Author responses to comments by Markus Schartau and Referee #2 are given in Sections 1 and 2 respectively. Reviewer comments are shown in italics. Line numbers given in the descriptions of manuscript changes refer to the marked up manuscript included at the end of this document.

1 Response to comments by Markus Schartau

Abstract, lines 1-3: This very first sentence is incorrect, as the parameters do not compensate for missing complexity. Rather, the plankton ecosystem models rely on parameterizations whose parameters have to represent a mixture of diverse plankton species that are pooled together and attributed to single state variables of plankton biomass.

The phrase “compensate for missing complexity" is perhaps unclear and could give the wrong impression regarding the role of the parameters. However, the aggregation of species that the reviewer refers to is only one aspect of the abstraction required in process models that prevents a direct mapping between model parameters and real-world biological quantities. We therefore defer from using the description suggested and instead talk about complexity in generic terms.

Manuscript change: Phrase re-worded to read “rely on adjustable parameters to capture the dominant biogeochemical dynamics of a complex biological system" (abstract, line 4).

Page 6329, lines 10-13: biota is part of nature! A plankton ecologist would disagree when stating that the biogeochemical models are mechanistic.

The term “mechanistic model" is widely used in the literature to describe process-based models, including biogeochemical models, because such models explicitly represent mechanisms by which the system operates. It does not imply that the modelled system is mechanical in any way. Nevertheless, “process-based" is probably a better term here and we have now used this instead, but we retain the use of the term “mechanistic" to describe our emulator. The term

"mechanistic" we feel emphasises more clearly the distinction between the type of process-based approach used here and the statistical approach more typically used in emulators. Also, its use avoids the need to use the phrase "process-based site-based emulator" which is a little awkward. It must of course be recognized that the process models are not purely mechanistic; they are semi-empirical because of the adjustable parameters. They are actually part of a spectrum from fully mechanistic to fully empirical. We have clarified this in the introduction. The term "dynamical emulator" was considered as another alternative but statistical emulators can be dynamical too, if they include a model of time dependence, so this was rejected.

Manuscript changes: Changed "mechanistic" to "process-based" (page 3, line 8); added text "The process-based models are often referred to as mechanistic, as distinct from statistical or data-based models. Yet they are also semi-empirical, incorporating adjustable parameters. Such parameters are important in process-based models of complex systems where incomplete knowledge and practical limits on the degree of complexity that can be resolved make it impossible to design a model that represents all relevant mechanisms." (page 3, lines 18 – 22).

6329, line 20: "...un-modelled..." I think is better to write "... unresolved..."

The sentence is now redundant.

Manuscript change: Sentence deleted (page 3, line 16).

6331, line 27 – 6332, line 10: The work of Nerger and Gregg (2008, Journal of Marine Systems) is a good reference here.

Although the ocean-colour data assimilation study of Nerger and Gregg (2008) is related in some ways to the other references cited, we deliberately restrict our text to parameter estimation studies. Nerger and Gregg's study focuses on bias-correction in a state estimation context, rather than estimating the parameters of the biogeochemical model so is not strictly relevant

here.

6334, lines 11-16: This paragraph should be revised according to the refined structure of the paper.

The structure of the paper has changed at a level which does not affect this paragraph so no revision is needed.

6334, line 19: I would rather talk of an “approximated", “simulation" or “emulation error", not "predicted error" in ff.

We are referring to an estimate of the probability distribution of the error. It is a prediction in the sense that it predicts a quantity that could subsequently be determined by running the target model. Agreed, the term “predicted error" is incorrect as it is the statistical description that is the prediction. The text has been clarified.

Manuscript changes: “statistical description of its predicted error" changed to “probabilistic prediction of its error" (page 8, line 14); “description of the predicted error" changed to “prediction of the error" (page 9, line 21).

6335, Equation (1): is not needed if Equation (2) is the one of primary interest for the authors.

We present Equation (1) as a generalization of Equation (2) to emphasize the importance of considering the covariance in future work, a point that we pick up in the discussion.

Manuscript change: Text has been changed to indicate explicitly that estimation of the mean and variance only is just a first step (page 9, line 25).

6335, Equation (2): The terms "R" and "P" are later again used but with different meaning. Therefore, a different terminology should be used. Actually "R" is part of Equation (1) and is not really needed here at all.

It is useful to retain the use of "$\mathbf{R}$" and "P" here for consistency with the cited reference (Stow et al, 2009). Later use of "$\boldsymbol{P}^{\star}$" is useful for consistency with Hemmings and Challenor (2012) so we are reluctant to change this either. We feel that the difference in notation between the scalar "P" and the vector "$\boldsymbol{P}^{\star}$" and the localized use of each in the text is sufficient to avoid confusion. We have now avoided re-use of "$\mathbf{R}$".

Manuscript change: "$\mathbf{R}_1$" and "$\boldsymbol{R}_2$" replaced with "$\mathbf{Q}_1$" and "$\boldsymbol{Q}_2$" respectively (page 22, Equation (16) and line 12).

6336, line 6: The full covariance is not used in the study. A reference to the covariance matrix "R" is not needed.

It is useful to retain this because it is referred to in the discussion (Section 5.2).

6336, lines 22-23: sentence is somehow clear to me but could be better formulated.

Text revised.

Manuscript change: This and the next sentence have been simplified and combined to read "1-D integrations of MEDUSA are performed in a 3-D context where physical and biogeochemical information from the target model provide environmental input data for the site-based simulations. " (page 10, lines 15 — 17).

6336, lines 25-28: I know what the authors want to say but the sentence is too difficult to understand. Since this is an important explanation at this point, the authors should consider a

better explanation.

To simplify the sentence we now avoid introducing the individual components of the biogeochemical environment here and also avoid introducing the 3-D ensemble. These are introduced later. This simplification allows a clearer focus on the key point that the biogeochemical environment, unlike the physical environment, is parameter-dependent.

Manuscript change: Text changed to read "The physical environment required by the 1-D simulator is independent of the biogeochemical model parameters. However, the biogeochemical environment is parameter-dependent making its representation in a site-based parametric analysis less straightforward." (page 10, lines 15 – 24).

6337, line 20 – 6338, line 26: The concept of distinguishing between primary and secondary tracers is fine, but it in the text it appears more complicated as it actually is. To write about a subset of tracers is confusing. Perhaps: "The perturbation term represents the effect of horizontal flux divergence. It is only applied to those state variables that determine nutrient concentrations and the biomass of the plankton. Other state variables (secondary tracers) like chlorophyll-a concentrations follow the flux corrections of the corresponding biomass (primary tracers). This way chlorophyll-a-to-nitrogen and biogenic silicon-to-nitrogen ratios are maintained."

The paragraph has been simplified, essentially following the reviewers suggestion but retaining the definitions of j for relating the description to Equation (3).

Manuscript change: Text changed to read "Tracer-specific perturbations are applied to tracers representing dissolved nutrients and the nitrogen content of the plankton. These are referred to as primary tracers. The phytoplankton chlorophyll and silicon tracers (secondary tracers) are affected indirectly, following the perturbations to the corresponding nitrogen tracers in such a way as to preserve the phytoplankton chlorophyll : nitrogen and silicon : nitrogen ratios. For a

primary tracer, $j = i$. For a secondary tracer, j indexes the relevant primary tracer." (page 11, line 21 – page 13, line 2).

6339, line 10: I would talk of "additional variability due to lateral advection effects" instead of "rate of change due to lateral advection of the transformed tracer."

The reviewer's comment indicates a problem with clarity of presentation. The purpose of the text was to introduce the use of transformed tracers, so the process being represented, i.e. lateral advection, was secondary. The original text reads "The primary tracer perturbations are applied in transformed concentration space. This means that the applied perturbation represents the rate of change due to lateral advection of the transformed tracer." In the new text, clarity has been improved by discussing the need for concentration dependencies before introducing the idea of transformed tracers to satisfy this need. (The phrase "rate of change due to lateral advection" is no longer used).

Manuscript changes: Re-arranged section of text. New section (page 13, line 12 – page 14, line 22) introduces transformed tracers with "This concentration dependency is introduced by using applied perturbations that represent rates of change of transformed tracers." (page 14, line 14).

6339, Equations (5, 6, and 7): Here a single equation would be enough. If Equation (6) were already included in Equation (5), then it would be more obvious why the factor of two is needed. Equation (7) is not needed because this had been explained before (see comment above).

It is useful to keep original Equations (5) and (6) separate because otherwise there would be no clear definition of $p^\star$, which is an important variable that describes elements of the environmental input vector for the simulator. It is also useful to retain original Equation (7) to ensure a complete mathematical description. However, there is an issue of clarity in the way

these equations were presented.

Manuscript changes: The text has been re-ordered so that the original Equation (5) is now given last to avoid breaking up the description of how the concentration dependency works in the simulator. New text, that now follows text justifying the use of the square root transformation, reads “A square root transformation was therefore chosen for all primary tracers at all sites so that a perturbation $p^\star$ specifies the rate of change of $\sqrt{c}$, where c is the tracer concentration. The implied concentration tendency is then

$$p = 2\sqrt{c}p^\star.$$

For secondary tracers the tendency is

$$p_i = \frac{c_i}{c_j} p_j$$

where i is the secondary tracer index and j indexes the associated primary tracer. The applied perturbation diagnosed from 3-D model output is

$$p^\star = -\boldsymbol{u}_\mathrm{h} \cdot \nabla_\mathrm{h} \sqrt{c}$$

where the subscript h denotes vectors in the horizontal plane and $\boldsymbol{u}_\mathrm{h}$ is the current velocity." (page 14, line 24 – page 15, line 12).

6340, line 1: "... it must increase towards zero..." This is not clear. Do you mean "... must be set zero..."?

No, it is not set to zero. The emphasis has been changed to improve clarity.

Manuscript changes: Changed “If we have a negative advective tendency it must increase towards zero as the concentration approaches zero to avoid the concentration becoming negative." to read “If we have a negative advective tendency it should increase towards zero as the concentration approaches zero, otherwise the concentration will become negative." (page

14, lines 7 – 9).

6340, line 15: Some operational oceanographers would disagree as it is not really "predictive skill". Better: "mapping" or "emulation" skill.

The sentence is redundant following reorganisation of Subsections 2.2 and 2.3 as described below.

pages 6340-6348: Subsection 2.2 "Informed and uninformed simulators" and Subsection 2.3 "The uninformed emulator". These subsections definitely require revision. First of all, the authors talk of an "uninformed simulator" first and then use the term "uninformed emulator". Many details in these subsections are confusing. Although correct it is difficult to follow as a reader. Here are my suggestions: • start with the description of the "uninformed emulator" since this is the ultimate emulator; introduce Equation (9) here and explain "uninformed simulator/emulator residual" and "uninformed simulator/emulator error". • introduce Equation (11) and explain the residual. I am not happy with the term "parametric environment residual". Better "parameter induced residual" • then derive Equation (12) from the combination of Equation (9) and (11). • now explain Equation (13). The equation is formally correct but cannot be assessed exactly. You therefore consider an informed emulator and derive its error information instead. Equation (8) would then become Equation (14). And "ϵ_S" would be set equal to "ϵ_1" of the former Equation (8). • At this point the different concepts of direct (with reference to the relevant Equation shown before) and indirect uncertainty quantifications are introduced and it is explained how they are computed

We agree that the clarity is greatly improved by delaying the introduction of the informed simulator until it is required for the estimation of ϵ_S. The structure has been changed accordingly but differs from that suggested by the reviewer. The sequence is: start with the description of the uninformed simulator; introduce the original Equation (9) and explain uninformed simulator residual/error; introduce the uninformed emulator; introduce the direct and indirect UQ

methods; describe the direct UQ computation procedure; describe the indirect UQ computation procedure, including original Equations (11), (12), (13) and (8), Equation (8) being the definition for the informed emulator.

The new structure differs from the reviewer's suggested sequence for two reasons. Firstly, we start with the uninformed simulator rather than the uninformed emulator because it is the simulator that is fully described by Equation (9). The uninformed emulator is then introduced as an extension, leading naturally to the introduction of the UQ methods. Secondly, Equations (11) - (13) are specific to the indirect UQ procedure and we feel it is not useful to introduce them in advance, out of context. The reasoning is the same as the reason for delaying the introduction of the informed simulator until the description of the indirect UQ procedure.

We retain the term "parametric environment residual" since it is important to indicate that the residual is related to the prescribed environmental input and is parameter-dependent. It is not really a "parameter-induced residual" as the reviewer suggests, rather the associated error is an environment-induced error as a consequence of the environment not being parameter-specific.

Manuscript changes: Removed informed simulator description (page 14, line 24 – page 16, line 4); renamed Subsection 2.2 from "Informed and uninformed simulators" to "The uninformed simulator and biogeochemical environment model"; inserted definition of informed simulator and its residual in "Indirect method for uncertainty quantification" section (page 20, lines 11 – 18); swapped notation for informed and uninformed simulator residuals ϵ_1 and ϵ_2 throughout the manuscript to avoid them being out of numerical sequence on first appearance. The meaning of the term "parametric environment error" is clarified by the addition of the text "Reliance on $\overline{\boldsymbol{B}}$ introduces a parameter-dependent source of environment-induced error into the simulation. The resulting contribution to simulation error is referred to as the parametric environment error." (page 18, line 26 – page 19, line 2). Removed later text describing the parametric environment residual, now redundant (page 19, lines 11 – 16).

Figures 1 and 2: These figures are elaborate already are helpful, but they would become even better if the most important mathematical symbols (as used in the corresponding equations)

are also added to the boxes.

Agreed.

Manuscript changes: Figures changed as suggested. Note that the path for informed emulator evaluation in Figure 2 is now redundant and has been removed.

6345, line 20: "For each data set..."? This is not clear enough. Why constructing it only for each data set? I thought that it has to be derived for all (primary) tracers. Please clarify.

It is indeed derived for all primary tracers. The data sets we refer to are multivariate. The phrase "each data set" refers to the initial state data set and the lateral flux perturbation data set. Each data set includes all relevant tracers. This is now clarified.

Manuscript change: "For each data set ..." changed to "For each of these two data sets ..." (page 21 line 18).

6346, line 2: "... containing the n available instances..." Shouldn't it mean " ... containing the m available instances...". If not, then it is needed to better explain the elements of the Y3D matrix.

The original text is correct. m is the length of the state vector. Now clarified.

Manuscript change: Included the text "(m is the number of elements in $\boldsymbol{S}$)" at page 21, line 28.

Equation (16): this is very nice and helpful! But, here R1 and R2 are not covariances; it could be confused with Equation (1). This comment is obsolete if the authors have agreed on removing Equation (1).

Agreed. Equation (1) was retained so alternative symbols have been substituted.

Manuscript change: "$\mathbf{R}_1$" and "$\boldsymbol{R}_2$" replaced with "$\mathbf{Q}_1$" and "$\boldsymbol{Q}_2$" respectively (page 22, Equation (16) and line 12).

6346, line15: Why is p=5 and not 4 or 6?

The choice of p value was largely subjective. No attempt was made to find an optimal value. The impact of using 4 or 6 EOFs instead would be expected to be small but was not tested.

6348, line 16: Add how large the ensemble was here.

Added.

Manuscript change: "small" changed to "10 member" (page 24, line 9).

6353, line 7: "... log-transformed 5 day mean..." Do you mean: five day mean of log-transformed or are have five day means been log-transformed? Please explain.

The five day means have been log-transformed. Text clarified.

Manuscript change: Added the phrase "applied to the 5 day means" (page 29, line 7).

6353, lines 8-10: "The log transformation..." If log-transformed variables are used then the variance does increase with increasing chlorophy-a concentration.

The reviewer's assertion regarding the error variance is not correct. Variance in untransformed chlorophyll concentration tends to increase as the concentration increases in such a way that the relative error in chlorophyll is much less variable than the absolute error. Consistent with

this, variance in log-transformed chlorophyll does NOT show a general tendency to increase as log-transformed chlorophyll increases. The error variance is thus stabilized.

Figure 3: The figure hardly resolves the seasonal variations. The upper scale should be set to 10 instead of 100; this might already help.

Agreed.

Manuscript change: Figure 3 modified as suggested.

6354, line 17: "Results are shown for ... with and without lateral flux perturbations." I suggest that the comparison with/without flux perturbation can be removed from the results section. Figure 4 can be omitted and Figure 6 (a very nice figure) can be discussed and shown in the discussion section. Also, I recommend considering only three sites from Figure 5 for the discussion section. It is good to have additional figures in discussion sections.

The reviewers suggestions are helpful and we have re-structured the manuscript accordingly, although not following the suggestions exactly. The informed simulator results shown in the original Figure 4 have been omitted as suggested and the with/without flux perturbation comparison has been moved. We prefer to keep the Figures in the results section but have introduced a new results subsection just before the discussion. Given this choice, it is then appropriate then to retain all sites in the original Figure 5.

Manuscript changes: Removed original Section 4.1 “Simulator performance with known parameter specific environment" (including Figure 4). Renamed original Section 4.2 “Emulation with parametric uncertainty in the biogeochemical environment" to become Section 4.1 “Emulator prediction of target model output" (page 31). This section has been re-worded to take the place of the initial results section and to omit text comparing uninformed simulator results with and without flux perturbations. The original Figure 6a that does not include flux

perturbations is no longer included in this section. Removed text comparing the uninformed and informed simulator results (page 34 line 22 – page 35 line 4). Added new Section 4.3 "The importance of lateral advection" (page 40) comprising the comparison of uninformed simulator results with and without flux perturbations that was omitted from the new Section 4.1 (including original Figure 6a, now Figure 8) plus relevant results from the original Section 4.1 (including original Figure 5, now Figure 9).

6355, lines 6-9: The results of parameter sets 10 and 6 are surprising. From parameter set 6 I would expect an early increase in biomass due to the high light-sensitivity. Together with the low half-saturation coefficients the net growth rates should be fairly high and I wonder why this is not expressed in the solution of 6. Parameter set 10 has some of the highest loss rates apart from grazing. Thus, the reduced grazing pressure might be the main reason for parameter set 10 to reveal enhanced net growth conditions.

The results shown are for the third and fourth years of the 4 year parameter perturbation experiment and can be counter-intuitive because of the time scale. The results for Parameter Set 6, for example, can be explained by nutrient depletion occurring as a consequence of strong phytoplankton growth early in the experiment. Since the results for this parameter vector are particularly extreme and relevant to the analysis, we have included an explanation.

Manuscript change: Added text "With this parameter vector, the phytoplankton light-response controlled by α_{Pn} is exceptionally strong and nutrient-limitation is reduced by low half-saturation concentrations $k_{\mathrm{N,Pn}}$ and $k_{\mathrm{Fe,Pn}}$. As a result, the phytoplankton can achieve very high growth rates. This can cause blooms that lead to long-term nutrient depletion as a consequence of organic material sinking out of the euphotic zone. Subsequent growth is then inhibited. At 4 sites (5, 10, 30 and 35N), nitrogen depletion during the 2 year spin-up period results in very low chlorophyll concentrations at the start of 1997 which remain relatively low throughout 1997 and 1998." (page 31, lines 5 – 11).

6357, lines 1-4: Note that parameter 6 includes the largest sensitivity to light and it will induce high net growth rates earlier than the other parameter combinations. Therefore a greater sensitivity to the initial "winter" concentrations can be expected.

This comment is no longer relevant following re-structuring of the section.

6358, lines 3-4: Parameter set 6 "These were excluded on the basis..." Does that mean the parameter range for α_{Pn} has an upper limit that was too high?

No. Parameter Set 6 results were not used in the uncertainty quantification because they appeared to be unrepresentative of the simulator array's performance, given the performance information shown by the 10 member sample. If we had a larger sample, they might not appear so. It is a separate issue from whether the parameter ranges are too large to be biologically meaningful. We could explore the emulator performance over any parameter space that we wished to.

6360, line 17: Both parameter sets (1 and 4) have a very low light-sensitivity but at the same time are very susceptibility to any small grazing pressure. Up to this point I have learned that some of the emulator errors are most apparent when special (extreme) parameter combinations are involved. Can't this information be introduced to the Latin Hyper Cube sampling?

The comment regarding Parameter Sets 1 and 4 is useful and this point has now been mentioned in the paper. It is true that we could include additional prior information in the LHS sampling by changing the parameter ranges or rejecting areas believed to correspond to biologically un-realistic combinations. We would not then learn about the limitations of the emulator performance in the "extreme" regions. If these regions are truly areas we are not interested in (in say a subsequent calibration exercise) then that is fine. However, given the levels of abstraction involved in modelling biogeochemistry, there is an argument for using only weakly informative priors.

Manuscript changes: An explanation of the extreme behaviour associated with Parameter Sets 1 and 4 has been included in the introduction to the results section (page 31, lines 12 — 19) and we refer back to this, changing the text "parameter vectors (Parameter Sets 1 and 4) that lead to unusually low winter-time chlorophyll concentrations at the start of 1997 in the target simulation (Fig. 3)" to "the two parameter vectors that were seen to cause unusually low winter-time chlorophyll concentrations at the start of 1997 in the target simulation (Fig 3., Parameter Sets 1 and 4)." (page 37, line 5).

6361, line 18 – 6362, line 9: Is this test really expedient? First, will we know the "true" degree of freedom when using real chlorophyll-a data, after these data had been processed with the same algorithm. Second, what do we learn about the "true" degree of freedom when the chlorophyll-a model results are "automatically" autocorrelated. To me, this part of the analysis can be skipped. It will become more relevant with new insight once real data are used and when a best parameter set is found. Figures 9 through 11 are just fine.

The performance evaluation test based on prediction intervals has been removed.

Manuscript changes: Removed text describing the method (page 38, line 6 – page 39, line 4). Removed reference to prediction interval test results in the presentation of the Table 3 results (page 39, line 5 – page 40, line 14). Removed the relevant columns from Table 3.

6364, Discussion: It would be good to have the discussion on the role of the flux perturbation (as explained before). The flux perturbation, as proposed by the authors, is of great importance and its potential to improve 1D model behaviour in general, can be discussed. The nice Figures (original 6) and a simplified version (with fewer locations) of Figure (original 5) could be placed here. This would put them into a more prominent position.

We have chosen to leave this information in the results section as we feel it inappropriate to include detailed experiment-specific data in the discussion. It has been made more prominent

by separating it out into its own section at the end of the results.

Manuscript changes: See response to earlier comment.

6365, Subsection 5.1 Mechanistic emulator performance rather Performance of a process based, dynamical emulator

Since we have chosen to retain the term “mechanistic emulator" for reasons discussed earlier, we continue to use this in the subsection title here for consistency.

6365, lines 17-19: "In a practical application..." How critical is this? How large must an inflation factor be?

The size of inflation factor has been added.

Manuscript change: Added text “The optimal factor would be the normalized error variance from the evaluation experiments (i.e. 1.28, based on the standard deviation of 1.12 for the 9 trial parameter vector experiments in Table 3)." (page 43, line 21).

6372, line 16: But then the transport matrix would be the actual target model and not the original 3D model for which the transport matrix model had been constructed.

This is a good point. The original text is misleading. The most straightforward scenario for using new steady state estimation methods like the Transport Matrix Method with our emulation technique is to incorporate the steady state estimation method into the target model. Running the target model would include its initialization via the TMM or another fast steady state estimation method. Initialization would be the same for simulations used in the science application and for simulations used in emulator construction.

Manuscript changes: Replaced “The new steady state estimation methods would be beneficial in site-based emulator construction too, where their ouse in generation the required 3-D reference ensemble would improve its representativeness of the target model. " with “Incorporating these new steady state estimation techniques into the target model prior to site-based emulation would be particularly advantageous." (page 47, line 27). The whole paragraph has been moved to Section 5.2 “Application of the emulation scheme" and re-worded to introduce spin-up issues and the motivation for fast steady state estimation more explicitly.

6375, line 1: " A dynamic, process- and site-based emulator..."

Since we have chosen to retain the term “mechanistic emulator" for reasons discussed earlier, this change is not applicable.

lines 18-28: this will still hold even if the results section is revised.

True.

How about an explicit statement or conclusion based on the findings learned from sampling the parameter space and how extreme parameter combinations affected your analyses?

We have included a short discussion of these findings.

Manuscript change: Added paragraph “The most notable instances of poor emulator performance occur for parameter vectors associated with the more extreme behaviour in the target model. This raises the question of whether it is really necessary to emulate the target model over such large parameter ranges. Certainly, restricting the parameter space further should help to make our reference sample more representative. In principle, comparisons with observational data at an early stage could be used to identify implausible target model behaviour and suggest ways in which the parameter space might be reduced. However, any

such constraints based on the sparse sampling of parameter space achieved by the target model ensemble could greatly increase the risk of excluding promising parameter combinations and should be undertaken with care. Modifications of the parameter space that are consistent with our biological understanding of the parameters are the most easily justified but, acknowledging the high level of abstraction involved in modelling a system of such complexity, we should avoid over-reliance on subjective priors. Increasing the sample size or improving the emulation methods may be preferable." (page 44, lines 3 – 15).

6381, Appendix A3: The parameter selection process is interesting and relevant for the experimental design (e.g. eventually excluding extreme parameter sets in some analyses). This section should be shortened and incorporated into section 3.2 Model parameter space. In lieu thereof the paragraphs on the Student's t-distribution can be omitted.

The sensitivity analysis for parameter selection in Appendix A3 (original numbering) is part of the parameter space definition process. Emulation issues, including the handling of results given by "extreme parameter sets", are separate from the definition of the parameter space. This is because the definition of the parameter space is part of the definition of the problem to be solved, rather than part of the emulation technique for solving it. The latter is the focus of the paper, while the definition of the problem can largely be considered as incidental, so is appropriately located in the appendix. We have therefore not made the change suggested by the reviewer, which we feel would have lengthened the main text unnecessarily, despite the removal of the paragraphs on the Student's t-distribution and prediction interval tests.

2 Response to comments by Reviewer #2

The manuscript presents a technique to quantify the errors associated with using 1D biogeochemical (BGC) models to approximate 3D models. This involves a comparison of the 1D model with output from a 3D model, with either identical or different parameters. The error

quantification is used to modify the 1D model, bringing it closer to the behaviour of the 3D model.

This summary may be misleading. All of our comparisons between 1-D and 3-D simulations are for identical trial parameter vectors. In runs with the uninformed simulator, it is only the environmental input data that are not specific to the trial parameter vector and they are not specific to any other parameter vector either. It is also worth clarifying that only part of the error information, the bias estimate, is used to modify the 1-D model results. The error quantification as a whole has a wider role in providing uncertainty estimates for the final emulator predictions of 3-D model output.

I think the technique is good, and the paper should be published after corrections. I have to admit that I found the paper extremely heavy going, which partly explains my slow response. I think it would be improved by a reduction in length wherever possible, and a stronger focus on the key points (i.e. how and how well the basic emulator works). Echoing the comments by Markus Schartau, are the comparisons between the uninformed and informed emulators, and direct vs indirect uncertainties really necessary?

The comparison between uninformed and informed emulators are not strictly necessary and have been removed.

Manuscript changes: Removed text comparing the uninformed and informed simulator results (page 34, line 22 – page 35, line 4). Removed text presenting the informed emulator results and their comparison with the uninformed emulator (page 40, line 15 – page 42, line 22). Removed associated table (original Table 4).

The discussion should focus more on the technique, and how well it works. It is not necessary to speculate at length on how the technique might (or might not) be improved. Additionally, the usefulness of the technique as a means to speed up a wide range of optimisation problems is

quite clear, and it is not necessary to go into any details of what those various techniques are.

The length of the discussion regarding how the technique might be improved has been reduced, while retaining a number of points that are potentially valuable for others who might wish to apply similar methods. With regard to the usefulness of the technique to speed up a range of optimisation problems, we feel that it is helpful to discuss the role of a mechanistic emulator in some detail to put our study in context, make clear our own motivation for the work and set out a roadmap for possible future developments. However, it has been possible to shorten the text a little.

Manuscript changes: Removed 2 paragraphs discussing possible causes of uncertainty underestimation (page 45, lines 1 – 19). Removed text describing speculative effects of simulator improvements (page 46, lines 3 – 9; page 46, lines 14 – 20). Removed paragraph describing potential use of mechanistic emulator with a surrogate-based optimisation scheme (page 51, lines 6 – 13).

P6330 L18 and 22 - It would be helpful to introduce the terms variational and sequential, when comparing the two types of inverse methods.

Done.

Manuscript change: Modified text to include a brief explanation of the terms (page 4, lines 16 – 24).

P6339 L11 - The explanation of the choice of transformation function should come here, not on the next page.

The text has been reorganized accordingly: the need for concentration dependency in the advective tendencies is introduced, followed by the idea of implementing concentration depen-

dency via tracer transformations and the explanation of choice of function, before presenting the equations.

Manuscript change: Section of text reorganized (page 13, line 12 – page 15, line 12).

P6341 - Please say how large the small ensemble is at this stage. Were any tests done to see if the ensemble was truly "representative".

It is not appropriate to give an ensemble size here; the method being described at this point is intended to be generic and the ensemble size is a specific detail of its implementation. In our implementation, an ensemble size of 10 was used, as described later. No tests were done to assess its representativeness. The resulting distributions for model outputs could in principle be compared with those from larger ensembles to get some indication. However, such experiments would be computationally expensive and possibly inconclusive. The ensemble is designed to be as representative as possible for a given sample size by the way in which the parameter space is sampled.

P6342 to P6348 - This section is quite hardcore. Can it not be simplified a bit?

There is very little scope for simplification here without losing important information. However, part of the section has changed as a consequence of the re-structuring suggested by Markus Schartau and this should improve the overall clarity of the presentation.

P6348 L12 - The use of satellite data doesn't seem immediately consistent with the idea of using 1D sites. Why where the locations of real time-series not used instead?

There is no conflict between the use of satellite data and the idea of using 1-D sites if we think of our use of 1-D sites as a sampling tool. Our 12 site array can be thought of as a prototype for a larger, global array. Satellite data provides the best resource for global coverage. Extending

the approach to make use of data from time series sites, in addition to satellite data, is seen as essential in the longer term. However, exploiting satellite data is seen as an effective way of creating a flexible baseline calibration system.

P6350 L12 to L19, P6351 L5 - Please explain the reasoning behind these modifications. It does not seem to reduce the number of free parameters, so what is the point?

Explanation added.

Manuscript change: Added text “The changes allow us to consider the effects of a phytoplankton rate parameter or a zooplankton rate parameter on the system without having to consider the impact of directly changing the relationship between rates for closely related plankton types. It is then easier to interpret parameter effects at a high level of abstraction which facilitates comparison with simpler models where parameters represent rates for more aggregated plankton compartments." (page 26, lines 10 – 14).

P6365 L19 - How would one know what inflation factor to apply?

This is now explained.

Manuscript change: Added text “The optimal factor would be the normalized error variance from the evaluation experiments (i.e. 1.28, based on the standard deviation of 1.12 for the 9 trial parameter vector experiments in Table 3)." (page 43, line 21).

Manuscript prepared for Geosci. Model Dev. Discuss.
with version 2014/05/30 6.91 Copernicus papers of the LaTeX class copernicus.cls.
Date: 26 January 2015

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)

J. C. P. Hemmings[1,2], **P. G. Challenor**[3,1], **and A. Yool**[1]

[1]National Oceanography Centre, Southampton, SO14 3ZH, UK
[2]Wessex Environmental Associates, Salisbury, UK
[3]College of Engineering, Mathematics and Physical Sciences, University of Exeter, Exeter, EX4 4QF, UK

Correspondence to: J. C. P. Hemmings (jhemmings@wessexenv.mail1.co.uk)

Abstract

Biogeochemical ocean circulation models used to investigate the role of plankton ecosystems in global change rely on adjustable parameters to ~~compensate for missing biological complexity~~capture the dominant biogeochemical dynamics of a complex biological system. In principle, optimal parameter values can be estimated by fitting models to observational data, including satellite ocean colour products such as chlorophyll that achieve good spatial and temporal coverage of the surface ocean. However, comprehensive parametric analyses require large ensemble experiments that are computationally infeasible with global 3-D simulations. Site-based simulations provide an efficient alternative but can only be used to make reliable inferences about global model performance if robust quantitative descriptions of their relationships with the corresponding 3-D simulations can be established.

The feasibility of establishing such a relationship is investigated for an intermediate complexity biogeochemistry model (MEDUSA) coupled with a widely-used global ocean model (NEMO). A site-based mechanistic emulator is constructed for surface chlorophyll output from this target model as a function of model parameters. The emulator comprises an array of 1-D simulators and a statistical quantification of the uncertainty in their predictions. The unknown parameter-dependent biogeochemical environment, in terms of initial tracer concentrations and lateral flux information required by the simulators, is a significant source of uncertainty. It is approximated by a mean environment derived from a small ensemble of 3-D simulations representing variability of the target model behaviour over the parameter space of interest. The performance of two alternative uncertainty quantification schemes is examined: a direct method based on comparisons between simulator output and a sample of known target model "truths" and an indirect method that is only partially reliant on knowledge of target model output.

In general, chlorophyll records at a representative array of oceanic sites are well reproduced. The use of lateral flux information reduces the 1-D simulator error considerably, consistent with a major influence of advection at some sites. Emulator robustness is assessed by comparing actual error distributions with those predicted. With the direct uncertainty quantification scheme, the emulator is reasonably robust over all sites. The indirect uncertainty quantification scheme

is less reliable at some sites but scope for improving its performance is identified. The results demonstrate the strong potential of the emulation approach to improve the effectiveness of site-based methods. This represents important progress towards establishing a robust site-based capability that will allow comprehensive parametric analyses to be achieved for improving global models and quantifying uncertainty in their predictions.

3 Introduction

A need for better understanding of the role marine biota will play in influencing the nature and rate of global change in response to human activities has led to the inclusion of ~~mechanistic~~ process-based models of ocean biogeochemistry in ocean circulation models (Sarmiento et al., 1993) and more recently in models of the whole Earth system (Séférian et al., 2013). They are designed to capture the dominant responses of complex ecosystems to variability in the physical environment. The biogeochemistry models vary in complexity from simple models in which the biota are represented by single phytoplankton and zooplankton types (e.g. Six and Maier-Reimer, 1996; Palmer and Totterdell, 2001) to more complex functional type models in which a much larger range of different planktonic groups are represented (e.g Moore et al., 2004; Gregg et al., 2003; Le Quéré, 2005; Aumont and Bopp, 2006). ~~Adjustable parametersin the models compensate for un-modelled biological complexity and incomplete knowledge~~

The process-based models are often referred to as mechanistic, as distinct from statistical or data-based models. Yet they are also semi-empirical, incorporating adjustable parameters. Such parameters are important in process-based models of complex systems where incomplete knowledge and practical limits on the degree of complexity that can be resolved make it impossible to design a model that represents all relevant mechanisms. Predictions given by each model are thus affected by structural uncertainty, associated with the model's design, and parametric uncertainty, associated with its chosen parameter values. The equivalent parameters in nature are typically highly variable in space and time and among different organisms present in any assemblage, making the optimal values particularly elusive. Effective use of ocean observations to constrain model parameters and reduce parametric uncertainty is necessary to improve

the predictive skill of particular models and to gain a better understanding of inadequacies in model design.

Any rigorous exploration of a biogeochemical model's parameter space is computationally intensive, requiring many thousands of simulations. This has generally dictated the use of fast site-based experiments for parametric analyses, following the pioneering work of Fasham and Evans (1995) and Matear (1995). Parameters are optimized to fit observations at individual sites (e.g. Losa et al., 2004; Fasham et al., 2006; Friedrichs et al., 2006, 2007; Dowd, 2011; Kidston et al., 2011; Fiechter et al., 2013; Prieß et al., 2013a; Ward et al., 2013) or at multiple sites simultaneously (Hurtt and Armstrong, 1999; Schartau and Oschlies, 2003; Hemmings et al., 2004; Kane et al., 2011; Xiao and Friedrichs, 2014). In these experiments, the biogeochemistry model is integrated in a 1-D or 0-D framework representing a single water column at each site, and a local approximation of the physical environment is used as forcing data to drive the simulation.

In the site-based study of Dowd (2011), a sequential data assimilation method with a stochastic configuration of a biogeochemistry model was used to estimate the models' static parameters in combination with its time varying state (i.e. its prognostic variables). ~~New joint parameter~~ Sequential methods use a series of analysis cycles in which analysis steps combine observations with model forecasts, taking into account the uncertainties in each. The forecast for each step is initialized from the previous analysis. Dowd (2011) estimated new joint probability distributions for state and parameters ~~were thus estimated~~ at each observation time on the basis of the new observations and a previous analysis. However, in most cases variational inverse methods are used, the aim being to constrain the parameters of the deterministic free-running model. Parameter values are varied with the objective of minimizing or maximizing some function of the model-data differences. The solution is then the best fit to the complete observational data set that satisfies the model equations exactly (ignoring error introduced by time discretization in the numerical solver). An exception is made in the inverse approach of Losa et al. (2004) where the model equations are used as a weak constraint and both parameters and state are estimated. This allows for sources of simulation uncertainty that are not associated with the adjustable parameters, such as structural error or error in the forcing data.

Sequential data assimilation approaches are particularly useful in short-term forecasting, where the forecast is highly dependent on the initial state and state estimation is the primary goal. However, for long-term future projections that must rely on free-running models, the estimation of model parameters is paramount. Methods that preserve the integrity of the model dynamics are inherently better suited to this problem but simulation error impacting on the state variables cannot be ignored and a more rigorous treatment of simulation uncertainty is needed before the potential of these methods can be fully realized (Hemmings and Challenor, 2012).

In this study, we focus specifically on simulation uncertainty introduced by the use of 1-D simulations to approximate 3-D model behaviour. The uncertainty is primarily associated with differences in the representation of the physical environment and differences in the horizontal fluxes and initial values of biogeochemical properties. Despite this uncertainty, site-based calibrations have been shown to improve the predictive skill of 3-D models (Oschlies and Schartau, 2005; Kane et al., 2011; McDonald et al., 2012). However, the relationship between 1-D and 3-D simulations is not well understood in quantitative terms. Parameter vectors that are optimal in one context are unlikely to be optimal in the other, inevitably compromising the utility of established parameter estimation methods.

The lack of information about biogeochemical fluxes associated with horizontal advection and diffusion is an obvious source of uncertainty. Some consideration has been given to this problem. Losa et al. (2004) introduced their weak constraint approach primarily to allow for the neglect of horizontal transport. Fasham et al. (2006) parametrized diffusive fluxes based on the analysis of a passive tracer release associated with an iron fertilization experiment, while Friedrichs et al. (2007) included an advective flux divergence term for nutrients based on 3-D model output. Fasham et al. (1999) took a different approach, optimizing parameters in a Lagrangian framework to fit data from a survey of the North Atlantic spring bloom. The survey followed the track of a drogued buoy to minimize the impact of horizontal advection on the biogeochemical system under study. More typically though, horizontal fluxes are ignored in site-based calibration studies.

In a relatively small number of studies, parameters have been optimized for the biogeochemistry model within its host 3-D circulation model. This is practical for limited time and space

domains: Garcia-Gorriz et al. (2003) and Huret et al. (2007) estimated parameters for regional models by assimilating satellite-derived chlorophyll data over periods of order 1 month. Doron et al. (2013) assimilated these data at a single point in time into an eddy-permitting model of the North Atlantic using an adapted Kalman filter analysis with a perturbed parameter ensemble simulation. The ensemble simulation was similarly of 1 month duration. Fan and Xianqing (2009) estimated spatially varying parameters for the global domain but with an assimilation window limited to 5 days. In contrast, Tjiputra et al. (2007) performed much longer global experiments, assimilating seasonal maps of surface chlorophyll and nitrate into a global model of the annual cycle, but relied on a coarse resolution model (3.5° horizontal resolution) and, in common with a number of other studies, only optimized locally in parameter space.

The type of compromises imposed on parametric analyses of 3-D biogeochemical models by limited computer resources are generic to many different fields in which computer models are used. This problem has motivated the development of statistical emulation techniques that allow more comprehensive investigations of parameter space to be achieved. A good introduction is given by O'Hagan (2006). An emulator provides a prediction of a chosen model output, or a metric used in its assessment, for any setting of the parameter values, together with a measure of uncertainty in that prediction. A relatively small ensemble of model runs is required to provide training data for emulator construction, although this is still a significant overhead for 3-D models.

Statistical emulation techniques have been applied to the estimation of marine biogeochemical model parameters in regional studies. Leeds et al. (2013) used emulators for computational efficiency in a Bayesian hierarchical framework that linked spatially distributed 1-D simulations. In other work, emulators were constructed for relatively expensive 3-D simulations to allow the required coverage of parameter space to be achieved: Hooten et al. (2011) used 50 ensemble members to represent a 7-dimensional parameter space, while Mattern et al. (2012) used a similar ensemble size in a 2 parameter study.

Although, to the authors' knowledge, the application of statistical emulators to ocean biogeochemistry has so far been limited to regional studies, they are starting to be used at the global scale for parametric analyses of other Earth system model components, including the

coupled ocean–atmosphere system (Williamson et al., 2013) and atmospheric aerosol concentrations (Lee et al., 2012). These studies involved the use of perturbed parameter ensemble simulations with global 3-D models. Williamson et al. (2013) investigated a 30-dimensional parameter space, benefitting from a very large ensemble generated using climateprediction.net, a distributed computing project in which personal computers are volunteered by members of the public. Lee et al. (2012) used a much smaller ensemble (80 members) to investigate parametric uncertainty over an 8-dimensional parameter space. The ensemble size was computationally practical owing to the coarse resolution of the model and the limited duration of the runs (4 months).

The application of statistical emulators to global ocean biogeochemical models would make investigation of the models' predictive potential more tractable. However, achieving sufficiently large training ensembles for periods that fully capture the seasonal variability at an appropriate spatial resolution will be challenging. Mesoscale and sub-mesoscale dynamics are known to have a strong impact on biogeochemical processes in the upper ocean (Lévy, 2008), yet global simulations that resolve the ocean mesoscale require considerable computing resources, severely limiting ensemble size.

Given the potential for improving the representation of biogeochemical cycles by increasing model resolution, avoidance of unnecessary trade-offs between resolution and ensemble size is desirable. Improving 1-D modelling capabilities is a potential solution. The goal would be to produce a set of site-based simulators that could serve as an efficient and reliable surrogate model for emulating arbitrary 3-D model outputs with quantified uncertainty. The number of sites could be adapted according to the required ensemble size and the resources available. Like a statistical emulator, the system would provide a prediction of model output and a measure of uncertainty in that prediction. We refer to the proposed system as a mechanistic emulator to distinguish it from statistical site-based emulators (Leeds et al., 2013) that treat the target model as a black box. For some parametric analyses, a mechanistic emulator of this type would be sufficient. Where more comprehensive analyses are required it would be used to bridge the gap between the 3-D target model and one or more statistical emulators of model outputs or metrics.

Here we introduce an experimental mechanistic site-based emulator and use it to explore the feasibility of establishing a robust relationship between 1-D and 3-D simulations. The emulator predicts annual cycles of surface chlorophyll output produced by a target model of the global ocean. The aim is to provide a way of exploiting satellite chlorophyll or related ocean colour products for making reliable inferences about the target model performance for arbitrary trial parameter vectors, without having to run the corresponding 3-D simulations.

Section 4 describes the components of the mechanistic emulator and the method for its construction and Sect. 5 gives the experimental method used to evaluate its performance. The results are presented in Sect. 6. In Sect. 7 the findings are discussed with regard to the potential of the emulation scheme as an enabling tool for improved parametric analyses of global models, using satellite ocean colour data in combination with in situ observations. A summary of the work is given in Sect. 8.

4 The mechanistic emulator

The site-based emulator combines a surrogate model with a ~~statistical description of its predicted~~ probabilistic prediction of its error with respect to the 3-D target model. The surrogate model takes the form of an array of 1-D simulators. Variation of the predicted error distribution of surface chlorophyll output from the surrogate model over its time and space domain is fully described. The intention is to establish a form of traceability between the surrogate model and the target model that allows robust inferences about target model skill to be made from analyses of surrogate model output.

Inferences about model performance are often made on the basis of a cost function, summarizing the misfit of a simulation to observational data. The cost function typically takes the form

$$J(\boldsymbol{y}_{\mathrm{P}}) = (\boldsymbol{y}_{\mathrm{P}} - \boldsymbol{y}_{\mathrm{O}})^{\mathrm{T}} \mathbf{R}^{-1} (\boldsymbol{y}_{\mathrm{P}} - \boldsymbol{y}_{\mathrm{O}}) \quad (1)$$

where $\boldsymbol{y}_{\mathrm{O}}$ is a vector of n observations, $\boldsymbol{y}_{\mathrm{P}}$ is the corresponding vector of predicted values and $\mathbf{R}^{-1}$ is the inverse of the $n \times n$ error covariance matrix (Stow et al., 2009). The superscript T is

the transpose operator. The error covariance matrix describes the predicted error structure of the model output. It weights the contributions of individual model-data misfits according to their significance, taking into account prior expectations of uncertainty.

It is commonly assumed that the individual misfits are independent. The off-diagonal elements of $\mathbf{R}$ are then zero and the cost function can be written

$$J(\boldsymbol{y}_{\mathrm{P}}) = \frac{1}{n}\sum_{i=1}^{n}\frac{(P_i - O_i)^2}{\sigma_{ii}^2} \tag{2}$$

where P_i and O_i are the elements of $\boldsymbol{y}_{\mathrm{P}}$ and $\boldsymbol{y}_{\mathrm{O}}$ respectively and σ_{ii}^2 represents the diagonal elements of $\mathbf{R}$.

If both observation and simulation error are relevant in an analysis, the error variance σ_{ii}^2 is the predicted variance of the combined error from both sources. When using a surrogate model, the simulation error includes the surrogate model error with respect to the target model. It may also include error from other sources such as target model input data or structural error, depending on the objective of the analysis. Hemmings and Challenor (2012) discuss cost function design for different analyses in more detail.

Predicted surrogate model error statistics can be used in a cost function to make the function more informative about the likely misfit between the target model and the observations. They do this by increasing the weight given to model-data misfit where the surrogate model error is expected to be small and decreasing the weight elsewhere. The cost function can then be used to evaluate the goodness-of-fit of the target model simulation to the observations, given the surrogate model output.

In the experimental emulator presented here, the statistical ~~description of the predicted~~ prediction of the error with respect to the target model is restricted to its mean and variance at individual data points. If the emulator were used in a cost function-based analysis, the predicted error variance would contribute directly to σ_{ii}^2 and the predicted mean error would be used to give bias-corrected values for P_i. ~~A~~Estimation of the mean and variance is a first step towards a more complete uncertainty quantification that would include the error covariance structure required ~~for a full specification of~~ to fully specify $\mathbf{R}$.

The target model in the present study is NEMO-MEDUSA, combining the MEDUSA 1.0 biogeochemistry model (Model for Ecosystem Dynamics, carbon Utilisation, Sequestration and Acidification) described by Yool et al. (2011) with the NEMO ocean model (Nucleus for European Modelling of the Ocean; Madec, 2008).

4.1 The biogeochemical simulator

The 1-D simulator incorporates a representation of the biogeochemistry that is identical to that in the target model. MEDUSA is an intermediate complexity model, representing the plankton ecosystem by 11 compartments in the form of biogeochemical tracers. These include 6 nitrogen pools for two phytoplankton groups (diatoms and non-diatoms), two zooplankton groups (micro- and meso-zooplankton), slow-sinking detritus and dissolved inorganic nitrogen. The remaining compartments represent two additional dissolved nutrients required by the phytoplankton (silicon and iron), the chlorophyll concentrations associated with the two phytoplankton types and the silicon concentration associated with the diatoms. The effect of fast-sinking detritus is represented by instantaneous vertical redistribution of material in the water column.

1-D integrations of MEDUSA are performed in a 3-D context ~~defined with reference to~~ where physical and biogeochemical information from the target model ~~. This information provides the required~~ provide environmental input data for the site-based simulations. The physical environment ~~is not parameter dependent so is defined by a single 3-D simulation. The biogeochemical environment, comprising initial concentrations and horizontal flux divergences~~ required by the 1-D simulator is independent of the biogeochemical ~~tracers, is defined with reference to a small ensemble of 3-D simulations representing variation over the parameter space to be investigated~~model parameters. However, the biogeochemical environment is parameter-dependent making its representation in a site-based parametric analysis less straightforward. The 1-D simulator for MEDUSA is configured using the Marine Model Optimization Testbed facility described by Hemmings and Challenor (2012). The testbed software, MarMOT 1.1, is open source and ~~available from~~ freely available as detailed in Appendix A.

The MEDUSA state variables are the biogeochemical tracer concentrations at each model grid point. The evolution equation for the concentration c_{ik} of the ith ~~MEDUSA~~ biogeochemical tracer at depth level k in the 1-D simulator is

$$\frac{\mathrm{d}c_{ik}}{\mathrm{d}t} = -(w_p + w_i)\frac{\partial c_i}{\partial z} + \frac{\partial}{\partial z}\left(K_\rho \frac{\partial c_i}{\partial z}\right) + SMS_{ik}(\boldsymbol{C}, \boldsymbol{F}) + p_{ik}(\boldsymbol{C}_k, p^\star_{jk}). \quad (3)$$

The first two terms represent the tendencies (i.e. rates of change) due to vertical flux divergence. w_p is the vertical velocity of the water, w_i is the active vertical velocity of the biological material relative to the water and K_ρ is the turbulent diffusion coefficient. SMS_{ik} is the source-minus-sink term from the MEDUSA plankton model. It is a function of the state vector $\boldsymbol{C}$ and a forcing vector $\boldsymbol{F}$ comprising temperature, downwelling solar radiation at the sea surface and input of soluble iron from atmospheric dust deposition. SMS_{ik} is depth-dependent because the light available for phytoplankton photosynthesis and the nutrient sources from the remineralization of fast-sinking detritus depend on tracer concentrations at $k-1$ shallower levels. w_i is assigned a constant sinking rate for the detritus tracer, corresponding to the MEDUSA sinking rate parameter for slow-sinking detritus. It is zero for all other tracers. Values for w_p, K_ρ and $\boldsymbol{F}$ are provided by the physical environment from the target model.

The final term in Eq. (3) is a perturbation term used to represent the effect of horizontal flux divergence. The divergence tendency for the ith tracer p_{ik} depends on the local state ~~and an applied perturbation $p^\star_{jk}$.~~ $\boldsymbol{C}_k$ ~~is a~~ (a vector containing the subset of tracer concentrations ~~in $\boldsymbol{C}$~~ at depth level k) and an applied perturbation $p^\star_{jk}$. Tracer-specific perturbations are applied to tracers representing dissolved nutrients and the nitrogen content of the plankton. ~~For most tracers , $j = i$.~~ These are referred to ~~in MarMOT~~ as primary tracers. ~~MEDUSA's primary tracers~~

~~are:~~

Nitrogen in non-diatom phytoplankton

Nitrogen in diatom phytoplankton

Nitrogen in microzooplankton

Nitrogen in mesozooplankton

Nitrogen in slow-sinking detritus

Nitrogen nutrient

Silicon nutrient (i.e. silicic acid)

Iron nutrient

~~Each of the remaining tracers~~The phytoplankton chlorophyll and silicon tracers (secondary tracers) are affected indirectly, following the perturbations to the corresponding nitrogen tracers in such a way as to preserve the phytoplankton chlorophyll : nitrogen and silicon : nitrogen ratios. For a primary tracer, ~~referred to as secondary tracers , is linked to a particular primary phytoplankton tracerby a time-varying ratio describing the plankton's internal composition. For these tracers~~$j = i$. For a secondary tracer, j indexes the ~~appropriate primary tracerso that the rate of change of the secondary tracer depends on the perturbation applied to the primary tracer. The secondary tracers are:~~

Chlorophyll in non-diatom phytoplankton

Chlorophyll in diatom phytoplankton

Silicon in diatom phytoplankton

~~Tendencies for the secondary tracers are derived from nitrogen tracer tendencies using the composition ratios implied by present tracer concentrations: the ratios of non-diatom~~

~~chlorophyll to non-diatom nitrogen, diatom chlorophyll to diatom nitrogen and diatom silicon to diatom nitrogen.~~ relevant primary tracer.

The ~~parameter-dependent~~ input data set required to define the biogeochemical environment ~~for 1-D simulations~~ comprises the initial state and the applied perturbations controlling the tracers' horizontal flux divergence tendencies. This is the biogeochemical environment vector

$$\boldsymbol{B} = \{\boldsymbol{C}(t_\mathrm{o}), \boldsymbol{P}^\star\}. \tag{4}$$

$\boldsymbol{C}(t_\mathrm{o})$ is the initial state vector containing the concentrations of the 11 tracers at each depth level on the model grid at time t_o and the vector $\boldsymbol{P}^\star$ contains applied perturbations at each depth level for the 8 primary tracers at 5 day period mid-points for $t > t_\mathrm{o}$. Perturbations represent the effect of lateral advection inferred from an analysis of local currents and upstream property gradients in the 3-D model output. The effect of horizontal diffusion is ignored.

The ~~primary tracer perturbations are applied in transformed concentration space. This means that the applied perturbation represents the rate of change due to lateral advection of the transformed tracer. A square root transformation was chosen for all primary tracers at all sites so that a perturbation $p^\star$ specifies the rate of change of $\sqrt{c}$, where c is the tracer concentration. This rate of change is~~

$$p^\star = -\boldsymbol{u}_\mathrm{h} \cdot \nabla_\mathrm{h} \sqrt{c}$$

~~where the subscript h denotes vectors in the horizontal plane and $\boldsymbol{u}_\mathrm{h}$ is the current velocity. The resulting tracer tendencies are concentration-dependent. For primary tracers the tendency is~~

$$p = 2\sqrt{c} p^\star.$$

~~For secondary tracers it is~~

$$p_i = \frac{c_i}{c_j} p_j$$

where i is the secondary tracer index and j indexes the associated primary tracer.

The need for concentration-dependent tendencies arises from a need to preserve co-variation of tendencies and concentrationsadvective tendencies of individual tracers are dependent on their upstream gradients and often tend to co-vary with their local concentrations. It is important to give some attention to preserving such relationships that are prevalent in the 3-D simulation as far as possible. A particular example of such a prevailing relationship occurs when tracer concentrations are low. If we have a negative advective tendency it must should increase towards zero as the concentration approaches zeroto avoid the concentration becoming , otherwise the concentration will become negative. In the 3-D simulation, this happens naturally because the upstream gradient driving it tends towards zero (assuming the upstream concentration cannot be negative). In the 1-D simulation, adaptation of tendencies to the local concentration is necessary to counter any inconsistencies between the two.

The extent to which horizontal concentration gradients of the tracerstend to co-vary with their local concentrations in the target model This concentration dependency is introduced by using applied perturbations that represent rates of change of transformed tracers. The choice of transformation determines the form of the dependency and is an important consideration in simulator design.

Analysis of 3-D simulations indicate that the concentration dependency of horizontal gradients varies temporally and spatially and between different tracers. Use of the square root transformation protects against the evolution of negative concentrations and was found by Hemmings and Challenor (2012) to be a reasonable compromise between using untransformed and log-transformed concentrations.

4.2 Informed and uninformed simulators

To quantify the predictive skill of A square root transformation was therefore chosen for all primary tracers at all sites so that a perturbation $p^\star$ specifies the rate of change of $\sqrt{c}$, where c is the site-based simulator under idealized conditions where its biogeochemical environment $\boldsymbol{B}$ is known, a version of the simulator is configured using parameter-specific environmental input data. This is referred to as the *informed simulator*. It is applicable only to parameter vectors for

~~which a corresponding~~ tracer concentration. The implied concentration tendency is then

$$p = 2\sqrt{c}p^{\star}. \tag{5}$$

For secondary tracers the tendency is

$$p_i = \frac{c_i}{c_j} p_j \tag{6}$$

where i is the secondary tracer index and j indexes the associated primary tracer. The applied perturbation diagnosed from 3-D ~~simulation with the target model is available. The true 3-D simulation output for any variable of interest is known for these parameter vectors , so emulation would not normally be necessary. The purpose of analyzing the informed simulator is to quantify the expected error associated with the basic 1-D simulation method~~model output is

$$p^{\star} = -\boldsymbol{u}_{\mathrm{h}} \cdot \nabla_{\mathrm{h}} \sqrt{c} \tag{7}$$

where the subscript h denotes vectors in the horizontal plane and $\boldsymbol{u}_{\mathrm{h}}$ is the current velocity.

~~In the informed simulation scenario, the relationship between an output value given by the target model with parameter vector $\boldsymbol{x}_{\mathrm{o}}$, denoted $f(\boldsymbol{x}_{\mathrm{o}})$, and the corresponding value given by the 1-D simulator $g(.,\boldsymbol{x}_{\mathrm{o}})$ is expressed by~~

$$f(\boldsymbol{x}_{\mathrm{o}}) = g[\boldsymbol{B}(\boldsymbol{x}_{\mathrm{o}}), \boldsymbol{x}_{\mathrm{o}}] + \epsilon_1$$

~~where $\boldsymbol{B}(\boldsymbol{x}_{\mathrm{o}})$ is the environment data derived from the 3-D simulation output for $\boldsymbol{x}_{\mathrm{o}}$ and ϵ_1 is a stochastic residual referred to as the *informed simulator residual*. Its negated value is the *informed simulator error* which is the consequence~~ Differences between the simulator output and that of the target model arise due to the combined effects of a number of ~~basic simulation errors~~sources of simulation error. Specifically these are approximation error in the physical

environment variables due to ~~averaging~~temporal averaging of the 3-D target model data on which they are based, error in the advective flux divergence tendencies, error introduced by ignoring horizontal diffusion and differences in solver numerics. ~~The simulator output may have biases so the residual ϵ_1 is not assumed to have zero mean~~Any differences between the initial state $\boldsymbol{C}(t_o)$ and the target model state at time t_o will contribute an additional source of error.

4.2 The uninformed simulator and biogeochemical environment model

In a calibration exercise or other parametric analysis, the 1-D simulator is used to learn about the likely behavior of 3-D target model simulations that have not been performed. For an arbitrary trial parameter vector $\boldsymbol{x}_o$, the parameter-specific biogeochemical environment $\boldsymbol{B}(\boldsymbol{x}_o)$ is typically unknown~~, introducing an additional source of simulation error. This error will be referred to as *parametric environment error*.~~ . Instead we use an environment vector derived from a statistical model. The corresponding 1-D simulator is referred to as the *uninformed simulator* ~~.~~ indicating that it is not informed by parameter-specific environment data. Our surrogate model consists of an array of uninformed simulators at different sites, spanning a range of oceanic conditions.

The ~~input~~ statistical model used to define the biogeochemical environment for the uninformed simulator is ~~defined by a statistical model of the parameter-dependent environmental input data. The model is~~ constructed with reference to a small ensemble of 3-D simulations, designed to be representative of the infinite set of 3-D simulations covering a parameter space of interest χ. ~~The relationship between the 3-D simulation and the uninformed simulator is given by~~ If we denote an output value from the simulator with biogeochemical environment vector $\boldsymbol{B}$ and parameter vector $\boldsymbol{x}$ by $g(\boldsymbol{B}, \boldsymbol{x})$ and the corresponding output from the target model by $f(\boldsymbol{x})$, then for parameter vector $\boldsymbol{x}_o$

$$f(\boldsymbol{x}_o) = g\left(\overline{\boldsymbol{B}}, \boldsymbol{x}_o\right) + \epsilon_{21} \tag{8}$$

where $\overline{\boldsymbol{B}}$ is an estimate of the expected environment $\mathrm{E}[\boldsymbol{B}(\boldsymbol{x})] : \boldsymbol{x} \in \chi$ and ~~ϵ_2 is a new stochastic residual, possibly having a non-zero mean~~ϵ_1 is a stochastic residual. This is the *uninformed simulator residual* and its negated value is the *uninformed simulator error*. The simulator output may have biases so the residual ϵ_1 is not assumed to have zero mean.

The environment model consists of a model for $\mathrm{E}[\boldsymbol{B}(\boldsymbol{x})]$, referred to as the *mean environment model*, and a *stochastic environment generator* that is used in quantifying the uncertainty of the simulator output. The environment model assumes multi-variate Gaussian probability distributions for a vector $\boldsymbol{S}(t_o)$ that specifies the initial state and for the applied advective flux perturbation vector $\boldsymbol{P}^\star$. $\boldsymbol{S}$ is an alternative description of the state $\boldsymbol{C}$. It comprises elements $\sqrt{c}$ for each primary tracer concentration c in $\boldsymbol{C}$ and composition ratios c_i/c_j for each secondary tracer concentration c_i in $\boldsymbol{C}$. c_j is the concentration of the associated primary tracer at the same depth level. An estimate of $\mathrm{E}[\boldsymbol{B}(\boldsymbol{x})]$ is given by the ensemble means of $\boldsymbol{S}(t_o)$ and $\boldsymbol{P}^\star$ from the 3-D ensemble.

4.3 The uninformed emulator

If an array of 1-D simulators is to be used to make robust inferences about the target model, it must be combined with uncertainty estimates for its predictions of target model output in the form of predicted error statistics. The combination of the uninformed simulator array with its predicted error statistics is referred to here as the *uninformed emulator*. This is the complete mechanistic emulator for the target model.

Two different methods are used in this study for quantifying uncertainty in the uninformed simulator output: a *direct method* and an *indirect method*. In the direct method, statistics for ~~ϵ_2~~ ϵ_1 are estimated by comparing simulator and target model output for matching parameter vectors, using the target model output available from our small 3-D ensemble. In the indirect method, the uncertainty ~~associated with parametric environment error~~ introduced by using the mean environment vector $\bar{\boldsymbol{B}}$ in place of the unknown environment vector $\boldsymbol{B}(\boldsymbol{x}_o)$ is treated separately from that due to ~~basic simulator error and~~ other simulator error sources. It is quantified by an uncertainty analysis, using the stochastic environment generator to create multiple realizations of the unknown environment. Uncertainty ~~arising from basic simulation error is quantified~~

with reference to the performance of the informed simulatorfrom other sources is estimated by applying the direct method to $g[\boldsymbol{B}(\boldsymbol{x}_o), \boldsymbol{x}_o]$, referred to as the *informed simulator*. The indirect method is more complicated to apply than the direct method but is less dependent on the small target model ensemble. This means that the indirect method could be more robust than the direct method in situations where the ensemble poorly represents the variability of target model solutions over the parameter space χ.

4.3.1 Direct method for uncertainty quantification

In the direct method, values of ~~ϵ_2~~ ϵ_1 for the variable of interest at each point in space and time are determined from matching pairs of uninformed simulator and target model output values using Eq. (8). Statistics for ~~ϵ_2~~ ϵ_1 are then estimated from this sample. A conceptual overview of the data flow in the emulator construction and evaluation process is given in Fig. 1.

The processing is divided into a *construction phase* and an *application phase*. In a practical application, the construction phase is intended for single execution, whereas the application phase must be executed for each trial parameter vector. The procedure for assessment of the uninformed emulator against a known truth is shown as an extension to the application phase.

Error statistics must be determined using target model data that are independent from those used in the simulation. This means that, in the construction phase, target model ensemble members used to determine ~~ϵ_2~~ ϵ_1 for the simulator output must be different from those used to construct the mean environment model for the simulator input. Furthermore, any target model ensemble member used to assess the uninformed emulator performance must be different from any ensemble member used in the construction phase.

4.3.2 Indirect method for uncertainty quantification

The indirect method requires an explicit quantification of the uncertainty associated with ~~parametric environment error. To define the parametric environment error~~use of the mean environment vector $\overline{\boldsymbol{B}}$ in lieu of unavailable parameter-specific environment information. Reliance on $\overline{\boldsymbol{B}}$ introduces a parameter-dependent source of environment-induced error into

the simulation. The resulting contribution to simulation error is referred to as the *parametric environment error*. To define it, we consider a perfect simulator $g_{\mathrm{T}}(.,.)$, such that

$$f(\boldsymbol{x}_{\mathrm{o}}) = g_{\mathrm{T}}[\boldsymbol{B}_{\mathrm{T}}(\boldsymbol{x}_{\mathrm{o}}), \boldsymbol{x}_{\mathrm{o}}] \tag{9}$$

where $\boldsymbol{B}_{\mathrm{T}}$ is the complete and accurate description of the local biogeochemical environment in the 3-D simulation, including advective and diffusive flux perturbations. The simulator is perfect in the sense that it exactly reproduces the results of the 3-D simulation. Introducing parametric uncertainty in the biogeochemical environment and representing the environment by its expectation then gives

$$f(\boldsymbol{x}_{\mathrm{o}}) = g_{\mathrm{T}}\{\mathrm{E}[\boldsymbol{B}_{\mathrm{T}}(\boldsymbol{x})], \boldsymbol{x}_{\mathrm{o}}\} + \epsilon_{\mathrm{B}} : \boldsymbol{x} \in \chi. \tag{10}$$

where ϵ_{B} is a stochastic residual, possibly with a non-zero mean~~, that~~. This is the negated parametric environment error ~~. This~~ or *parametric environment residual* ~~is that part of the departure of the target model output from the uninformed simulator output that is associated with the impact of biogeochemical environment error on the simulator. The biogeochemical environment error considered is specifically that associated with parametric uncertainty and does not include error associated with differences between the environment data used in the simulator $\boldsymbol{B}$ and the environment in the target model $\boldsymbol{B}_{\mathrm{T}}$. The latter is treated as a basic simulation error.~~.

It is important to note that many different designs are possible for a perfect simulator satisfying Eq. (9), having different formulations for concentration dependency in the flux divergence tendencies. Variants of the applied perturbation $\boldsymbol{P}^{\star}$ will give different results for the simulator term in Eq. (10), where the environment is not consistent with the simulation state, and therefore different residuals. The parametric environment error is therefore not just a property of the target model but depends also on the simulator design.

Combining Eqs. (8) and (10), the residual for the target model output with respect to the uninformed simulator output can be expressed as

$$\epsilon_{21} = \epsilon_{\mathrm{S}} + \epsilon_{\mathrm{B}} \tag{11}$$

where ϵ_{S} is a stochastic residual given by

$$\epsilon_{\mathrm{S}} = g_{\mathrm{T}}\{\mathrm{E}[\boldsymbol{B}_{\mathrm{T}}(\boldsymbol{x})], \boldsymbol{x}_{\mathrm{o}}\} - g\left(\overline{\boldsymbol{B}}, \boldsymbol{x}_{\mathrm{o}}\right). \tag{12}$$

ϵ_{S} is the departure of the hypothetical output of the perfect simulator with the true mean environment from the output of the uninformed simulator. ~~This is referred to~~ The first term describes a perfect *mean environment simulation*, while the second term described its approximation by the simulator. In this context, we can refer to the uninformed simulator as a mean environment simulator. We refer to ϵ_{S} as the *mean environment simulation residual*. ~~It is closely related to the informed simulator residual ϵ_1,~~ Mean environment simulation error (the negated residual) is caused by basic simulation errors that are not associated with parametric uncertainty in the environment.

It is not possible to evaluate the perfect simulator term in Eq. (12) and directly determine values for ϵ_{S}. However, we can get a handle on the impact of basic simulation errors from analysing the informed simulator. The relationship between the target model output for $\boldsymbol{x}_{\mathrm{o}}$ and that of the corresponding informed simulator is given by

$$f(\boldsymbol{x}_{\mathrm{o}}) = g[\boldsymbol{B}(\boldsymbol{x}_{\mathrm{o}}), \boldsymbol{x}_{\mathrm{o}}] + \epsilon_2 \tag{13}$$

where $\boldsymbol{B}(\boldsymbol{x}_{\mathrm{o}})$ is the environment data derived from 3-D simulation output for $\boldsymbol{x}_{\mathrm{o}}$ and ϵ_2 is a stochastic residual, possibly having non-zero mean, referred to as the *informed simulator residual*. Its negated value is the *informed simulator error*.

The residuals ϵ_2 and ϵ_S are closely related, in that the input ~~$\overline{\boldsymbol{B}}$~~ $\boldsymbol{B}(\boldsymbol{x}_{\mathrm{o}})$ in the informed simulator is intended to approximate the ~~perfect simulator input $\mathrm{E}[\boldsymbol{B}_{\mathrm{T}}(\boldsymbol{x})]$~~ true parameter-specific environment in the same way that ~~$\boldsymbol{B}(\boldsymbol{x}_{\mathrm{o}})$ in the informed simulator~~ $\overline{\boldsymbol{B}}$ in the uninformed simulator (or mean environment simulator) is intended to approximate the ~~true parameter-specific environment~~perfect simulator input $\mathrm{E}[\boldsymbol{B}_{\mathrm{T}}(\boldsymbol{x})]$. Both residuals are affected by basic simulation errors. The difference is that the environment in Eq. (12) is ~~no longer~~ not specific to the parameter vector $\boldsymbol{x}_{\mathrm{o}}$.

The uninformed simulator is one of a set of generic simulators, in which the constraint that the input environment is intended to represent the parameter-specific environment does not apply. In generic simulators, inconsistencies between the environment and the simulation state are likely to be greater than in the informed simulator. The mean environment simulation residual ϵ_S may therefore be more sensitive to the concentration-dependency formulation than the informed simulator residual ~~ϵ_1. However, it is not possible to evaluate the perfect simulator term in Eq. (12), so to~~ ϵ_2. Nevertheless, to model ϵ_S we make the pragmatic assumption that it is identically distributed to ~~ϵ_1~~ϵ_2. Statistics for ~~ϵ_1, defined by Eq. (13),~~ ϵ_2 are determined by direct comparison of informed simulator output with true output records from the target model.

The model for the parametric environment residual ϵ_B is derived from a parametric uncertainty analysis, following Hemmings and Challenor (2012). The environment corresponding to the trial parameter vector is unknown so we examine the distribution of the residual over many possible environments, aiming to achieve adequate coverage of the environment space that maps to the parameter space of interest. The method involves running a 1-D ensemble simulation based on a sample of environment realizations. These are generated using the mean environment model and stochastic environment generator introduced in Sect. **??**4.2.

The environment generator uses independent statistical models for generating the initial state and the input flux perturbations. For each ~~data set~~of these two data sets, separate multi-variate Gaussian models are constructed using Empirical Orthogonal Functions that capture the dominant modes of variability in the target model ensemble output at each site. The statistical models for the initial state preserve spatial covariances (in the vertical) and covariances between the biogeochemical properties, as characterized by the first 5 EOFs of the sample anomalies, anomalies being determined with respect to the ensemble means. The statistical models for the advective flux perturbations preserve temporal and spatial covariances and covariances between the 8 primary tracers, again as characterized by the first 5 EOFs of the anomalies.

To derive the statistical model for a simulator's initial state from a target model ensemble of size n, an $n \times m$ matrix $\mathbf{Y}_{3d}$ is constructed containing the n available instances of the initial state, as defined by the alternative state vector $\boldsymbol{S}$. (m is the number of elements in $\boldsymbol{S}$.) If $y_{.j}$ is

the mean and s_j^2 the variance of the jth column of $\mathbf{Y}_{3d}$, then the matrix $\mathbf{Z}_{3d}$ with elements

$$z_{ij} = \frac{y_{ij} - y_{\cdot j}}{s_j} \tag{14}$$

is the normalized form of $\mathbf{Y}_{3d}$ for which each column has zero mean and unit variance.

The environment generator uses the eigenvalues and eigenvectors obtained from the spectral decomposition of the correlation matrix for $\mathbf{Z}_{3d}$:

$$\mathbf{\Sigma} = \mathbf{Z}_{3d}^{\mathrm{T}} \mathbf{Z}_{3d} = \mathbf{V} \mathbf{\Lambda} \mathbf{V}^{\mathrm{T}}. \tag{15}$$

$\mathbf{\Lambda}$ is a diagonal matrix with diagonal elements $\lambda_1 \geq \lambda_2 \ldots \geq \lambda_m$ containing the eigenvalues of $\mathbf{\Sigma}$. Columns of $\mathbf{V}$ are the corresponding eigenvectors.

A data set containing N realizations of the alternative state vector is generated by

$$\mathbf{Z}_{1d} = \mathbf{R}\mathbf{Q}_1 \mathbf{\Lambda}_p^{\frac{1}{2}} \mathbf{V}_p^{\mathrm{T}} + \boldsymbol{Q}_2 \sqrt{1 - \mathrm{tr}(\mathbf{\Lambda})} \tag{16}$$

where the subscript p is used to indicate the first p rows and columns of $\mathbf{\Lambda}$ and columns of $\mathbf{V}$. (Here $p = 5$.) ~~$\mathbf{R}_1$~~ $\mathbf{Q}_1$ is an $n \times p$ matrix of random values and ~~$\boldsymbol{R}_2$~~ $\boldsymbol{Q}_2$ is a column vector of random values. The random variates are independent and normally distributed with zero mean and unit variance. $\mathbf{Z}_{1d}$ is back-transformed (re-arranging Eq. 14) to obtain an $N \times m$ matrix containing N realizations of the state vector $\boldsymbol{S}(t_o)$ for the 1-D environment ensemble. The same analysis is applied to the n available instances of the advective flux perturbation vectors from the 3-D ensemble to generate N realizations of the $\mathbf{P}^\star$ vector.

Each of the N randomly generated environment realizations is used to provide a separate estimate of the parametric environment residual corresponding to a possible truth. For the ith ensemble member this is

$$\epsilon_{Bi} = g(\boldsymbol{B}_i, \boldsymbol{x}_o) - g\left(\overline{\boldsymbol{B}}, \boldsymbol{x}_o\right) \tag{17}$$

where $\boldsymbol{B}_i$ is the ith environment realization generated by the environment model. For the true environment, $\boldsymbol{B}_i$ would be $\boldsymbol{B}(\boldsymbol{x}_o)$, as in the informed simulator. The environment residual statistics $\mathrm{var}(\epsilon_B)$ and $\mathrm{E}(\epsilon_B)$ are approximated by $\mathrm{var}(\epsilon_{Bi})$ and $\mathrm{E}(\epsilon_{Bi}) : i \in \{1, \ldots, N\}$.

In Eq. (17), we rely on the simulator $g(.,.)$ to provide estimates for the terms $f(\boldsymbol{x}_o)$ and $g_T(E[\boldsymbol{B}_T(\boldsymbol{x})], \boldsymbol{x}_o)$ in Eq. (10). Thus, the estimated environment residual statistics are to some extent affected by basic simulation errors and will not be strictly independent of the statistics for the mean environment simulation residual ϵ_S.

It should be noted that the residual ϵ_B and its predicted distribution are dependent on the trial parameter vector $\boldsymbol{x}_o$. Hemmings and Challenor (2012) demonstrated that the dependency of environment error variance estimates on variations in the simulation trajectory over the parameter space is potentially important in the context of a parametric analysis. For this reason, estimation of the environment residual statistics must be performed for each trial parameter vector in the analysis, so is a significant overhead.

If the underlying distributions of the residuals ϵ_S and ϵ_B are taken to be Gaussian then they are fully described by their means and variances. Statistics for the uninformed simulator residual ~~ϵ_2~~ ϵ_1 are obtained under the assumption that ϵ_S and ϵ_B can be considered only weakly dependent such that

$$E(\epsilon_1) = E(\epsilon_S) + E(\epsilon_B) \tag{18}$$

and

$$\text{var}(\epsilon_1) \approx \text{var}(\epsilon_S) + \text{var}(\epsilon_B). \tag{19}$$

Any indirect dependency between ϵ_S and ϵ_B that might arise from their dependencies on the simulator design are ignored. The uninformed simulator statistics are determined by substituting our estimates for the residual statistics for each error component in Eqs. (18) and (19). In doing so, we also ignore potential dependency arising from the effect of basic simulation errors on $\text{var}(\epsilon_{Bi})$.

A conceptual overview of the data flow for the indirect method is given in Fig. 2. Once again, the processing is divided into a construction phase intended for single execution and an application phase to be applied with each trial parameter vector. The procedure for assessment of the uninformed emulator is included in the application phase. ~~Also included is a corresponding~~

procedure for assessing the *informed emulator*. That is the array of informed simulators with its own predicted bias and error variance. Its assessment provides an indication of the quality of informed simulator statistics used in constructing the uninformed emulator.

5 Experimental method

Anticipating the use of satellite ocean colour data for model calibration, an emulator was constructed for the NEMO-MEDUSA surface chlorophyll output at an array of oceanic sites. The surface chlorophyll concentration is the sum of the surface level chlorophyll concentrations for the two phytoplankton types. Data for defining the biogeochemical environment were provided by a ~~small~~ 10 member reference ensemble of global 3-D simulations with the NEMO-MEDUSA target model. For emulator assessment, the known "truth" for a given trial parameter vector is defined by chlorophyll output from a target model simulation with that parameter vector.

5.1 1-D experimental framework

To provide a representative range of oceanic conditions for the experiments, 12 sites were selected, located on a meridional transect along 20° W in the North Atlantic at 5° intervals from 5 to 60° N. This spans the sub-tropical gyre and temperate regions further north where large spring blooms are typical, extending into the sub-polar gyre south of Iceland. To the south, it also crosses a high productivity region off the East African coast between the shelf break and the Cape Verde Islands.

Physical forcing data for the 1-D experiments, in the form of vertical velocity w_p, the vertical diffusion coefficient K_ρ and temperature are taken from 5 day mean output common to all of the 3-D NEMO-MEDUSA simulations. 5 day mean time series of downwelling solar radiation at the sea surface and the soluble iron flux from dust deposition are likewise taken from 5 day data common to all reference simulations.

Biogeochemical environment vectors for the 1-D experiments are based on initial state vectors and applied perturbation vectors from 1 or more 3-D simulations. Initial concentrations are taken from NEMO-MEDUSA restart files. Approximate values for the applied perturbation $p^\star$ are derived from the target model's 5 day mean current vector and primary tracer concentration fields using Eq. (7).

1-D simulations use the same vertical grid as the 3-D NEMO-MEDUSA simulations. The dynamics of interest are largely confined to the upper ocean where the seasonal signal is most pronounced. A depth threshold of 1000 m was therefore chosen for the simulations, reducing the number of model levels from 63 to 37 with consequent computational savings. Level 36 spans the 1000 m threshold and Level 37 is included purely to act as a sink for detritus entering from above. In the target model, sinking detritus is re-mineralized at the bottom of the water column. In the simulator it is re-mineralized in Level 37 instead and the vertical velocity and diffusion at the bottom of Level 36 are set to zero to prevent any interaction between Level 37 and the water column above. Zeroing the vertical velocity does have the effect of introducing an anomalous divergence in the vertical flow but the effect on the overall simulation is negligible. The upper ocean levels have boundaries at depths 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 and 1093 m.

The schemes used for vertical tracer transport are the same as those used in the target model and are described by Madec (2008). The diffusion scheme is an implicit scheme and the advection scheme is the Monotonic Upstream Scheme for Conservative Laws (Van Leer, 1977; Hourdin and Armengaud, 1999), introduced into NEMO for use in biogeochemical modelling studies by Lévy et al. (2001). A 1 h forward Euler time step is used.

5.2 Model parameter space

Full details of the derivation of the parameter space for the emulation experiments are given in Appendix B. Initially, a 28-dimensional parameter space of interest was defined; 28 parameters of particular relevance to the seasonal plankton dynamics in the upper ocean were selected from a set of 60 potential input parameters in the MarMOT 1.1 implementation of MEDUSA. The

parameter bounds were defined according to a set of rules designed to ensure that parameter values within the bounds are biologically plausible with respect to their defined roles.

The set of adjustable input parameters differs from the set of internal model parameters defined by Yool et al. (2011) due to a number of modifications made to facilitate parametric analyses. For example, where pairs of parameters such as rate parameters are used in the model for the two different phytoplankton types, the diatom parameter has been replaced in the input vector by the ratio of the two internal parameters. The input non-diatom parameter then scales both of the internal phytoplankton parameter values without affecting their relationship, while the new input parameter controls the relationship. The zooplankton parameters are treated similarly. The changes allow us to consider the effects of a phytoplankton rate parameter or a zooplankton rate parameter on the system without having to consider the impact of directly changing the relationship between rates for closely related plankton types. It is then easier to interpret parameter effects at a high level of abstraction which facilitates comparison with simpler models where parameters represent rates for more aggregated plankton compartments.

The dimensionality of the initial parameter space was reduced further with reference to a sensitivity analysis, performed at the experimental sites, to identify parameters that are influential with respect to annual primary production and sinking particle flux outputs from the model (see Appendix B). Improving the reliability of these outputs in the target model will be important for understanding and predicting change in the global carbon cycle. 8 model parameters were chosen on the basis of the findings. The corresponding parameter space is defined by Table 1.

One finding of the sensitivity analysis was that the input parameters controlling the relationship between associated internal parameters for different plankton types were less influential than the input parameters exerting control over the different plankton types jointly. None of the input parameters from the first set were selected. The mapping of input parameters to internal parameters means that varying any of the 5 non-diatom phytoplankton parameters in Table 1 will also change the corresponding internal diatom parameters in proportion. The non-diatom density-independent loss rate and half-saturation concentration for density-dependent loss will additionally affect the corresponding internal parameters for both zooplankton types

in proportion and the microzooplankton grazing half-saturation concentration will affect the corresponding internal parameter for mesozooplankton in the same way.

5.3 3-D reference simulations

A 10 member ensemble of 3-D simulations was used to create a reference sample of NEMO-MEDUSA output data that is representative of variability in the target model solution over the defined parameter space. The 10 parameter vectors are distributed in parameter space according to a Latin hypercube design (McKay et al., 1979). For improved coverage, a “maximin" criterion (Johnson et al., 1990) was applied to 1000 randomly generated hypercubes: the hypercube design is selected that maximizes the smallest Euclidean distance between parameter vector pairs in terms of their positions on a parameter space grid with an equal number of intervals in each dimension. Grid intervals are in log units for rate parameters and half-saturation concentrations.

The chosen parameter vectors are given in Table 2. NEMO-MEDUSA integrations were performed for each of the 10 parameter vectors to provide representative output for a 2 year period, beginning in 1997. The second year, 1998, is the first complete year for which satellite ocean colour data from the SeaWiFS sensor are available (although these data are not used in the present study). The integrations, at 1° horizontal resolution, were initialized from the NEMO-MEDUSA simulation of Yool et al. (2011) at the beginning of 1995 and integrated for 4 years with their respective modified parameter sets, thereby allowing a 2 year spin-up period prior to any analysis to attenuate the worst effects of transient behaviour with respect to the seasonal cycle in the upper ocean. A longer spin-up time would normally be envisaged for a practical application, consistent with the intended use of the target model.

The 3-D reference sample is used in two ways. Chlorophyll records are used for evaluating 1-D simulation error, while the initial concentrations and horizontal gradients of the biogeochemical tracers are used to provide parameter-specific environment information for 1-D simulator construction.

5.4 Emulator construction and assessment

Performance of the basic 1-D simulator array is evaluated, with respect to surface chlorophyll, for a ~~-~~set of trial parameter vectors ~~drawn from the parameter space under two different scenarios. The first is an informed simulator scenario in which~~ for which the true target model output is known. The performance of emulators constructed using the two uncertainty quantification methods are then assessed. Finally, to explore the importance of the lateral flux perturbations, we assess the performance of simulator arrays in which these are omitted. In this context, the behaviour of an alternative array employing informed simulators is examined in addition to that of the uninformed simulator array used in the ~~parameter-specific environment is known from a NEMO-MEDUSA reference simulation and used as input to the 1-D simulator. The second is an uninformed simulator scenario in which the parameter-specific environment is taken to be unknown. The robustness of the corresponding emulators is then assessed~~emulator. Doing this allows us to see the impact of omitting lateral flux perturbations in a scenario where other error sources are minimized. The experimental methods for the assessments are as follows.

5.4.1 Simulator assessment

~~The skill of the informed simulator is described here~~ Informed simulator skill is described by error statistics calculated from a set of 10 experiments with the representative parameter vectors defined in Table 2, so that each experiment corresponds to one of the available 3-D reference simulations. In each experiment, the informed simulator is initialized at the start of 1997 and run for 2 years. ~~Experiments were performed with and without the application of perturbations representing the effects of lateral advection.~~ If the set of representative parameter vectors is denoted by $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots \boldsymbol{x}_{10}\}$ then the trial parameter vector for the ith experiment is $\boldsymbol{x}_i$ and the environment is defined by the 3-D ensemble member with parameter vector $\boldsymbol{x}_i$.

The error statistics describing the skill of the uninformed simulator were determined from 10 ~~further~~ similar experiments, covering the same time period~~, were performed to investigate the skill of the uninformed simulator with lateral advection perturbations~~. One experiment was performed for each parameter vector in $\boldsymbol{X}$ but simulator construction was performed on a leave-

one-out basis: in the ith experiment, the trial parameter vector is $\boldsymbol{x}_i$ and the mean environment is derived from the 9 NEMO-MEDUSA ensemble members with $\boldsymbol{x} \neq \boldsymbol{x}_i$, $\boldsymbol{x} \in \boldsymbol{X}$, leaving the NEMO-MEDUSA output $f(\boldsymbol{x}_i)$ as independent data for validation. Thus, each experiment uses a slightly different version of the simulator, constructed by applying the same method to a different 9 member ensemble.

Error statistics are calculated with respect to the log-transformed 5 day mean chlorophyll output. The log transformation applied to the 5 day means acts to stabilize the error variance which otherwise tends to increase with increasing chlorophyll concentration. Its use in the analysis of surface chlorophyll variability is strongly supported by theoretical considerations and empirical data (Campbell, 1995).

5.4.2 Assessment of the full emulator ~~robustness~~

Validation of the complete uninformed emulator for surface chlorophyll is by analysis of the results from the 10 leave-one-out experiments, taking into account the predicted simulator error statistics to determine the emulator robustness. These uncertainty estimates are, like the simulator itself, required to be independent of parameter-specific environment information. Thus, for the ith experiment, they are derived using the 9 NEMO-MEDUSA ensemble members with $\boldsymbol{x} \neq \boldsymbol{x}_i$. The uninformed emulator uncertainty is quantified using the direct and indirect methods.

When the indirect method is used, the 9 NEMO-MEDUSA ensemble members are used to derive statistics for the two component residuals ϵ_{S} and ϵ_{B}. In the estimation of the statistics for the mean environment simulation residual ϵ_{S} (assumed identically distributed to the informed simulator residual ~~ϵ_1~~ϵ_2), the 3-D ensemble members are required for comparison with the corresponding informed simulators to determine informed simulator error. In the estimation of the statistics for the parametric environment residual ϵ_{B}, the 3-D ensemble is required for building the environment model used in the parametric uncertainty analysis.

When the direct method is used, the 9 NEMO-MEDUSA ensemble members are used to derive statistics for the uninformed simulator residual ~~ϵ_2~~ϵ_1. Each of the 9 corresponding uninformed simulators require independent data for their mean environment input. In the ith exper-

iment, the mean environment for the uninformed simulator with parameter vector $\boldsymbol{x}_j$ is derived from the 8 NEMO-MEDUSA members with $\boldsymbol{x} \neq \boldsymbol{x}_i \cap \boldsymbol{x} \neq \boldsymbol{x}_j$. As a result, simulators must be constructed with 90 different mean environment estimates to calculate the uncertainty estimates for the 10 experiments.

For the uncertainty quantification analyses, Gaussian error distributions in log-transformed chlorophyll are assumed so that the resulting probability density functions for the residuals are fully described by their mean and variance, both of which are allowed to vary in time and between sites. The residuals are defined with respect to log-transformed 5 day mean chlorophyll concentrations. Their predicted distributions are described by their monthly means and variances, interpolated to 5 day intervals. Appendix C gives the estimation method for the residual statistics and the resulting time series.

6 Results

The surface chlorophyll records from the 3-D NEMO-MEDUSA reference ensemble at each of the experimental sites are shown in Fig. 3. This shows the spatial variation in chlorophyll from values a little above 0.001 $\mathrm{mg\,m^{-3}}$ in the oligotrophic gyre at 30° N and 35° N for Parameter Set 6 to seasonal highs associated with the spring bloom in temperate regions (45–60° N), approaching $10\,\mathrm{mg\,m^{-3}}$ for a number of the parameter vectors. It also illustrates the variability in the seasonal response of the plankton dynamics which is generally stronger at the more northerly sites.

The variation between records produced by different parameter vectors is large compared with the seasonal variability. At some sites, particularly 5–10° N and 25–35° N, the parameter dependency manifests primarily as a control on the overall chlorophyll concentration level in the surface layer, throughout the annual cycles. These are generally the more oligotrophic sites, where concentrations remain below or very close to $1\,\mathrm{mg\,m^{-3}}$ for all parameter vectors. At other sites, particularly in the north, the different parameters also have a notable influence on the dynamic range and there is some evidence of an impact on the characteristics of the spring bloom. ~~Some parameter sets~~

Some parameter vectors tend to have the same effect on overall surface chlorophyll levels at all sites. For example, Parameter Set 10 gives elevated levels over the whole data set. However, this is not generally the case. Parameter Set 6, for example, shows a strong tendency to give low chlorophyll concentrations at ~~most~~ many of the sites but gives some of the higher concentrations at 55 and 60° N. With this parameter vector, the phytoplankton light-response controlled by α_{Pn} is exceptionally strong and nutrient-limitation is reduced by low half-saturation concentrations $k_{\mathrm{N,Pn}}$ and $k_{\mathrm{Fe,Pn}}$. As a result, the phytoplankton can achieve very high growth rates. This can cause blooms that lead to long-term nutrient depletion as a consequence of organic material sinking out of the euphotic zone. Subsequent growth is then inhibited. At 4 sites (5, 10, 30 and 35° N), nitrogen depletion during the 2 year spin-up period results in very low chlorophyll concentrations at the start of 1997 which remain relatively low throughout 1997 and 1998.

Parameter Sets 1 and 4 also lead to some interesting site-specific impacts. They are associated with very low winter-time chlorophyll concentrations at the most northerly sites, particularly in 1997, although are associated with some of the highest concentrations throughout 1997 and 1998 at the most southerly sites. These parameter vectors combine low α_{Pn} values with low values for the grazing half-saturation concentration k_{μ}, reducing phytoplankton production at low light levels and making them more susceptible to zooplankton grazing. This makes the phytoplankton less well-suited to over-wintering at the high latitude sites where light availability is very low due to the combination of low surface irradiance and deep winter mixing.

The strong variation between parameter vectors indicates the potential for significant constraints on the parameter values to be realized by the assimilation of satellite chlorophyll data.

6.1 ~~Simulator performance with known parameter-specific environment~~Emulator prediction of target model output

Chlorophyll concentrations given by the ~~informed~~ uninformed simulator at all sites are compared against the corresponding values from the matching 3-D experiment in Fig. ~~??~~4. Data are shown for the 1998 annual cycle only so are representative of the simulator performance one year on from its initialization year, during which errors have had time to develop. ~~Results are shown for simulations with and without lateral flux perturbations.~~

The ~~introduction of lateral flux perturbations improves the performance of the simulator array considerably. Without these perturbations,~~ correlation between simulator and target model values is good. Pearson's correlation coefficient r for the simulator and target model output is ~~0.75~~0.91, indicating that ~~just 56~~83 % of the variance in the log-transformed surface chlorophyll from the simulator array is explained by the target model output. ~~With the parameter-specific perturbations, the variance explained increases to 86($r = 0.93$). In general, surface chlorophyll is then well reproduced by the simulators.~~ There are some notable examples of poor performance though. In particular, the ~~simulator~~ results for Parameter Set 6 indicate a ~~very strong negative biasat low concentrations~~ strong positive bias, with the simulator array ~~underestimating some~~ overestimating some surface chlorophyll values by an order of magnitude. ~~Similarly, for other parameter vectors, there are some large positive biases at higher concentrations, although the problem is less systematic.~~

~~Error statistics for the informed simulator results, with and without lateral flux perturbations, are given for each site in Fig. 9. The error for each log-transformed 5day mean chlorophyll concentration is defined by~~

~~$$d_1 = g[\boldsymbol{B}(\boldsymbol{x}_o), \boldsymbol{x}_o] - f(\boldsymbol{x}_o)$$~~

~~where $\boldsymbol{B}(\boldsymbol{x}_o)$ is the appropriate set of environmental input data, either including or not including perturbations.~~

~~The use of lateral flux perturbations leads to strong reductions in bias and r.m.s. error at most of the low and mid-latitude sites to 40N, and at 50N from the summer of 1998 onwards. The improvement is particularly notable at~~ There are some fairly large negative biases for other parameter sets, notably Parameter Sets 7 and 10 ~~N, 25N, 35N and 40N, where the addition of these perturbations correct a long-term drift very effectively, albeit with slight over-correction of the positive bias at 10N. Performance is a little more equivocal at 20N where perturbation of the simulation leads to a relatively large over-correction of a negative bias but the overall r.m.s. error is still reduced.~~

~~The perturbed simulator does not perform better everywhere. The main exception is seen at 60N, where the simulator shows a tendency to over estimate chlorophyll in the summer of 1998.~~

Another exception is an over correction of the positive bias at 50N in 1997 which leads to a bias of larger magnitude over some parts of the year. These detrimental effects are minor compared with the overall improvement achieved.

6.2 Emulation with parametric uncertainty in the biogeochemical environment

Chlorophyll concentrationsgiven by the uninformed simulator array in 1998 are compared against the corresponding values from the matching 3-D experiment in Fig. 4. As before, results are shown for simulations with and without lateral flux perturbations.

The results obtained without flux perturbations are very similar to the informed simulator results obtained when flux perturbations are not used. Of the variance in the uninformed simulator output, just 54is explained by the target model ($r = 0.73$). Comparison with Fig. **??**a shows differences for some parameter vectors, most notably Parameter Set 6 for which the positive bias at low concentrations is an order of magnitude greater than that of the informed simulator array . These differences are due only to differences in the initial state at the beginning of 1997, specifically the replacement of the parameter-specific initial state by a mean state. Although Parameter Set 6 simulations appear sensitive to the change in initial state under certain conditions, this is the exception rather than the rule. In fact, for most parameter vectors, comparisons between Figs. 4 and **??** show very little difference, suggesting that sensitivity to variations in initial state consistent with its expected uncertainty range is generally very low.

With the addition of lateral flux perturbations, the overall performance of the uninformed simulator array is greatly improved, although less so than that of the informed simulator array. The correlation between simulator and target model values is less than that for the corresponding informed simulators, with 83of the simulator variance explained by the target model ($r = 0.91$) compared with 86for the informed simulator. Once again, performance is poor for at mid-range concentrations, although these are less systematic. Also, the simulator array poorly reproduces the relatively low variability in chlorophyll associated with Parameter Set 6, although now the simulated chlorophyll values for this parameter vector have a positive bias. The large change in bias with the change from a parameter-specific environment to a mean environment is consistent

~~with the corresponding change in the absence of flux perturbations, so the change in initial conditions is likely to be the dominant factor.~~ 1.

The chlorophyll output from the uninformed emulator includes a bias correction term which depends on the uncertainty quantification method. (This corrects for spatio-temporal biases rather than for parameter-related biases.) When using the direct uncertainty quantification method, the bias-corrected error in log-transformed 5 day mean chlorophyll is

$$d_{\mathrm{Ud}} = g\left(\overline{\boldsymbol{B}}, \boldsymbol{x}_{\mathrm{o}}\right) + \overline{u}_{21} - f(\boldsymbol{x}_{\mathrm{o}}) \tag{20}$$

where ~~$\overline{u}_2$~~ $\overline{u}_1$ is our estimate of E~~(ϵ_2)~~(ϵ_1). When using the indirect method, the bias correction includes corrections for both the mean environment simulation bias and the bias associated with parametric environment uncertainty. The bias-corrected error is then

$$d_{\mathrm{Ui}} = g\left(\overline{\boldsymbol{B}}, \boldsymbol{x}_{\mathrm{o}}\right) + \overline{u}_{\mathrm{S}} + \overline{u}_{\mathrm{B}}(\boldsymbol{x}_{\mathrm{o}}) - f(\boldsymbol{x}_{\mathrm{o}}) \tag{21}$$

where $\overline{u}_{\mathrm{S}}$ and $\overline{u}_{\mathrm{B}}$ are our estimates of $\mathrm{E}(\epsilon_{\mathrm{S}})$ and $\mathrm{E}(\epsilon_{\mathrm{B}})$ respectively and $\overline{\boldsymbol{B}}$ is our estimate of the mean environment. The estimates ~~$\overline{u}_2$~~ $\overline{u}_1$ and $\overline{u}_{\mathrm{S}}$ were determined without reference to results for Parameter Set 6. These were excluded on the basis of the unrepresentative simulator performance, to avoid excessive influence from a single outlier. Time series of $\overline{u}_2$ and $\overline{u}_{\mathrm{S}}$ are therefore based on an ensemble size of 8 (or 9, when Parameter Set 6 is the trial parameter vector).

Error statistics for the uninformed emulator results are given in Fig. 5. Results are presented for the basic simulator array with no bias correction (~~$\overline{u}_2 = \overline{u}_{\mathrm{S}} = \overline{u}_{\mathrm{B}} = 0$~~$\overline{u}_1 = \overline{u}_{\mathrm{S}} = \overline{u}_{\mathrm{B}} = 0$) and for the full emulator with bias correction. There are only minor differences between the mean and r.m.s. values for d_{Ud} and d_{Ui}.

~~A comparison between the uninformed simulator results without bias correction and the corresponding results given by the informed simulators (Fig. 9) show major differences in early 1997. The relatively poor uninformed simulator results for this period are the consequence of transient behaviour associated with error in the initial conditions. This error source seems to influence the model primarily in the early half of the year before the local dynamics start to~~

dominate over the environmental influences. The lack of parameter-specific information about the lateral fluxes appears to be less serious. Nevertheless, in 1998, the errors typically remain slightly larger at most sites than the corresponding informed simulator errors. Much larger errors are seen at some sites, particularly at 5N, between 25 and 35N and at 50N.

Omitting lateral flux perturbations altogether can lead to particularly large biases associated with serious drifts as seen in Fig. 9. However, examination of the uninformed simulator results in Fig. 5 show that even at the sites where the error is relatively large, the biases are not. This indicates that a scheme based on average flux perturbations for the parameter space (i.e. the mean environment) can solve the problem of drift to a large extent, even though the environment information is not parameter specific.

Biases are further Biases are reduced by the emulator's bias correction scheme, irrespective of the method used. Time series of simulator bias before and after correction (Fig. 5) show that in both cases the bias correction is effective at all sites, with the possible exception of 20° N where d_{Ui} shows the introduction of a negative bias in the summer of 1998 when using the indirect uncertainty quantification method. In particular, note that the summer 1998 bias at 60° N is largely removed and the correction is particularly effective in removing negative bias at some of the more oligototrophic sites (5° N and 25–30° N) and at 50° N in 1997.

The relatively high r.m.s. errors for early 1997 at most sites are the consequence of transient behaviour associated with error in the initial conditions. This source of error seems to influence the model primarily in the early half of the year, before the local dynamics start to dominate over the environmental influence. The lack of parameter-specific information about the lateral fluxes appears to be a less dominant source of simulation error. Nevertheless, it does contribute strongly to the relatively large 1998 errors at 5° N and at 50° N.

6.2 Robustness of the emulator

The robustness of the uninformed emulator is assessed by comparing the MarMOT-MEDUSA chlorophyll records with the NEMO-MEDUSA results for the matching parameter sets, taking into account the quantified emulator uncertainty in terms of the predicted bias and error variance. The results are presented here in terms of the normalized emulator error, which is the

error in the bias-corrected simulator output scaled by the reciprocal of its predicted standard deviation. The scaling factor ensures that the predicted normalized error distribution for both versions of the emulator is Gaussian with zero mean and unit standard deviation at all times and locations.

The normalized uninformed emulator error for each log-transformed 5 day mean surface chlorophyll concentration depends on the uncertainty quantification method. For the direct method, it is given by

$$D_{\mathrm{Ud}} = \frac{g\left(\overline{\boldsymbol{B}}, \boldsymbol{x}_{\mathrm{o}}\right) + \overline{u}_2 - f(\boldsymbol{x}_{\mathrm{o}})}{s_2} \frac{g\left(\overline{\boldsymbol{B}}, \boldsymbol{x}_{\mathrm{o}}\right) + \overline{u}_1 - f(\boldsymbol{x}_{\mathrm{o}})}{s_1} \tag{22}$$

where ~~s_2^2~~ s_1^2 is our estimate for ~~$\mathrm{var}(\epsilon_2)$~~ $\mathrm{var}(\epsilon_1)$. For the indirect method, it is

$$D_{\mathrm{Ui}} = \frac{g\left(\overline{\boldsymbol{B}}, \boldsymbol{x}_{\mathrm{o}}\right) + \overline{u}_{\mathrm{S}} + \overline{u}_{\mathrm{B}}(\boldsymbol{x}_{\mathrm{o}}) - f(\boldsymbol{x}_{\mathrm{o}})}{\sqrt{s_{\mathrm{S}}^2 + s_{\mathrm{B}}^2(\boldsymbol{x}_{\mathrm{o}})}} \tag{23}$$

where s_{S}^2 and s_{B}^2 are our estimates for $\mathrm{var}(\epsilon_{\mathrm{S}})$ and $\mathrm{var}(\epsilon_{\mathrm{B}})$ respectively. ~~s_2^2~~ s_1^2 and s_{S}^2, like the residual mean estimates ~~$\overline{u}_2$~~ $\overline{u}_1$ and $\overline{u}_{\mathrm{S}}$, were determined without reference to the results for Parameter Set 6, so were likewise based on a sample size of 8 (or 9 when Parameter Set 6 is the trial parameter vector). The denominator in Eq. (22) varies between 0.014 and 0.62 $\log_{10}$ units and that in Eq. (23) varies between 0.015 and 0.50 $\log_{10}$ units (with chlorophyll concentration in $\mathrm{mg\,m^{-3}}$). Further details of the residuals' statistics and their variation in time and between sites can be found in Appendix C.

The normalized uninformed emulator errors for each experiment are shown in Fig. 6. In Experiment 6 (pertaining to trial Parameter Set 6), the positive errors already noted are extreme, relative to the predicted error variance. This is a consequence of the unusually large simulator errors associated with Parameter Set 6. The atypical behaviour associated with this parameter vector may be truly representative of the model dynamics over a significant region of parameter space. However, such detail is not resolved with our small sample so is not represented in

the data used for emulator construction. Large normalized error values in Experiment 6 are therefore unsurprising.

When the indirect uncertainty quantification method is used, Fig. 6 shows that there are also very large extremes associated with the post-initialization phase, particularly at 55° N and 60° N. These high D_{Ui} values occur for experiments with the two parameter vectors ~~(Parameter Sets 1 and 4) that lead to~~ that were seen to cause unusually low winter-time chlorophyll concentrations at the start of 1997 in the target simulation (Fig. 3, Parameter Sets 1 and 4). Fortunately, at these sites, the extreme error appears fairly transient, lasting only a few months. At other sites, in particular at 5° N, D_{Ui} remains correlated to some extent with its early 1997 value over the whole 2 year period, suggesting that parametric error in the initial state may be introducing ~~a persistent bias~~persistent biases. This pattern seems to be a common feature of the more oligotrophic sites, being reflected also at latitudes from 25 to 35° N. At more northerly sites, there is a tendency for persistent biases over long time periods where relatively large errors occur (e.g. at 45° N and 50° N for the indirect method) but this pattern develops later with no obvious connection to initialization error.

Comparing the two uncertainty quantification methods, it is seen that D_{Ui} initially tends to be larger than D_{Ud} at all sites. The post-initialization D_{Ud} values are more consistent with their predicted distribution. In particular, the extreme positive D_{Ui} values seen in early 1997 are not replicated in D_{Ud}. From these observations, it is clear that the indirect method is generally less effective at quantifying initial uncertainty. Furthermore, at the oligotrophic sites where the early 1997 biases tend to persist, there is a general tendency for D_{Ui} to be larger than D_{Ud} over the 2 year period.

The normalized error distributions for the uninformed emulators are compared with the predicted distribution in Fig. 7. Results, including 1998 data only, are shown for each site. Experiment 6 is excluded to allow the results for the remaining experiments to be more clearly represented. The emulator with direct uncertainty quantification appears fairly robust with D_{Ud} distributions broadly similar to the predicted distribution at all sites. The worst performance is arguably at 30° N where there are a significant proportion of anomalously low values associated with persistent negative errors in the experiments with Parameter Sets 1 and 4 (Fig. 6). How-

ever, D_{Ui} shows a strong tendency to be larger than expected at a number of the sites. In general, these are the sites that have already been associated with persistent error in some of the experiments (5° N, 25–35° N, 45–50° N). A smaller proportion of the D_{Ud} values at 15 and 20° N are rather larger than predicted. These are associated with extreme negative biases occurring in Experiment 9 that persist only for a month or two.

The performance of the emulator is further evaluated by defining prediction intervals for the log-chlorophyll output and testing the proportion of the independent validation data that fall within these intervals. This test takes into account the expected error in the predicted means of the residuals, given the sample sizes used in emulator construction. For a perfect emulation scheme, we should expect approximately 95of the validation data from the target model to fall within the emulators' 95prediction intervals. This prediction interval is

$$g' \pm t_{n-1} s \sqrt{1 + \frac{1}{n}}$$

where g' is the bias-corrected simulator output, s is the standard deviation estimate for the simulator residual, n is the sample size on which it is based and t_{n-1} is the 97.5 percentile of Student's t-distribution with $n - 1$ degrees of freedom. The equivalent prediction interval for the normalized error D is

$$\pm t_{n-1} \sqrt{1 + \frac{1}{n}}$$

In practice, the effective sample size for the monthly residual statistics can be expected to vary with the degree of temporal autocorrelation between 5day samples which has not been quantified. When the direct method is used, the effective sample size is normally known to be in the range 9 to 54. (This range applies to the 11 out of every 12 pseudo-monthly bins that contain 6 5day samples, if using all 9 ensemble members.) This implies that $2.03 \leq |D| \leq 2.43$.

The effective sample size is less well-defined when the indirect method is used because of the way the parametric environment residual is derived: a relatively large ensemble size is used

~~for the parametric uncertainty analysis but the statistical model for its input data is based on a sample of just 9 biogeochemical environments. For the purposes of evaluating the uncertainty predictions for the emulator, the 95prediction interval for the normalized error is taken to be the mid-range value of ± 2.23 irrespective of which uncertainty quantification method is used.~~

Table 3 summarizes the uninformed emulator results in terms of the mean and standard deviation of the normalized errors~~and the proportion of the validation data that fall within the prediction interval, which we describe as the prediction "success rate"~~. Statistics are given for all 10 experiments combined and in brackets for the 9 experiments excluding Experiment 6. The difference between the two sets of results illustrate to some extent the sensitivity of the evaluation statistics to sampling error.

When the emulator performance with direct uncertainty quantification is evaluated over all experiments and all sites, ~~there is a success rate of 92. This, together with the high~~ the D_{Ud} standard deviation ~~of~~ is rather high at 1.41, ~~suggests~~ suggesting that the emulator is a little over-confident. When Experiment 6 is excluded from the evaluation, the ~~success rate goes up to 94and the~~ standard deviation drops to 1.13. Whether or not this is a more appropriate measure of performance depends on the extent to which the model dynamics with Parameter Set 6 are representative of its behaviour over a significant region of parameter space. The performance with respect to the other parameter vectors is fairly reliable at all sites, with ~~success rates from 90 to 99,~~ standard deviations from 0.98 to 1.31 and very little sign of post-correction bias shown by D_{Ud} mean values. All but 2 of the standard deviations are above 1, indicating a slight tendency for the spread of the simulator residuals to be under-estimated. When Experiment 6 results are included in the evaluation data set, this tendency for over-confidence is more evident and there are notable positive biases at a number of sites (D_{Ud} mean greater than 0.3 at 10, 15 and 50° N). These are associated with relatively large D_{Ud} standard deviations (1.55 to 2.07). ~~Nevertheless, the site-specific success rate does not fall below 88.~~

The ~~overall success rate drops from 92 to 88if we use the indirect uncertainty quantification method instead of the direct method. This, together with the~~ high standard deviation in D_{Ui} of 1.82, is consistent with results already presented that show the emulator with indirect uncertainty quantification has a clear tendency towards over-confidence in its predictions. ~~The~~

overall success rate goes up to 90if If Experiment 6 is excludedand the , the overall standard deviation is less at 1.39, but the performance still leaves some room for improvement. The over-confidence is particularly notable at the highly oligotrophic site at 5° N, with a standard deviation of just over 2 reflecting the persistent parameter-specific biases already noted at this site (Fig. 6). There is also a tendency for the emulator with indirect uncertainty quantification to significantly under-estimate chlorophyll concentrations. In particular, the 9 parameter vector sample shows large negative biases of around -0.7 at some of the other oligotrophic sites (20, 25 and 35° N). Fairly large negative biases of -0.30 and -0.38 are also seen at 5 and 15° N respectively. Nevertheless, the performance at a number of the sites is good, with 7 out of 12 sites showing success rates of 90or more when Experiment 6 is excluded (10–20. The subset of 5 sites at 10° N, 40–45° N and 55–60° N). Other performance measures for these sites are quite reasonable too, although the negative normalized biases at 15 and 20N are a little large, approaching 0.7 at the latter site. has standard deviations in the range 0.74 to 1.32 with small biases (-0.1 to 0.2).

The indirect uncertainty quantification method relies in part on statistics estimated from analysis of the informed emulator. Specifically, the statistics for the mean environment simulator residual ϵ_S are equated to those for the informed simulator residual ϵ_1, under the pragmatic assumption that the two residuals are identically distributed. As a way of investigating the quality of the ϵ_1 statistics, it is useful then to assess the robustness of the informed emulator. The normalized error for the informed emulator is

$$D_I = \frac{g(\boldsymbol{B}(\boldsymbol{x}_o), \boldsymbol{x}_o) + \overline{u}_1 - f(\boldsymbol{x}_o)}{s_1}$$

where $\overline{u}_1$ and s_1^2 are our estimates of the expected simulation error $E(\epsilon_1)$ and the simulation error variance $\text{var}(\epsilon_1)$ ($\overline{u}_1 = \overline{u}_S$ and $s_1 = s_S$). The denominator in Eq. (??) varies between 0.012 and 0.25 $\log_{10}$ units.

6.3 The importance of lateral advection

Table ?? summarizes the results obtained with In the majority of site-based calibration studies, the informed emulator. The performance shows a similar degree of robustness to that of the uninformed emulator with direct uncertainty quantification. The overall success rate is effect of lateral advection is ignored. It is useful then to examine the extent to which the skill of our 1-D simulations is dependent on the explicit representation of the advective flux divergence term. Fig. 8 shows the chlorophyll values given by the uninformed simulator array compared with the matching target model output when the uninformed simulators are run with all lateral flux perturbations removed. Comparison with Fig. 4 shows that the omission of lateral flux perturbations degrades the performance of the simulator array considerably. Pearson's correlation coefficient r for the simulator and target model output drops from 0.91 to 0.75, indicating that just 56 % of the variance in the same at 92log-transformed surface chlorophyll from the simulator array is explained by the target model output, compared with 83 % , rising to 95when Experiment 6 is excluded. The impact of Experiment 6 on the overall D_{I} standard deviation of is greater than that on the standard deviation of D_{Ud}, increasing from a promising value of 1.17 to 2.0 when this one experiment is included. With Experiment 6 excluded, the informed emulator performance appears fairly reliable with a success rate of over 90at all sites. The normalized biases are small everywhere and the standard deviations are relatively close to 1. However, they are above 1 at all but one of the sites, suggesting that the informed emulator may also have some tendency towards over-confidence in its predictions. in the standard simulator array with lateral flux perturbations.

The tendency shown by the informed emulator is small compared with that exhibited by the uninformed emulator with indirect uncertainty quantification at the affected sites (impact of omitting lateral flux perturbations is most clearly seen in the performance of the informed simulator array, where removing the effects of the parametric environment uncertainty minimizes other sources of error. The initial state error for this simulator array is zero and the lateral flux perturbations are parameter-specific. The error for each log-transformed 5 day mean chlorophyll concentration is defined by

$$d_{\mathrm{I}} = g[\boldsymbol{B}(\boldsymbol{x}_{\mathrm{o}}), \boldsymbol{x}_{\mathrm{o}}] - f(\boldsymbol{x}_{\mathrm{o}}) \tag{24}$$

where $\boldsymbol{B}(\boldsymbol{x}_o)$ is the appropriate set of environmental input data, either including or not including lateral flux perturbations. Error statistics for the informed simulator results, with and without perturbations, are given for each site in Fig. 9.

The use of lateral flux perturbations leads to strong reductions in bias and r.m.s. error at most of the low and mid-latitude sites to 40° N, and at 50° N from the summer of 1998 onwards. The improvement is particularly notable at 10° N, ~~25–35~~25° N, ~~45–50~~35° N ~~), irrespective of whether Experiment 6 is excluded. This, together with~~ and 40° N, where the addition of these perturbations correct a long-term drift very effectively, albeit with slight over-correction of the positive bias at 10° N. Performance is a little more equivocal at 20° N where perturbation of the ~~lack of any clear correlation between the performance of the two emulator versions across sites, suggests that error in the ϵ_1 statistics is very unlikely to be the primary factor causing under-estimation~~ simulation leads to a relatively large over-correction of a negative bias but the overall r.m.s. error is still reduced.

The perturbed simulator does not perform better everywhere. The main exception is seen at 60° N, where the simulator shows a tendency to over estimate chlorophyll in the summer of 1998. Another exception is an over correction of the positive bias at 50° N in 1997 which leads to a bias of larger magnitude over some parts of the year. These detrimental effects are minor compared with the overall improvement achieved.

It is clear from Fig. 9 that omitting lateral flux perturbations altogether can lead to particularly large biases associated with serious drifts. Biases of magnitude 0.6 $\log_{10}$ units, representing a factor 4 error in surface chlorophyll, are not uncommon. Examination of the uninformed ~~emulator uncertainty.~~simulator results in Fig. 5 before any bias correction shows that even at the sites where the error is relatively large, the biases are not. The largest biases are of magnitude 0.3 $\log_{10}$ units, equivalent to a factor of 2. This indicates that a scheme based on average flux perturbations for the parameter space (i.e. the mean environment) can reduce the problem of drift to a large extent, even though the environment information is not parameter specific.

7 Discussion

In this section, the performance of the experimental mechanistic emulator is first examined and scope for its improvement identified. Practical application of the site-based emulation scheme is then considered and its envisaged role in enabling advances in the parametric analysis and calibration of global biogeochemical models is discussed.

7.1 Mechanistic emulator performance

Two alternative versions of a mechanistic emulator for surface chlorophyll from global NEMO-MEDUSA simulations have been evaluated. Each of these site-based emulators uses the same set of site-specific 1-D simulators. The two emulators differ in the method they employ to quantify uncertainty in the simulator predictions.

The site-based emulator with direct uncertainty quantification is able to predict the 1998 chlorophyll record for a given parameter vector to an accuracy broadly consistent with its uncertainty prediction at all sites. It should therefore serve as a reasonably reliable emulator of the target model for parametric analyses. There is a slight tendency to under-estimate the uncertainty, which is likely to be a consequence of the small target model ensemble size used to represent the known truth (8 or 9). This interpretation would be consistent with a parametric uncertainty analysis of a regional 3-D biogeochemical model by Fiechter (2012), spanning a similar parameter space, in which an ensemble size of 10 was found to give significantly low estimates for ensemble spread compared with 25, 50 and 100 member ensembles. In a practical application, the tendency towards over-confidence could be compensated for by a small inflation factor applied to the residual variance estimate. The optimal factor would be the normalized error variance from the evaluation experiments (i.e. 1.28, based on the standard deviation of 1.13 for the 9 trial parameter vector experiments in Table 3).

The emulator with indirect uncertainty quantification is able to predict the 1998 record to an accuracy consistent with its uncertainty prediction at about half of the experimental sites, so clearly has some potential. However, it shows a tendency to be over confident in its predictions at other sites, particularly at the more oligotrophic sites studied. Its performance therefore requires

some improvement before it can be considered generally robust over a wide range of oceanic conditions.

The most notable instances of poor emulator performance occur for parameter vectors associated with the more extreme behaviour in the target model. This raises the question of whether it is really necessary to emulate the target model over such large parameter ranges. Certainly, restricting the parameter space further should help to make our reference sample more representative. In principle, comparisons with observational data at an early stage could be used to identify implausible target model behaviour and suggest ways in which the parameter space might be reduced. However, any such constraints based on the sparse sampling of parameter space achieved by the target model ensemble could greatly increase the risk of excluding promising parameter combinations and should be undertaken with care. Modifications of the parameter space that are consistent with our biological understanding of the parameters are the most easily justified but, acknowledging the high level of abstraction involved in modelling a system of such complexity, we should avoid over-reliance on subjective priors. Increasing the sample size or improving the emulation methods may be preferable.

While the indirect uncertainty quantification method is currently less robust than the direct method, it has the advantage of being less reliant on the small target model ensemble. Simulator uncertainty due to basic simulation error and parametric environment error are quantified separately, the latter being the uncertainty due to substitution of the true parameter-specific environmental input by a mean environment. The quantification method for basic simulation uncertainty relies wholly on the target model ensemble. However, that for parametric environment uncertainty relies on it only for providing environmental data for a 1-D uncertainty analysis. The output uncertainty depends on the way in which these input data interact with the parameter-specific dynamics in the 1-D simulators and the 1-D ensemble size can be relatively large. As found by Hemmings and Challenor (2012), the output standard deviation can be highly dependent on the trial parameter vector (see Appendix C). This parameter dependency cannot be accounted for by the direct method. For this reason, a refined version of the indirect method could prove to be more robust than the direct method, particularly if basic simulation errors can be reduced so that the uncertainty quantification for this error component becomes less critical.

There are a number of possible causes of the under-estimation of uncertainty by the indirect method: deficiencies in the statistical models for the informed simulator residual ϵ_I (associated with basic simulation errors) and the parametric environment residual ϵ_B, deficiencies in the statistical environment model used in estimating the ϵ_B statistics, under-sampling of the modelled distributions and violation of the assumptions required for quantifying the total emulator uncertainty. The ϵ_I statistics are believed to be fairly robust on the basis of the good performance of the informed emulator (i.e. that based on the informed simulator array). This suggests that the main problem is more likely to be associated with the characterization of uncertainty in ϵ_B or with violation of the total uncertainty quantification assumptions. These are that the mean environment simulation residual ϵ_S, attributable to basic simulation errors occurring when the mean environment is used, is identically distributed to ϵ_I and that ϵ_S and ϵ_B can be treated as independent for the purposes of combining error statistics.

Factors that might lead to under-estimation of the parametric environment uncertainty are under-sampling of the 1-D simulation space by the environment uncertainty ensemble or inadequate representation of the variability of the environment data over the parameter space by the statistical environment model. Under-sampling of the 1-D simulation space seems a less likely cause: a sample size of 100 was used. This is believed to be representative of the simulation distribution with scope for reduction in a practical application, although sensitivity to ensemble size has yet to be properly investigated.

The presence of very large normalized error values early in 1997 when the indirect uncertainty quantification method is used suggests that the environment model for the initial conditions should be improved, perhaps through the use of different variance-stabilizing transformations in the EOF analysis used to characterize the environmental uncertainty. Tracer-specific transformations should be considered in place of the square root transformations applied to all primary tracers. Another refinement that may improve performance in the post-initialization phase would be to include covariances between the initial state and the advective flux divergences of the transformed tracer concentrations, instead of modelling the two separately. The persistence of biases at some sites over the whole simulation period, in particular those associ-

ated with poor emulator performance, suggests that such improvements could improve robustness of the emulation of the 1998 chlorophyll records.

The potential for under-estimation of emulator uncertainty can be reduced further by improvements to the simulator. Reduction in basic simulation errors should reduce the magnitude of ϵ_1 and ϵ_S, thereby reducing the potential for bias in the ϵ_S statistics, as well as reducing reliance on the small target model ensemble. Simulator improvement should also reduce the impact of simulation error on estimates of ϵ_B from the environmental uncertainty analysis, reducing the risk of significantly violating the the assumption of independence between the estimates of ϵ_S and ϵ_B. A fairly simple way of improving the simulator may itself would be to provide physical forcing based on 3-D model output at higher temporal resolution for the experimental sites, as the impacts of important weather events are attenuated in the 5 day mean output.

Improvements in the representation of concentration dependency in the simulator's lateral flux divergence tendencies are also likely to be beneficial. They are likely to reduce variance in ϵ_B which is in part a consequence of inconsistencies between the tracer concentrations and their applied perturbations. In addition, they may similarly affect ϵ_S, which is expected to be more sensitive to the concentration-dependency formulation than ϵ_1. Certainly, improvements in the representation of concentration dependency should reduce the risk of uncertainty associated with ϵ_S being significantly greater than that associated with ϵ_1 and would tend to decrease any indirect dependency between ϵ_S and ϵ_B.

Concentration dependency in the 1-D simulations is controlled by the transformation applied to the tracer concentrations. A promising approach to improving its representation might be to introduce tracer-specific transformations, possibly varying in space and time, based on statistical analyses of 3-D model output. A key consideration will be the need to reduce the potential for positive feedback cases, where concentration errors reinforce error in the advective tendencies. This type of positive feedback can cause the growth of large positive errors, particularly in the dissolved nutrient tracers. It may also lead to excessive nutrient depletion rates where an initial tendency towards negative bias in nutrient concentrations is increased by reduction in lateral supply. Such errors are likely to have a greater impact on surface chlorophyll at oligotrophic

sites, where the phytoplankton dynamics are more sensitive to nutrient concentration~~and it~~. It is at these sites where the emulator with indirect uncertainty quantification appears least robust. However, an investigation of the surface nutrient records output by the simulator (not presented) did not show evidence of severe nutrient depletion that might be expected from positive feedback.

7.2 Application of the emulation scheme

For calibration of global ocean biogeochemical models against ocean colour data, the spatial extent of the simulator array can readily be extended to produce a mechanistic emulator with truly global coverage based on a larger set of representative sites. Similarly, the emulation procedure could be extended to records of the annual cycle from multiple years. Importantly, we expect the method to be applicable to models of much higher resolution than the 1° target model used in the present demonstration, with minimal adaptation. The requirement for a small ensemble of 3-D reference simulations is relatively modest, making useful parametric analyses feasible for eddy-permitting and eddy-resolving global models.

While the emulation scheme has the potential to make considerable reductions in the number of 3-D simulations required in a parametric analysis, it must be recognized that even a single 3-D simulation may be a large overhead if long spin-up periods are required. The 2 year spin-up period employed for producing the reference ensemble in our experiments is sufficient to demonstrate proof-of-concept. However, biogeochemical models and carbon cycle models in particular require long spin-up times, typically thousands of years, to reach equilibrium. This implies that in many practical applications much longer spin-up times would be needed. Fortunately, recent advances in the estimation of steady state annual cycles for global models (Khatiwala, 2007, 2008) promise to alleviate this problem. The efficient Transport Matrix Method of Khatiwala (2007) has recently been exploited in parametric analyses where simulations are evaluated against global nutrient data (Kriest et al., 2010, 2012) . It has also been combined with a surrogate-based optimization technique for practical parameter estimation (Prieß et al., 2013b) . Incorporating these new

steady state estimation techniques into the target model prior to site-based emulation would be particularly advantageous.

Continued development of the indirect uncertainty quantification method is motivated by its potential in situations where a known truth is unavailable. Such a situation arises if we want to emulate a target model for which we have no model-specific 3-D ensemble but must rely on results for a related model. For example, we might try to emulate a high-resolution model, for which we have perhaps just one simulation, by adapting the method to make use of biogeochemical information from lower resolution ensembles. In this scenario, the statistical environment model could be constructed using the high resolution flow field in combination with upstream gradient and initial state information from the low resolution model. Additional uncertainty in the gradient and state information associated with the change of resolution would be quantified with reference to the equivalent high resolution model fields. The effect of basic simulation errors would, of course, have to be quantified with reference to the single high resolution simulation but this is less likely to be a problem if the basic simulation errors can be made small compared with the parametric environment error.

In applying the emulator with indirect uncertainty quantification to each trial parameter vector, the requirement for a parameter-specific set of 1-D ensemble simulations in the environmental uncertainty analysis imposes a large overhead. The significance of this overhead depends on the experimental set up. For a 1° target model emulated by a global array of simulators at 10° intervals, the computational savings in replacing the 3-D simulation by the emulator array would be fairly limited if an ensemble size of 100 were used as in the present study (being largely those due to the reduced vertical domain and use of pre-calculated physical fields). However, for a 0.25° model with the same array, savings would be considerable. Moreover, it seems likely that the ensemble size could be reduced and investigation of the sensitivity of performance measures to ensemble size would certainly be worthwhile.

In a practical calibration exercise where the uncertainty statistics are required for weighting model-data misfit to account for simulation uncertainty, we should not ignore temporal covariance in simulation error. Although the covariance structure of the error has not been quantified in this study, the results are indicative of strong temporal correlation over long time scales at

some sites. This suggests that it will be important to extend the chosen uncertainty quantification procedure to predict the temporal error covariances for each site-specific simulator. Correlation between sites may also need to be considered, particularly if sites are relatively close together.

Although the emphasis of the present study has been on emulating surface chlorophyll, the method can in principle be used to emulate other observable variables associated with the target model. A full set of model outputs are available from the 1-D simulations at each site and simulation uncertainty measures can similarly be predicted for any of these variables, although the robustness of such predictions is as yet untested. Use of in situ observations in conjunction with the satellite ocean colour data will provide valuable additional constraints on parameter values, making this an important extension to the mechanistic emulator capability.

7.3 The role of a site-based mechanistic emulator

Thorough investigation of the large multi-dimensional parameter spaces associated with mechanistic biogeochemistry models like MEDUSA will inevitably place great demands on our computer resources. For most parametric analyses, it is envisaged that the mechanistic emulator would be used in combination with one or more statistical emulators for which it would provide the training data and associated uncertainty estimates. This would facilitate the use of rigorous Bayesian analysis techniques which would otherwise not be computationally feasible. Introducing mechanistic emulation as an intermediate step should greatly decrease the number of expensive 3-D simulations that are needed.

Modern Bayesian calibration methods, following Kennedy and O'Hagan (2001), provide a comprehensive statistical framework for addressing issues of parametric uncertainty as well as uncertainty from other sources. They allow estimation of joint posterior distributions for model parameters and model discrepancy. Model discrepancy, originally referred to as model inadequacy, quantifies error associated with the model design that cannot be corrected by parameter adjustment. Arhonditsis (2008) and Zhang and Arhonditsis (2009) demonstrate the application of Bayesian calibration methods to marine biogeochemical modelling in a 1-D framework using synthetic data, indicating the value of these methods for quantifying uncertainty associated with

model predictions. A capability for routine application of these methods to biogeochemistry at the global scale would contribute to more robust probabilistic predictions of global change.

A flexible alternative to full Bayesian calibration is the well-established history matching approach adopted by Williamson et al. (2013) in their coupled ocean–atmosphere model analysis. This relatively simple technique uses perturbed parameter ensembles in combination with an implausibility metric to rule out regions of parameter space. The implausibility function takes into account the relevant uncertainties and can be applied iteratively, introducing additional observational data at each stage, to rule out successive regions. The initial focus can be on simple model outputs that are easy to model statistically over the whole parameter space. Subsequent re-focussing of computational effort on smaller regions of parameter space can then be used to develop statistics for more complex outputs.

In this way, history matching can be used as a precursor to Bayesian calibration or, if the region of parameter space not ruled out by the history matching process is sufficiently small, further calibration may be omitted in favour of an averaged parameter vector. The emphasis on defining a “not-ruled-out-yet" region of parameter space, rather than finding the optimal parameter vector, is well-suited to ecosystem modelling where the “underdetermination problem" highlighted by Ward et al. (2010) is ubiquitous.

It is important to recognize that the site-based experimental framework is designed to investigate relatively short time-scale responses of the biogeochemistry to physical drivers. The efficiency of the method makes the corresponding output relatively easy to model statistically and so is well suited to the early stages of history matching. However, we cannot rule out the possibility of interactions with the ocean circulation that would compromise performance of particular parameter vectors in much longer simulations. Further tests would be needed in 3-D simulations to fully determine suitability.

~~The situation is further complicated by the fact that biogeochemical models and carbon cycle models in particular require long spin-up times, typically thousands of years, to reach equilibrium. However recent advances in the estimation of steady state annual cycles for global models (Khatiwala, 2007, 2008) promise to alleviate this problem. The efficient Transport Matrix Method of Khatiwala (2007) has recently been exploited in parametric analyses where~~

~~simulations are evaluated against global nutrient data(Kriest et al., 2010, 2012) . It has also been combined with a surrogate-based optimization technique for practical parameter estimation (Prieß et al., 2013b) . The new steady state estimation methods would be beneficial in site-based emulator construction too, where their use in generating the required 3-D reference ensemble would improve its representativeness of the target model.~~

~~In general, although the need for parametric analyses in relatively expensive 3-D experiments remains, a large site-based ensemble capability should allow us to achieve major reductions in the size of the prior parameter space for such experiments. Exploration of the reduced space then becomes much more tractable. Alternatively, for calibration purposes, a mechanistic site-based emulator might be used as the fast surrogate model in a surrogate-based optimization scheme such as that employed by (Prieß et al., 2013b) . It would then be used in a sequence of optimization loops in conjunction with single evaluations of the 3-D target model at each iteration.~~

In designing a calibration strategy for ocean biogeochemical models, we can take advantage of the relatively weak coupling between the upper ocean and the interior and the different time-scales associated with upper ocean processes and the sinking and remineralization of material in the deep ocean. Site-based methods are best suited to the optimization of parameters associated with seasonal productivity cycles in the upper ocean, occurring on short time scales compared with those for the redistribution of plankton by the large scale circulation. Parameters associated primarily with slow deep water processes that interact more strongly with the circulation can be optimized separately in 3-D experiments, without compromizing the seasonal dynamics.

There are parallels with an established system used in terrestrial carbon cycle modelling. This is the Carbon Cycle Data Assimilation System (Rayner et al., 2005), which uses a two stage process to calibrate a terrestrial biogeochemistry model. The first step involves optimization of parameters controlling phenology and soil moisture by assimilating satellite data related to vegetation activity. The second step then uses fields from the optimized model as input to a simpler model version, combined with a 3-D atmospheric transport model, for constraining the remaining model parameters to fit atmospheric CO_2 data.

7.4 Site-based process model analysis

As a final point, it should be stressed that we have focused here on enabling parametric analyses for a coupled model system, where the optimal parameter values are conditional on a particular representation of the physical ocean. This is important for applications of biogeochemistry models in specific host model configurations. However, there is also a need to be able to evaluate and improve the fidelity of the biogeochemistry model with respect to the processes it is designed to represent, independently of a particular physical simulation. This is emphasized by parameter optimization experiments of Friedrichs et al. (2006) which show that likely error in the physical forcing data can have a large effect on the biogeochemical simulations, leading to inappropriate posterior parameter values.

Site-based methods can be adapted to allow for such error by including a quantification of uncertainty in the physical environment in the analysis as suggested by Hemmings and Challenor (2012). By doing this, we aim to emulate the output that would be obtained from the biogeochemistry model if it were embedded in a perfect physical simulation. History matching could then be used to rule out areas of parameter space that are inconsistent with a plausible representation of the biogeochemical dynamics. Computing effort would be focused primarily on data-rich sites, including established biogeochemical time series observatories.

A statistical model of the biogeochemical environment would be required for the 1-D simulations at each site. The methods introduced here provide the basis for constructing such a model. However, they would need to be refined to allow for additional uncertainty involved in making inferences about a hypothetical perfect physics ensemble from analysis of a practical 3-D ensemble. The development of a robust method is more likely to be achievable if a good observationally-constrained statistical description of the local flow field can be established. Then, only the upstream tracer gradient and initial state information would need to be inferred from the 3-D model analysis. Furthermore, it should be possible to take initial state information from an observation-based statistical model of the real-world state, say from a climatology. Inferences about the model would then be restricted to its behaviour over relatively short time scales. However, this seems likely to be the most practical approach.

In principle, the site-based capability could be adapted for use in a Lagrangian framework allowing a Eulerian simulator array to be augmented by 1-D simulations following Argo floats or surface drifter trajectories. Physical data from Eulerian observatories and Lagrangian platforms, in combination with satellite Earth observation data could be used in conjunction with 3-D simulations to develop observationally-constrained statistical representations of the physical environment to which the biogeochemistry responds. Bringing these different components of the global observation system together in a robust statistical framework for model calibration and assessment will be an important step in developing a reliable predictive capability for the Earth system that accounts for the role of marine biogeochemistry in global change.

8 Summary and conclusions

A mechanistic site-based emulator for annual cycles of surface chlorophyll output from the global NEMO-MEDUSA model was presented. The emulation scheme introduces two fundamental improvements to our site-based biogeochemical modelling capabilities: an explicit representation of the lateral flux divergences of the model tracers, following Hemmings and Challenor (2012), and a quantification of output uncertainty with respect to the target model.

The emulator relies on an array of 1-D simulators of the target model dynamics. In the absence of parameter-specific 3-D model information about the environment at each site, the simulators use a mean environment provided by a small ensemble of target model simulations. This 3-D ensemble is designed to be representative of variability in the model dynamics over the parameter space of interest. It provides information about the local environment in the form of estimates of the required initial state and lateral flux divergences, together with their uncertainties. The use of lateral flux information reduces simulator error considerably, consistent with a major influence of advection at some sites, and this has been instrumental in achieving a promising level of performance.

Two different versions of the mechanistic emulator have been evaluated. One is constructed using a direct uncertainty quantification method, in which output uncertainty is quantified by comparison with a known truth. The other is constructed using an indirect method, in which

output uncertainty is inferred from separate analyses for two contributing factors: the set of basic simulation errors and the parametric environment error. Uncertainty due to basic simulation errors is quantified by applying the direct method to the simulator with a known parameter-specific environment. Parametric environment error is the error in the simulator output when an unknown parameter-specific environment is approximated by the mean environment (an estimate of the expectation of the environment over the parameter space of interest). Uncertainty associated with this error is quantified by 1-D uncertainty analyses.

The analysis for NEMO-MEDUSA indicates that the emulator with direct uncertainty quantification should provide a reasonably robust site-based emulation capability for the surface chlorophyll output from 3-D models. The indirect uncertainty quantification scheme, although more expensive in terms of the number of 1-D simulations required, has the advantage of accounting for the dependency of simulation uncertainty on the trial parameter vector. However, as implemented here, it was found to be less robust. Nevertheless, a number of improvements to the method have been suggested which are expected to improve its reliability. Irrespective of whether this leads to the performance of the indirect method exceeding that of the direct method in terms of robustness, the indirect method provides the basis for a more flexible approach that is less reliant on target model simulations. The potential of both versions of the emulation scheme to improve the effectiveness of site-based approaches to parametric analysis of ocean biogeochemical models is clear.

Our experimental mechanistic emulator serves as a prototype for an improved site-based capability. This facility would allow robust inferences to be made about the parameter-dependent behaviour of global biogeochemical models on the basis of analyses performed on representative arrays of 1-D simulators. It would thus enable the routine execution of relevant parameter perturbation ensembles with 100s of members. In conjunction with statistical emulators, this would enable comprehensive investigations of large parameter spaces to be performed.

In addition, the new developments in the treatment of lateral advection and quantification of environmental uncertainty for 1-D simulators will be important for performing analyses of biogeochemistry models that are based on their representation of the biogeochemical dynamics, rather than being conditional on a particular representation of the physical circulation. This

type of process-based analysis is essential for assessing and improving the fidelity of process representation in biogeochemical models.

Site-based analyses of both coupled and stand-alone biogeochemistry models promise to make important contributions to our ability to constrain model parameters and quantify biogeochemical uncertainty in ocean and Earth system model predictions.

Appendix A: Code availability

MarMOT 1.1 is open source software available under the CeCILL Free Software License Agreement. It is designed for use on UNIX-based systems, including LINUX and Mac OS X. The original code was released on 21st November 2013. The current version, MarMOT 1.1.1, released on 23rd January 2015, is functionally equivalent to the original but includes modifications to address a known portability issue and improve reliability. A tar archive containing the MarMOT 1.1.1 distribution can be downloaded from the National Oceanography Centre's web site at http://noc.ac.uk/project/marmot or supplied by the corresponding author on request. The software release includes a set of command line tools for handling MarMOT-compatible data tables. Full documentation and test data are included with the distribution.

The MEDUSA 1.0 code is available as a supplement to Yool et al. (2011) . A version of this original code with adaptations for interfacing with the MarMOT testbed is included in the MarMOT 1.1.1 distribution.

Appendix B: Defining the parameter space

The first step in parametric analysis of a model, whether for purposes of uncertainty analysis or calibration, is defining the parameter space to be investigated. Our primary interest here is in exploring uncertainty in the seasonal cycle and its impact on annual primary production and the export of material from the euphotic zone. We therefore want to investigate plankton system parameters that have a significant influence on these processes. These are identified by a formal

sensitivity analysis involving 28 relevant model parameters varied over ranges consistent with their defined roles in the model.

B1 Initial parameter selection

The MEDUSA 1.0 model as described by Yool et al. (2011) has over 60 parameters. Our focus is on the seasonal cycle in the euphotic zone with the ultimate aim of using satellite-derived chlorophyll data to constrain upper ocean plankton dynamics in the model. On this basis, a number of parameter groups are excluded from the model analysis. These are the parameters of the inorganic iron and carbonate systems and parameters associated with the remineralization of sinking particles that occurs mainly in the ocean interior. Parameters related to stoichiometry are, in general relatively well known compared with many of the other parameters and are also excluded from the analysis. However, this is largely a pragmatic decision to reduce the size of the parameter space; sensitivity to these parameters within their expected ranges should ideally be explored in future studies. The parameters referred to are the carbon : nitrogen and iron : nitrogen ratios for the organic components and the parameters controlling the variable chlorophyll : carbon ratios for the two phytoplankton types and the diatom silicon : nitrogen ratios.

The remaining set of parameters used in MEDUSA includes parameters that are conceptually related in such a way as to complicate the interpretation of parametric analyses in which they are varied independently. For example, the two phytoplankton types each have their own set of rate parameters, so adjusting a rate parameter for one phytoplankton type affects the relative rates for each type. There are no individual parameters controlling the overall rates associated with phytoplankton as an aggregated biotic group. To avoid problems of this kind, the input parameter set in the MarMOT 1.1 configuration of MEDUSA has been modified from the parameter set used internally.

The 37 input parameters relevant to this study and their relationships to the internal parameters specified in Yool et al. (2011) are shown in Tables 4 to 6. The standard values tabulated are those used in the standard simulation of Yool et al. (2011) or their equivalents. The standard simulation is referred to in the National Oceanography Centre's archive as EXP276 (available

on request from A. Yool; axy@noc.ac.uk). There are inconsistencies between values for 3 of the zooplankton density-dependent loss parameters in Table 6 ($f_{\mu 2,\mathrm{Z}\mu}$, $f_{\mathrm{kZ}\mu}$ and $f_{\mu 2,\mathrm{Zm}}$) and values appearing in Yool et al. (2011) since the latter were incorrect. The correct standard simulation values for the microzooplankton maximum loss rate and half saturation concentration are $\mu_{2,\mathrm{Z}\mu} = 0.1$ and $k_{\mathrm{Z}\mu} = 0.5$ respectively (in units of d^{-1} and $\mathrm{mmol\,N\,m^{-3}}$). These match the corresponding standard simulation values for phytoplankton. The correct value for the mesozooplankton maximum loss rate $\mu_{2,\mathrm{Zm}}$ is $0.2\,\mathrm{d}^{-1}$.

Pairs of rate or half-saturation concentration parameters for the different phytoplankton or zooplankton types have been replaced by a base value, pertaining to the smaller plankton type (non-diatoms or microzooplankton), and a relative value for the larger type (diatoms or mesozooplankton). This leads to new parameters that are non-dimensional factors. For the diatom growth process these are $f_{\alpha\mathrm{Pd}}$, f_{VPd}, $f_{\mathrm{kN,Pd}}$ and $f_{\mathrm{kFe,Pd}}$. For mesozooplankton growth we have f_{gm} and f_{km}. The new parameters for the diatom loss processes are $f_{\mu 1,\mathrm{Pd}}$, $f_{\mu 2,\mathrm{Pd}}$, f_{kPd}. For the zooplankton loss processes, the microzooplankton values $f_{\mu 2,\mathrm{Z}\mu}$ and $f_{\mathrm{kZ}\mu}$ are defined in terms of the non-diatom phytoplankton values and the mesozooplankton values $f_{\mu 2,\mathrm{Zm}}$, f_{kZm} are defined in terms of the microzooplankton values. This suite of modifications allow individual parameters, the base values, to be varied without affecting the relationships between closely associated parameters. The parameter relationships can be controlled independently using the new parameters.

A similar approach is taken for assimilation efficiencies and feeding preference parameters. The carbon assimilation efficiency for zooplankton grazers has been re-expressed in terms of their nitrogen assimilation efficiency by a non-dimensional offset parameter $a_{\beta\mathrm{C}}$. The value is the fraction of the maximum possible offset determined by the constraint that assimilation efficiencies must logically be within the range 0–1. Mesozooplankton feeding preferences have been re-expressed in a hierarchical way so that instead of preference factors for each individual food type, there is an overall preference for live food (as opposed to detritus) p_{mLive} and two conditional preferences: a preference for phytoplankton given live food $p_{\mathrm{c,\,mP}}$ and a preference for non-diatoms given phytoplankton $p_{\mathrm{c,\,mPn}}$.

Yool et al. (2011) used identical values for some parameter pairs and groups to avoid introducing arbitrary complexity. The new definition of the input parameter parameter set described here allows the values of associated internal parameters to be kept the same while varying their values via the base parameter. Adding additional complexity over that of the original model is not justified for the present calibration experiments so the relevant non-dimensional factors are fixed at 1 wherever identical parameter values were used by Yool et al. (2011), thereby further reducing dimensionality of the parameter space. By the same argument, $a_{\beta C}$ is fixed at 0.

The standard value for the fast detritus fraction of mesozooplankton losses $D2_{frac}$ is 1, implying that all mesozooplankton losses are treated as fast-sinking detritus. Adjusting this value would cause the losses to be divided between slow and fast sinking detritus adding a small amount of additional complexity to the model processes. Again, we chose to avoid introducing this new complexity and left this parameter fixed.

As a consequence of excluding less relevant parameter groups from the analysis and choosing to avoid the introduction new complexity, an initial parameter space of 28 dimensions was considered in the present study. The remaining parameters are constrained a priori to take their standard values; this constraint effectively becomes part of the model design. Further dimension reduction was performed objectively on the basis of a sensitivity analysis.

B2 Parameter ranges

Acceptable ranges for each of the parameters to be included in the analysis are defined according to a set of rules as follows.

Rule 1: for all positive parameters with no inherent upper limit, bounds are symmetric about the prior value on a geometric scale. This applies to rate parameters and half-saturation concentrations, whether expressed in absolute or relative units. Rate parameter bounds are set initially at half and double the prior. A factor of 5 is used for half-saturation concentrations.

Rule 2: for fractions, such as efficiencies and feeding preferences, limits are initially set at ± 0.25. Limits of 0.05 and 0.95 are imposed on the lower and upper bounds respectively and the bounds are adjusted if necessary.

Rule 3: the sign of differences between associated internal parameters is preserved. This is done for rates and half-saturation concentrations by imposing 1 as a lower or upper limit for the ranges of the parameters that are expressed as relative values, depending on whether their priors are greater than or less than 1. The relevant bound is adjusted if necessary.

Rule 4: if one or other bound is adjusted in applying Rule 3, then symmetry is used to reset the opposite bound. Geometric symmetry is applied to rates and half-saturation concentrations. This rule applies a constraint on the difference between associated parameters that is dependent on their difference in the prior parameter set.

The resulting parameter space is defined by Table 7. Log-transformed values are used for some parameters when dividing up the parameter space for sampling purposes. The dimensions to which this applies are indicated in the table.

B3 Parameter sensitivity analysis

Following the initial parameter selection, further reduction in the dimensionality of the parameter space to be explored in the calibration process is based on the potential impact of parameters on annual primary production and the ratio of annual particulate export to annual primary production, referred to as the pe-ratio. (The inorganic fraction of particulate carbon export associated with carbonate production is excluded.) The value of the pe-ratio at 207 m is used since this is the greatest depth at which photosynthesis can occur in the model.

Annual mean values for 1998 at 12 sites were determined for 5000 different parameter vectors in the 28 dimensional parameter space. The parameter vectors were distributed in parameter space using a Latin hypercube design (McKay et al., 1979) with a “maximin" criterion (Johnson et al., 1990) applied to 10 randomly generated hypercubes. For generating the design points, distance is defined in terms of positions on a parameter space grid with an equal number of intervals in each dimension. Grid intervals are in log units for rate parameters and half-saturation concentrations. The sensitivity analysis was performed using the 1-D experimental framework described in Sect. 5, with the time step increased to 2 h for efficiency. 1-D simulations were initialized from the standard 3-D simulation of Yool et al. (2011) at the start of 1997, allowing

one complete annual cycle for adjustment to the new parameter values and the 1-D context to reduce the impact of transient behaviour. Lateral fluxes were ignored.

The results of an initial sensitivity analysis for all 28 parameters were examined to identify parameters that have a clear impact on the primary production and the pe-ratio. Parameters that individually explained less than 5 % of the variance in both variables at all sites were then automatically excluded. The sensitivities of the two variables to the remaining parameters are summarized in Table 8 in terms of the number of sites out of 12 at which the parameter explains at least 5 % of the variance and the proportion of variance explained given by the squared Pearson correlation coefficient r^2.

There are 9 parameters that explain more that 5 % of the variance in both model outputs. Of these, k_{C} has a relatively weak effect on both and is excluded. Of the remaining 3 parameters, V_{Pn} is the only one with any stronger influence than k_{C} on either output, having some impact on primary production. However, its effect does not appear to be any greater than the least influential of the other parameters to be retained. Given its lack of influence on pe-ratio, it is discarded along with f_{km} and $\mu_{2,\mathrm{Pn}}$ leaving an 8-dimensional parameter space for the emulation experiments.

The sensitivity analysis was repeated in the 8 dimensional parameter space, again with a sample size of 5000 parameter vectors. Discarding the other 20 parameters reduced the total variance in primary production at each site by between 5 and 38 %. The reduction in the pe-ratio variance was generally less, varying from 6 to 19 %. The parametric uncertainty in primary production and pe-ratio associated with the final 8-dimensional parameter space is illustrated by the coefficient of variation (ratio of standard deviation to mean) for the two variables at each site. The coefficient of variation for primary production ranges from 0.29 (at 15° N) to 0.48 (at 55° N). That for the pe-ratio is generally greater, ranging from 0.38 (at 60° N) to 1.06 (at 30° N).

Appendix C: Quantification of simulator uncertainty

Uncertainty for the log-transformed 5 day mean chlorophyll output is quantified in terms of time series of the predicted monthly means and variances of the uninformed simulator residual.

In the direct uncertainty quantification method, these statistics are derived from differences between the 5 day uninformed simulator output and the corresponding target model output over all parameter vectors in the Construction Phase ensemble. In the indirect method, they are derived from the sums of the mean and variance estimates for the mean environment simulation residual ϵ_{S} and the parametric environment residual ϵ_{B}. The ϵ_{S} statistics are estimated from differences between the 5 day informed simulator output and the target model output over the parameter vectors in the Construction Phase ensemble. The ϵ_{B} statistics are estimated from the 5 day output of a parametric uncertainty analysis using 100 ensemble members.

For each residual, the mean and variance of the 5 day probability distributions are estimated from the relevant ensemble-based sample: u_i, $i \in \{1, \ldots, n\}$. The unbiased population variance estimator

$$s_u^2 = \frac{\sum_{i=1}^{n} (u_i - \overline{u})^2}{n-1} \tag{C1}$$

is used. The 5 day statistics are then used to derive monthly means and variances which are interpolated to give continuous time series $\overline{u}_{\mathrm{m}}(t)$ and $s_{\mathrm{m}}^2(t)$ respectively for uncertainty quantification. The procedure for calculating the time series from the 5 day statistics is as follows.

5 day samples are grouped in pseudo-monthly bins (intervals of 30.42 days) and the monthly mean residual $\overline{u}_{\mathrm{m}}$ is estimated from the k sample means in each bin using the unweighted average, so

$$\overline{u}_{\mathrm{m}} = \frac{1}{k} \sum_{i=1}^{k} \overline{u}_i \tag{C2}$$

where $\overline{u}_{\mathrm{i}}$ is the mean of the ith 5 day sample. $\overline{u}_{\mathrm{m}}$ is then linearly interpolated between monthly mid-points to obtain $\overline{u}_{\mathrm{m}}(t)$. Values for early January 1997 and late December 1998 are equated to those at the respective monthly mid-point. $\overline{u}_{\mathrm{m}}(t)$ is the estimate of the expected residual used for bias correction.

The true residual for the trial parameter vector $\boldsymbol{x}_{\mathrm{o}}$ can be expressed as

$$\psi_{\mathrm{o}}(t, \boldsymbol{x}_{\mathrm{o}}) = \overline{u}_{\mathrm{m}}(t) + \epsilon_{\mu} + \epsilon_{\psi} \tag{C3}$$

where ϵ_μ is the departure of the true residual mean from the estimated residual mean:

$$\epsilon_\mu = \mu(t) - \overline{u}_{\mathrm{m}}(t) \tag{C4}$$

and ϵ_ψ is the departure of the true residual from the true residual mean:

$$\epsilon_\psi = \psi_{\mathrm{o}}(t, \boldsymbol{x}_{\mathrm{o}}) - \mu(t). \tag{C5}$$

For the purposes of uncertainty quantification, these departures are assumed to be independent Gaussian random variables with zero means and variances $s^2_\mu(t)$ and $s^2_\psi(t)$ respectively, derived from the sample data. Variances s^2_μ and s^2_ψ are determined for each pseudo-monthly bin. The monthly variance estimate for the residual is then given by

$$s^2_{\mathrm{m}} = s^2_\mu + s^2_\psi. \tag{C6}$$

This is converted to a continuous time series by interpolation and end-point extrapolation, as for the residual means, to obtain $s^2_{\mathrm{m}}(t)$.

For each bin, s^2_μ is given by the monthly variance of the anomaly between the 5 day sample mean $\overline{u}$ and the expected residual estimate $\overline{u}_{\mathrm{m}}$ at the 5 day interval mid-point. So

$$s^2_\mu = \frac{\sum_{i=1}^{k} (a_i - \overline{a})^2}{k-1} \tag{C7}$$

where

$$a_i = \overline{u}_{\mathrm{i}} - \overline{u}_{\mathrm{m}}(t_{\mathrm{i}}). \tag{C8}$$

s^2_ψ is given by the pooled estimates of the residual variance

$$s^2_\psi = \frac{1}{k} \sum_{i=1}^{k} s^2_{\mathrm{u},i} \tag{C9}$$

where $s^2_{\mathrm{u},i}$ is the variance estimated from the ith 5 day sample.

Determination of monthly means and variances for the residuals from the 5 day samples is expected to give more robust estimates. However, the increase in effective sample size depends on the extent to which samples are temporally correlated over each pseudo-monthly bin. This is not quantified in the present study.

Time series of uninformed simulator residual statistics given by the direct and indirect uncertainty quantification methods are shown in Fig. 10. (Note that for an arbitrary residual ϵ_{X}, $\overline{u}_{\mathrm{m}}$ is denoted $\overline{u}_{\mathrm{X}}$ and s_{m} is denoted s_{X}.) For both methods, the time series determined for all 10 trial parameter experiments are shown. The statistics for the uninformed simulator residual ~~ϵ_2~~ ϵ_1 predicted by the direct method do not account for dependency of the true residual distributions on the trial parameter vectors. Thus, variation in the time series between experiments is due only to sampling uncertainty. The ~~ϵ_2~~ ϵ_1 statistics predicted by the indirect method do account for this parameter dependency and the variation between experiments is then in part due to the parameter-specific dynamics of the environment ensemble simulation used for the parametric uncertainty analysis.

Time series for the statistics of the component residuals contributing to the uninformed simulator statistics given by the indirect method are shown in Fig. 11. The statistics for the mean environment simulation residual ϵ_{S}, like the ~~ϵ_2~~ ϵ_1 statistics given by the direct method, differ between experiments only due to sampling uncertainty. They exhibit less variation between experiments than the ~~ϵ_2~~ ϵ_1 statistics, reflecting the lack of dependency of the true distribution of ϵ_{S} on the trial parameter vector. The statistics for the parametric environment residual ϵ_{B}, the componenent residual that explicitly accounts for the trial parameter vector dependency in the uninformed simulator uncertainty, show much greater variation between experiments.

The Supplement related to this article is available online at doi:10.5194/gmdd-0-1-2015-supplement.

Acknowledgements. The authors would like to thank Andrew Coward and the NEMO development team at the National Oceanography Centre, Southampton for their technical support of the global NEMO-

MEDUSA modelling activities. Thanks are also due to Markus Schartau and one anonymous referee for their helpful and constructive comments on the original manuscript. The financial support of the Natural Environment Research Council (NERC) is gratefully acknowledged. Support for the present study was provided via the National Centre for Earth Observation's Carbon Cycle research theme (grant reference: EARTH010003). NCEO is a NERC research centre.

References

Arhonditsis, G. B., Papantou, D., Zhang, W., Perhar, G., Massos, E., and Shi, M.:. Bayesian calibration of mechanistic aquatic biogeochemical models and benefits for environmental management, J. Marine Syst., 73, 8–30, 2008.

Aumont, O. and Bopp, L.: Globalizing results from ocean in situ iron fertilization studies, Global Biogeochem. Cy., 20, GB2017, doi:10.1029/2005GB002591, 2006.

Campbell, J. W.: The lognormal distribution as a model for bio-optical variability in the sea, J. Geophys. Res., 100, 13237–13254, 1995.

Doron, M., Brasseur, P., Brankart, J.-M., Losa, S. N., and Melet, A.: Stochastic estimation of biogeochemical parameters from Globcolour ocean colour satellite data in a North Atlantic 3-D ocean coupled physical-biogeochemical model, J. Marine Syst., 117–118, 81–95, 2013.

Dowd, M.: Estimating parameters for a stochastic dynamic marine ecological system, Environmetrics, 22, 501–515, doi:10.1002/env.1083, 2011.

Fan, W. and Xianqing, L.: Data assimilation in a simple marine ecosystem model based on spatial biological parameterizations, Ecol. Model., 220, 1997–2008, doi:10.1016/j.ecolmodel.2009.04.050, 2009.

Fasham, M. J. R. and Evans, G. T.: The use of optimization techniques to model marine ecosystem dynamics at the JGOFS station at 47° N 20° W, Philos. T. Roy. Soc. B, 348, 203–209, 1995.

Fasham, M. J. R., Boyd, P. W., and Savidge, G.: Modeling the relative contributions of autotrophs and heterotrophs to carbon flow at a Lagrangian JGOFS station in the Northeast Atlantic: the importance of DOC, Limnol. Oceanogr., 44, 80–94, 1999.

Fasham, M. J. R., Flynn, K. J., Pondaven, P., Anderson, T. R., and Boyd, P. W.: Development of a robust marine ecosystem model to predict the role of iron in biogeochemical cycles: a comparison of results for iron-replete and iron-limited areas, and the SOIREE iron-enrichment experiment, Deep-Sea Res. Pt. I, 53, 333–366, 2006.

Fiechter, J.: Assessing marine ecosystem model properties from ensemble calculations, Ecol. Model., 242, 164–179, doi:10.1016/j.ecolmodel.2012.05.016, 2012.

Fiechter, J., Herbei, R., Leeds, W., Brown, J., Milliff, R., Wikle, C., Moore, A., and Powell, T.: A Bayesian parameter estimation method applied to a marine ecosystem model for the coastal Gulf of Alaska, Ecol. Model., 258, 122–133, 2013.

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 Res. Pt. II, 53, 576–600, 2006.

Friedrichs, M. A. M., Dusenberry, J. A., Anderson, L. A., Armstrong, R. A., Chai, F., Christian, J. R., Doney, S. C., Dunne, J., Fujii, M., Hood, R., McGillicuddy Jr., D. J., Moore, K., Schartau, M., Spitz, Y., and Wiggert, J. D.: Assessment of skill and portability in regional marine biogeochemical models: role of multiple planktonic groups, J. Geophys. Res., 112, C08001, doi:10.1029/2006JC003852, 2007.

Garcia-Gorriz, E., Hoepffner, N., and Ouberdous, M.: Assimilation of SeaWiFS data in a coupled physical-biological model of the Adriatic Sea, J. Marine Syst., 40–41, 233–252, 2003.

Gregg, W., Ginoux, W. P., Schopf, P. S., and Casey, N. W.: Phytoplankton and iron: validation of a global three-dimensional ocean biogeochemical model, Deep-Sea Res. Pt. II, 50, 3143–3169, 2003.

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), Geosci. Model Dev., 5, 471–498, doi:10.5194/gmd-5-471-2012, 2012.

Hemmings, J. C. P., Srokosz, M. A., Challenor, P., and Fasham, M. J. R.: Split-domain calibration of an ecosystem model using satellite ocean colour data, J. Marine Syst., 50, 141–179, 2004.

Hooten, M. B., Leeds, W. B., Fiechter, J., and Wikle, C. K.: Assessing first-order emulator inference for physical parameters in nonlinear mechanistic models, J. Agric. Biol. Envir. S., 16, 475–494, doi:10.1007/s13253-011-0073-7, 2011.

Hourdin, F. and Armengaud, A.: The use of finite-volume methods for atmospheric advection of trace species. Part I: Test of various formulations in a general circulation model, Mon. Weather Rev., 127, 822–837, 1999.

Huret, M., Gohin, F., Delmas, D., Lunven, M., and Garçon, V.: Use of SeaWiFS data for light availability and parameter estimation of a phytoplankton production model of the Bay of Biscay, J. Marine Syst., 65, 509–531, 2007.

Hurtt, G. C. and Armstrong, R. A.: A pelagic ecosystem model calibrated with BATS and OWSI data, Deep-Sea Res. Pt. I, 46, 27–61, 1999.

Johnson, M., Moore, L., and Ylvisaker, D.: Minimax and maxmin distance designs, J. Stat. Plan. Infer., 26, 131–148, 1990.

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, J. Geophys. Res., 116, C06011, doi:10.1029/2009JC006005, 2011.

Kennedy, M. C. and O'Hagan, A.: Bayesian calibration of computer models, J. Roy. Stat. Soc. B, 63, 425–464, 2001.

Khatiwala, S.: A computational framework for simulation of biogeochemical tracers in the ocean, Global Biogeochem. Cy., 21, GB3001, doi:10.1029/2007GB002923, 2007.

Khatiwala, S.: Fast spin up of Ocean biogeochemical models using matrix-free Newton–Krylov, Ocean Model., 23, 121–129, doi:10.1016/j.ocemod.2008.05.002, 2008.

Kidston, M., Matear, R., and Baird, M. E.: Parameter optimisation of a marine ecosystem model at two contrasting stations in the Sub-Antarctic Zone, Deep-Sea Res. Pt. II, 58, 2301–2315, doi:10.1016/j.dsr2.2011.05.018, 2011.

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

Kriest, I., Oschlies, A., and S. Khatiwala, S.: Sensitivity analysis of simple global marine biogeochemical models, Global Biogeochem. Cy., 26, GB2029, doi:10.1029/2011GB004072, 2012.

Lee, L. A., Carslaw, K. S., Pringle, K. J., and Mann, G. W.: Mapping the uncertainty in global CCN using emulation, Atmos. Chem. Phys., 12, 9739–9751, doi:10.5194/acp-12-9739-2012, 2012.

Leeds, W. B., Wikle, C. K., Fiechter, J., Brown, J., and Milliff, R. F.: Modeling 3-D spatio-temporal biogeochemical processes with a forest of 1-D statistical emulators, Environmetrics, 24, 1–12, doi:10.1002/env.2187, 2013.

Le Quéré, C., Harrison, S. P., Prentice, I. C., Buitenhuis, E. T., Aumont, O., Bopp, L., Claustre, H., Cotrim da Cunha, L., Geider, R., Giraud, X., Klaas, C., Kohfeld, K. E., Legendre, L., Manizza, M., Platt, T., Rivkin, R. B., Sathyendranath, S., Uitz, J., Watson, A. J., and Wolf-Gladrow, D.: Ecosystem dynamics based on plankton functional types for global ocean biogeochemistry models, Glob. Change Biol., 11, 2016–2040, doi:10.1111/j.1365-2486.2005.1004.x, 2005.

Lévy, M.: The modulation of biological production by oceanic mesoscale turbulence, Lect. Notes Phys., 744, 219–261, doi:10.1007/978-3-540-75215-8_9, 2008.

Lévy, M., Estublier, A., and Madec, G.: Choice of an advection scheme for biogeochemical models, Geophys. Res. Lett., 28, 3725–3728, 2001.

Losa, S. N., Kivman, G. A., and Ryabchenko, V. A.: Weak constraint parameter estimation for a simple ocean ecosystem model: what can we learn about the model and data?, J. Marine Syst., 45, 1–20, 2004.

Madec, G.: NEMO Reference Manual, Ocean Dynamic Component: NEMO-OPA, Note du Pole de modélisation, l'Institut Pierre-Simon Laplace, Paris, France, No. 27., ISSN 1288–1619, 2008.

Matear, R. J.: Parameter optimization and analysis of ecosystem models using simulated annealing: a case study at Station P, J. Mar. Res., 53, 571–607, 1995.

Mattern, J. P., Fennel, K., and Dowd, M.: Estimating time-dependent parameters for a biological ocean model using an emulator approach, J. Marine Syst., 96–97, 32–47, 2012.

McDonald, C. P, Bennington, V., Urban, N. R., and McKinley, G.: A 1-D test-bed calibration of a 3-D Lake Superior biogeochemical model, Ecol. Model., 225, 115–126, 2012.

McKay, M. D., Conover, W. J., and Beckman, R. J.: A comparison of three methods for selecting values of input variables in the analysis of output from a computer code, Technometrics, 21, 239–245, 1979.

Moore, J. K., Doney, S. C., and Lindsay, K.: Upper ocean ecosystem dynamics and iron cycling in a global three-dimensional model, Global Biogeochem. Cy., 18, GB4028, doi:10.1029/2004GB002220, 2004.

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

O'Hagan, T.: Bayesian analysis of computer code outputs: a tutorial, Reliab. Eng. Syst. Safe., 91, 1290–1300, 2006.

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

Prieß, M., Koziel, S., and Slawig, T.: Marine ecosystem model calibration with real data using enhanced surrogate-base optimization, J. Comput. Sci., 4, 423–437, doi:10.1016/j.jocs.2013.04.001, 2013a.

Prieß, M., Piwonski, J., Koziel, S., Oschlies, A., and Slawig, T.: Accelerated parameter identification in a 3-D marine biogeochemical model using surrogate-based optimization, Ocean Model., 68, 22–36, doi:10.1016/j.ocemod.2013.04.003, 2013b.

Rayner, P. J., Scholze, M., Knorr, W., Kaminski, T., Giering, R., and Widmann, H.: Two decades of terrestrial carbon fluxes from a carbon cycle data assimilation system (CCDAS), Global Biogeochem. Cy., 19, GB2026, doi:10.1029/2004GB002254, 2005.

Sarmiento, J. L., Slater, R. D., Fasham, M. J. R., Ducklow, H. W, Toggweiler, J. R., and Evans, G. T.: A seasonal three-dimensional ecosystem model of nitrogen cycling in the North Atlantic Euphotic Zone, Global Biogeochem. Cy., 7, 417–450, doi:10.1029/93GB00375, 1993.

Schartau, M. and Oschlies, A.: Simultaneous data-based optimization of a 1-D-ecosystem model at three locations in the North Atlantic: Part I – Method and parameter estimates, J. Mar. Res., 61, 765–793, 2003.

Séférian, R., Bopp, L., Gehlen, M. Orr, J. C., Ethé, C., Cadule, P., Aumont, O., Salas y Mélia, D., Voldoire, A., and Madec, G.: Skill assessment of three earth system models with common marine biogeochemistry, Clim. Dynam., 40, 2549–2573, doi:10.1007/s00382-012-1362-8, 2013.

Six, K. D. and Maier-Reimer, E.: Effects of plankton dynamics on seasonal carbon fluxes in an ocean general circulation model, Global Biogeochem. Cy., 10, 559–583, 1996.

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

Tjiputra, J. F., Polzin, D., and Winguth, A. M. E.: Assimilation of seasonal chlorophyll and nutrient data into an adjoint three-dimensional ocean carbon cycle model: sensitivity analysis and ecosystem parameter optimization, Global Biogeochem. Cy., 21, GB1001, doi:10.1029/2006GB002745, 2007.

Van Leer, B.: Towards the ultimate conservative difference scheme IV: a new approach to numerical convection, J. Comput. Phys., 23, 276–299, 1977.

Ward, B. A., Friedrichs, M. A. M, Anderson, T. R., and Oschlies, A.: Parameter optimisation techniques and the problem of underdetermination in marine biogeochemical models, J. Marine Syst., 81, 34–43, 2010.

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, Prog. Oceanogr., 116, 49–65, 2013.

Williamson, D., Goldstein, M., Allison, L., Blaker, A., Challenor, P., Jackson, L., and Yamazaki, K.: History matching for exploring and reducing climate model parameter space using observations and a large perturbed physics ensemble, Clim. Dynam., 41, 1703–1729, doi:10.1007/s00382-013-1896-4, 2013.

Xiao, Y. and Friedrichs, M.: The assimilation of satellite-derived data into a one-dimensional lower trophic level marine ecosystem model, J. Geophys. Res.-Oceans, 119, 2691–2712, doi:10.1002/2013JC009433, 2014.

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

Zhang, W. and Arhonditsis, G. B.: A Bayesian hierarchical framework for calibrating aquatic biogeochemical models, Ecol. Model., 220, 2142–2161, doi:10.1016/j.ecolmodel.2009.05.023, 2009.

Table 1. 8-dimensional MEDUSA parameter space for target model emulation.

Parameter	Description and units	Lower bound	Upper bound
α_{Pn}	chlorophyll-specific initial slope of P-I curve for non-diatoms $\mathrm{g\,C\,(g\,Chl)^{-1}\,(W\,m^{-2})^{-1}\,d^{-1}}$	7.5	30
$k_{\mathrm{N,Pn}}$	N nutrient uptake half-saturation concentration for non-diatoms $\mathrm{mmol\,N\,m^{-3}}$	0.1	2.5
$k_{\mathrm{Fe,Pn}}$	Fe nutrient uptake half-saturation concentration for non-diatoms $\mathrm{mmol\,Fe\,m^{-3}}$	0.000066	0.0017
k_{μ}	microzooplankton grazing half-saturation concentration $\mathrm{mmol\,N\,m^{-3}}$	0.16	4
ϕ	zooplankton grazing inefficiency –	0.05	0.45
$\mu_{\mathrm{1,Pn}}$	non-diatom phytoplankton density-independent loss rate $\mathrm{d^{-1}}$	0.01	0.04
k_{Pn}	non-diatom phytoplankton half-saturation concentration for density-dependent loss $\mathrm{mmol\,N\,m^{-3}}$	0.1	2.5
w_{g}	detrital sinking rate $\mathrm{m\,d^{-1}}$	1.5	6

Table 2. Representative sample from 8-dimensional MEDUSA parameter space.

Parameter set	α_{Pn}	$k_{\mathrm{N,Pn}}$	$k_{\mathrm{Fe,Pn}}$	k_{μ}	ϕ	$\mu_{1,\mathrm{Pn}}$	k_{Pn}	w_{g}
1	12.2	1.54	0.00104	0.19	0.27	0.0325	0.31	1.61
2	10.6	1.12	0.00021	0.94	0.39	0.0283	0.22	4.87
3	18.5	2.13	0.00011	0.36	0.23	0.0123	1.54	5.60
4	8.0	0.31	0.00145	0.26	0.15	0.0246	0.81	3.22
5	14.0	0.81	0.00055	0.68	0.35	0.0107	0.43	2.12
6	28.0	0.12	0.00008	1.79	0.11	0.0214	0.12	2.44
7	9.2	0.43	0.00015	3.41	0.19	0.0187	0.16	1.85
8	21.2	0.22	0.00076	1.30	0.07	0.0141	0.59	3.69
9	24.4	0.16	0.00039	0.49	0.43	0.0162	1.12	4.24
10	16.1	0.59	0.00028	2.47	0.31	0.0373	2.13	2.80

Table 3. Uninformed emulator robustness evaluation for all 10 experiments and for the 9 experiments excluding Experiment 6.

	Direct UQ Method		Indirect UQ Method	
Site	D_{Ud} mean	D_{Ud} std. dev.	D_{Ui} mean	D_{Ui} std. dev.
60° N	0.03 (0.08)	1.17 (1.16)	0.15 (0.20)	0.98 (0.89)
55° N	0.01 (−0.03)	1.17 (1.07)	−0.02 (−0.07)	1.10 (0.97)
50° N	0.48 (0.05)	1.88 (0.98)	0.58 (−0.15)	3.06 (1.43)
45° N	0.16 (0.03)	1.19 (0.99)	0.08 (−0.07)	1.48 (1.32)
40° N	−0.19 (−0.07)	1.41 (1.29)	−0.12 (−0.10)	0.98 (0.99)
35° N	0.16 (0.04)	1.06 (1.02)	−0.47 (−0.72)	1.89 (1.77)
30° N	0.07 (−0.04)	1.15 (1.15)	0.36 (0.03)	1.88 (1.53)
25° N	−0.07 (−0.03)	1.08 (1.12)	−0.76 (−0.70)	1.63 (1.69)
20° N	−0.04 (−0.07)	1.33 (1.31)	−0.64 (−0.66)	1.04 (1.01)
15° N	0.47 (−0.01)	2.07 (1.21)	0.23 (−0.38)	2.60 (1.29)
10° N	0.36 (0.01)	1.55 (1.03)	0.19 (−0.10)	1.21 (0.74)
5° N	0.03 (−0.07)	1.23 (1.22)	−0.09 (−0.30)	2.10 (2.01)
ALL	0.12 (−0.01)	1.41 (1.13)	−0.04 (−0.25)	1.82 (1.39)

~~Informed emulator robustness for all 10 experiments and for the 9 experiments excluding Experiment 6. Site D_{I} mean D_{I} std. dev. $|D_{\mathrm{I}}| \leq 2$~~

~~60N 0.04 (0.04) 1.18 (1.09) 94(96)55N −0.02 (−0.02) 1.18 (1.07) 94(95)50N −0.12 (0.02) 1.40 (1.03) 93(97)45N −0.14 (−0.00) 1.18 (0.94) 97(98)40N −0.58 (−0.12) 2.41 (1.34) 87(93)35N −0.22 (−0.02) 1.42 (1.06) 94(97)30N −0.34 (−0.06) 1.51 (1.20) 89(94)25N −0.03 (−0.00) 1.12 (1.17) 96(95)20N 0.13 (0.09) 1.43 (1.43) 93(93)15N −0.81 (−0.06) 2.93 (1.22) 86(95)10N −1.26 (−0.06) 4.06 (1.14) 86(95)5N 0.03 (−0.05) 1.20 (1.24) 93(92)~~

~~ALL −0.28 (−0.02) 2.00 (1.17) 92(95)~~

Table 4. MEDUSA phytoplankton parameters (MarMOT 1.1 configuration).

Symbol	Description and units	Standard value
α_{Pn}	chlorophyll-specific initial slope of P-I curve for non-diatoms $\mathrm{g\,C\,(g\,chl)^{-1}\,(W\,m^{-2})^{-1}\,d^{-1}}$	15
$f_{\alpha\mathrm{Pd}} = \frac{\alpha_{\mathrm{Pd}}}{\alpha_{\mathrm{Pn}}}$	chlorophyll-specific initial slope of P-I curve for diatoms relative to that for non-diatoms –	0.75
V_{Pn}	maximum non-diatom growth rate at 0 °C $\mathrm{d^{-1}}$	0.53
$f_{\mathrm{VPd}} = \frac{V_{\mathrm{Pd}}}{V_{\mathrm{Pn}}}$	maximum growth rate at 0 °C of diatoms relative to that of non-diatoms –	0.9434
$k_{\mathrm{N,Pn}}$	N nutrient uptake half-saturation concentration for non-diatoms $\mathrm{mmol\,N\,m^{-3}}$	0.5
$f_{\mathrm{kN,Pd}} = \frac{k_{\mathrm{N,Pd}}}{k_{\mathrm{N,Pn}}}$	N nutrient uptake half-saturation concentration for diatoms relative to that for non-diatoms –	0.5
k_{Si}	Si nutrient uptake half-saturation concentration for diatoms $\mathrm{mmol\,Si\,m^{-3}}$	0.75
$k_{\mathrm{Fe,Pn}}$	Fe nutrient uptake half-saturation concentration for non-diatoms $\mathrm{mmol\,Fe\,m^{-3}}$	0.00033
$f_{\mathrm{kFe,Pd}} = \frac{k_{\mathrm{Fe,Pd}}}{k_{\mathrm{Fe,Pn}}}$	Fe nutrient uptake half-saturation concentration for diatoms relative to that for non-diatoms –	2.03

Table 5. MEDUSA zooplankton parameters (MarMOT 1.1 configuration).

Symbol	Description and units	Standard value
g_μ	maximum microzooplankton grazing rate d^{-1}	2
$f_{\mathrm{gm}} = \frac{g_{\mathrm{m}}}{g_\mu}$	maximum grazing rate of mesozooplankton relative to that of microzooplankton –	0.25
k_μ	microzooplankton grazing half-saturation concentration $\mathrm{mmol\,N\,m^{-3}}$	0.8
$f_{\mathrm{km}} = \frac{k_{\mathrm{m}}}{k_\mu}$	grazing half-saturation concentration for mesozooplankton relative to that of microzooplankton –	0.375
ϕ	zooplankton grazing inefficiency –	0.2
β_{N}	zooplankton N assimilation efficiency –	0.69
$a_{\beta\mathrm{C}} = \frac{\beta_{\mathrm{C}}-\beta_{\mathrm{N}}}{\beta_{\mathrm{N}}}, \beta_{\mathrm{C}} \leq \beta_{\mathrm{N}}$ $a_{\beta\mathrm{C}} = \frac{\beta_{\mathrm{C}}-\beta_{\mathrm{N}}}{1-\beta_{\mathrm{N}}}, \beta_{\mathrm{C}} > \beta_{\mathrm{N}}$	offset of zooplankton C assimilation efficiency from that of N as a fraction of maximum offset possible –	0
k_{C}	zooplankton net C growth efficiency –	0.8
$p_{\mu\mathrm{Pn}}$	microzooplankton grazing preference for live food (non-diatom phytoplankton) –	0.75
$p_{\mathrm{mLive}} = p_{\mathrm{mPn}} + p_{\mathrm{mPd}} + p_{\mathrm{mZ}\mu}$	mesozooplankton grazing preference for live food (phytoplankton or microzooplankton) –	0.85
$p_{\mathrm{c,\,mP}} = \frac{p_{\mathrm{mPn}}+p_{\mathrm{mPd}}}{p_{\mathrm{mPn}}+p_{\mathrm{mPd}}+p_{\mathrm{mZ}\mu}}$	mesozooplankton conditional grazing preference for phytoplankton, given live food –	0.5882
$p_{\mathrm{c,\,mPn}} = \frac{p_{\mathrm{mPn}}}{p_{\mathrm{mPn}}+p_{\mathrm{mPd}}}$	mesozooplankton conditional grazing preference for non-diatoms, given phytoplankton –	0.3

Table 6. MEDUSA plankton loss-related parameters (MarMOT 1.1 configuration).

Symbol	Description and units	Standard value
$\mu_{1,\mathrm{Pn}}$	non-diatom phytoplankton density-independent loss rate d^{-1}	0.02
$f_{\mu 1,\mathrm{Pd}} = \frac{\mu_{1,\mathrm{Pd}}}{\mu_{1,\mathrm{Pn}}}$	density-independent loss rate of diatoms relative to that of non-diatom phytoplankton –	1
$f_{\mu 1,Z\mu} = \frac{\mu_{1,Z\mu}}{\mu_{1,\mathrm{Pn}}}$	density-independent loss rate of microzooplankton relative to that of non-diatom phytoplankton –	1
$f_{\mu 1,\mathrm{Zm}} = \frac{\mu_{1,\mathrm{Zm}}}{\mu_{1,Z\mu}}$	density-independent loss rate of mesozooplankton relative to that of microzooplankton –	1
$\mu_{2,\mathrm{Pn}}$	non-diatom phytoplankton maximum density-dependent loss rate d^{-1}	0.1
k_{Pn}	non-diatom phytoplankton half-saturation concentration for density-dependent loss $\mathrm{mmol\,N\,m^{-3}}$	0.5
$f_{\mu 2,\mathrm{Pd}} = \frac{\mu_{2,\mathrm{Pd}}}{\mu_{2,\mathrm{Pn}}}$	maximum density-dependent loss rate of diatoms relative to that of non-diatom phytoplankton –	1
$f_{\mathrm{kPd}} = \frac{k_{\mathrm{k_{Pd}}}}{k_{\mathrm{Pn}}}$	density-dependent loss half-saturation concentration of diatoms relative to that of non-diatom phytoplankton –	1
$f_{\mu 2,Z\mu} = \frac{\mu_{2,Z\mu}}{\mu_{2,\mathrm{Pn}}}$	maximum density-dependent loss rate of microzooplankton relative to that of non-diatom phytoplankton –	1
$f_{\mathrm{k}Z\mu} = \frac{k_{Z\mu}}{k_{\mathrm{Pn}}}$	density-dependent loss half-saturation concentration of microzooplankton relative to that of non-diatom phyto-plankton –	1
$f_{\mu 2,\mathrm{Zm}} = \frac{\mu_{2,\mathrm{Zm}}}{\mu_{2,Z\mu}}$	maximum density-dependent loss rate of mesozooplankton relative to that of microzooplankton –	2
$f_{\mathrm{kZm}} = \frac{k_{\mathrm{Zm}}}{k_{Z\mu}}$	density-dependent loss half-saturation concentration of mesozooplankton relative to that of microzooplankton –	1.5
$\mathrm{D1}_{\mathrm{frac}}$	fast detritus fraction of diatom losses –	0.75
$\mathrm{D2}_{\mathrm{frac}}$	fast detritus fraction of mesozooplankton losses –	1
Diss	diatom frustule dissolution rate d^{-1}	0.006
w_{g}	detrital sinking rate $\mathrm{m\,d^{-1}}$	3

Table 7. MEDUSA parameter space for 28-dimensional sensitivity analysis.

Parameter	Standard value	Lower bound	Upper bound	Transformation
α_{Pn}	15	7.5	30	log
$f_{\alpha\mathrm{Pd}}$	0.75	0.56	1	log
V_{Pn}	0.53	0.27	1.1	log
f_{VPd}	0.9434	0.89	1	log
$k_{\mathrm{N,Pn}}$	0.5	0.1	2.5	log
$f_{\mathrm{kN,Pd}}$	1.5	1	2.3	log
k_{Si}	0.75	0.15	3.8	log
$k_{\mathrm{Fe,Pn}}$	0.00033	0.000066	0.0017	log
$f_{\mathrm{kFe,Pd}}$	2.03	1	4.1	log
g_{μ}	2	1	4	log
f_{gm}	0.25	0.13	0.5	log
k_{μ}	0.8	0.16	4	log
f_{km}	0.375	0.14	1	log
ϕ	0.2	0.05	0.45	
β_{N}	0.69	0.44	0.94	
k_{C}	0.8	0.55	0.95	
$p_{\mu\mathrm{Pn}}$	0.75	0.5	0.95	
p_{mLive}	0.85	0.6	0.95	
$p_{\mathrm{c,\,mP}}$	0.5882	0.34	0.84	
$p_{\mathrm{c,\,mPn}}$	0.3	0.05	0.55	
$\mu_{1,\mathrm{Pn}}$	0.02	0.01	0.04	log
$\mu_{2,\mathrm{Pn}}$	0.1	0.05	0.2	log
k_{Pn}	0.5	0.1	2.5	log
$f_{\mu 2,\mathrm{Zm}}$	2	1	4	log
f_{kZm}	1.5	1	2.3	log
$\mathrm{D1}_{\mathrm{frac}}$	0.75	0.5	0.95	
Diss	0.006	0.003	0.012	log
w_{g}	3	1.5	6	log

Table 8. Parameter sensitivity of annual mean model output from 28 dimensional analysis, showing parameters that explain 5 % or more of the variance in either variable at 1 or more sites.

Parameter	Primary Production		Particulate Export Ratio at 207 m		Selected ?
	No. of sites with $r^2 \geq 0.05$	Maximum r^2	No. of sites with $r^2 \geq 0.05$	Maximum r^2	
α_{Pn}	8	0.44	4	0.08	yes
V_{Pn}	4	0.15	0	< 0.05	no
$k_{\mathrm{N,Pn}}$	5	0.22	2	0.10	yes
$k_{\mathrm{Fe,Pn}}$	9	0.34	3	0.08	yes
k_{μ}	5	0.17	2	0.13	yes
f_{km}	0	< 0.05	1	0.05	no
ϕ	4	0.15	7	0.17	yes
k_{C}	3	0.07	2	0.10	no
$\mu_{\mathrm{1,Pn}}$	6	0.11	11	0.10	yes
$\mu_{\mathrm{2,Pn}}$	3	0.06	0	< 0.05	no
k_{Pn}	4	0.22	7	0.15	yes
w_{g}	6	0.17	11	0.38	yes

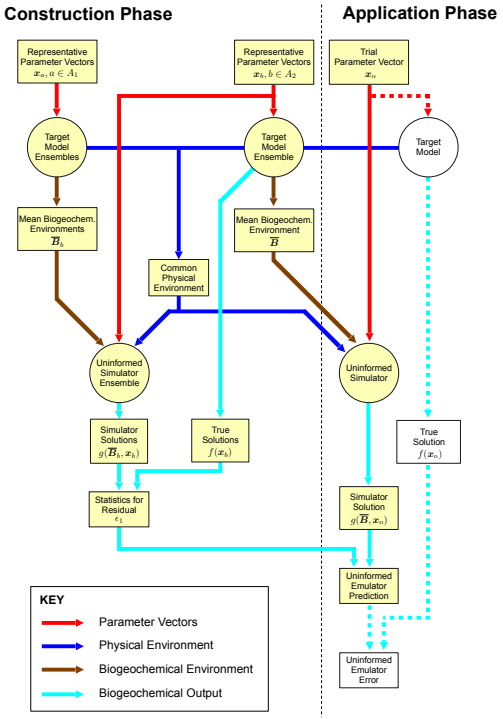

Figure 1. Data flow for emulator construction and application to the prediction of target model output where simulator uncertainty is quantified by the direct method. A_1 and A_2 are arbitrary sets of indices satisfying $A_1 \cap A_2 = \emptyset$. Simulation steps are indicated by circles. The dotted lines and uncoloured boxes indicate data flow for validating emulator performance against a known truth. They are not part of the practical application procedure, where the truth would be unknown.

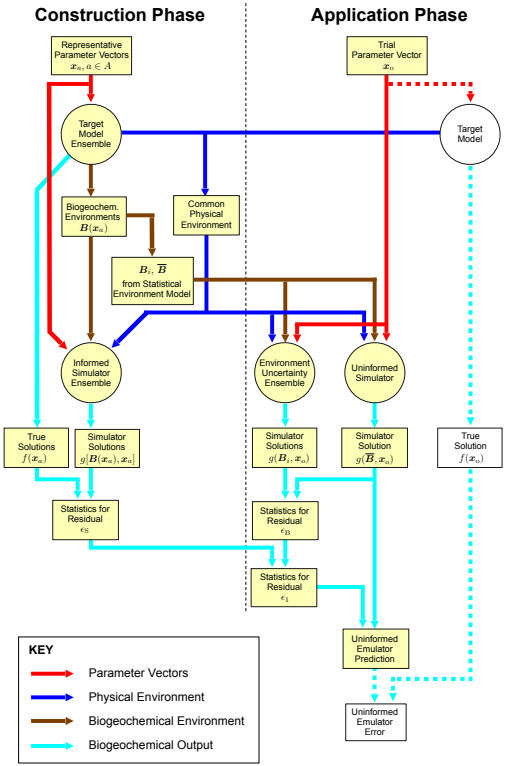

Figure 2. Data flow for emulator construction and application to the prediction of target model output where simulator uncertainty is quantified by the indirect method. A is an arbitrary set of indices. Simulation steps are indicated by circles. The dotted lines and uncoloured boxes indicate data flow for validating emulator performance against a known truth. They are not part of the practical application procedure, where the truth would be unknown.

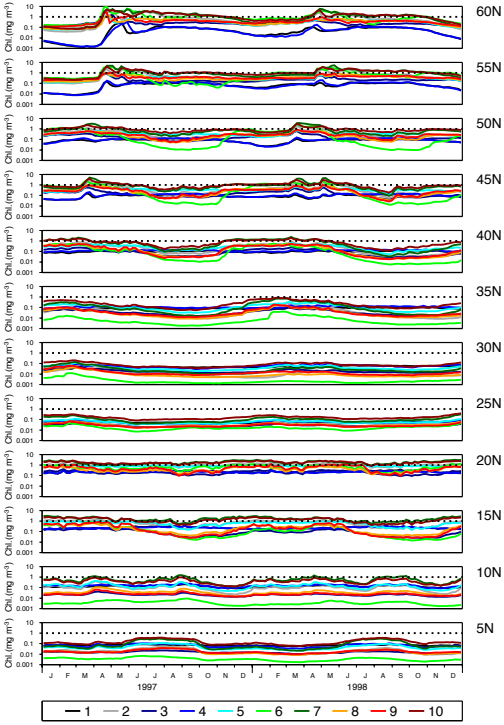

Figure 3. 5 day mean surface chlorophyll output from 3-D NEMO-MEDUSA simulations for the 10 parameter vectors in Table 2, colour coded by Parameter Set number.

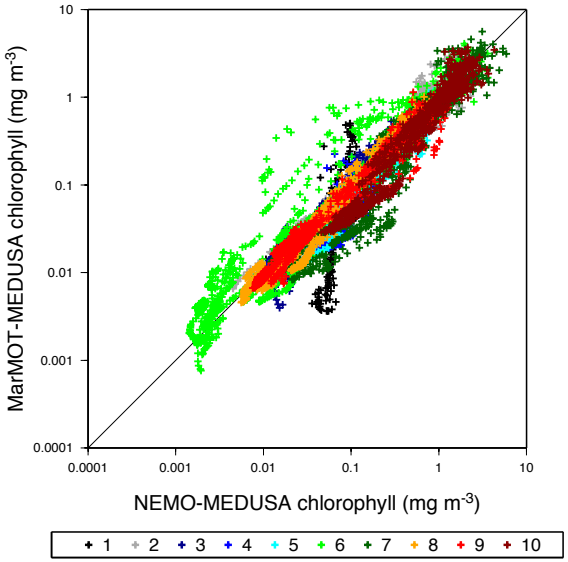

Figure 4. ~~5 day mean surface chlorophyll output for 1998 at all 12 sites from the informed simulator **(a)** without lateral flux perturbations and **(b)** with lateral flux perturbations, compared with that from the matching 3-D NEMO-MEDUSA reference simulation, Results are shown for the 10 different parameter vectors in Table 2, colour coded by Parameter Set number. Informed simulator error statistics for $\log_{10}$(surface chlorophyll) () over 10 experiments, one experiment for each of the parameter vectors in Table 2. **(a)** bias and **(b)** r.m.s. error. The statistics are shown for informed simulators with and without lateral flux perturbations.~~ 5 day mean surface chlorophyll output for 1998 at all 12 sites from the uninformed simulator~~**(a)** without lateral flux perturbations and **(b)** with lateral flux perturbations~~, compared with that from the matching 3-D NEMO-MEDUSA reference simulation. Results are shown for the 10 different parameter vectors in Table 2, colour coded by Parameter Set number.

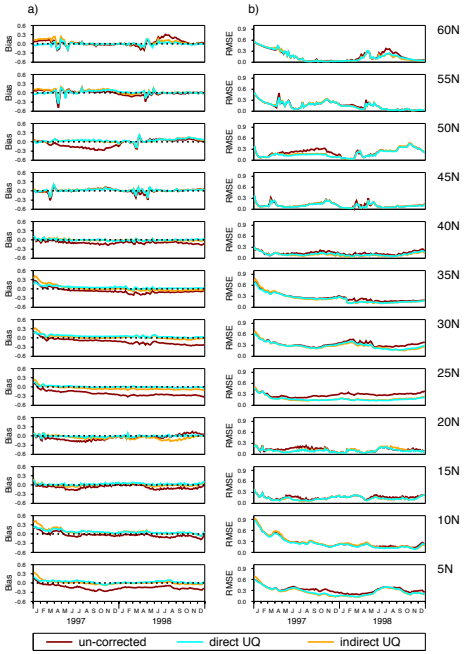

Figure 5. Uninformed emulator error statistics for $\log_{10}$(surface chlorophyll) ($\mathrm{mg\,m^{-3}}$) over 10 experiments, one experiment for each of the parameter vectors in Table 2: **(a)** bias and **(b)** r.m.s. error. The statistics are shown for the simulators without bias correction and for the bias-corrected simulator array, which is the uninformed emulator. The emulator statistics are given for emulator versions constructed using direct and indirect uncertainty quantification methods (i.e. for errors d_{Ud} and d_{Ui}).

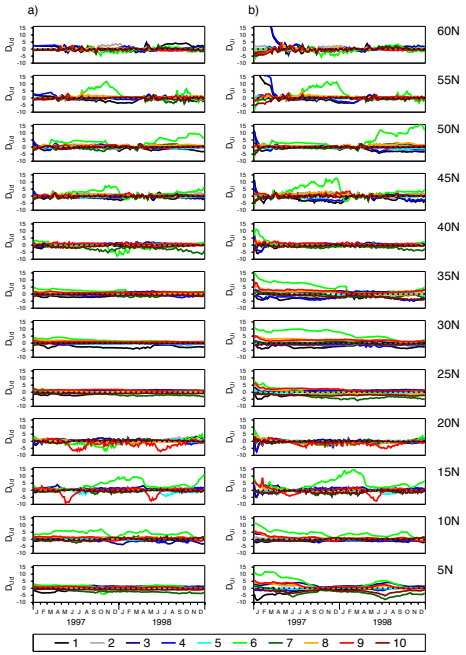

Figure 6. Normalized uninformed emulator error for emulator versions constructed using **(a)** the direct uncertainty quantification method (D_{Ud}) and **(b)** the indirect uncertainty quantification method (D_{Ui}). Errors are shown for the 10 different parameter vectors in Table 2, colour coded by Parameter Set number. Off scale D_{Ui} values not shown at the beginning of 1997 go up to about 26 at 55° N and about 35 at 60° N.

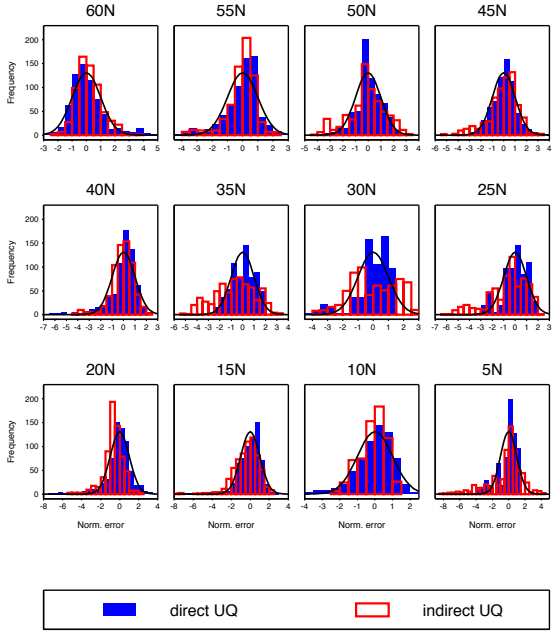

Figure 7. 1998 distributions of the normalized error for the uninformed emulator constructed using the direct and indirect uncertainty quantification methods: D_{Ud} and D_{Ui}. Results for 9 of the 10 parameter vector experiments are combined. Experiment 6, for which large extremes occur, is excluded. The predicted normalized error distribution, over-plotted for reference, is Gaussian with zero mean and unit standard deviation at all times and locations.

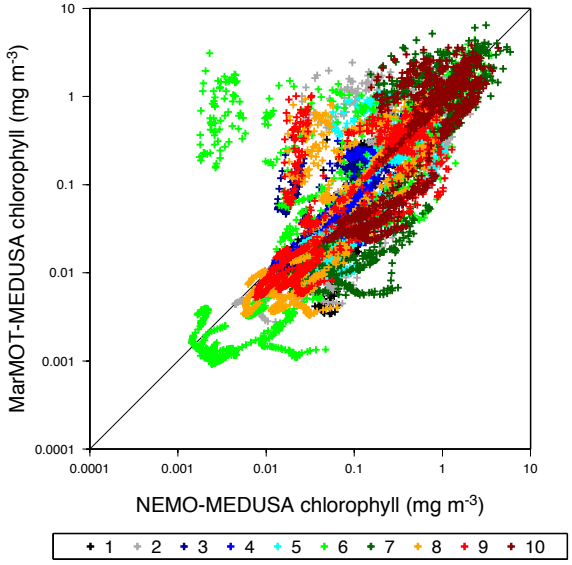

Figure 8. 5 day mean surface chlorophyll output for 1998 at all 12 sites from the uninformed simulator with lateral flux perturbations set to zero, compared with that from the matching 3-D NEMO-MEDUSA reference simulation. Results are shown for the 10 different parameter vectors in Table 2, colour coded by Parameter Set number.

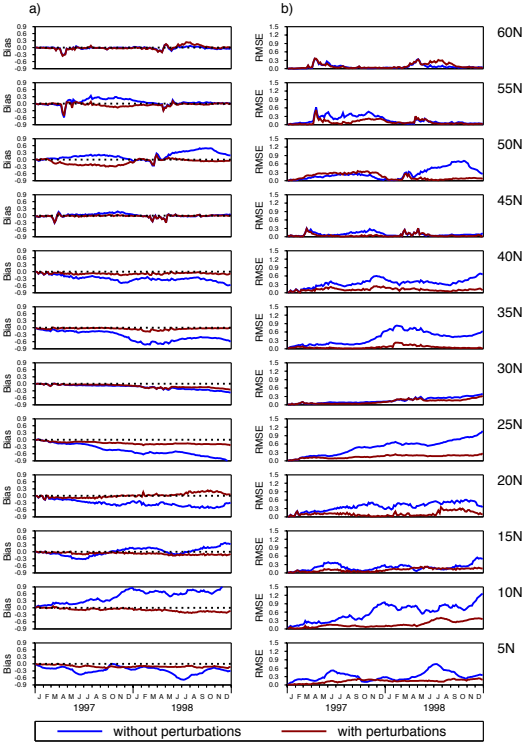

Figure 9. Informed simulator error statistics for $\log_{10}$(surface chlorophyll) ($\mathrm{mg\,m^{-3}}$) over 10 experiments, one experiment for each of the parameter vectors in Table 2. **(a)** bias and **(b)** r.m.s. error. The statistics are shown for informed simulators with and without lateral flux perturbations.

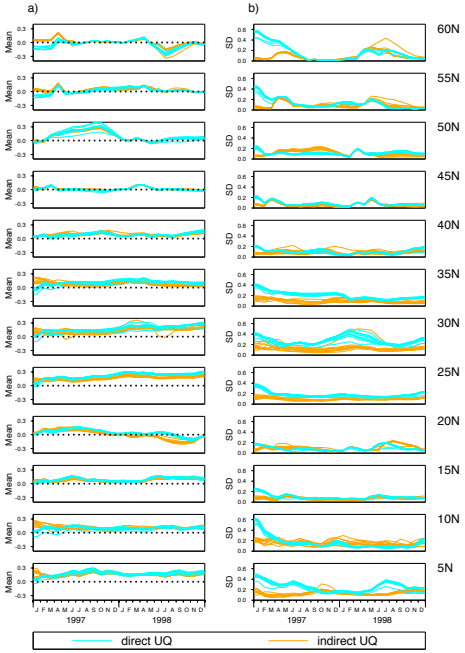

Figure 10. Statistics for the uninformed simulator residual ~~ϵ_2~~ $\underset{\sim}{\epsilon_1}$, predicted by the direct and indirect uncertainty quantification methods for all 10 experiments: **(a)** residual means ~~$\overline{u}_2$~~ $\underset{\sim}{\overline{u}_1}$ and $\overline{u}_\mathrm{S} + \overline{u}_\mathrm{B}$; **(b)** residual standard deviations ~~s_2~~ $\underset{\sim}{s_1}$ and $\sqrt{s_\mathrm{S}^2 + s_\mathrm{B}^2}$. Values are in $\log_{10}$(chlorophyll) units with chlorophyll in $\mathrm{mg\,m^{-3}}$.

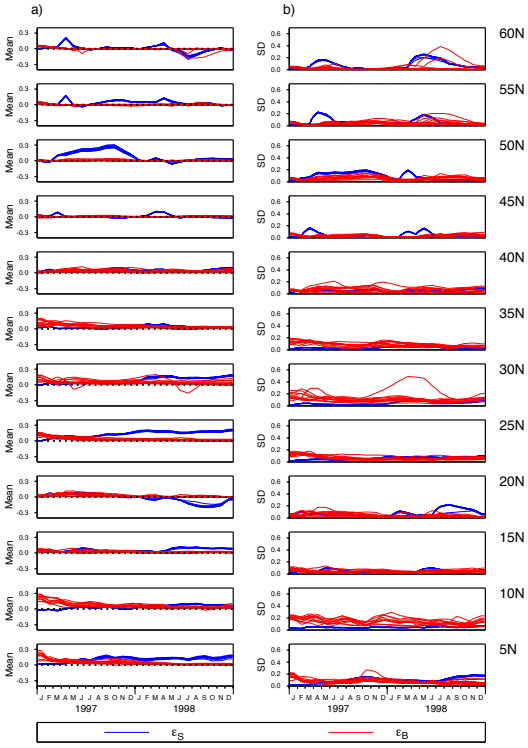

Figure 11. Predicted statistics for the mean environment simulation residual ϵ_{S} and the parametric environment residual ϵ_{B} for all 10 experiments: **(a)** residual means $\overline{u}_{\mathrm{S}}$ and $\overline{u}_{\mathrm{B}}$; **(b)** residual standard deviations s_{S} and s_{B}. Values are in $\log_{10}$(chlorophyll) units with chlorophyll in $\mathrm{mg\,m^{-3}}$.