemble ~~with a wide~~

range of predictions covering the in situ observations, is is also 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 , that has not been specifically tuned to one station. However for chlorophyll concentration the default run outperforms has higher correlation coefficient and lower bias than 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 contribute to this high bias, especially those which uses ρ_2 ρ_l and G_2.

Nonetheless, in stationsEven though at some 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 , RMSE is higher for the monthly means (Fig. 5(c))of , the in situ or satellite-derived chlorophyll values (during months of high chlorophyll) are within the ensemble range. Similar monthly-mean chlorophyll patterns have also been observed in For example at PAP, ALOHA, Cariaco and L4 (with some exceptions in summer monthmonths) (Fig. 5(b)), whereby the in situ chlorophyll in and satellite derived chlorophyll are 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 the 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 ecosystemse. g. importance of . For example the timing of phytoplankton blooms for affect the survival of zooplankton and fish larvae(Cushing, 1990), as observed by 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 observationin phytoplankton phenology against observations. In situ initiation, bloom timing, and duration in at L4 are earlier than in most of the ensemble members, although still lying within the ensemble rangeand some . Some ensemble mean timings (termination and peak bloom time) in this station 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), which makes the in situ timing falls in the lower end of the ensemble range. 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 the 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 can result in rapid nutrient turn-over, which results

~~in a single peak event~~leads to a single bloom peak, but if the selection is based on the stoichiometry of C:N, the nutrients would regenerate ~~slower, and therefore result in~~ more slowly leading to 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~~ blooms starting ~50 days late, and ~~in southern hemisphere , model estimates subtropical bloom start date~~ southern hemisphere subtropical blooms starting ~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~~ is typically >1 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~~ bloom timing, e.g. PAP and L4, are more than one month. Additionally, the whole ensemble range produced by this study ~~can help to provide~~ provides 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~~ biases between in situ observations and ~~large number of process ensemble~~ sometimes the entire ensemble of results, such as ~~in the oligotrophic stations for DIN and L4 for~~ the phytoplankton peak timings ~~, emphasise~~ at L4, emphasises that the inclusion of some missing processes, such as active prey selection, ~~and species would~~ may be needed to improve the performance of the model (Friedrichs et al., 2007; Kriest et al., 2010; Sailley et al., 2014)~~and~~ . Additionally 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)~~, that are used in MEDUSA. We did not include equations that allow such selection or species, as ~~in this paper we tried~~ this paper tries 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~~ can result in different plankton and nutrient dynamics. ~~Therefore, comparing~~ Comparing the performance of greater model complexity and the ensemble method ~~was~~ is beyond the scope of this study. Furthermore, the mismatch of the phenology between the ensemble and the in situ observation, such as that in station Cariaco, may largely caused by the physical input, which drives the upwelling and mixing process, therefore controlling the seasonal pattern of the phytoplankton (Doney, 1999; Sinha et al., 2010).

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. This study emphasises that small perturbations in the 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~~ an ensemble of a single biogeochemical model resulting from perturbing the model ~~structure~~processes, 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, 20120 500, 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~~ inter-annual 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_h) $\frac{n}{n+k}$	Sigmoidal (U_s) $\frac{n^2}{n^2+k^2}$	Exponential (U_e) $1-\exp(-\frac{n}{k})$	Trigonometric (U_t) $\frac{2}{\pi}\arctan\left(\frac{n}{k}\right)$
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_h,ξ_h) $\mu\frac{P}{P+k_M}P$	Linear (ρ_l,ξ_l) μP	Quadratic (ρ_q,ξ_q) μP^2	Sigmoidal (ρ_s,ξ_s) $\mu\frac{P^2}{P^2+k_M^2}P$
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	0.10
	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	0.19	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 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 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~~U_h). 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 DIN profiles from ensemble mean ((a) and (c) respectively), in situ observations ((b) and (d) for chlorophyll and DIN respectively), and 75^{th} and 25^{th} quartile range of concentrations at each depth ((e) for chlorophyll and (f) for 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 75^{th} (top) and 25^{th} (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 75^{th} (top) and 25^{th} (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, 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 DIN concentration from all depth is used instead.

For station L4, in situ 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 75^{th}(top) and 25^{th}(bottom) quartiles. Red line is the median, and the whiskers are the ensmeble minimum and maximum of the averaged DIN. In station L4 and PAP data for DIN is only available from 2000-2007 and 2002-2004 respectively.

Figure 7. 10-year monthly mean of DIN averaged over 200 m from all the study sites ((a)-(e)), showing the seasonal dynamics of DIN (mmol m^{-3}). For station PAP, the DIN shown is the overall profile, and in L4, the in situ 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 75^{th} (top) and 25^{th} (bottom) quartiles. The red line is the median. The whiskers are the ensemble minimum and maximum mean of 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 DIN concentrations at oligotrophic stations. Station ALOHA is shown on (a)-(f) and BATS is shown on (g)-(l). White solid line in (a) and (g) represents mixed layer depth. (e), (f), (k), and (l) are the 75^{th} and 25^{th} percentile range of chlorophyll ((e) for ALOHA and (k) for BATS) and DIN ((f) for ALOHA and (l) BATS) over the depth. The range is obtained by averaging the chlorophyll and DIN concentrations of each ensemble members over the time series at each depth. Black dots in (e), (f), (k), and (l) are the mean of the ensemble. Ensemble mean chlorophyll profiles (shown on (a) and (g)) and DIN ((c) and (l)) are obtained from all of the ensemble members. in situ chlorophyll are shown in (b) and (h), and DIN are shown in (d) and (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~~,U_b,U_s,U_e, and ~~U_4~~U_t) 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 DIN (mmol m^{-3}), at five stations from all ensemble members. The first panel ((a)-(e)) shows 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~~,U_b,U_s,U_e, and ~~U_4~~U_t) 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 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 (c) show chlorophyll and DIN ensemble mean at Cariaco respectively. White solid line in (a) is the mixed layer depth. (e) and (f) shows the 75^{th} and 25^{th} percentile of chlorophyll and 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 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 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 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

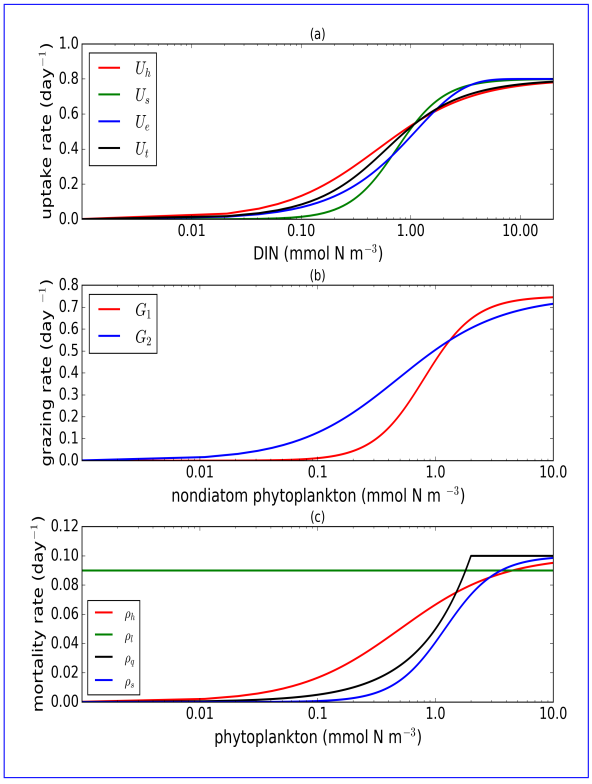

Figure 1. Nearly identical curves which describes resource uptake (a), zooplankton grazing (b), and phytoplankton mortality (c). Figure (a) shows four uptake functions, which have been optimised to the default uptake function, monod (~~U_1~~$\underset{\sim}{U_h}$). Figure (b) shows two grazing functional 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.

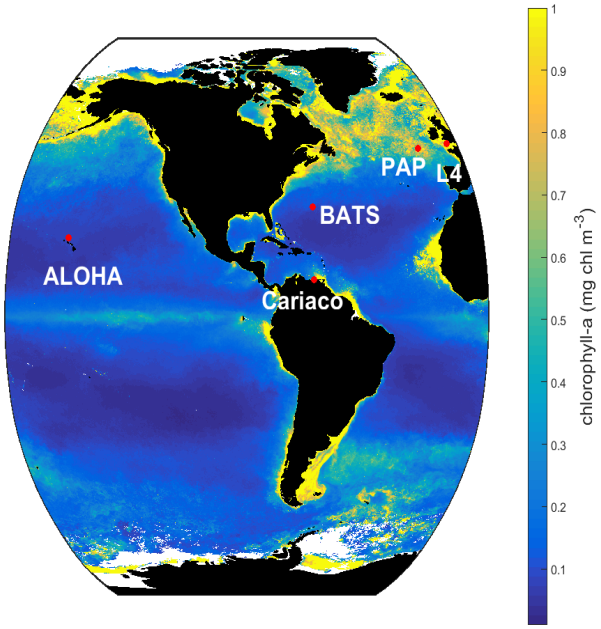

Figure 2. SeaWIFs-derived mean 1998 chlorophyll-a (mg m^{-3}) overlain with the 5 oceanographic stations time series site (Red dots). These stations are located in different oceanic regions: oligotrophic (ALOHA and BATS), coastal (L4 and Cariaco), and abyssal plain (PAP).

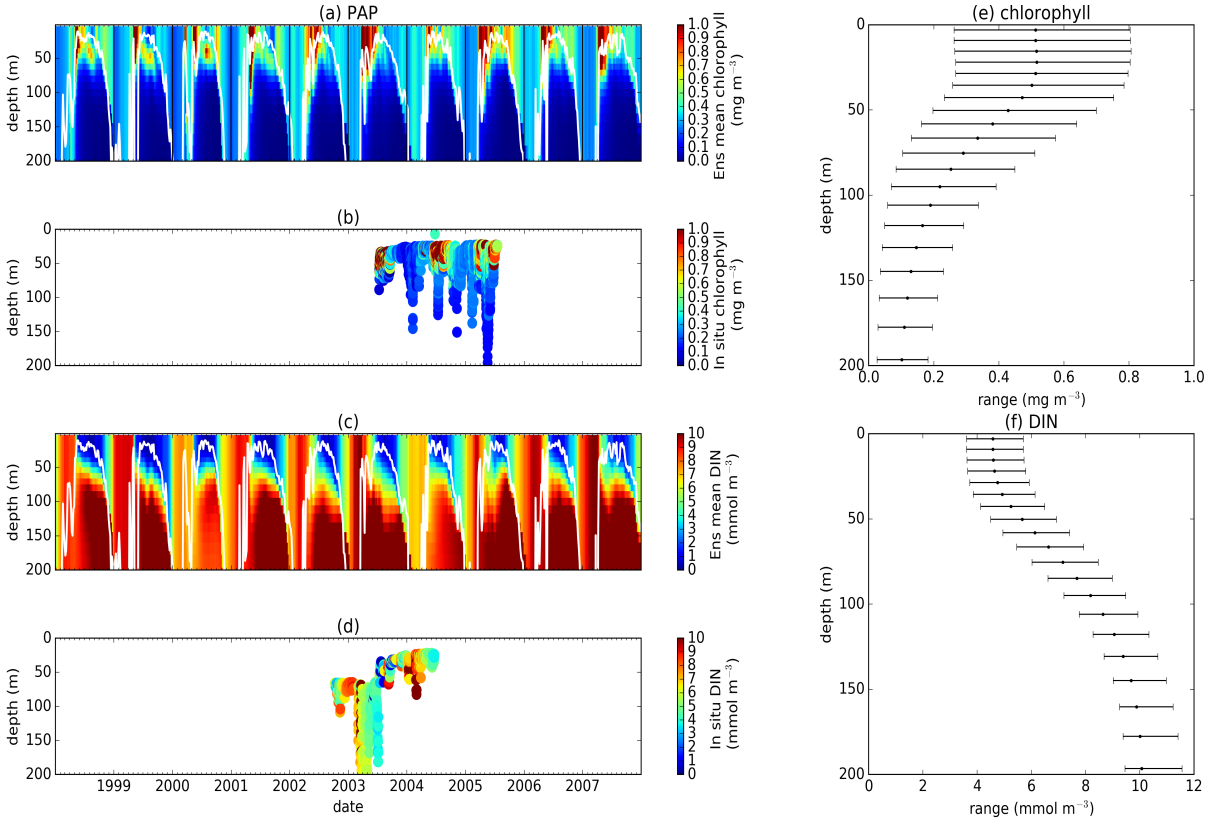

Figure 3. Chlorophyll and DIN profiles from ensemble mean ((a) and (c) respectively), in situ observations ((b) and (d) for chlorophyll and DIN respectively), and 75th and 25th quartile range of concentrations at each depth ((e) for chlorophyll and (f) for DIN) at station PAP. The range are obtained 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.

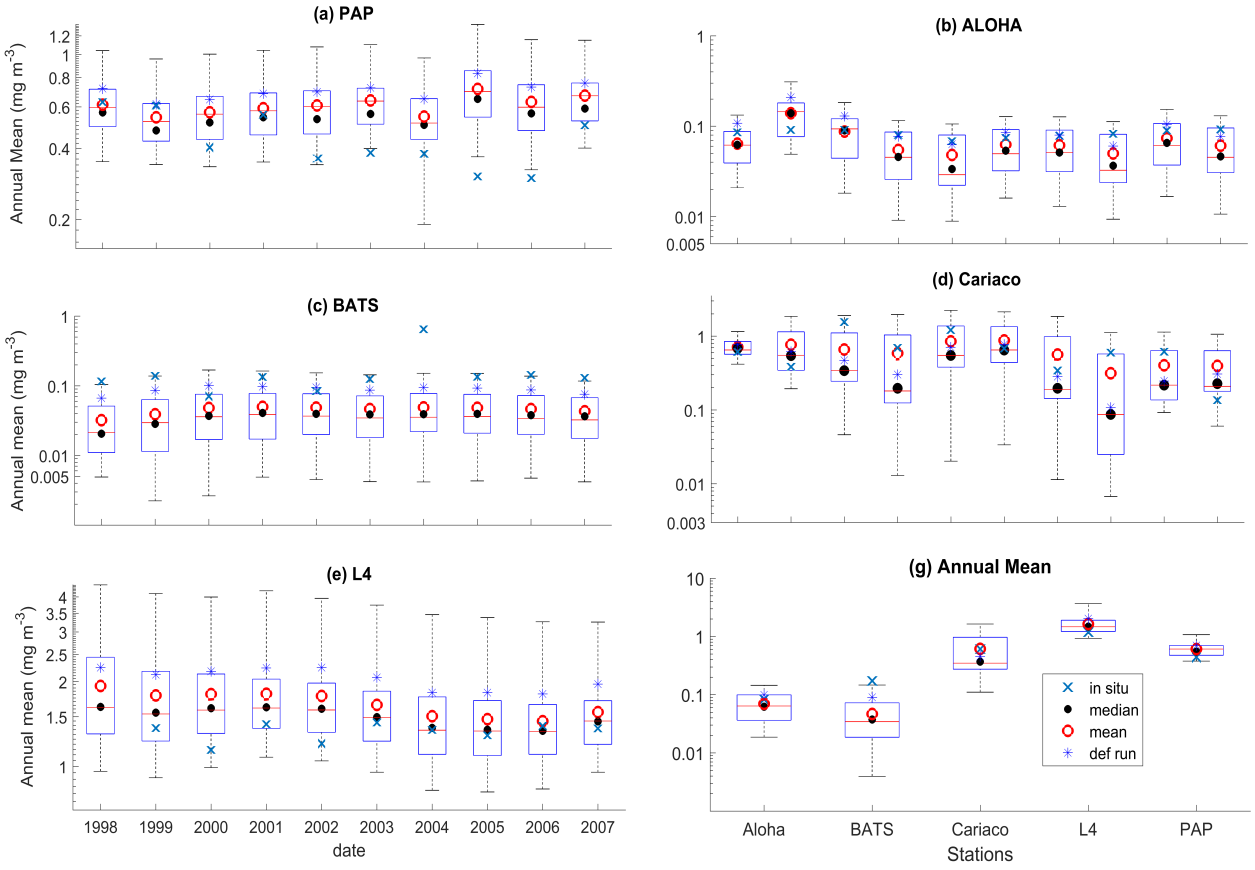

Figure 4. Inter-annual mean of surface chlorophyll from all the study sites ((a)-(e)) and the 10-year annual mean (g), all measured in mg m^{-3}. The boxplots show the ensemble annual means. Blue cross is the in situ observation, red open circle, black dot, and blue stars are the ensemble mean, median, 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.

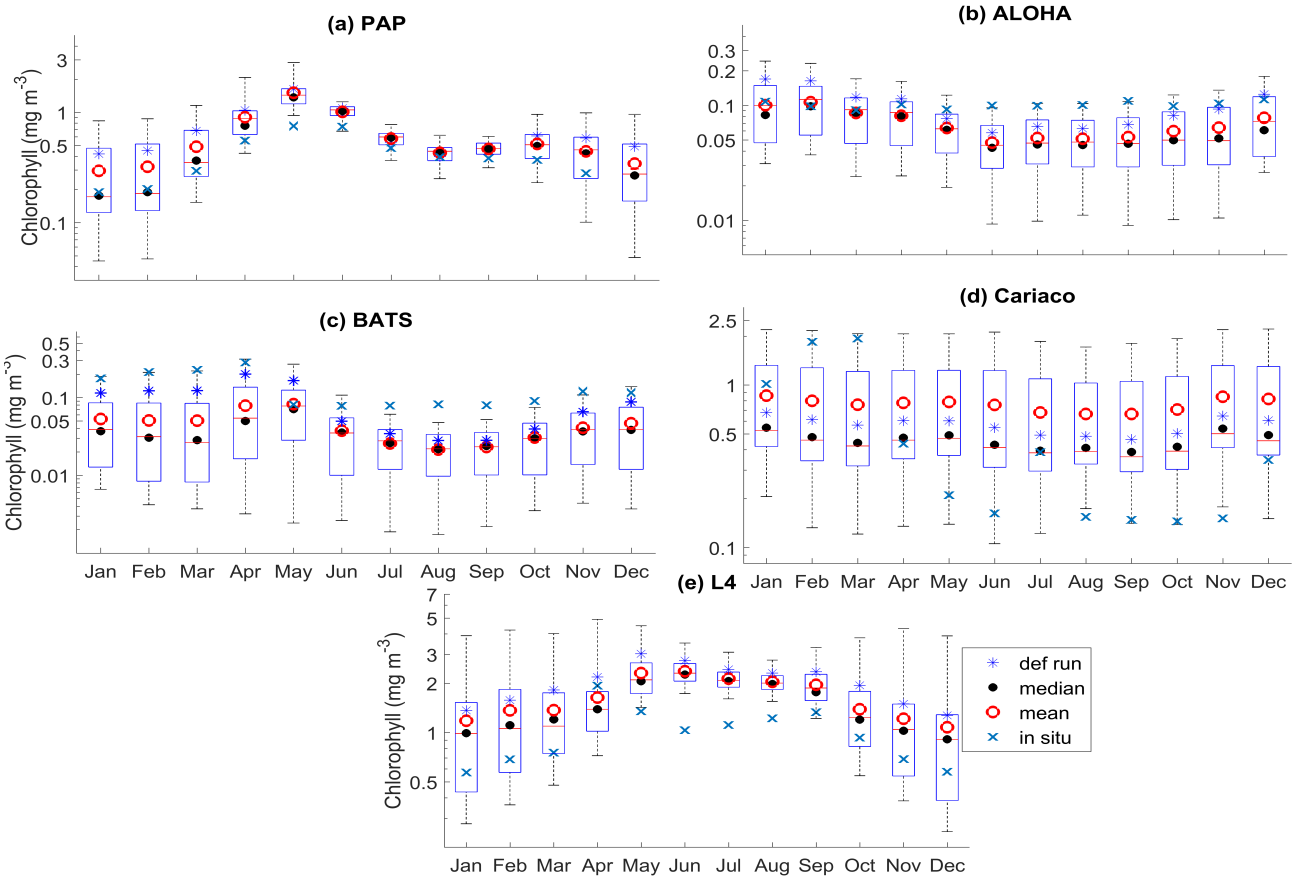

Figure 5. 10-year monthly mean surface chlorophyll from all the study sites ((a)-(e)), showing the seasonal dynamics of surface chlorophyll (mg m^{-3}). The boxplots show the ensemble seasonal means. Blue cross is the in situ observation, red open circle, black dot, and blue stars are the ensemble mean, 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.

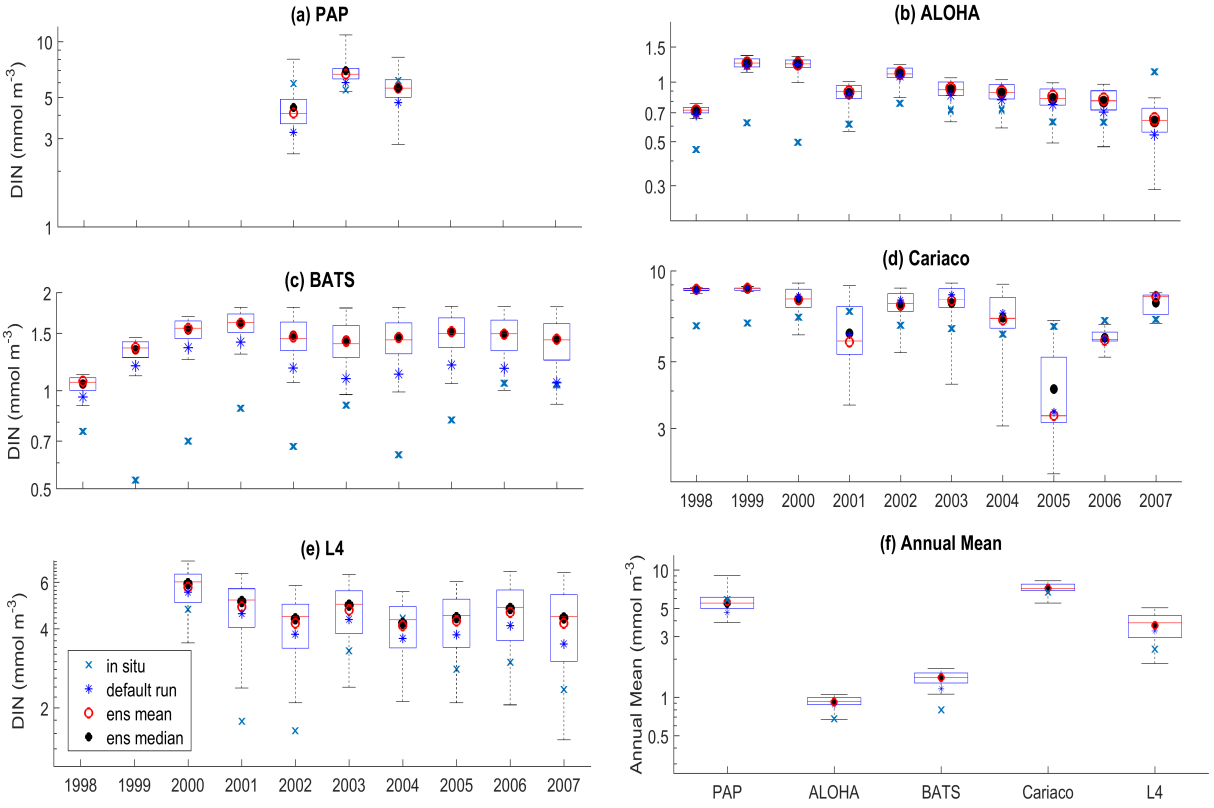

Figure 6. Inter-annual variability of DIN averaged over 200 m, from all the study sites ((a)-(e)), and the annual mean (f). Since the in situ data for PAP does not always cover the first 200m, the overall mean DIN concentration from all depth is used instead. For station L4, in situ DIN is only collected 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 averaged DIN. In station L4 and PAP data for DIN is only available from 2000-2007 and 2002-2004 respectively.

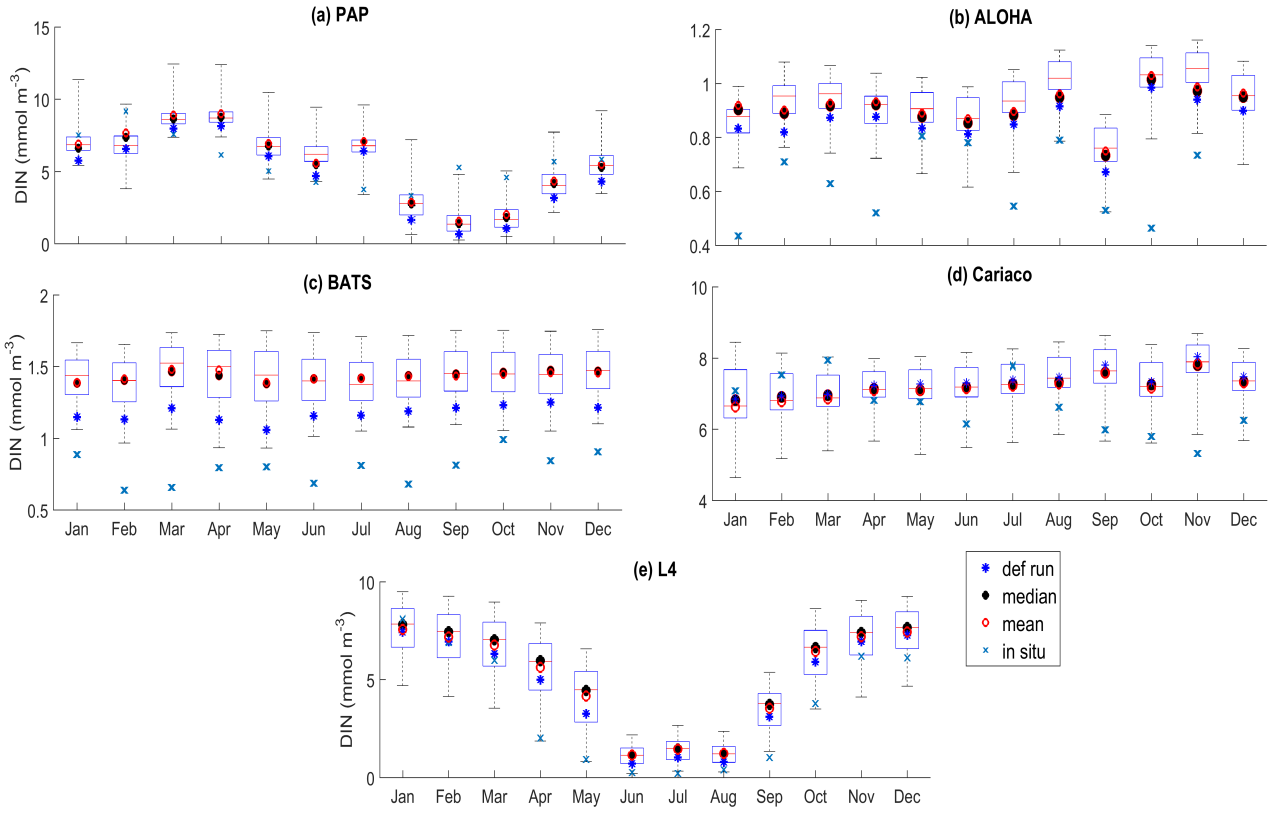

Figure 7. 10-year monthly mean of DIN averaged over 200 m from all the study sites ((a)-(e)), showing the seasonal dynamics of DIN (mmol m^{-3}). For station PAP, the DIN shown is the overall profile, and in L4, the in situ DIN concentration is only available at the surface. The boxplot shows the ensemble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

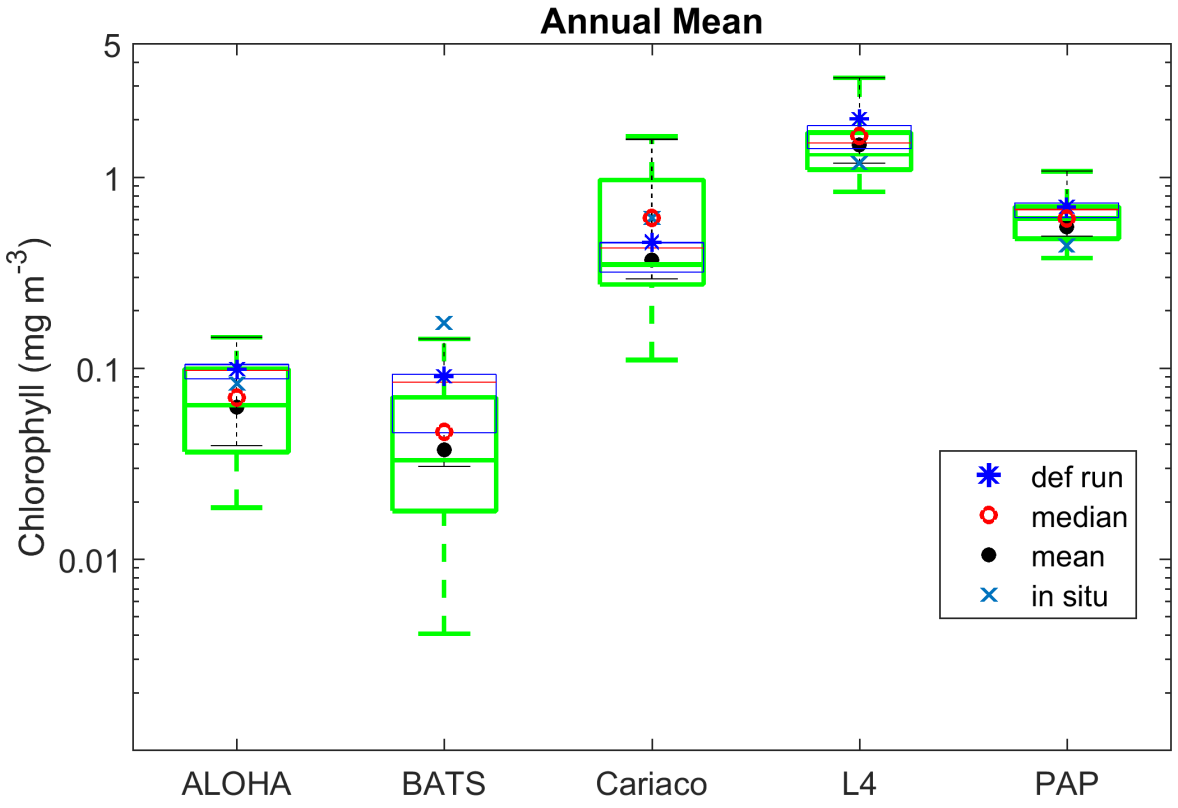

Figure 8. Annual mean of surface chlorophyll when changing only one process at a time (blue box), overlain with annual mean of all ensemble members (green box) at five oceanographic stations. Ensemble mean and median plotted in the figure (shown in red open circle and black closed circle), are the from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

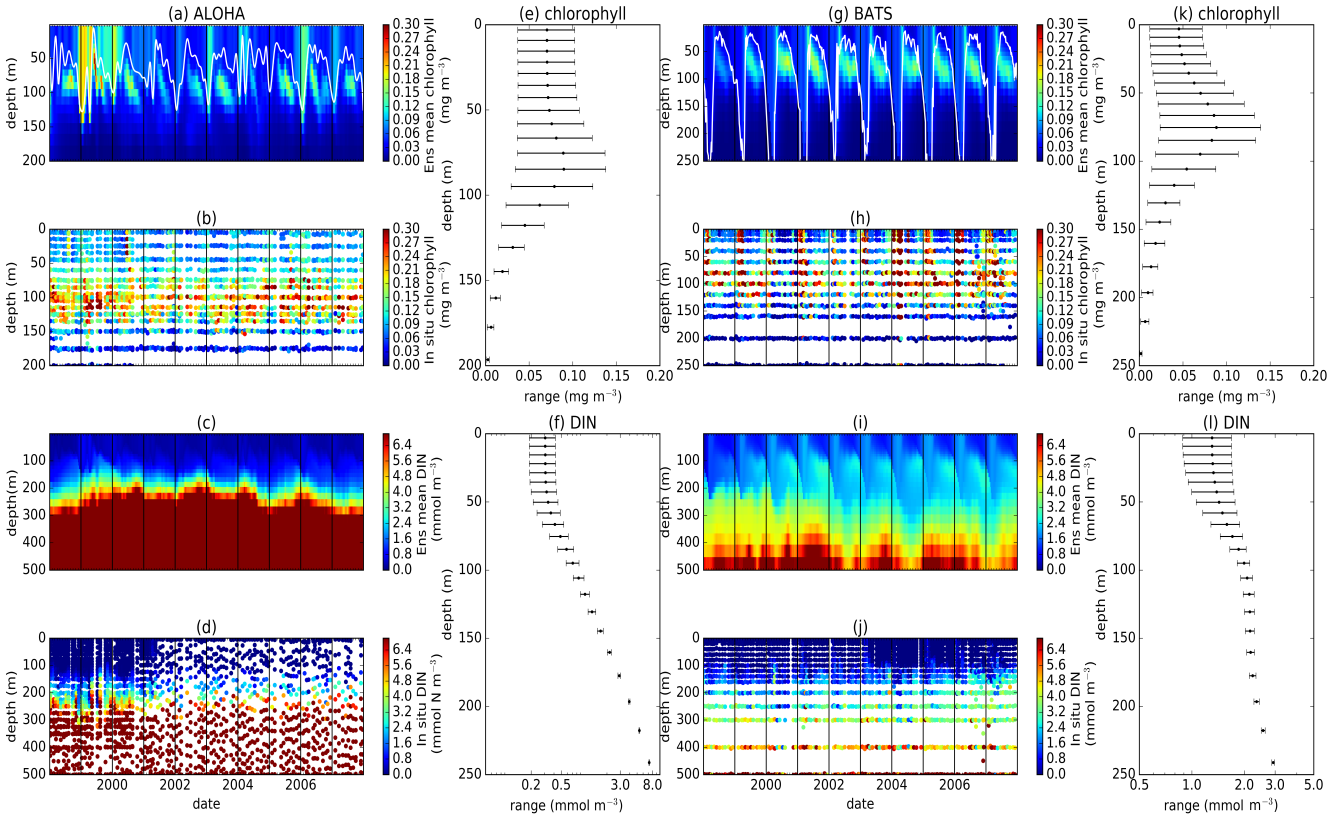

Figure 9. Time series (from January 1998-December 2007) of ensemble mean and in situ, and range of chlorophyll and DIN concentrations at oligotrophic stations. Station ALOHA is shown on (a)-(f) and BATS is shown on (g)-(l). White solid line in (a) and (g) represents mixed layer depth. (e), (f), (k),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

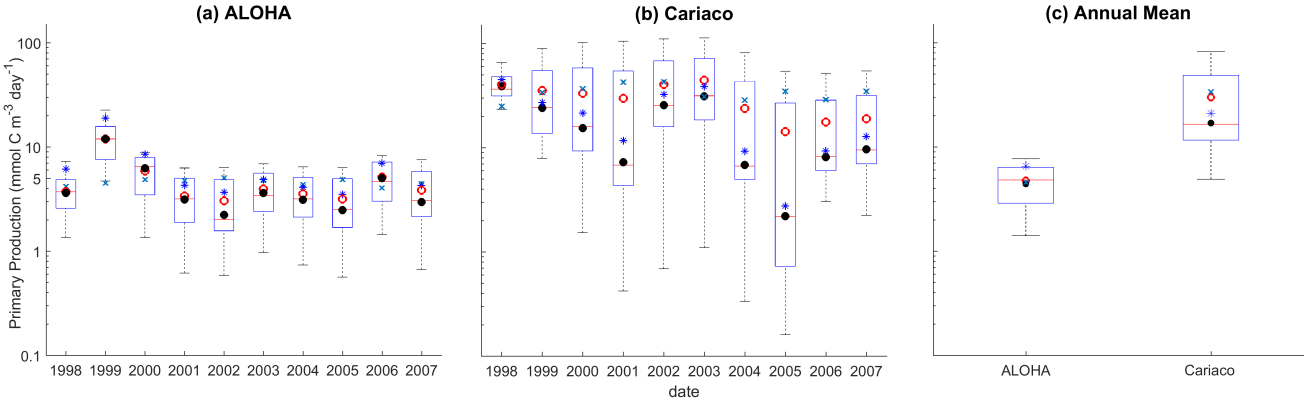

Figure 10. Mean integrated primary production averaged over 200m that are available in (a) ALOHA and (b) Cariaco, and (c) the annual mean. The NRR for ALOHA and Cariaco are 1.12 and 0.80 respectively.

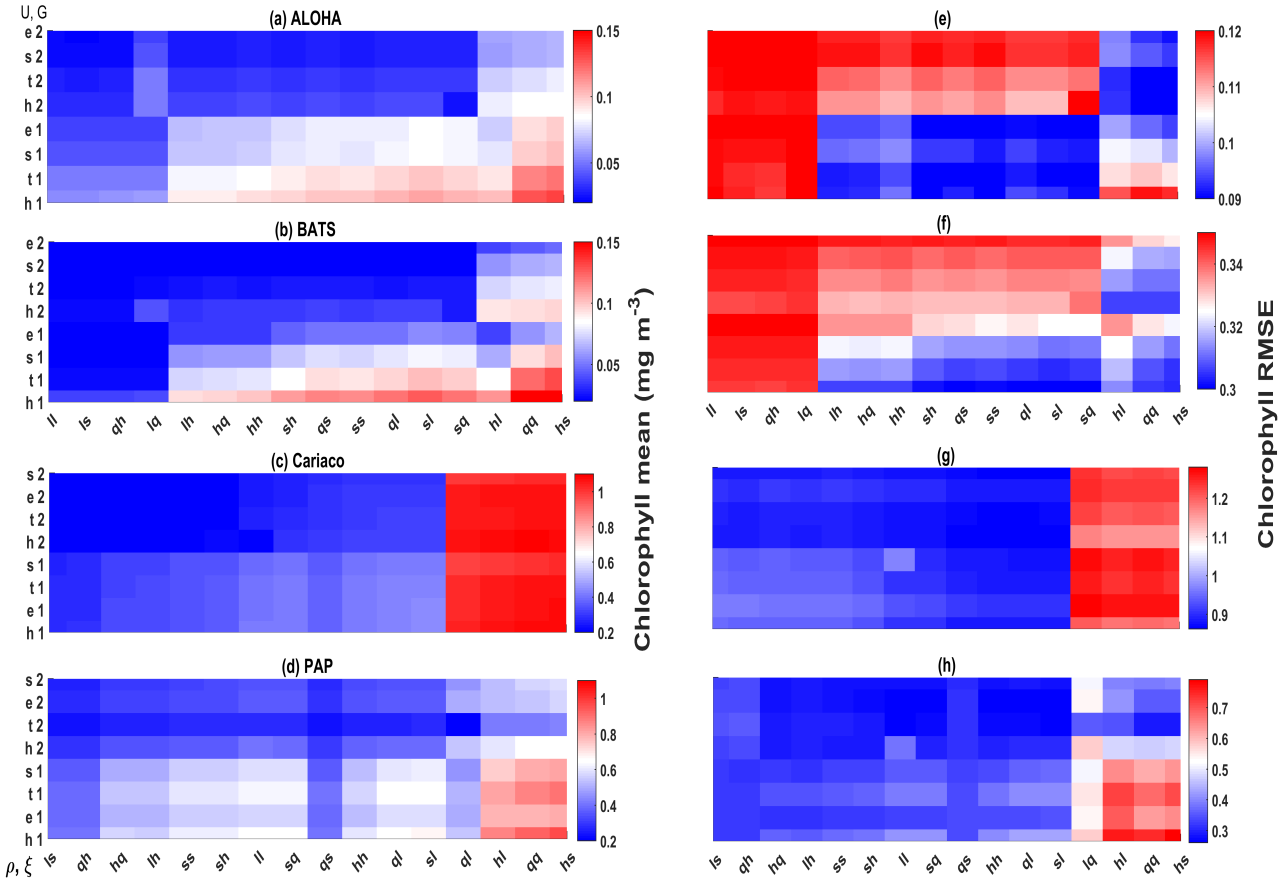

Figure 11. Chlorophyll profile 10-year means ((a)-(d)) and its RMSEs ((e)-(h)) at four oceanographic station from all of the ensemble members. Station L4 is not included as chlorophyll data is only taken at the surface. These are arranged by the lowest chlorophyll (top left) mean to the highest (bottom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

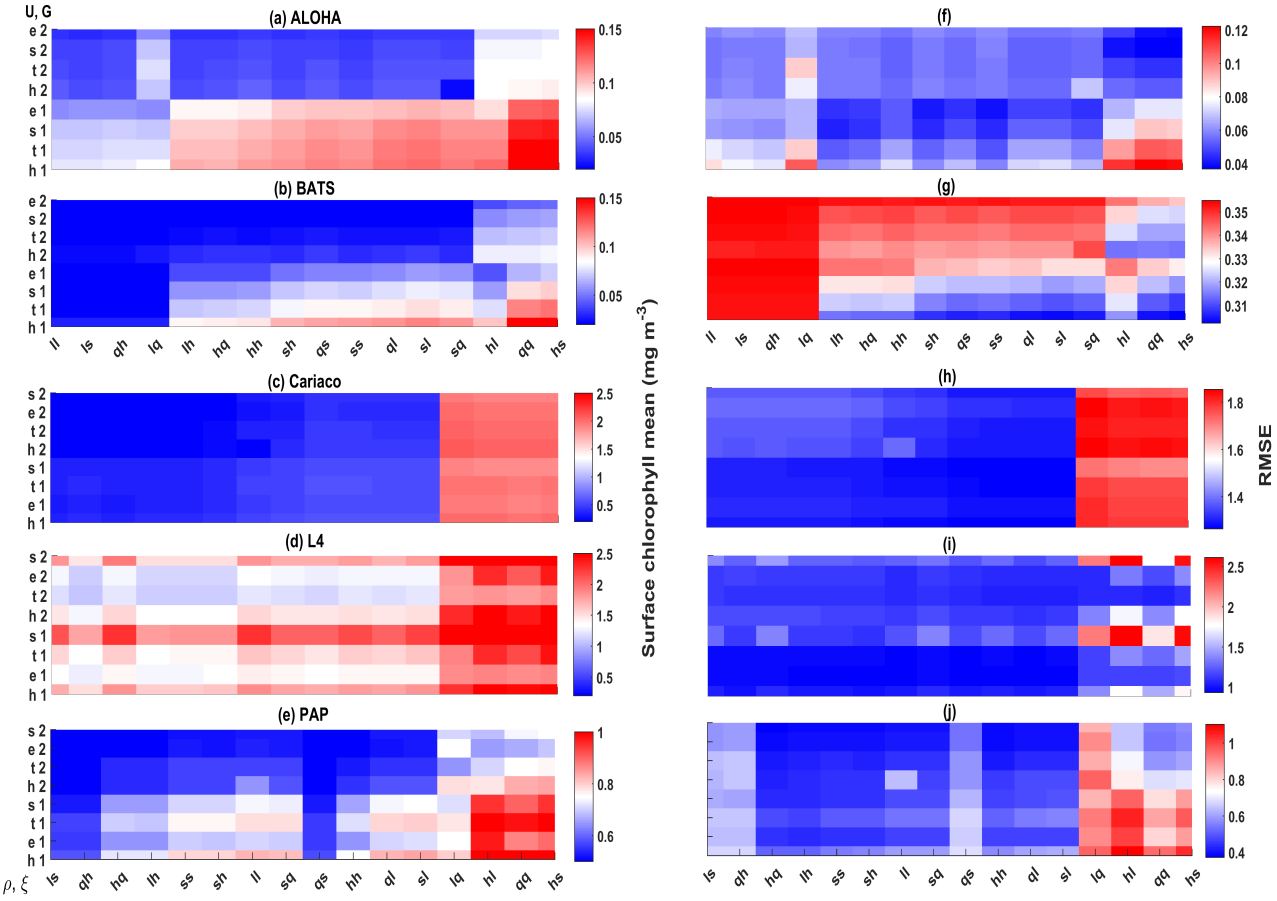

Figure 12. 10-year mean and RMSE of surface chlorophyll (mg m^{-3}) at five stations from all ensemble members. The first panel ((a)-(e)) shows surface chlorophyll mean and RMSEs are shown on the second panel ((f)-(j)). Concentrations and RMSEs are arranged by the lowest chlorophyll (top left) mean to 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 (~~$U_1, U_2, U_3,$~~ $U_h, U_s, U_e,$ and ~~U_4~~ U_t) and grazing (G_1 and G_2), and x-axis shows combinations of phytoplankton (ρ) and zooplankton (ξ) mortalities.

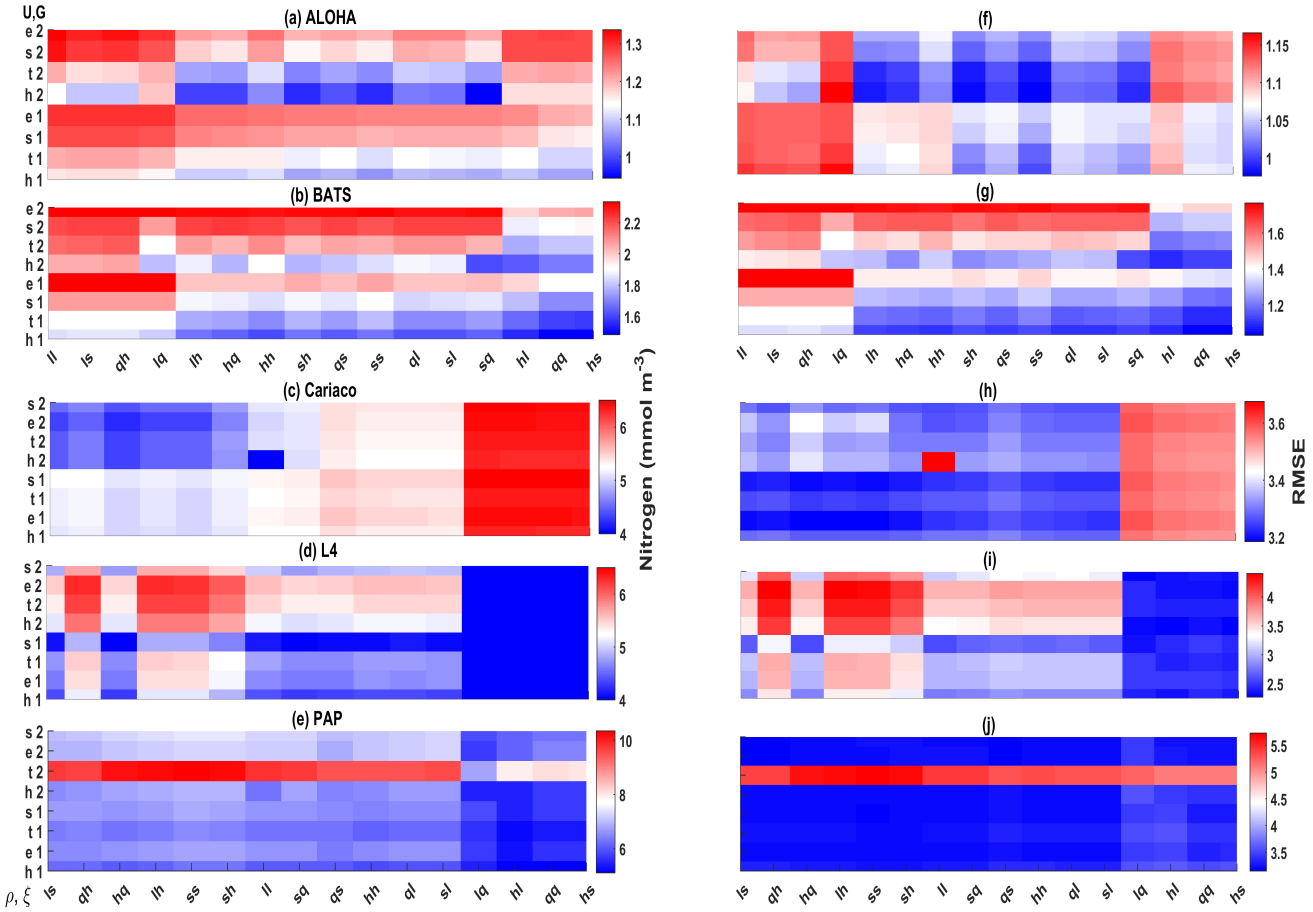

Figure 13. 10-year mean and RMSE of DIN (mmol m^{-3}), at five stations from all ensemble members. The first panel ((a)-(e)) shows DIN mean and RMSEs are shown on the second panel ((f)-(j)). Concentrations and RMSEs are arranged by the lowest chlorophyll (top left) mean to the highest (bottom right), 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~~,U_h,U_s,U_e, and ~~U_4~~U_t) and grazing (G_1 and G_2), and x-axis shows combinations of phytoplankton (ρ) and zooplankton (ξ) mortalities.

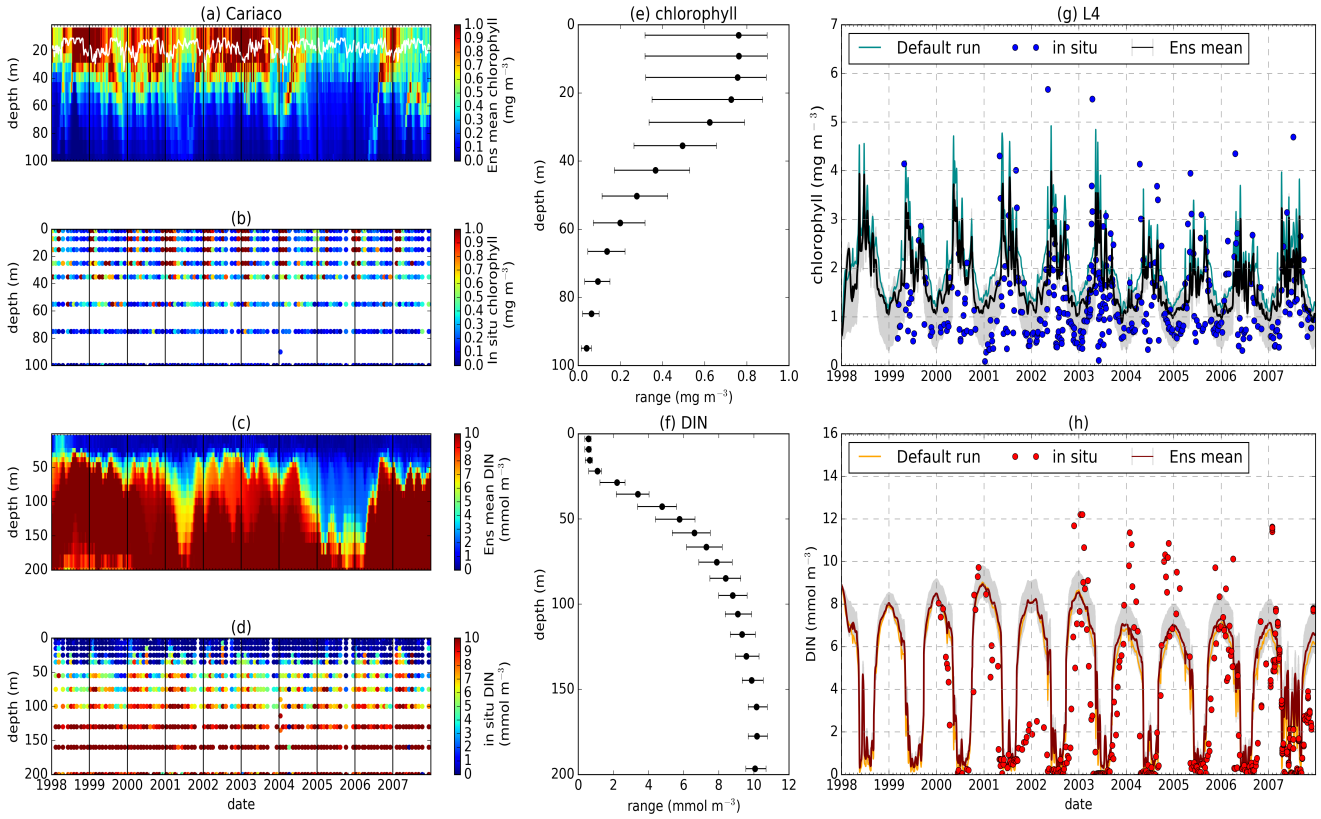

Figure 14. Time series of chlorophyll and DIN profile of ensemble mean, their range, and in situ concentrations at the coastal stations Cariaco (a-f) and L4 (g-h) from January 1998-December 2007. (a) and (c) show chlorophyll and DIN ensemble mean at Cariaco respectively. White solid line in (a) is the mixed layer depth. (e) and (f) shows the 75th and 25th percentile of chlorophyll and 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 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 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 DIN respectively.

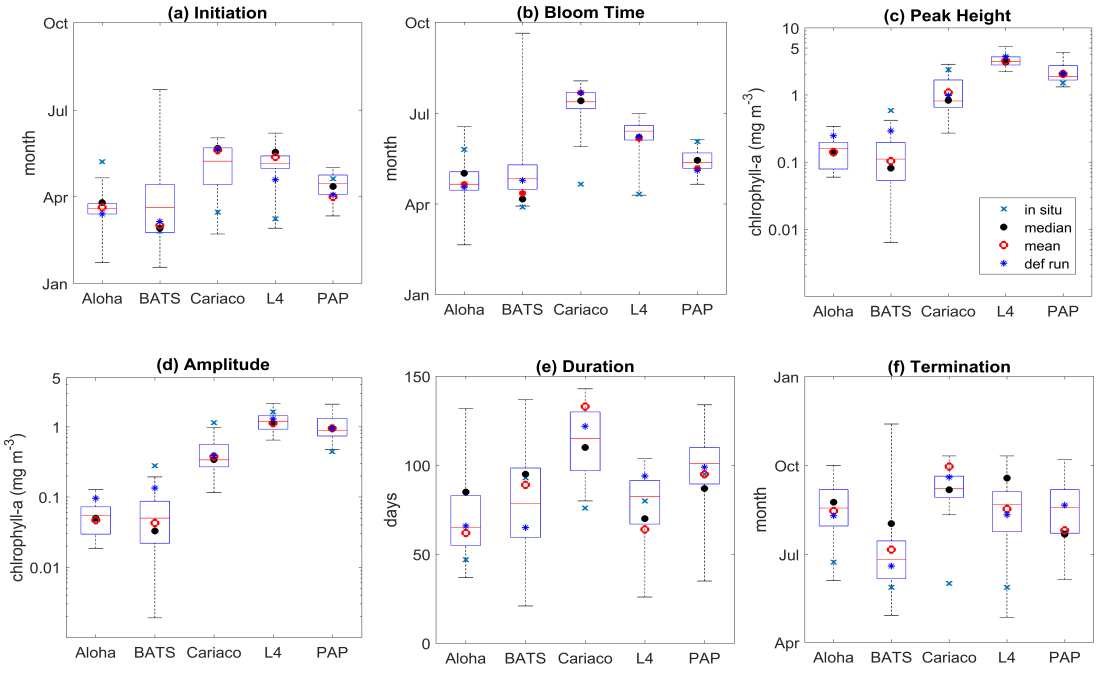

Figure 15. Phytoplankton phenology metrics at the five stations. Blue cross is the in situ, red, black, and blue dots are the ensemble mean, median, and the default run respectively. The timings and concentrations are averaged annually from January 1998 to December 2007.

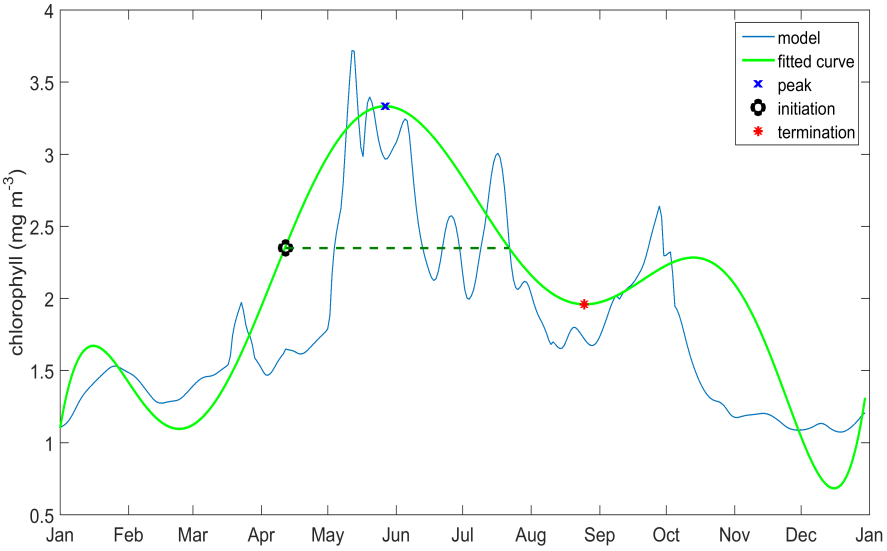

Figure A1. Determining phenology using a combination of threshold method and curve fit at station L4, here the initiation is when the fitted curve is above 50% of the maximum peak, however the termination is on the first valley.