National
Oceanography Centre
NATURAL ENVIRONMENT RESEARCH COUNCIL

National Oceanography Centre

Research & Consultancy Report No. 41

Traceability of performance between two ocean biogeochemistry models of differing complexity

J C P Hemmings

2013

National Oceanography Centre, Southampton
University of Southampton Waterfront Campus
European Way
Southampton
Hants SO14 3ZH UK

[a]Author contact details:
Tel: +44 (0)23 8059 7793
Email: j.hemmings@noc.ac.uk

DOCUMENT DATA SHEET

AUTHOR HEMMINGS, J C P	*PUBLICATION DATE* 2013

TITLE

Traceability of performance between two ocean biogeochemistry models of differing complexity.

REFERENCE

Southampton, UK: National Oceanography Centre, 57pp.
(National Oceanography Centre Research and Consultancy Report, No. 41)

ABSTRACT

Executive Summary

Purpose and scope

A diverse range of candidate ocean biogeochemistry models exists for addressing scientific questions of societal importance in an Earth system context. Limitations imposed by computer resources favour the use of simpler models. However, there are recognized benefits of supporting different levels of complexity, not least because the appropriate level of complexity for a given application is an open research question. An important aim is to ensure that when simplifications are made there is a traceable link between models so that the implications are understood.

A pilot study in traceability of model performance is presented in which the ability of a simple surrogate model, based on HadOCC, to emulate the behaviour of the intermediate complexity MEDUSA model is investigated. Adjustable HadOCC parameter values are optimized to fit MEDUSA output for an array of sites representing a range of oceanic conditions.

Method

Parameter optimization experiments were performed in an experimental framework, comprising an array of 12 North Atlantic sites, using the Marine Model Optimization Testbed (MarMOT), a NERC software tool for computationally intensive analyses of biogeochemistry models. Synthetic data for calibration were taken from MEDUSA output for a single year. Predictive skill of the calibrated surrogate model was determined with reference to independent annual cycles at the calibration sites and on a meridional transect along 20W.

Results

The calibration process substantially improved the surrogate model's performance as an emulator of MEDUSA. There was a high degree of parameter redundancy: the number of adjustable parameters in the surrogate model was reduced from 17 to 10 with minimal effect on the emulation performance. Despite weak prior parameter constraints, posterior parameter values were broadly consistent with a mechanistic interpretation of the simpler model, although some deviated strongly from accepted values.

The calibrated model performs well in terms of annual cycles of primary production as well as their annual means and inter-annual variability. Performance for sinking particle flux is generally good although seasonal cycles were less well replicated than for primary production. The most notable deficiencies were biases in primary production and sinking flux, which although small were consistent over large geographic regions, and a tendency to underestimate inter-annual variability in sinking particle flux. Both are potentially significant but should be judged relative to the performance of the more complex model against real-world data

Recommendations

Replacement of plankton functional type models by a simpler surrogate with objectively constrained parameters should be considered for representing biogeochemical cycles in Earth system simulations where computational resources are limiting. Output from the reference model should be used to constrain parameters and the emulation uncertainty quantified using independent output so that results from the simpler model can be used to make inferences about the expected behaviour of the reference model.

The parameters of the surrogate model should not be subject to strong prior constraints. Posterior values outside accepted ranges should be tolerated if they are shown to improve predictive skill but treated as an indicator of deficiency in the model design, the correction of which, if feasible, could lead to more reliable predictions in the long term.

Different levels of biogeochemical complexity are best supported within a single community model in the form of a traceable hierarchy. Careful consideration should be given to the design of such a hierarchy to strengthen traceable links by ensuring that common process formulations are applicable at different complexity levels wherever possible and maximize the number of equivalent parameters between the alternative model configurations.

Present assessments of model design are compromised by parametric uncertainty. A capability for objective evaluation of design is needed that allows adequate exploration of large multi-dimensional spaces associated with models' adjustable parameters. This introduces computational demands that directly oppose those required for realistic 3-D simulations. Progress will depend on the use of 1-D modelling capabilities and statistical emulators in conjunction with 3-D modelling tools.

Establishing a traceable link between biogeochemistry models and reality will require process-based assessment of model response to physical drivers as well as assessment of performance when coupled to other ESM components. A global testbed facility is needed, comprising co-incident biogeochemical and physical observations. Robust statistical treatment of uncertainty in the physical environment will be a pre-requisite for reliable calibration of the biogeochemistry component.

KEYWORDS:

ISSUING ORGANISATION **National Oceanography Centre**
University of Southampton Waterfront Campus
European Way
Southampton SO14 3ZH
UK

Not generally distributed - please refer to author

Contents

1 Introduction

A diverse range of candidate models exists for representing ocean biogeochemistry and ecosystem function in Earth system models. These sub-component models, referred to as mechanistic models, are designed to capture the dominant response of the biogeochemical system to its physical and chemical drivers. A representative group of models is currently being examined by the i-MarNet research network to inform a decision by NERC and the Met Office on the baseline model that will be used as the ocean biogeochemistry component of the next generation UK Earth System Model. They range from the Hadley Centre Ocean Carbon Cycle model (HadOCC), a simple "NPZD" model based on a 4 compartment nitrogen cycle, through to the European Regional Seas Ecosystem Model (ERSEM) and the PlankTOM10 model, which are relatively complex plankton functional type models.

NPZD models like HadOCC represent the flow of nitrogen between nutrient (dissolved inorganic nitrogen), phytoplankton, zooplankton and detritus reservoirs. In the more complex models, the 2 plankton pools are replaced by up to 10 functional groups. Additional tracers represent other nutrients and bacteria as well as different elements within the organic pools for modelling stoichiometric variability. As a consequence of the large number of tracers, these models are too computationally demanding to be integrated routinely in ESM simulations. Nevertheless, complex models have a role for two reasons. Firstly, they can be used to address questions that cannot be tackled by the simpler models where the latter do not include the relevant outputs. Secondly, it is possible that they may have greater predictive skill by virtue of a more accurate mechanistic representation of important biogeochemical processes. Some tentative evidence of this is provided by Friedrichs et al (2007) in a particular experimental context. However, the generality of their conclusions remains unproven; the optimal level of complexity required to represent ocean biogeochemistry for answering particular scientific questions is an open research question.

All of the mechanistic ocean biogeochemistry models are necessarily semi-empirical, relying on adjustable parameter values to compensate for missing biogeochemical complexity and variability and incomplete ecological knowledge. Parametric uncertainty thus complicates any comparative assessment of model design. Uncertainty in the physical environment, to which many biogeochemical processes are particular sensitive, is a further barrier to model assessment since a poor fit to biogeochemical observations can be the result of interaction of errors in ocean biogeochemistry and physics models on a range of time scales. Acknowledging these issues, i-MarNet will need to develop a strategy for assessing suitability of different biogeochemistry sub-component models for addressing priority science questions in an ESM context.

There are recognized benefits of supporting different levels of complexity within a new biogeochemistry module for UKESM. The approach will give the flexibility for running more biologically complex simulations for limited regions or time periods or at higher resolution for particular applications or for detailed development work. Results from such simulations will inform the interpretation of global simulations with simpler biogeochemistry. An important aim is to ensure that where simplifications are made there is a clear, traceable link between models of different complexity so that the implications of these simplifications can be understood. Although we cannot assume a priori that a particular, more complex representation will necessarily give better predictive skill than a simpler one, the ability to explore the sensitivity of predictions to model structure and process formulations is important.

The idea of traceability between models of different complexity embraces different concepts. We can easily establish a basic level of structural traceability if it is possible to map compartments of one model onto those of another. Attempting to establish traceability between process formulations affecting the inter-compartmental flows is less straightforward but can in principle be addressed by model design; we can make design decisions that make it easier to see how one model's representation of a process relates to another's, using common formulations wherever possible. However, if we want to assess traceability between the performance of model designs we need to investigate model behaviour and how this varies with the models' adjustable parameters. Establishing traceability to nature is a related but more challenging problem.

A pilot study in traceability of model performance is presented here, investigating the ability of a simple surrogate model to emulate the dynamics of an intermediate complexity model, the Model of Ecosystem Dynamics, nutrient Utilisation, Sequestration and Acidification (MEDUSA 1.0; Yool et al 2011). A version of HadOCC, described in Appendix A, is used as the surrogate model. HadOCC parameter values are first optimized in an attempt to obtain a best fit to MEDUSA output over a set of North Atlantic sites representing a wide range of oceanic conditions. The experimental details are preceeded by a comparison between the two models in Section 2.

2 Comparison of Model Designs

2.1 Structural Traceability

MEDUSA 1.0 has 11 tracers of which 6 are nitrogen tracers. These can be mapped onto the 4 nitrogen compartments represented in HadOCC: dissolved inorganic nitrogen N, phytoplankton P, zooplankton Z and detritus D. DIN and detrital nitrogen are common to both models. The phytoplankton pool in HadOCC corresponds to the sum of MEDUSA's non-diatom phytoplankton and diatom pools and zooplankton corresponds to the sum of MEDUSA's microzooplankton and mesozooplankton pools. The remaining tracers in MEDUSA are 2 chlorophyll tracers representing the chlorophyll content of the 2 phytoplankton types, one representing diatom silicon and 2 more representing dissolved silicon and iron. HadOCC models local variation in chlorophyll composition of the phytoplankton but does not carry chlorophyll as a tracer. It has no representation of silicon or iron cycles.

2.2 Process Formulations

The focus of the study is to examine traceability between different model structures. Some minor modifications were therefore made to the HadOCC design and default parameter values to remove differences that are unrelated to the differences in model structure and would unnecessarily complicate interpretation of the results. In particular, MEDUSA-like temperature dependency was introduced in phytoplankton growth and remineraliztion rates, carbon:nitrogen ratios were made uniform across all organic components and a common light transmission and photosynthesis sub-model was used (with differences from the native sub-models in both HadOCC and MEDUSA). The version of HadOCC used in the optimization experiments is described in Appendix A. The remaining differences in process representation between the two model designs are outlined here.

Temperature dep introduced to allow HadOCC to avoid major differences between the models induced purely by physical forcing

2.2.1 Photosynthesis

In MEDUSA, photosynthetic rate is a function of PAR E_{d}, temperature T, DIN N and dissolved iron. In the case of diatoms it is also affected by the concentration of dissolved silicon in the form of silicic acid. Iron concentrations can directly limit total primary production while silicate concentration affects the partitioning of photosynthesis between diatoms and non-diatoms. This has a more subtle effect on total production via the differential growth rates of diatoms and non-diatoms and the interactions between different ecosystem compartments. Neither iron or silicon are represented in the simpler HadOCC model so nutrient limitation is a function of DIN only and the range of environmental factors to which HadOCC can respond is reduced.

Ignoring iron and silicon limitation, photosynthesis for each phytoplankton type in MEDUSA is given in terms of the initial P-E curve slope α, the maximum growth rate V_{P} and the DIN limitation factor Q_{N} by

$$\mu_{\mathrm{P}} = \bar{J}[\alpha(\alpha_{\mathrm{chl}}, \theta_{\mathrm{chl}}), V_{\mathrm{P}}(V_0, T), E_{\mathrm{d}}] \cdot Q_{\mathrm{N}}(k_{\mathrm{N}}) \tag{1}$$

where model parameters α_{chl}, V_0 and k_{N} are dependent on the phytoplankton type and θ_{chl} is the applicable C:Chl ratio. This contrasts with the HadOCC formulation

$$\mu_{\mathrm{P}} = \bar{J}[\alpha(\alpha_{\mathrm{chl}}, \theta_{\mathrm{chl}}), V_{\mathrm{P}}(V_0, T, Q_{\mathrm{N}}(k_{\mathrm{N}})), E_{\mathrm{d}}] \tag{2}$$

The HadOCC functions $\bar{J}$, α, V_{P} and Q_{N} are given in Appendix A. The MEDUSA forms are identical except for the absence of the nitrogen limitation factor Q_{N} factor from V_{P} which defines the maximum of the photosynthesis-PAR response curve (P-E curve) for saturating light. In MEDUSA, Q_{N} applies as a scaling factor to the overall light response, whereas in HadOCC it only limits the maximum growth rate.

This difference in co-limitation by light and nitrogen means that HadOCC would tend to exhibit higher photosynthetic rates than MEDUSA if both models had the same photosynthesis parameters. The effect is compensated for by lower prior parameter values for the chlorophyll-specific slope α_{chl} and the half saturation concentration k_{N}: $\alpha_{\mathrm{chl}} = 5.56$ mg C (mg Chl)$^{-1}$ (E m^{-2})$^{-1}$, compared with 37.8 and 28.4 mg C (mg Chl)$^{-1}$ (E m^{-2})$^{-1}$ for non-diatoms and diatoms respectively in MEDUSA (taking 1 E d^{-1} to be 2.52 W for the PAR spectrum integral); $k_{\mathrm{N}} = 0.1$, compared with 0.5 and 0.75 mmol N m^{-3} for non-diatoms and diatoms. The prior V_0 in HadOCC is chosen to be the same as the non-diatom value 0.53 d^{-1}. For diatoms it is 0.5 d^{-1}. The difference in formulation means that there is no equivalence of parameters between models so comparison of alternative formulations requires that parameters are subjected to external constraints. This is a good example of where design might be rationalized by either adopting a single formulation that is judged to be preferable on theoretical grounds or by making both options available in both models to facilitate sensitivity studies and inter-comparison.

2.2.2 Grazing

The MEDUSA grazing formulation for zooplankton of either type grazing on food type X is

$$G_{\mathrm{X}} = g_{\max} p_{\mathrm{X}} \cdot \frac{X^2}{k_{\mathrm{F}}^2 + \Sigma(p_{\mathrm{X}} X^2)} \cdot Z \tag{3}$$

where p_{X} is the prescribed preference for food type X and Σ represent summation over all food types.

The equivalent HadOCC formulation is

$$G_{\mathrm{X}} = g_{\max} \cdot \frac{X}{\Sigma X} \cdot \frac{F^2}{k_{\mathrm{F}}^2 + F^2} \cdot Z \tag{4}$$

where

$$F = \max(0, \Sigma X - F_{\mathrm{threshold}}) \tag{5}$$

The squared half-saturation constant k_{F} can be expressed as $\frac{g_{\max}}{\epsilon_{\mathrm{F}}}$ where ϵ_{F} is a prey capture rate parameter.

For $F_{\mathrm{threshold}} = 0$, the HadOCC grazing formulation would be identical to the MEDUSA formulation for a single food source but not when both phytoplankton and detritus are present. The parameters $g_{\max}$ and ϵ_{F} are functionally equivalent in both models but the MEDUSA values differ between the two zooplankton types so there are no equivalent HadOCC values. The prior HadOCC value for $g_{\max}$ is 0.8 d^{-1}, compared with 2 and 0.5 in MEDUSA for micro- and meso-zooplankton respectively. For ϵ_{F} it is 3.2 d^{-1} (mmol N $\mathrm{m}^{-3})^{-2}$), compared with 3.12 for microzooplankton and 5.56 for mesozooplankton.

The partitioning of grazed material between zooplankton biomass, DIN and detritus can be made identical between the two models by setting the relevant HadOCC parameters to the corresponding MEDUSA values, i.e. $\phi_{\mathrm{I}} = 0.8$, $\beta_{\mathrm{P}} = 0.69$, $\beta_{\mathrm{D}} = 0.69$ and $\phi_{\mathrm{mfN}} = 1$. (For HadOCC priors, see Table 5.)

2.2.3 Phytoplankton Losses

Both HadOCC and MEDUSA have linear and density-dependent loss terms for the phytoplankton. However, the formulation for density dependency differs as does the way in which losses are partitioned between DIN and detritus.

In HadOCC, the total phytoplankton loss is $M_{\mathrm{P}} + \eta P$ where the density-dependant mortality term

$$M_{\mathrm{P}} = mP^2. \tag{6}$$

$m = m_\mathrm{o}$ for $P >= 0.01$ mmol N m^{-3} (otherwise mortality is suppressed). All of the linear loss is associated with metabolism and goes to DIN. Density dependent losses are intended to represent mortality. A fraction ϕ_MPN of the density-dependent mortality goes to DIN and the remainder to detritus.

MEDUSA has an equivalent linear loss term, with $\eta = 0.02$ d^{-1}, for both phytoplankton types, but the density-dependent mortality is

$$M_\mathrm{P} = 0.1\frac{P^2}{0.5 + P}. \tag{7}$$

As in HadOCC, all of the linear loss goes to DIN but the density-dependent mortality goes exclusively to detritus.

The prior value for ϕ_MPN in HadOCC is very low at 0.01, so the parameterizations are superficially very similar. The similarity is greatest at low phytoplankton concentrations. However, the MEDUSA formulation for density dependency diverges from a P^2 HadOCC-like term with $m \approx 0.2$ d^{-1}(mmol N m^{-3})$^{-1}$ as the concentration increases, tending towards a linear term at high concentrations relative to the half-saturation concentration of 0.5 mmol N m^{-3}. The effect is that at high concentrations linear losses dominate in MEDUSA with fractions going to both DIN and detritus. In the limit, the DIN fraction of linear losses tends to 0.17 as P^2 losses tend to zero. In HadOCC, P^2 losses dominate, leading ultimately to greater losses, with ϕ_MPN effectively controlling the partitioning. Once again the differences mean that there is no direct equivalence of parameters in the phytoplankton loss formulations between the two models.

2.2.4 Zooplankton Losses

As for phytoplankton, both models have linear and density-dependent loss terms. Total losses in HadOCC are

$$M_\mathrm{Z} = m_1 Z + m_2 Z^2 \tag{8}$$

with a fraction ϕ_MZN going to DIN and the remainder to detritus. Linear loss rates for both zooplankton types in MEDUSA are the same as those for the phytoplankton

(0.02 d^{-1}). The formulation of density-dependant mortality is also the same but with different coefficients for the mesozooplankton. For microzooplankton

$$M_{\mathrm{Z}} = 0.1\frac{P^2}{0.5 + P}. \tag{9}$$

For mesozooplankton

$$M_{\mathrm{Z}} = 0.2\frac{P^2}{0.75 + P}. \tag{10}$$

As with the phytoplankton, there is no direct equivalence of parameters in the loss formulations between the two models.

2.2.5 Sinking and Remineralization of Detritus

Although HadOCC and MEDUSA have the same detritus pools, part of the detritus production in MEDUSA by-passes this pool. Two types of detritus are represented in MEDUSA: slow-sinking detritus, which enters the detritus pool and fast-sinking detritus which is parameterized by removing material at one model level and instantly re-distributing it among deeper levels as DIN according to a remineralization scheme based on the ballasting model of Dunne et al. (2007). The slow-sinking fraction of detritus production comprises all material egested by zooplankton, all of the density-dependent mortality of the smaller plankton types (non-diatom phytoplankton and microzooplankton) and 25% of the density-dependent diatom mortality. The fast-sinking fraction comprises all of the density-dependent mesozooplankton mortality and the remaining 75% of the density-dependant diatom mortality. The parameterization of sinking and remineralization for detritus in HadOCC is identical to that for slow-sinking detritus but the prior value for the sinking velocity parameter is higher than the MEDUSA value: 10 m d^{-1} compared with 3 m d^{-1}.

The additional complexity of the MEDUSA parameterizations is motivated primarily by the importance of accurately modelling the vertical flux of carbon that drives the biological pump. The requirement for separate representation of small slow-sinking detritus and large fast-sinking detritus is, in turn, a key motivating factor for subdividing phytoplankton and zooplankton types.

2.2.6 Carbon:Chlorophyll Ratio

Photo-acclimation in both models is introduced by varying the biomass-specific chlorophyll concentration in response to the ratio of realized photosynthesis μ_P to αE_d. In MEDUSA, this is done by adjusting the chlorophyll tracers associated with each phytoplankton type. The implementation in HadOCC is very different in that an iterative approximation to a steady state model of the carbon:chlorophyll ratio is used (with one iteration per time step). Chlorophyll is not handled as a tracer so the ratio is based purely on local conditions. The differences in implementation of photo-acclimation inhibit direct comparison between the two parameterizations. Experiments are performed here with both variable and fixed C:Chl ratios in HadOCC.

3 Method

3.1 Experimental Framework

To provide a range of oceanic conditions for the experiments, 12 sites were selected located on a meridional transect along 20W in the North Atlantic from 5N to 60N. 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 Canary Islands.

The experiments were performed in a 1-D framework using the Marine Model Optimization Testbed (MarMOT) system (Hemmings and Challenor, 2012). MarMOT was developed with NERC support via the National Centre for Earth Observation as a flexible, user friendly software tool to enable computationally intensive biogeochemical model analyses for which 3-D tools like NEMO are unsuited.

The evolution of a biogeochemical tracer concentration C_i in a MarMOT water column simulation is given by

$$\begin{aligned}\frac{\mathrm{d}C_i}{\mathrm{d}t} &= -(w_\mathrm{p}+w_i)\frac{\partial C_i}{\partial z} - \frac{\partial w_i}{\partial z}C_i + \frac{\partial}{\partial z}\left(K_\rho \frac{\partial C_i}{\partial z}\right) \\ &\quad +\mathrm{SMS}_i(\vec{C},\vec{F}) + p_i(C_i, p_i^\star) + r_i(C_i^\mathrm{ref} - C_i). \qquad (11)\end{aligned}$$

The first three terms represent the tendencies 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 (if any) and K_ρ is the turbulent diffusion coefficient. Note that any vertical divergence in w_i changes the concentration, whereas vertical divergence in the flow is balanced by fluid continuity so that the associated concentration tendency is zero, assuming homogeneity of tracer concentration in the horizontal. SMS_i is the source-minus-sink term from the selected biogeochemistry model which is a function of the state vector $\vec{C}$ and a forcing vector $\vec{F}$. In this study, $\vec{F}$ comprises local photosynthetically available radiation and temperature. The biogeochemistry model also provides w_i. In both MEDUSA and HadOCC, w_i is non-zero for detritus only and the detrital sinking rate is constant so term 2 dissappears. p_i is a perturbation term (potentially concentration dependent) driven by an applied perturbation $p^\star$. The final term is a relaxation term given by the product of a rate r_i and the deviation of C_i from a reference concentration C_i^{ref}.

Physical forcing data for the experiments, in the form of vertical velocity, vertical diffusivity and temperature are taken from 5 day mean output of a 3-D NEMO-MEDUSA simulation running at 1° horizontal resolution (Yool et al., 2011). 5 day mean time series of downwelling solar radiation at the sea surface are taken from the same simulation. In addition, mixing layer depth is required for the variable carbon:chlorophyll parameterization in HadOCC. This was taken to be the 5 day mean NEMO turbocline depth. 10 years of output from 1996-2005 were used.

Comparisons are shown in Figure 1 for DIN between MEDUSA running in the 3D general circulation model and in the 1D MarMOT testbed with and without corrections for lateral advection. Without these effects there are major differences at some sites, notably 40N, 35N, 20N and especially at 10N where DIN remains depleted in the 3D simulation, contrasting with the occurrence of periodically high concentrations in the absence of advective effects. The missing advective tendencies are the product of the horizontal current velocity and the upstream tracer gradients. By adding advective flux divergences diagnosed from NEMO-MEDUSA output as tracer perturbations $p_i = p_i^\star(z,t)$ we can correct for the differences between 1D and 3D simulations to a large extent at most of the sites. This suggests that much of the discrepancy is due to the absence of horizontal advection in the 1D simulations. However, adding advective perturbations seems to over-compensate at 20N and 60N. The failure at these sites requires further investigation but may be due to other sources of error such as the absence of horizontal diffusion. The reduced time resolution of the forcing is another potential source of error.

In skill assessments where model results are to be compared with real-world data, it would be necessary to properly account for the impact of lateral fluxes. However, applying the NEMO-MEDUSA flux divergence tendencies in HadOCC simulations is not generally appropriate because such tendencies depend on the upstream biogeochemical tracer gradients which typically co-vary to some extent with the local concentration (Hemmings and Challenor, 2012). The desired tendencies are thus

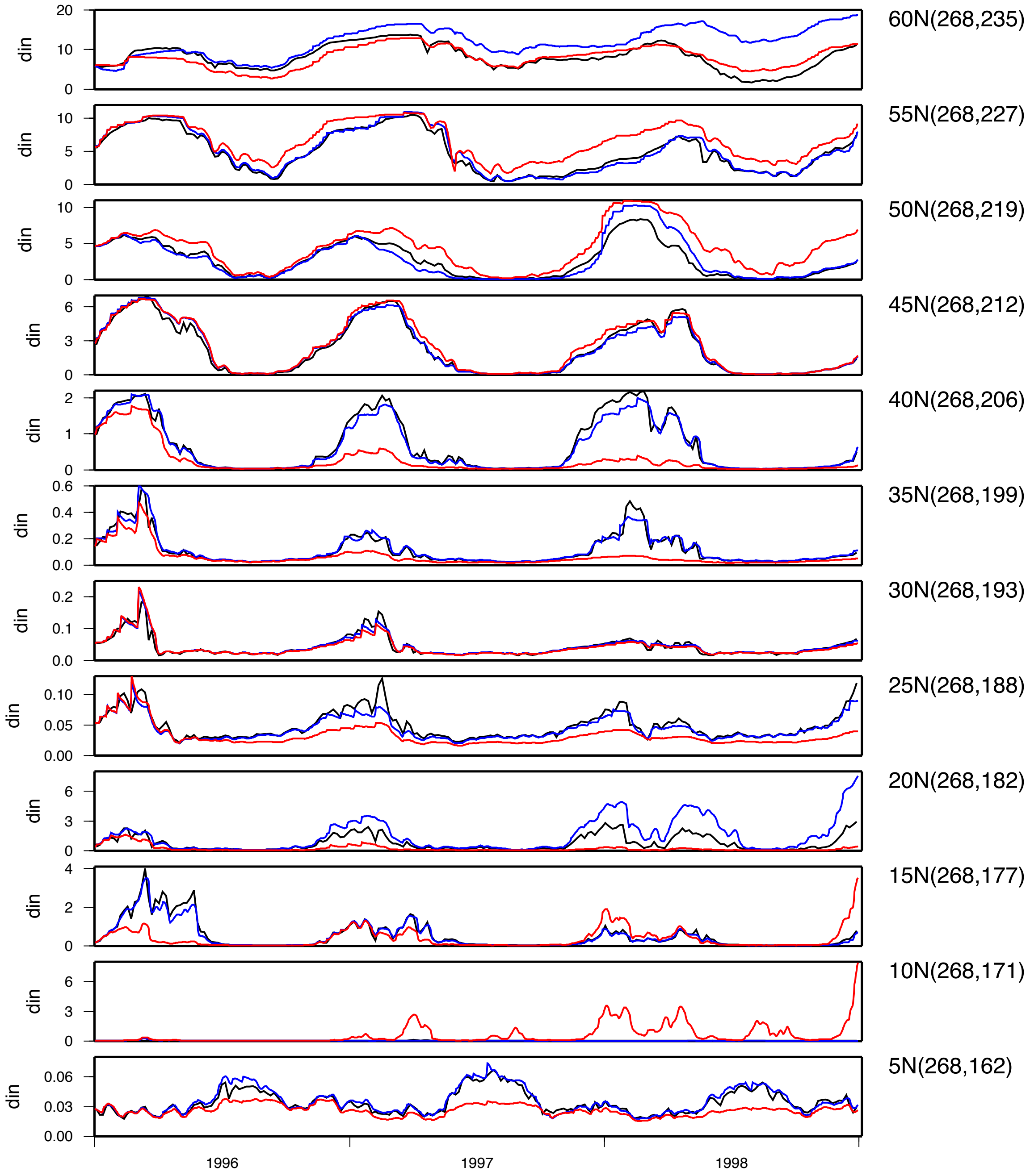

Figure 1: Surface level DIN concentration (mmol N m^{-3}) from the 3-D MEDUSA simulation in NEMO (black) and 1-D simulations in MarMOT with advective tendencies from the 3-D simulation (blue) and without (red). Plots are for each calibration site on the 20W transect. The ORCA1 grid reference is shown in brackets.

model-dependent. The NEMO-MEDUSA tendencies could not therefore be considered compatible with the HadOCC simulation results. Relatively small concentration differences can potentially introduce drift as a result of inadequate handling of concentration dependencies in biogeochemical gradients. This may explain the problem exhibited in the MEDUSA simulation at 60N. We would expect horizontal DIN gradients to be reduced at very high and very low concentrations associated with the gyre interiors. The sustained increase in DIN is a possible consequence of initially small positive errors occurring with no corresponding reduction in advective tendency.

Research is on-going to quantify uncertainty in lateral fluxes associated with variation in model biogeochemistry with a view to effectively managing this source of uncertainty in model calibration and assessment procedures. The introduction of such uncertainty in the context of the idealized synthetic data experiments required for the present study would be an unnecessary complication. No tendencies associated with lateral processes were therefore applied. As a consequence, there is significant drift at some sites when the simulation is run for multiple years. This could in principle be corrected for by relaxation to a MEDUSA climatology. However, this is unnecessary here and would complicate the interpretation of the HadOCC results.

3.2 Simulation Details

Parameter optimization experiments are computationally intensive, requiring many thousands of model integrations. While longer simulations are desirable for removing transient effects, shorter simulations are desirable for reducing run-time. They are also less susceptible to the effects of long-term drift. As a compromise, the simulation period was chosen to be 2 years with the synthetic observations taken from the second year.

1-D simulations use the same vertical grid as the 3-D NEMO simulation. The dynamics of interest are largely confined to the upper ocean. The deeper nutrient distribution is affected by the flux of sinking particles and their remineralization rate but these effects tend to be correlated down the water column so that the inclusion of deep observations is unlikely to provide useful independent constraints. On this assumption, a depth threshold of 1000 m was chosen for the simulations, reducing the number of model levels from 63 to 37 with consequent computational savings. The bottom of level 36 is below the threshold. The vertical velocity and diffusion at the bottom of this level were set to zero to prevent any interaction between level 37 and the water column above. Sinking detritus is remineralized in the bottom level so level 37 was included as a sink for detritus entering from 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 approximate 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 and 991 m. An implicit scheme is used for diffusion and the MUSCL scheme for advection. The time step is 2 hours (forward Euler).

3.3 Data Constraints

A calibration data set of synthetic observations was created from 1997 MEDUSA output taken from a 2 year simulation initialized from the 3-D NEMO-MEDUSA simulation at the beginning of 1996. The data set comprises concentrations of DIN, total phytoplankton nitrogen and total zooplankton nitrogen at each of the 35 model levels (spanning 0 - 1000 m). Additionally, the flux of particulate organic nitrogen (PON) at the bottom of each level was used as an additional constraint in a subset of experiments. This includes contributions from both slow-sinking and fast-sinking detritus and is directly related to the flux of particulate organic carbon (POC) by the Redfield ratio (6.625). Each variable was sampled at 5 day intervals, taking a daily mean value.

Any assessment of predictive skill requires validation against independent data not used in parameter fitting. For validation purposes, a series of 2 year simulations was used with data taken from the second year to create a validation data set comprising data from years 1998-2005. Column integrated primary production was included as an additional validation variable.

3.4 Parameter Optimization

3.4.1 Cost Function

The parameter optimization procedure used to calibrate HadOCC as an emulator of MEDUSA relies on a cost function that summarizes the misfit between the HadOCC simulation and the synthetic observation data set. The MarMOT system supports a generic cost function for multiple variable types and multiple simulation cases of the form

$$J = \frac{1}{N}\sum_{k=1}^{C}\sum_{j=1}^{m}\sum_{i=1}^{n_k} p_{ijk} w_{ijk}(x_{ijk} - y_{ijk})^2 \tag{12}$$

$$N = \sum_{k=1}^{C}\sum_{j=1}^{m}\sum_{i=1}^{n_k} p_{ijk} \tag{13}$$

where C is the number of cases (here equivalent to calibration sites), n_k is the number of observation points for case k (in space and time) and m is the number of observed variables; x_{ijk} is the simulated value of the j-th variable at the i-th observation point and y_{ijk} is its observed value. We refer to the squared residual $(x_{ijk} - y_{ijk})^2$ as the model misfit. The coefficient p_{ijk} is 1 if the variable is present in the observation set or 0 otherwise. Since the present study is based on synthetic data, we have values for all calibration set variables at all observation points so p_{ijk} is always 1. w_{ijk} is a weighting factor to be applied to the misfit. If w_{ijk} is the reciprocal of the expected residual variance for a perfect simulation then the cost function value for a perfect simulation should approach 1 for large N.

In the present study where we are calibrating a surrogate model with reference to a known true solution, so can consider both observation and simulation error to be zero. The residual variance for a perfect simulation will therefore be that associated with model discrepancy, originally introduced as model inadequacy (Kennedy and O'Hagan, 2001). Model discrepancy is defined as the difference between the simulation output for the best possible parameter set, which is unknown, and reality (in our case the MEDUSA output).

Prior knowledge of model discrepancy for the surrogate model here is considered insufficient to justify differential weighting for individual observations. In general though, different weights are appropriate for observed variables having different units. A widely used approach, the Characteristic Scale method (e.g. Friedrichs et al., 2007), uses the overall variance of the variable in the observation set as a scaling factor. This can be particularly useful for balancing misfit contributions from different variables to avoid results that depend on arbitrary relationships between units. However, it is essentially a pragmatic solution as the variance in the data set is not directly related to the expected error variance. The r.m.s errors obtained here with respect to the sinking particle flux data are of a similar magnitude numerically to those for the state variables without any scaling. For simplicity, a uniform weighting $w_{ijk} = 1$ was therefore applied to each variable.

For real-world calibration exercises, Hemmings and Challenor (2012) recommend the eventual replacement of the Characteristic Scale method by a more explicit treatment of observation and simulation uncertainties. The recommended method is demanding in terms of the physical data and uncertainty quantification effort it will require but is strongly supported by a demonstration of its potential in experiments with synthetic data.

3.4.2 Optimization Procedure

The MarMOT optimizer combines a genetic algorithm for identifying promising areas of a bounded parameter space and a non-gradient direction set algorithm for bounded or unbounded local minimization. The genetic algorithm is a global method in the sense that it is able to locate multiple minima in the cost function by sampling the global parameter space. However, it searches the parameter space in discrete intervals, limiting the accuracy with which it can locate a particular minimum. In contrast, the direction set algorithm navigates towards a local minimum from a given starting point, making it unsuited to finding the global minimum in a cost function with complex topography, but can give greater accuracy.

The genetic algorithm is a micro-genetic algorithm (μGA) (Krishnakumar, 1989), based on an implementation by Carroll (1996). It has been applied to the problem of plankton model optimization by Schartau and Oschlies (2003) and subsequently by other workers including Ward et al. (2010) who compared its performance with the local variational adjoint technique employed by Friedrichs et al. (2007).

The direction set algorithm was designed by Powell (1964) to locate a cost function minimum in a continuous unbounded free parameter space. The implementation of bounded minimization is described by Hemmings and Challenor (2012). The version of Powell's algorithm used is that described in Press et al. (1992), with reference to Acton (1970). Line minimization is performed using Brent's method (Brent, 1973). No gradient information is used so it does not require the provision of an adjoint code for calculating the cost function gradient with respect to the model parameters. It is therefore more straight-forward to apply than the variational adjoint method in situations where the formulation of the plankton model is not fixed. The algorithm has been applied in a number of plankton model calibration studies (e.g. Fasham and Evans, 1995; Hemmings et al., 2004; Dadou et al., 2004; Fasham et al., 2006).

The optimization procedure was identical for each set of optimization experiments. Initial optimization was performed with the μGA which was run for a minimum of 1000 generations to provide a pre-conditioned set of parameter vectors for local searches with the direction set algorithm; the best 5 parameter vectors from the population were selected. On any convergence in the parameter vector population, defined by uniformity across the population in at least 95 % of the bits in the binary code describing the parameter vectors, a new random population is generated, retaining the best individual. Additional generations after Generation 1000 were run until the next convergence. The algorithm was configured with a single-point cross-over between bit strings at a probability of 1. Each parameter was represented by 8 bits in the μGA, giving 256 possible values prior to refinement by the local searches.

The population size for the μGA was chosen to match the number of free parameters following the recommendation of Schartau and Oschlies (2003). Initial parameter vectors in the original population were 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 500 randomly generated hypercubes: the hypercube design is selected that maximizes the smallest Euclidean distance between parameter vector pairs in terms of their positions on the 256^n grid.

The direction set algorithm was applied to each unique parameter vector in the final population returned by the μGA and the lowest cost result selected. To investigate the sensitivity of the result to the initial parameter vectors, each application of the optimizer was repeated for 10 alternative designs, choosing those with the largest minimum Euclidean distances from a sample of 500 randomly generated hypercubes.

3.5 Parameter Ranges

Parameter bounds are required for the μGA and can be prescribed or omitted in the MarMOT implementation of the Powell algorithm. Some parameters must necessarily be restricted: rate parameters must be positive; fractions must be between 0 and 1. Ranges are also often used to ensure that parameters have biologically meaningful values. Whether this is desirable depends on the the experimental goals. Useful information about the limitations of a model design can be gained from results in which posterior parameter values lie outside the range of plausible values consistent with the parameter’s description. Out-of-range values are problematical if we want to interpret the model mechanistically but should not be ruled out if the purpose is to use the model as an emulator to best predict the results of another model. Neither should we necessarily rule out such values in the context of a real world experiment if they lead to better predictive skill, although a change of model design would be preferable.

In the present study, the aim is to find a HadOCC parameter set that allows it to best emulate the behaviour of MEDUSA. Broad parameter ranges are chosen to avoid introducing constraints at an early stage that could compromise this goal. The MEDUSA data are used as the primary constraint. For each of the rate parameters,

Table 1: HadOCC parameter ranges for optimizer

Parameter	Symbol	Prior	Lower bound	Upper bound	Log flag
photmax0	V_0	0.53	0.05	5	1
alphachl	α_{chl}	5.56	0.5	500	1
kdin	k_{N}	0.1	0.01	10	1
presp	η	0.05	0.005	0.5	1
pmortdd	m_{o}	0.05	0.005	0.5	1
fpmortdin	ϕ_{MPN}	0.01	0	1	0
gmax	$g_{\max}$	0.8	0.1	10	1
epsfood	ϵ_{F}	3.2	0.3	30	1
fingest	ϕ_{I}	0.77	0.1	1	0
betap	β_{P}	0.9	0.1	1	0
betad	β_{D}	0.65	0.1	1	0
fmessydin	ϕ_{mfN}	0.1	0	1	0
zmort	m_1	0.05	0.005	0.5	1
zmortdd	m_2	0.3	0.03	3	1
fzmortdin	ϕ_{MZN}	0.67	0	1	0
dsink	w_{D}	10.0	0	30	0
remin0	λ_0	0.016	0.0015	0.15	1

with the exception of detrital sinking rate, parameter optimization was performed in log space to avoid negative values, the transformation being applied before discretization. For these parameters, the lower and upper bounds for the μGA were initially set a factor of 10 lower or higher than the prior parameter value and the bounds were removed for subsequent local optimization with the Powell algorithm to avoid imposing artificial constraints. Detrital sinking rate was constrained to remain below 30 m d^{-1} to avoid numerical instability. Two parameters, α_{chl} and k_{N} tended to be forced outside the prescribed range in preliminary experiments leading to rather higher final values. For these parameters, the upper bound was increased by another order of magnitude. Identical ranges were used in all subsequent experiments. These are shown in Table 1. (Refer to Table 5 for parameter descriptions and units.)

Table 2: Posterior parameter values.

Parameter	PRIOR	OPT17-NPZ	OPT13-NPZ	OPT13CC-NPZ	OPT13CC-NPZF	OPT10CC-NPZF
photmax0	0.53	**0.222**	**0.133**	**0.269**	**0.273**	**0.268**
alphachl	5.56	**87.2**	**106**	**17.9**	**18.4**	**16.3**
kdin	0.1	**0.62**	**0.494**	**0.856**	**0.923**	**1.33**
presp	0.05	**0.0901**	0.05	0.05	0.05	0.05
pmortdd	0.05	**0.148**	0.05	0.05	0.05	0.05
fpmortdin	0.01	**0.27**	**1.75e-13**	**0.184**	**0.156**	**0.299**
gmax	0.8	**0.828**	**0.878**	**0.805**	**0.786**	**0.817**
epsfood	3.2	**0.326**	**0.264**	**0.403**	**0.381**	**0.254**
fingest	0.77	**0.905**	**0.56**	**1**	**0.979**	0.8
betap	0.9	**1**	**1**	**0.86**	**0.995**	0.69
betad	0.65	**0.173**	**0.286**	**0.1**	**0.115**	0.69
fmessydin	0.1	**0.849**	0.1	0.1	0.1	1
zmort	0.05	**0.0231**	**0.0166**	**0.0137**	**0.0146**	**0.00538**
zmortdd	0.3	**0.123**	**0.0887**	**0.172**	**0.189**	**0.149**
fzmortdin	0.67	**0.509**	**0.721**	**0.613**	**0.572**	**0.245**
dsink	10	**7.45**	**6.41**	**6.54**	**6.62**	**6.98**
remin0	0.016	**0.0181**	0.016	0.016	0.016	0.016

4 Results

4.1 Parameter Optimization Experiments

4.1.1 Sequence of Experiments

Some preliminary experiments were performed with 17 adjustable parameters. In these experiments, clear correlations were noted between certain pairs of parameters in the posterior parameter set over the 10 optimization runs. In particular, phytoplankton metabolic loss rate η (presp) and density-dependent mortality m_{o} (pmort) were both positively correlated with the base phytoplankton maximum growth rate V_0 (photmax0) and there was a strong positive correlation between detrital sinking rate w_{D} (dsink) and the base remineralization rate λ_0 (remin0). The two phytoplankton loss parameters and the remineralization rate were excluded from the optimization procedure in subsequent experiments. In general, the parameters controlling partitioning of losses and the by-products of grazing between DIN and detritus appeared to be poorly constrained. Three separate parameters ϕ_{MPN} (fpmortdin), ϕ_{MPZ} (fzmortdin) and ϕ_{mfN} (fmessydin) all perform essentially the same role, albeit associated with different processes. The potential for parameter interactions was reduced by also excluding ϕ_{mfN}, the DIN fraction of material resulting from messy feeding, in subsequent experiments.

The posterior parameter sets corresponding to the experimental results presented here are given in Table 2 with the values of the optimized parameters shown in bold. One of the preliminary 17 parameter experiments is included for comparison. This is referred to as OPT17-NPZ, the suffix "-NPZ" indicating that the observational constraints comprised DIN, phytoplankton and zooplankton. The equivalent 13 parameter experiment is OPT13-NPZ. These two experiments were performed on the model version with fixed C:Chl ratio. In Experiment OPT13CC-NPZ, identical constraints to those in OPT13-NPZ are applied to the variable C:Chl ratio version. The altered model design led to some important differences in posterior parameters. In particular, the value of α_{chl} (alphachl) is very much greater in both experiments with fixed C:Chl. This is explained by the lack of potential in the fixed C:Chl version to increase the quantum efficiency of the photosynthetic response at low light levels by increasing the biomass-specific chlorophyll. To compensate, the chlorophyll-specific response is increased by increasing α_{chl}. There appears to be an interaction with the base temperature-dependent maximum growth rate V_0 (photmax0): a particularly low value of 0.133 d^{-1} in Experiment OPT13-NPZ is apparently required to avoid the high α_{chl} value causing excessive production at higher light levels. V_0 is reduced by 75% from the prior. The remaining experiments were all performed on the variable C:Chl model which is more directly comparable with MEDUSA.

In experiment OPT13CC-NPZF, the effect of introducing the sinking particle flux data as an additional constraint was explored. The resulting parameter set was surprisingly similar, suggesting strong coupling between the particle flux and the state variables used.

In the 13 parameter experiments, many parameters still appear to be poorly constrained (see Appendix B) consistent with an under-determined problem. The apparent scope for arbitrary parameter adjustment prompted a final experiment to investigate the possibility of achieving similar or improved emulator performance by using MEDUSA-specific values. In OPT10CC-NPZF, the parameters controlling the partitioning of grazed material between zooplankton biomass, DIN and detritus were set to their MEDUSA equivalents accordingly. These are parameters ϕ_{I} (fingest), β_{P} (betap), β_{D} (betad) and ϕ_{mfN} (fmessydin). The experiment is otherwise identical to OPT13CC-NPZF.

The performance with respect to the calibration variables in the independent validation data set is shown in Table 3 which gives the mean intra-annual r.m.s values for the years 1998-2005. The overall misfit for the nitrogen concentration variables are also shown for comparison with the flux values and the two sets of values are broadly comparable despite the different units. The misfits of these variables to the calibration data were likewise similar indicating that the optimization process in the last two experiments was not unduly distorted by the relative contribution of variables with different units.

Table 3: Goodness-of-fit to MEDUSA data (1998-2005): calibration set variables.

Experiment	N r.m.s.e.	P r.m.s.e.	Z r.m.s.e.	N,P,Z r.m.s.e.	PON flux r.m.s.e.
	mmol N m^{-3}	mmol N m^{-3}	mmol N m^{-3}	mmol N m^{-3}	mmol N m^{-2} d^{-1}
PRIOR	1.27	0.15	0.132	0.74	0.66
PRIORCC	1.18	0.143	0.125	0.692	0.621
OPT17-NPZ	0.238	0.0542	0.0714	0.147	0.284
OPT13-NPZ	0.265	0.0688	0.0654	0.163	0.307
OPT13CC-NPZ	0.237	0.0685	0.0471	0.145	0.298
OPT13CC-NPZF	0.24	0.07	0.0495	0.147	0.292
OPT10CC-NPZF	0.257	0.0824	0.0534	0.159	0.305

The performance against the validation data is similar for all posterior parameter sets. There is a substantial improvement in r.m.s. error for all variables over the corresponding results for the uncalibrated model, indicating improved predictive skill. Errors for DIN are generally high relative to the plankton concentrations, although this difference is reduced from an order of magnitude to a factor of 4 or 5 after calibration. It is interesting to note that there is little difference in performance with respect to the particle flux when it was added as an additional constraint. This suggests that the extra constraint is largely redundant, given perfect observations of the state variables.

4.1.2 Posterior Parameter Sets

A number of patterns are common to all experiments. The values of α_{chl} (alphachl) and k_{N} (kdin) are always much higher in the posterior parameter set, reducing light limitation and nitrogen limitation respectively. The value of V_0 (photmax0) is correspondingly low, avoiding excessive production.

On the other hand, the maximum ingestion rate $g_{\max}$ (gmax) seems better behaved, remaining close to the prior value a little below the geometric mean of the MEDUSA values for the different zooplankton types. However, the prey capture rate ϵ_{F} (epsfood) is an order of magnitude lower than the prior indicating that grazing is much more strongly limited by food concentration in the calibrated model. This is compensated for by low values of the zooplankton mortality parameters (zmort and zmortdd).

Another interesting pattern is the consistent tendency for the optimization procedure to cause divergence in the values of the different assimilation efficiencies for the different food types betap for phytoplankton and betad for nitrogen. This seems unrealistic, especially since differences in food quality as represented in the model have been removed by the use of a uniform C:N ratio for both.

Table 4: Goodness-of-fit to MEDUSA data (1998-2005): production and export (Convert PON to POC)

Experiment		R.m.s. error			Bias	
	P. Production	100 m PON flux	531 m PON flux	P. Production	100 m PON flux	531 m PON flux
	mmol C m^{-2} d^{-1}	mmol N m^{-2} d^{-1}	mmol N m^{-2} d^{-1}	mmol C m^{-2} d^{-1}	mmol N m^{-2} d^{-1}	mmol N m^{-2} d^{-1}
PRIOR	22.1	0.733	0.209	-3.38	-0.262	0.0645
PRIORCC	20.9	0.686	0.192	-2.18	-0.242	0.0693
OPT17-NPZ	11.2	0.278	0.13	6.58	0.00424	0.00683
OPT13-NPZ	8.36	0.336	0.134	-0.118	-0.015	-0.00115
OPT13CC-NPZ	8.58	0.336	0.139	1.77	-0.0329	0.00124
OPT13CC-NPZF	9.11	0.335	0.138	2.68	-0.0352	0.00427
OPT10CC-NPZF	8.51	0.366	0.142	-0.0535	-0.047	0.0145

In summary, calibration of the surrogate model has introduced a number of anomalies in parameter values that are inconsistent with the model's intended mechanistic interpretation. The anomalously low value of photmax0 was improved to some extent by using the variable C:Chl version of the model. A further anomaly was removed by fixing the assimilation efficiencies to MEDUSA values. This was possible with minimal effect on the overall performance of the resulting emulator, although it does lead to some increases in a tendency towards positive DIN bias in the oligotrophic regions. (See Figure 14 in Appendix C for more details)

4.2 HadOCC Performance as an Emulator

To assess the performance of HadOCC as an emulator of MEDUSA we focus on primary production and export of organic material associated with the sinking particle flux. Overall statistics for the set of experiments are given in Table 4. All of the calibrated model results show substantial improvements over the results given by the prior parameter sets. The one exception is that OPT17-NPZ shows a larger bias in primary production, suggesting that there is little advantage in optimizing as many as 17 parameters. The OPT10CC-NPZF parameter set gives an extremely small primary production bias, although it also gives a positive bias in the sinking particle flux at 531 m which is relatively large compared with the other posterior parameter sets.

4.2.1 Annual Cycles of Primary Production and Sinking Particle Flux

More detail of the performance is shown by analysis of the seasonal variability. This is shown for the years 1997-2005 in Figures 2, 3 and 4 for the variable C:Chl version of the model.

The seasonal patterns of primary production (Figure 2) are well represented by HadOCC after calibration with only very minor differences between the individual posterior parameter sets. This contrasts with generally poor representation before calibration, particularly at lower latitudes from 10-20N, although primary production in the sub-tropical gyre (25-35N) is better reproduced by the prior parameter set. Despite exceptionally good emulation of the seasonal signal south of the sub-tropical gyre and for most of the year in the north too, one notable deficiency is evident: there is a tendency for the surrogate model not to capture the early phase of the temperate spring bloom in MEDUSA. This difference in timing is most evident at 55 and 60N and there is clearly a negative bias in primary production at these sites. It is compensated for by an obvious positive bias in the subtropical gyre.

Patterns in the 100 m particle flux (Figure 3) are strongly linked to the productivity and are well reproduced by HadOCC after calibration, although there are some exceptions at lower latitudes where there is a tendency to underestimate the annual peaks. There is no obvious bias in the sub-tropics, despite the positive bias in primary production. However, there is a tendency for a lagged response, consistent with the lack of fast-sinking detritus in the surrogate model. The effect is greater and more widespread at 531 m (Figure 4) and the performance of HadOCC as an emulator for the particle flux at this depth is generally worse, particularly at the lower latitudes south of the gyre (5-20N). Nevertheless, the signal is well reproduced in the sub-tropical gyre where calibration leads to a major reduction in bias. If we ignore the lag, the overall seasonal pattern is fairly well reproduced at high latitudes where the effect of calibration is to remove the sharp peaks and spread the particle flux out more over the year. The lagged response of the surrogate model provides a likely explanation for the relatively high r.m.s. values for DIN, compared with those for the plankton.

In general, calibration has a major effect on the seasonal patterns of production and export. The patterns are apparently well determined by the data constraints even though there are differences in final parameter values, particularly between OPT13CC-NPZF and OPT10CC-NPZF (suggesting that parameter redundancy rather than lack of data constraints is the cause of poorly constrained parameters). The figures do not show any difference in performance for the calibration year, 1997, so there is no evidence of over-fitting.

4.2.2 Annual Mean Primary Production and Sinking Particle Flux

Figures 5 and 6 summarize the performance for the annual mean primary production and the particle flux at 3 depth levels (100 m, 207 m and 531 m). The data are from the calibration sites for years 1997-2005. The presentation in Figure 5 focusses on

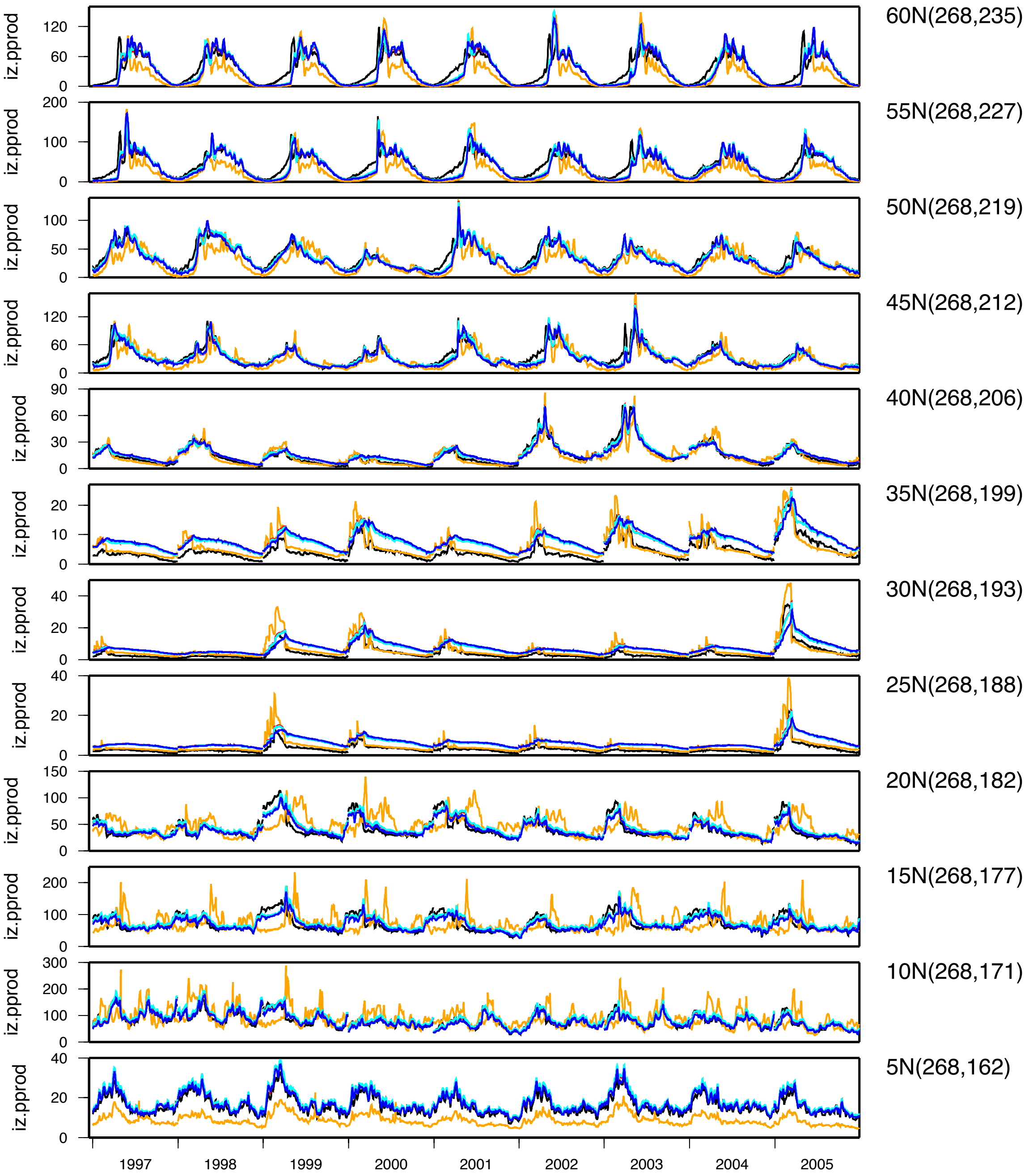

Figure 2: Primary production (mmol C m^{-2} d^{-1}) from second year of 2 year simulations: MEDUSA (black); HadOCC with prior parameters (orange), OPT13CC-NPZ parameters (red), OPT13CC-NPZF parameters (cyan), OPT10CC-NPZF parameters (blue). Plots are for each calibration site on the 20W transect. The ORCA1 grid reference is shown in brackets.

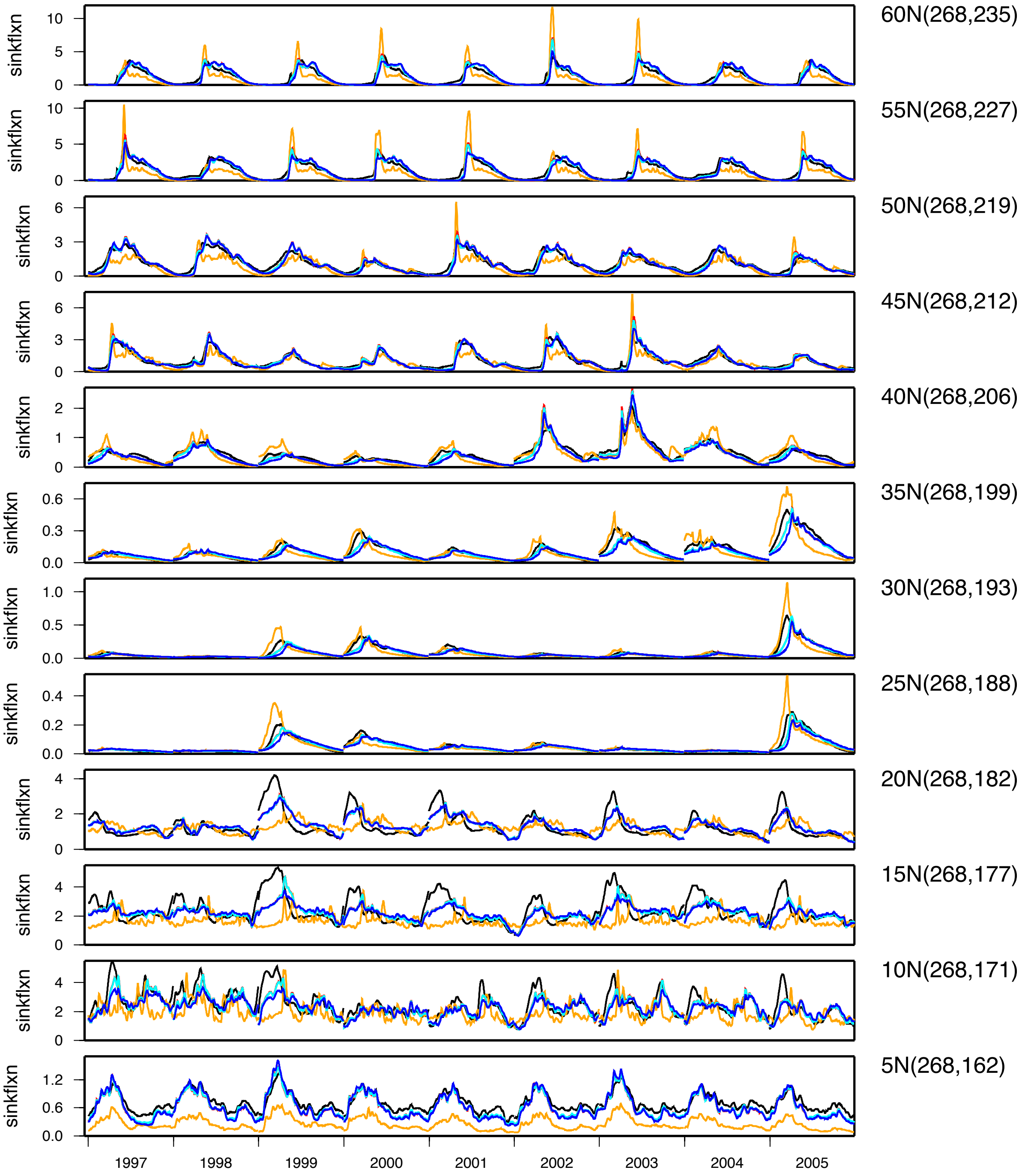

Figure 3: PON flux at 100 m (mmol N m^{-2} d^{-1}) from second year of 2 year simulations: MEDUSA (black); HadOCC with prior parameters (orange), OPT13CC-NPZ parameters (red), OPT13CC-NPZF parameters (cyan), OPT10CC-NPZF parameters (blue). Plots are for each calibration site on the 20W transect. The ORCA1 grid reference is shown in brackets.

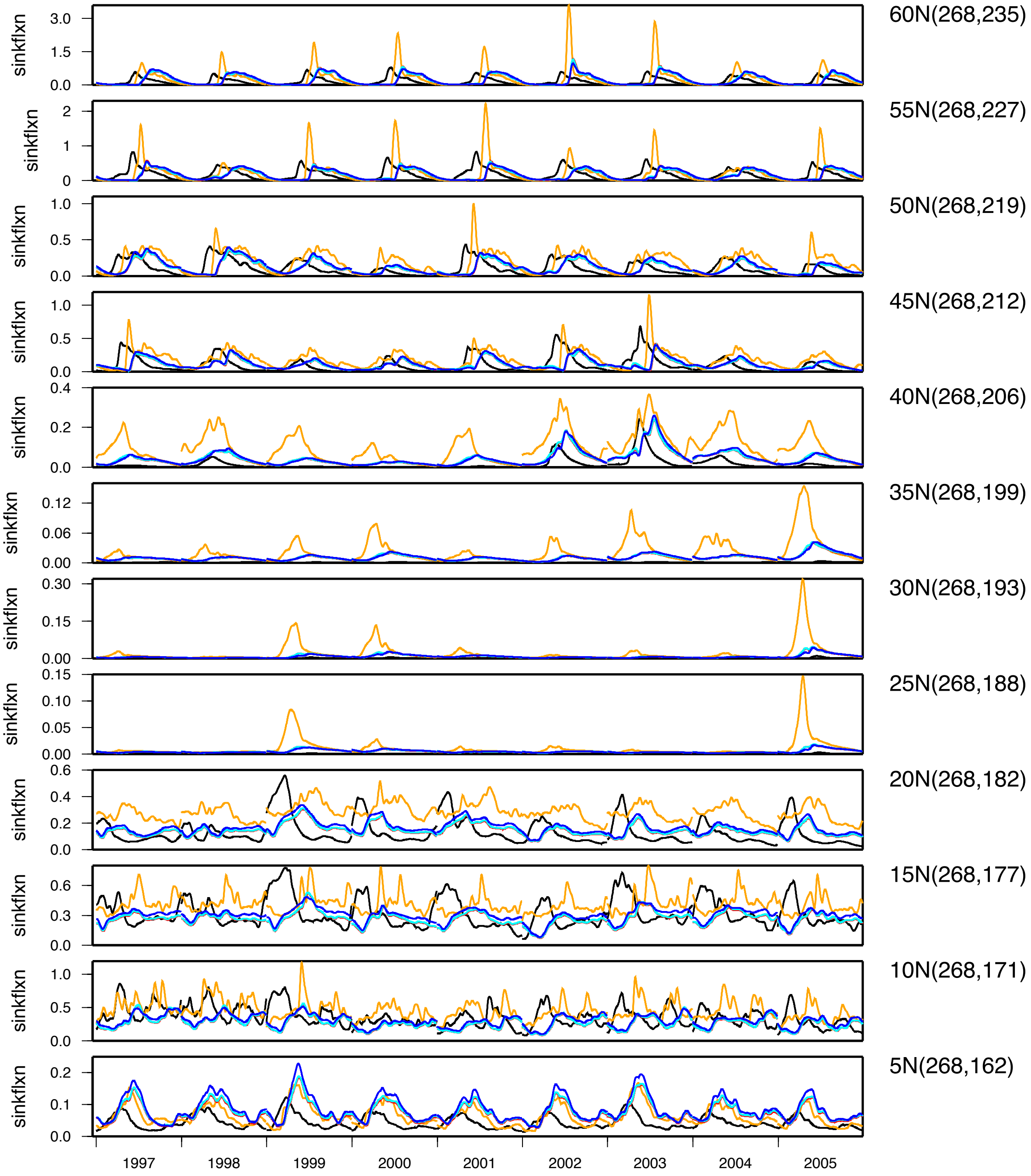

Figure 4: PON flux at 531 m (mmol N m^{-2} d^{-1}) from second year of 2 year simulations: MEDUSA (black); HadOCC with prior parameters (orange), OPT13CC-NPZ parameters (red), OPT13CC-NPZF parameters (cyan), OPT10CC-NPZF parameters (blue). Plots are for each calibration site on the 20W transect. The ORCA1 grid reference is shown in brackets.

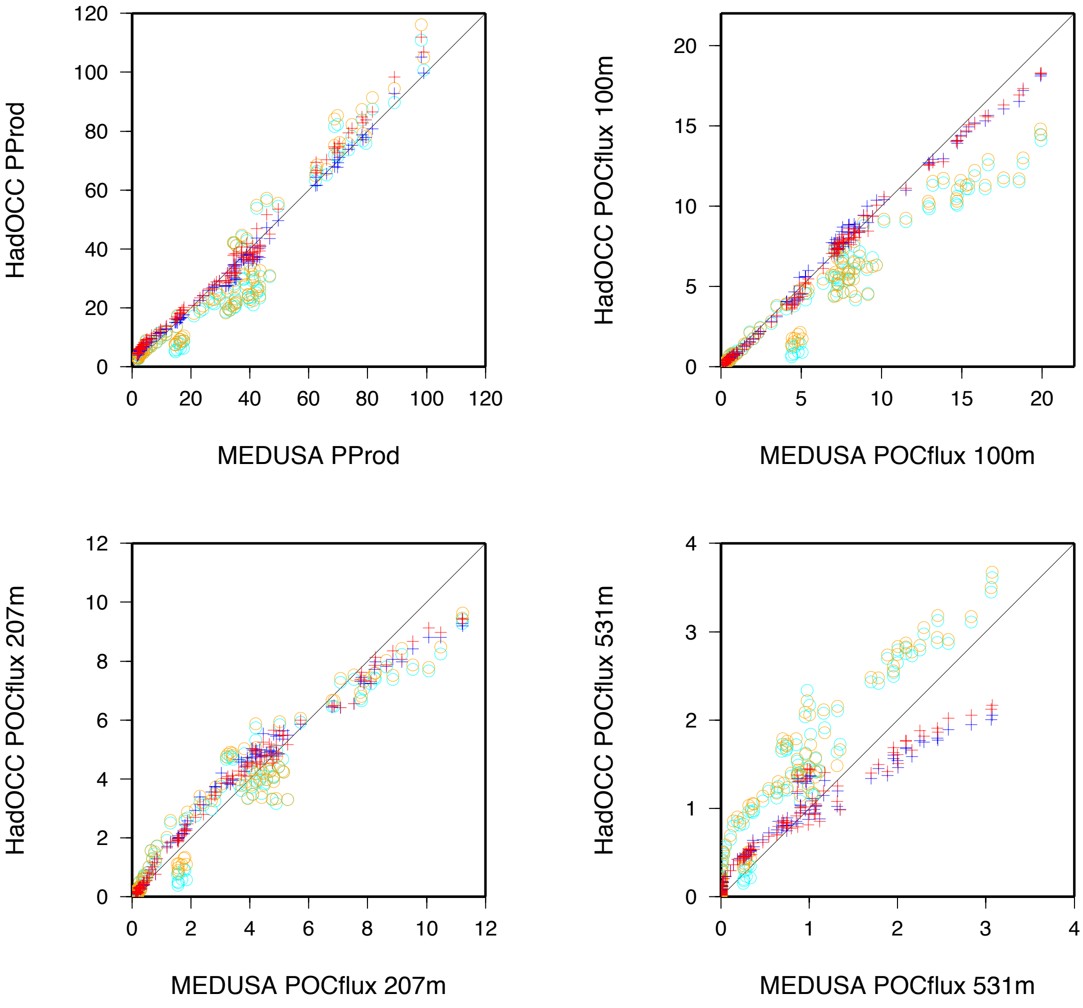

Figure 5: Comparison between 1997-2005 HadOCC and MEDUSA output for primary production and POC flux at 100 m, 207 m and 531 m (mmol C m^{-2} d^{-1}) for parameter sets: PRIOR (cyan circle), PRIORCC (orange circle), OPT13-NPZ (blue '+'), OPT13CC-NPZ (red '+'). Data are from second year of 2 year simulations at the calibration sites.

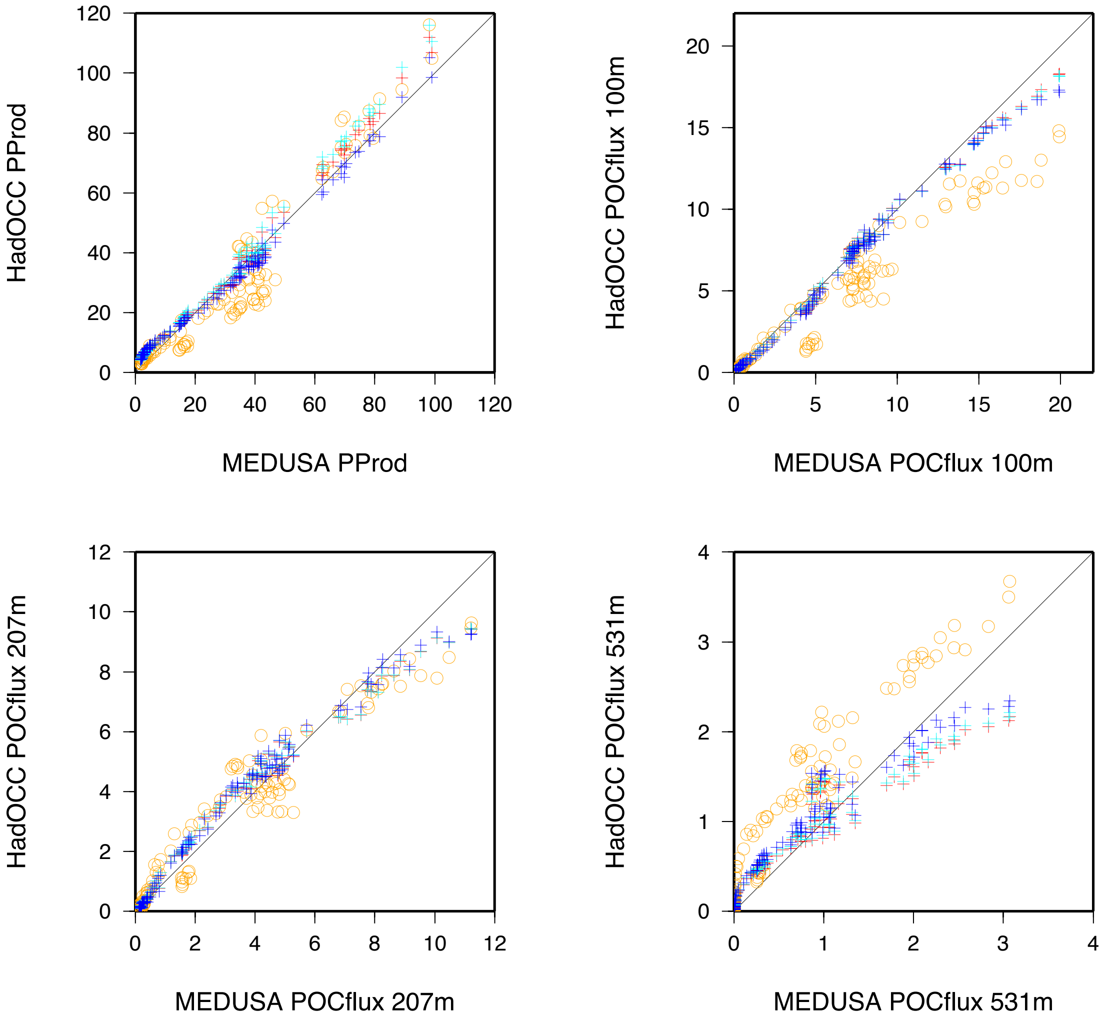

Figure 6: Comparison between 1997-2005 HadOCC and MEDUSA output for primary production and POC flux at 100 m, 207 m and 531 m (mmol C m^{-2} d^{-1}) for parameter sets: PRIORCC (orange circle), OPT13CC-NPZ (red '+'), OPT13CC-NPZF (cyan '+'), OPT10CC-NPZF (blue '+'). Data are from second year of 2 year simulations at the calibration sites.

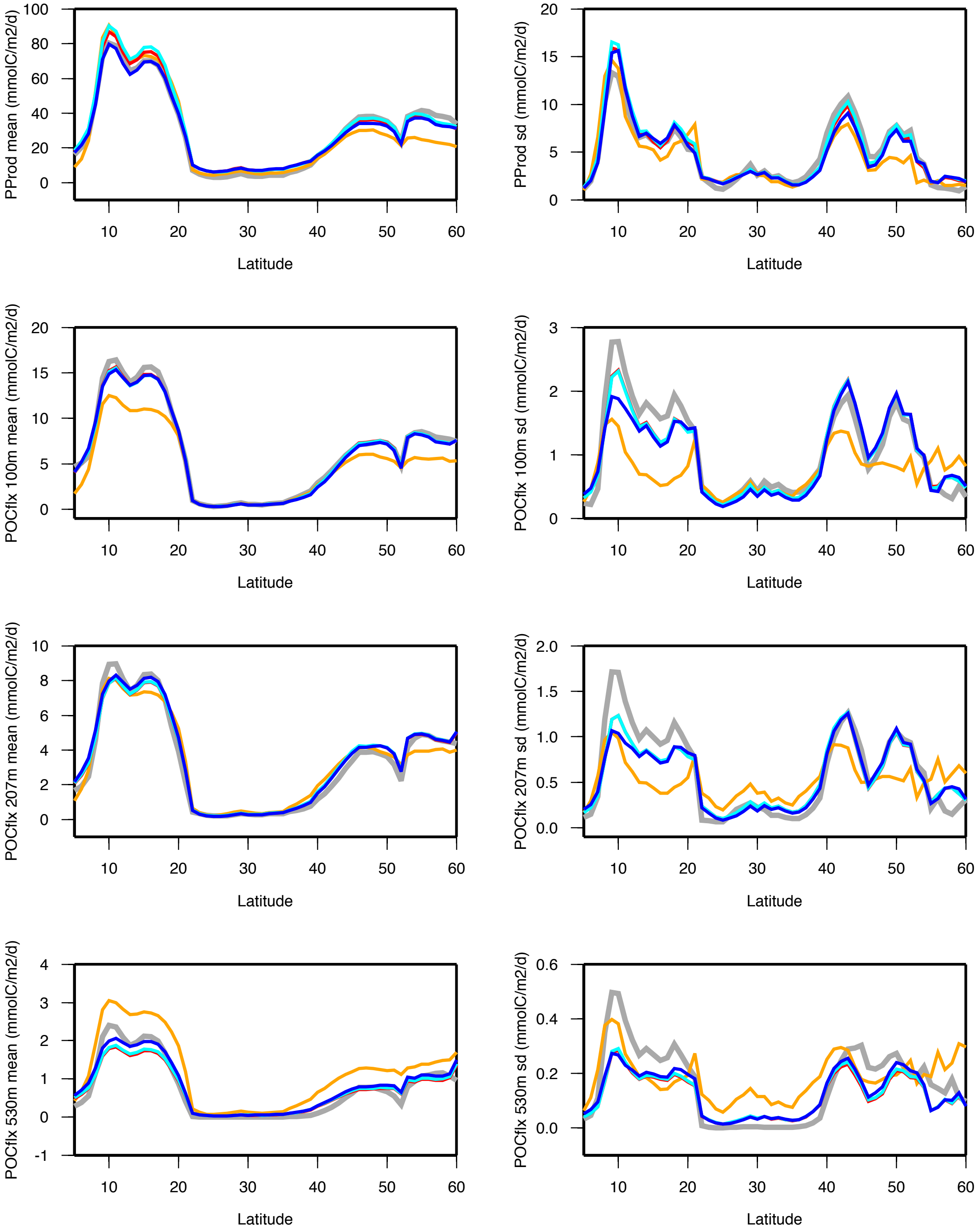

Figure 7: 9 year mean and standard deviation (1997-2005) of annually averaged primary production and POC flux at 100 m, 207 m and 531 m on 20W meridion for parameter sets: PRIORCC (orange), OPT13CC-NPZ (red), OPT13CC-NPZF (cyan), OPT10CC-NPZF (blue). Data are from second year of 2 year simulations at 1 degree intervals.

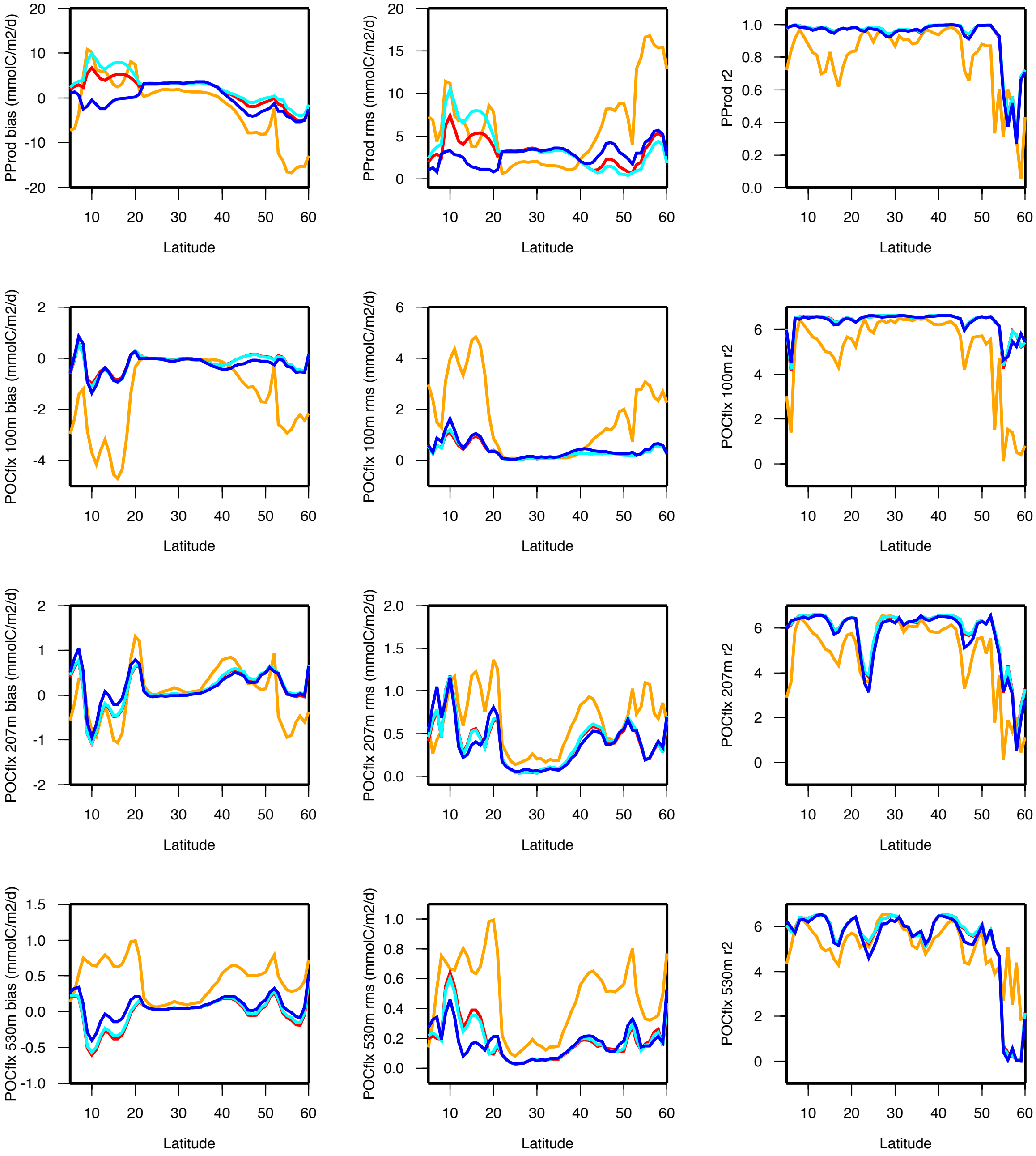

Figure 8: 9 year bias, r.m.s. error and R^2 (1997-2005) for annually averaged primary production and POC flux at 100 m, 207 m and 531 m on 20W meridion for parameter sets: PRIORCC (orange), OPT13CC-NPZ (red), OPT13CC-NPZF (cyan), OPT10CC-NPZF (blue). Data are from second year of 2 year simulations at 1 degree intervals.

the difference between the fixed and variable C:Chl versions of the model. Figure 6 presents the results for the variable C:Chl version under the different experimental constraints.

The overall performance of the calibrated emulator appears good, with only minor differences between the different parameter optimization experiments. However, some deficiencies remain after calibration. There is a persistent over-estimation of the 531 m flux at lower flux values. For all the posterior parameter sets there is a contrasting negative bias for higher values despite the existence of a strong positive bias in both versions of the uncalibrated model. Using the MEDUSA parameters controlling the fate of grazed material (OPT10CC-NPZF) improves the situation for the higher values (above about 1.5 mmol C m^{-2} d^{-1}) without having much effect on the lower values, suggesting that further improvement might be possible with a more exhaustive exploration of parameter space. Note that the higher overall bias in Table 4 for Experiment OPT10CC-NPZF is actually the result of reduction in the negative bias, rather than the indication of poorer performance, highlighting the risks of any reliance on summary statistics.

Figures 7 and 8 show the performance of HadOCC, with variable C:Chl, against MEDUSA in a range of simulations at 1° intervals along the 20W transect from 5-60N. The data analyzed are annual means for each year from 1997-2005 (taking the second year of a set of 2 year simulations as before). Even without calibration, HadOCC does a reasonable job at reproducing the dominant meridional patterns of primary production and POC flux, although the inter-annual variability of the POC flux is less well represented that the primary production (as shown by the divergence in the standard deviations and the lower R^2 values). While bias in POC tends to be relatively small at 207 m, at the more productive latitudes it tends to be strongly negative at 100 m and strongly positive at 531 m before calibration. (The negative bias appears linked to a similar bias in production at high latitudes but not in the low latitude high productivity region.) The change with depth indicates a tendency for lesser attenuation in POC flux from grazing and remineralization in HadOCC than in MEDUSA between 100 and 531 m. The biases are largely corrected for by calibration, as is the negative bias in production.

The performance of the calibrated model over the transect is generally good with only small biasses remaining in primary production and sinking particle flux. However, these biases tend to have large spatial extents, persisting across regions of similar climatic conditions so they cannot be ignored, even though they are greatly improved as a result of calibration. There are some notable differences in primary production biasses between simulations with different posterior parameter sets. In particular, the application of the MEDUSA parameter values affecting the partitioning of grazed material in Experiment OPT10CC-NPZF tends to reduce primary production, leading to a reduction in bias at low latitudes and an increase in negative bias at high latitudes. However, it has the opposite effect on the 531 m particle flux, leading to a reduction in the negative bias present over most of the regions affected.

The emulator's ability to simulate the inter-annual variability is an important determinant of its ability to respond correctly in a climate change scenario. Examination of the standard deviations (Figure 7) and R^2 values (Figure 8) shows that the calibrated model generally performs well with respect to both the magnitude and patterns of variability. The uncalibrated model also represents the inter-annual variability in primary production fairly well but does a poor job at capturing the spatial pattern in the scale of the sinking flux variability between years. This deficiency is largely corrected by calibration. However, HadOCC cannot reproduce the magnitude of inter-annual variability in the particle flux in the most productive waters at low latitudes where the transect passes close to the North African coastal region. This is true at all 3 depth levels. The flux variability at 530 m is also underestimated in the high productivity region above 45N, although in the northernmost part of this region (55-60N) it tends to be over-estimated in the shallower levels, probably as a result of high inter-annual variability in production relative to MEDUSA in this area. The R^2 values are generally much lower here for both production and particle flux, indicating poor representation of the inter-annual variability by the surrogate model. At 531 m they are actually lower for the calibrated model.

5 Discussion

The feasibility of using a simple biogeochemistry model as an emulator to reproduce the behaviour of a more complex model has been demonstrated in this study, showing how a relatively fast model suited to long ESM integrations might be linked to a more detailed model. Good performance of such a surrogate model can be achieved by calibration, using output from the more detailed model to constrain parameter values. The specific results for the HadOCC model should be seen as provisional. The emulation performance would need to be assessed over a wider range of environmental conditions to assess traceability of performance in the context of global simulations. Also, it must be recognised that the experimental framework is designed to constrain relatively short time-scale responses to physical drivers, in particular the seasonal response of the system, we cannot rule out the possibility of interactions with the ocean circulation that would compromise performance in longer simulations. Further tests would be needed, within a 3-D NEMO simulation or ideally the fully coupled ESM, to fully determine suitability.

Some deficiencies were noted in the performance of the calibrated model. Probably the most important for Earth system modelling applications are (i) the distribution of biases in primary production and sinking flux which were consistent over large geographic regions, apparently correlated with regional climatic conditions, and (ii) the underestimation of inter-annual variability. Nevertheless, these anomalies are relatively minor and the associated uncertainty can be quantified and its impact on predictions assessed. The significance of this source of error should ultimately be judged relative to the performance of the more complex model against real-world data.

5.1 Interpretation of Parameter Values

Another observation was that calibration led to parameter values that in some cases deviated strongly from accepted values. This is not unusual in the calibration of biogeochemistry models if strong prior constraints on parameters are omitted, irrespective of whether the data constraints are from another model or from observations of nature, but is it a problem?

Brynjarsdóttir (2013, submitted) differentiate between physical parameters and tuning parameters. Physical parameter are those which have meaning within the science underlying the mechanistic model, while tuning parameters do not. They are typically used to approximate some more complex un-modelled process. Their 'true' values are then interpreted as the values which give the best fit to reality. They can play an important role in interpolation for predicting system responses to observable conditions but, unlike physical parameters, they do not help us to extrapolate to new conditions for which we have no observations. In ecosystem modelling, most parameters are somewhere in-between. They do not correspond to values that are constant in nature. The equivalent values in nature are usually highly variable in space and time and across different taxanomic groups, but most are expected to have fairly well-defined ranges determined by what is biologically possible. Outside of these ranges their meaning becomes lost and they can no longer be considered as physical parameters. They are then of limited use in extrapolation.

While the ability to extrapolate is desirable for ESM predictions, the lack of well-defined physical parameters in biogeochemistry models will inevitably make reliable extrapolation elusive. However, given the wide range of climatic conditions across the present day Earth system from which we can sample, much can be achieved by interpolation using a model with well-calibrated tuning parameters particularly for medium-term predictions (years to decades) where we might not expect to see widespread environmental conditions that have no present-day analogue. With this in mind, it can be argued that it is preferable for posterior parameters to have physically meaningful values but not essential. Form a pragmatic viewpoint, different values should be tolerated if they are shown to empirically improve predictive skill. However, if out-of-range value are necessary, a deficiency in the model design is indicated and correcting this deficiency is likely to lead to more reliable predictions in the long term.

5.2 Model Design Considerations

Parameter adjustment is one way to improve traceable performance between models. Another way is to improve direct traceability between the designs in terms of structure and process formulation. Structural traceability between the two models in this study is good, there being a direct correspondence between the aggregated phytoplankton and zooplankton compartments in MEDUSA and the the two plankton compartments in HadOCC. However, direct traceability in process representation is severely compromised because of differences in process formulations. This limits the number of parameters that can be usefully compared between the two models.

One of the aims of i-MarNet is to develop a new community model supporting different levels of biogeochemical complexity as options within a traceable hierarchy. This should provide an opportunity to develop common formulations wherever possible. In some cases, where there is significant uncertainty regarding the best theoretical formulation, more than one formulation could be included but should be made available to all model structures to allow proper inter-comparison. This approach should reduce the number of parameters that require adjustment if we set parameters in the simple surrogate model to their equivalents in the more complex model. It cannot necessarily be assumed that this approach will give the best performing emulator but the limited application of the method in this pilot study did not appear to have significant adverse effects. Importantly, it should make it much easier to understand the inter-model relationships. Furthermore, any reduction in the number of parameters to be considered leads to major computational savings and should result in better constrained posterior values as the potential for parameter interactions is reduced.

5.3 Handling Parametric Uncertainty

The results presented here emphasize the extent to which model behaviour can be modified by adjusting parameters with reference to appropriate data constraints, either from another model or from observations of nature. However, they also highlight some of the issues that must be addressed if we are to be able to use model's adjustable parameters effectively. The most obvious of these is the demands that investigation of large multi-dimensional parameter spaces place on our computer resources. The 5 parameter optimization experiments in the present study involved a total of approximately 139 000 simulations. Despite this, coverage of the parameter space is generally poor considering that for a 17 parameter space 2^{17} or approximately 130 000 simulations would be required to examine all parameter interactions on the basis of just one low and one high value. There is no way of knowing whether we have found a global minimum in the cost function and even if there were it is not clear how significant the global minimum would be without a more systematic exploration of the shape of the cost function in parameter space.

Modern Bayesian calibration methods, following Kennedy and O'Hagan (2001), provide a more comprehensive statistical framework for addressing issues of parametric uncertainty as well as uncertainty from other sources. These allow joint posterior distributions for model parameters and the model discrepancy term to be estimated at the expense of increasing computational requirements still further in terms of the number of simulations. Nevertheless, they have been taken up widely in many fields including modelling of aquatic ecosystems (e.g. Arhonditsis, 2008; McDonald et al., 2012).

The computational demands of investigating parameter space are in direct opposition to those for realistic 3-D simulations which require high resolution and long spin-up times, particularly for the carbon cycle. A practical solution is to improve the realism of our 1-D modelling capabilities. This has been the driving force behind the development of MarMOT. Fast statistical emulators of model outputs or cost functions can be used for more comprehensive exploration of parameter space (O'Hagan, 2006). Such emulators have been constructed for HadOCC outputs using MarMOT ensemble simulations to provide the required training data (Oxlade, 2012). A useful resource introducing Bayesian calibration, statistical emulators and other state-of-the-art methods used to analyze the uncertainty in outputs produced by models of complex processes has been created under the RCUK project "Managing Uncertainty in Complex Models". This is the MUCM Toolkit at http://mucm.ac.uk/.

5.4 Establishing a Traceable Link to Reality

Here we have examined the idealized problem of how to establish traceability between one model and another. Ultimately our aim is to establish traceability between a hierarchy of models and the Earth system. Inadequacy of the global observing system and the consequent uncertainties make this a much more challenging problem. Capabilities are needed to assess both the fidelity of model response to its physical drivers and its performance when coupled to an ocean, climate or Earth system model. Assessing the model response to physical drivers against biogeochemical observations is particularly difficult because of the sensitivity of biogeochemical processes to highly uncertain physical variables, in particular those controlling access to light and nutrients required for primary production.

Developing a capability for assessing the model representation of biogeochemistry separately from that of ocean physics will require a synthesis of a diverse set of biogeochemical and physical observations from many different locations. The most useful of these are time-series data. Both Eulerian time-series from fixed observatories and Lagrangian time-series from Argo/Bio-Argo floats can be used, in conjunction with satellite data. The uncertain physical environment would ideally be represented by the best statistical description that can be achieved using the observations available in combination with high resolution physical simulations. Hemmings and Challenor (2012) describe how this type of information can be used in the MarMOT system to drive ensembles of water column simulations for model assessment and calibration in the presence of environmental uncertainty. Using 1-D simulations, with a realistic 3-D context provided by analysis of auxiliary data allows computational effort to be focused on data-rich sites or float tracks to address environmental and parametric uncertainties. A global testbed capability constructed in this way would provide a skill assessment facility for model representations of upper ocean biogeochemistry. Posterior parameter distributions obtained would serve as prior distributions to be refined by a limited number of 3-D experiments with reference to subsurface nutrient distributions before application in the ESM.

Acknowledgements The author would like to thank Tom Anderson, Andrew Yool and Katya Popova for their advice on this work. The HadOCC model code has been made available by the Met Office in support of collaborative work under the Joint Weather and Climate Research Programme and thanks are also due to Ian Totterdell and Rosa Barciela for help with the model configuration. The present work has been funded by the Natural Environment Research Council via the i-MarNet project (Integrated Marine Biogeochemical Modelling Network to Support UK Earth System Research).

A HadOCC Model Description

The HadOCC model described here is a modified version of the model of Palmer and Totterdell (2001) incorporating subsequent developments to the Met Office version and some minor modifications introduced specifically for the present study. The nitrogen tracers are phytoplankton P, zooplankton Z, detritus D and dissolved inorganic nitrogen N. The main differences from the original version are (i) the application of nitrogen limitation to the photosynthesis-PAR curve maximum, rather than as a scaling factor for the whole curve, reducing its effect at low light levels, (ii) introduction of a variable carbon:chlorophyll ratio and (iii) changes to the pathways of material originating from grazing and mortality (Totterdell, personal communication, 2005).

For the present study, MEDUSA-like temperature dependency was introduced in phytoplankton growth and remineralization rates. In the case of remineralization, for which the standard HadOCC version uses a constant depth-varying profile, the dependency on temperature replaces depth dependency. MEDUSA and HadOCC use different light-transmission and photosynthesis sub-models in their respective standard configurations. These were replaced by a common formulation for the purposes of the traceability experiments. The two models also differ in their handling of stoichiometry. These differences are removed here by using uniform Redfield carbon:nitrogen ratios of 6.625 for all organic components in both models, as in the reference MEDUSA simulation (Yool et al., 2011).

Process parameterizations and source-minus-sink terms are defined below. Refer to Table 5 for parameter values.

Phytoplankton Growth

The photosynthesis sub-model gives the level mean biomass-specific growth rate $\bar{J}$ as the depth integral over each model level of the photosynthesis-PAR response curve

$$J = \frac{V_{\mathrm{P}} \alpha E_{\mathrm{d}}}{\sqrt{V_{\mathrm{P}}^2 + (\alpha E_{\mathrm{d}})^2}} \tag{14}$$

where α is the low-light response, dependent on the carbon:chlorophyll ratio θ_{chl}:

$$\alpha = \frac{\alpha_{\mathrm{chl}}}{\theta_{\mathrm{chl}}}. \tag{15}$$

The maximum photosynthetic rate V_{P} is given by the product of the base growth rate at 0° C, an exponential function of temperature T and a nitrogen limitation factor Q_{N}:

$$V_{\mathrm{P}} = (V_0 \cdot 1.066^T) Q_{\mathrm{N}} \tag{16}$$

where

$$Q_{\mathrm{N}} = \frac{N}{N + k_{\mathrm{N}}}. \tag{17}$$

Downwelling PAR is determined by the light transmission sub-model

$$E_{\mathrm{d}}(z) = E_{\mathrm{d}}(0) \exp(-(k_{\mathrm{water}} + k_{\mathrm{pig}} \cdot 1.25\,Chl)z) \tag{18}$$

A ratio of chlorophyll to total pigment concentration of 0.8 is assumed and $E_{\mathrm{d}}(0)$ is taken to be 43 % of total downwelling solar radiation at the sea surface.

Zooplankton Grazing

Phytoplankton and detritus losses due to herbivorous zooplankton activity are $G_{\mathrm{P}} = hP$ and $G_{\mathrm{D}} = hD$ respectively, where h is the grazing rate per unit food concentration:

$$h = \frac{Z}{F_{\mathrm{tot}}} g_{\mathrm{max}} \frac{F^2}{F^2 + k_{\mathrm{F}}^2}; \tag{19}$$

$F = \max(0, F_{\mathrm{tot}} - F_{\mathrm{threshold}})$, where $F_{\mathrm{tot}} = P + D$ and $F_{\mathrm{threshold}} = 0.01\,\mathrm{mmol\,N\,m^{-3}}$.

$$k_{\mathrm{F}}^2 = \frac{g_{\mathrm{max}}}{\epsilon_{\mathrm{F}}} \tag{20}$$

Phytoplankton Mortality

$M_{\mathrm{P}} = mP^2$; $m = 0$ for $P <= 0.01$ mmol N m^{-3}, otherwise $m = m_{\mathrm{o}}$.

Zooplankton Mortality

$M_{\mathrm{Z}} = m_1 Z + m_2 Z^2$.

Detrital Remineralization

$\lambda = \lambda_0 \cdot 1.066^T D$

Nitrogen Equations

$$\begin{aligned}
\mathrm{SMS_P} &= \bar{J}P - M_\mathrm{P} - \eta P - G_\mathrm{P} && (21)\\
\mathrm{SMS_Z} &= \phi_\mathrm{I}(\beta_\mathrm{P} G_\mathrm{P} + \beta_\mathrm{D} G_\mathrm{D}) - M_\mathrm{Z} && (22)\\
\mathrm{SMS_D} &= (1-\phi_\mathrm{MPN})M_\mathrm{P} + (1-\phi_\mathrm{MZN})M_\mathrm{Z} \\
&\quad + a_\mathrm{PD} G_\mathrm{P} + (a_\mathrm{DD} - 1)G_\mathrm{D} - \lambda D && (23)\\
\mathrm{SMS_N} &= \phi_\mathrm{MPN} M_\mathrm{P} + \eta P + \phi_\mathrm{MZN} M_\mathrm{Z} \\
&\quad + \phi_\mathrm{mfN}(1-\phi_\mathrm{I})(G_\mathrm{P} + G_\mathrm{D}) + \lambda D - \bar{J}P && (24)
\end{aligned}$$

where $a_\mathrm{PD} = (1-\phi_\mathrm{mfN})(1-\phi_\mathrm{I}) + (1-\beta_\mathrm{P})\phi_\mathrm{I}$ and $a_\mathrm{DD} = (1-\phi_\mathrm{mfN})(1-\phi_\mathrm{I}) + (1-\beta_\mathrm{D})\phi_\mathrm{I}$.

The active vertical velocity of detritus relative to the water is equal to the sinking velocity parameter w_D. It is zero for all other tracers.

Phytoplankton Carbon:Chlorophyll ratio

For simulations with fixed C:Chl ratio $\theta_\mathrm{chl} = 40$ gC (gChl)$^{-1}$. In variable C:Chl ratio simulations

$$\begin{aligned}
\theta_\mathrm{chl} &= \min\left(\sqrt{\theta_\mathrm{min}\frac{\alpha_\mathrm{chl} E_\mathrm{d}}{J_\mathrm{cc}(\theta_\mathrm{chl})}}, \theta_\mathrm{max}\right) && (25)\\
J_\mathrm{cc}(\theta_\mathrm{chl}) &= V_\mathrm{P}\left[1 - \exp\left(-\frac{\alpha_\mathrm{chl} E_\mathrm{d}}{\theta_\mathrm{chl} V_\mathrm{P}}\right)\right] && (26)
\end{aligned}$$

where θ_min and θ_max are the minimum and maximum C:Chl ratios and J_cc is the carbon-specific photosynthetic rate. J and J_cc are approximately equivalent representations of the same P-E curve. θ_min is 20 gC (gChl)$^{-1}$ and θ_max is 2000 gC

Table 5: HadOCC model parameters.

Parameter	Symbol	Prior value
Maximum photosynthetic rate at 0° C	V_0	0.53 d^{-1}
Initial slope of photosynthesis-PAR curve	α_{chl}	5.56 mg C (mg Chl)$^{-1}$ (E m^{-2})$^{-1}$
Half-saturation conc. for DIN uptake	k_{N}	0.1 mmol N m^{-3}
Phytoplankton metabolic loss rate	η	0.05 d^{-1}
Phytoplankton density-dependent mortality	m_{o}	0.05 d^{-1}(mmol N m^{-3})$^{-1}$
Fraction of phytoplankton mortality to DIN	ϕ_{MPN}	0.01
Maximum grazing rate	$g_{\max}$	0.8 d^{-1}
Prey capture rate	ϵ_{F}	3.2 d^{-1} (mmol N m^{-3})$^{-2}$
Fraction of grazed material ingested	ϕ_{I}	0.77
Assimilation efficiency for phytoplankton	β_{P}	0.9
Assimilation efficiency for detritus	β_{D}	0.65
Fraction of messy feeding to DIN	ϕ_{mfN}	0.1
Zooplankton linear mortality	m_1	0.05 d^{-1}
Zooplankton density-dependent mortality	m_2	0.3 d^{-1}(mmol N m^{-3})$^{-1}$
Fraction of zooplankton mortality to DIN	ϕ_{MZN}	0.67
Detrital sinking velocity	w_{D}	10 m d^{-1}
Detrital remineralization rate at 0° C	λ_0	0.016 d^{-1}

(gChl)$^{-1}$. The upper limit is raised from the standard HadOCC value of 200 gC (gChl)$^{-1}$ to improve compatibility with MEDUSA which does not impose a maximum.

The model is based on the steady state solution of the Geider *et al.* (1997) photoacclimation model, describing the light dependency of the C:Chl ratio under balanced growth conditions. Such conditions are rarely achieved in the upper boundary layer because of the interaction between acclimation and vertical mixing. Conceptually, the upper boundary layer defined by the variable is treated as fully-mixed. So, for levels wholly within the boundary layer

$$\theta_{\mathrm{chl}} = \min\left(\sqrt{\theta_{\min}\frac{\alpha_{\mathrm{chl}}E_{\mathrm{d}}}{J_{\mathrm{cc}}(\langle\theta_{\mathrm{chl}}\rangle_{1..\mathrm{k_{mld}}})}}, \theta_{\max}\right) \tag{27}$$

where $\langle . \rangle_{1..\mathrm{k_{mld}}}$ indicates averaging over the levels above the turbocline depth.

B Cost Function Evaluations and Repeatability

Figures 9 - 13 show trial cost function evaluations for each experiment on parameter axes. Each plot is a window on the multi-dimensional parameter space showing cost function values calculated by both the μGA and Powell algorithms during the optimization process. The final results for each of 10 different optimizer initializations are shown to give an indication of repeatability.

C Emulation Results for HadOCC State Variables

Figures 14 - 18 show the annual cycles for the model state variables in the surface level for the years 1997-2005.

References

Acton, F. S.: Minimum methods, in: Numerical Methods That Work, Fourth printing, Mathematical Association of America, Washington DC, 448–476, 1990.

Arhonditsis, G.B., Papantou, D., Zhang, W., Perhar, G., Massos, E., Shi, M., 2008. Bayesian calibration of mechanistic aquatic biogeochemical models and benefits for environmental management. J. Mar. Syst. 73, 830.

Brent, R. P.: An algorithm with guaranteed convergence for finding the minimum of a function of one variable, in: Algorithms for Minimization without Derivatives, Prentice-Hall, Englewood Cliffs, N. J., 61–80, 1973.

Brynjarsdóttir, J. and O'Hagan, A. submitted. Learning about physical parameters: the importance of model discrepancy. J. Uncertainty Quantification.

Carroll, D. L.: Chemical laser modelling with genetic algorithms, Amer. Inst. Aero. Astro., 34, 338–346, 1996.

Dadou, I., Evans, G. and Garçon, V.: Using JGOFS in situ and ocean color data to compare biogeochemical models and estimate their parameters in the subtropical North Atlantic Ocean, J. Mar. Res., 62, 565–594, 2004.

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

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., 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

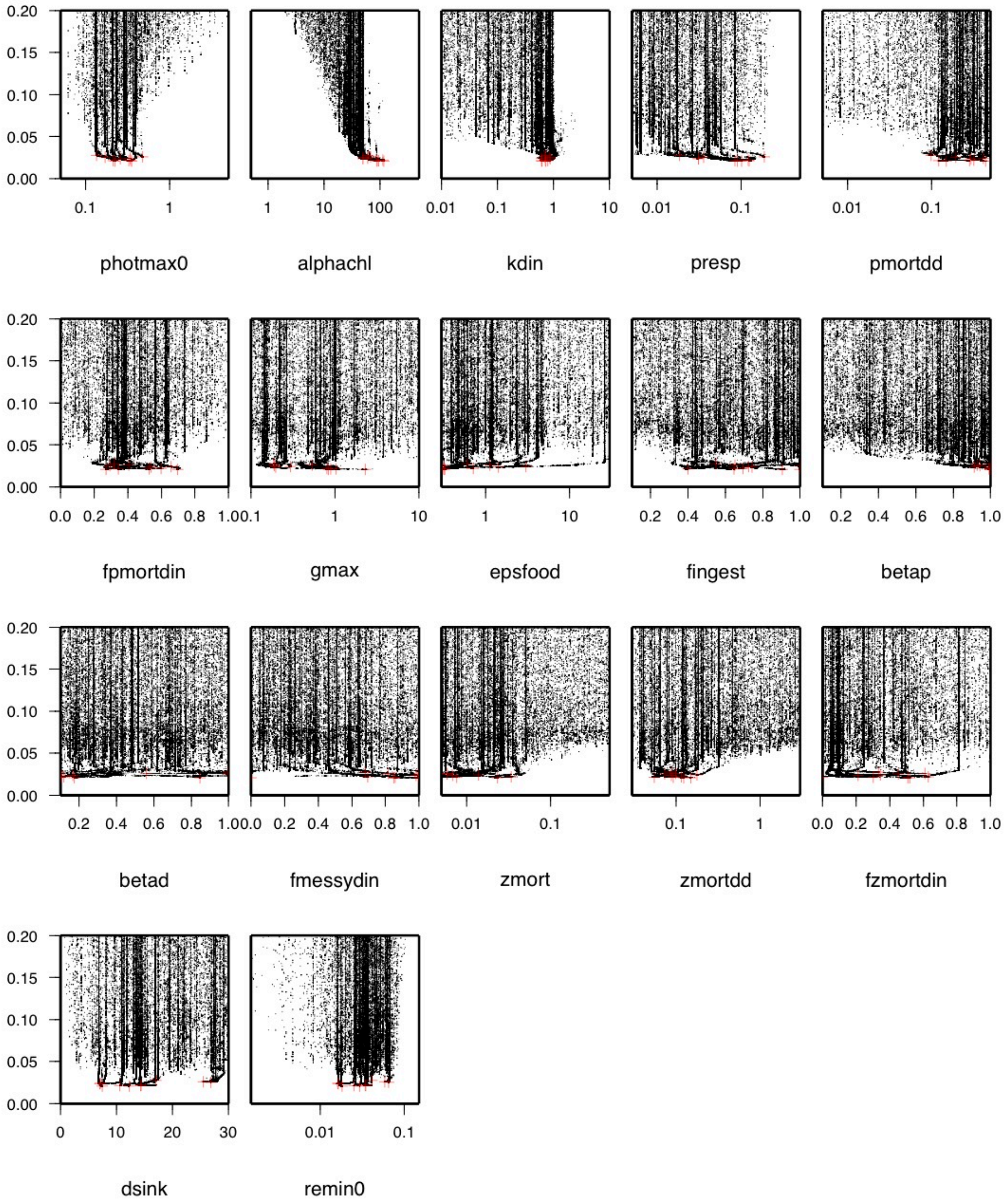

Figure 9: Cost function evaluations for Experiment OPT17-NPZ. Final results for 10 different optimizer initializations are shown by red crosses. See Table 5 for parameter units.

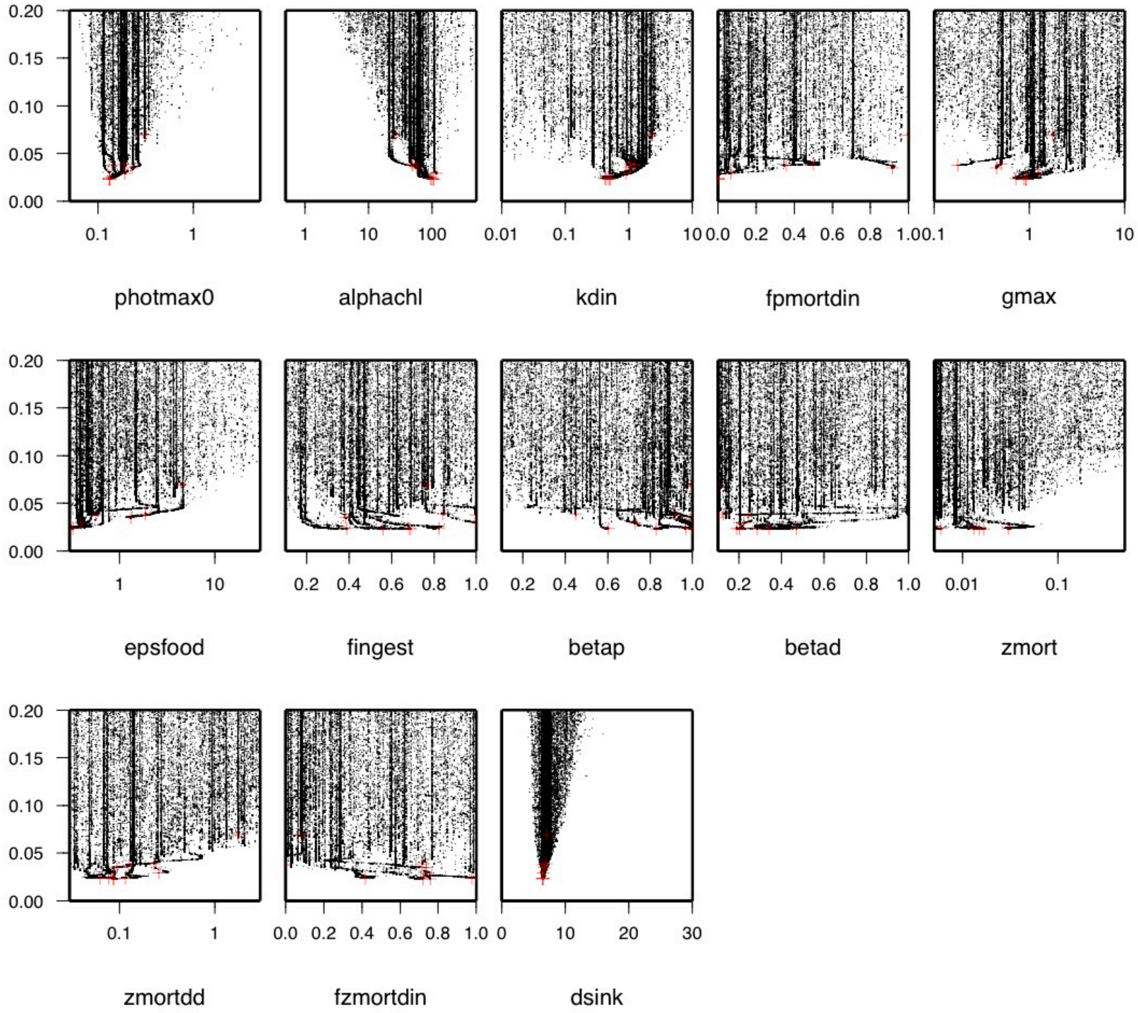

Figure 10: Cost function evaluations for Experiment OPT13-NPZ. Final results for 10 different optimizer initializations are shown by red crosses. See Table 5 for parameter units.

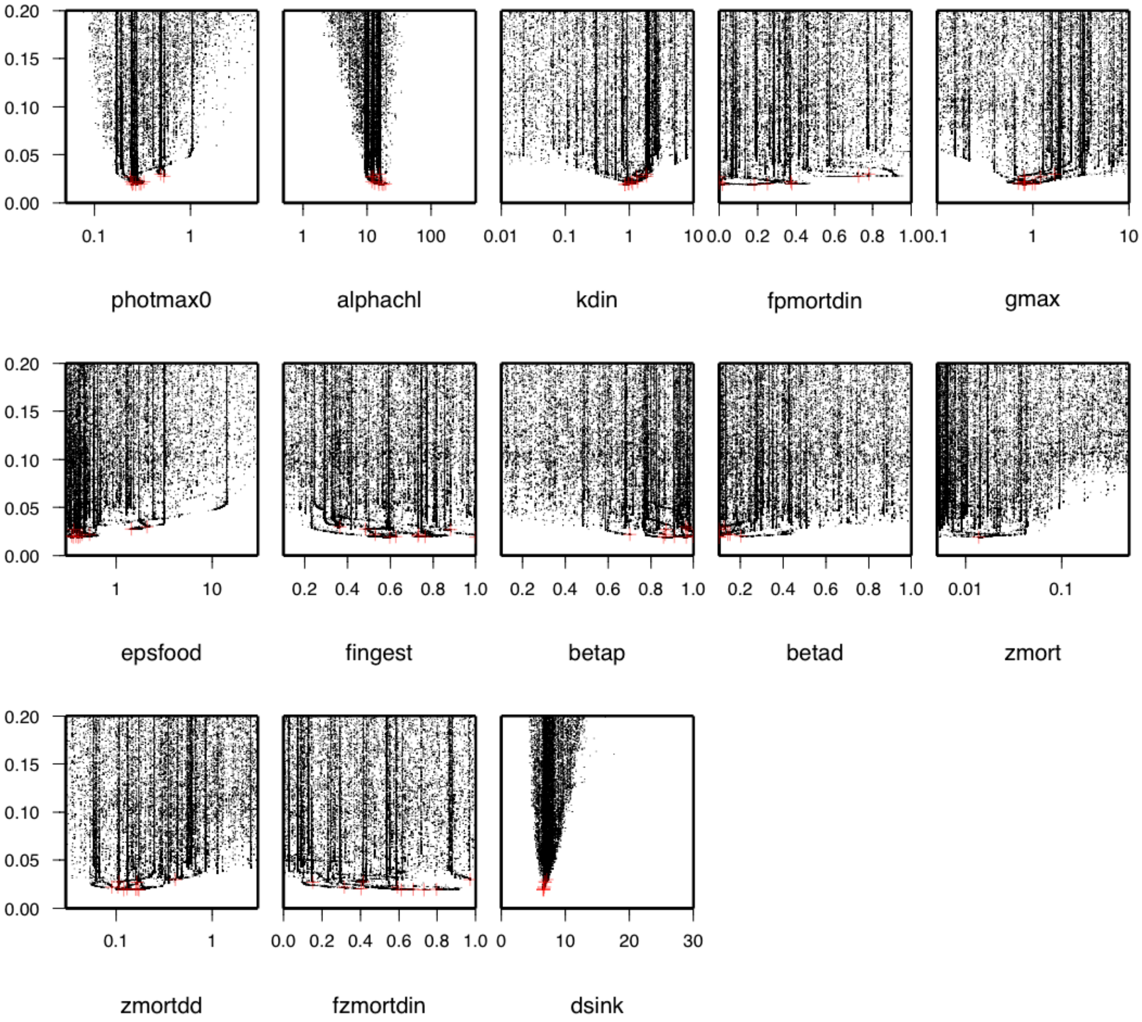

Figure 11: Cost function evaluations for Experiment OPT13CC-NPZ. Final results for 10 different optimizer initializations are shown by red crosses. See Table 5 for parameter units.

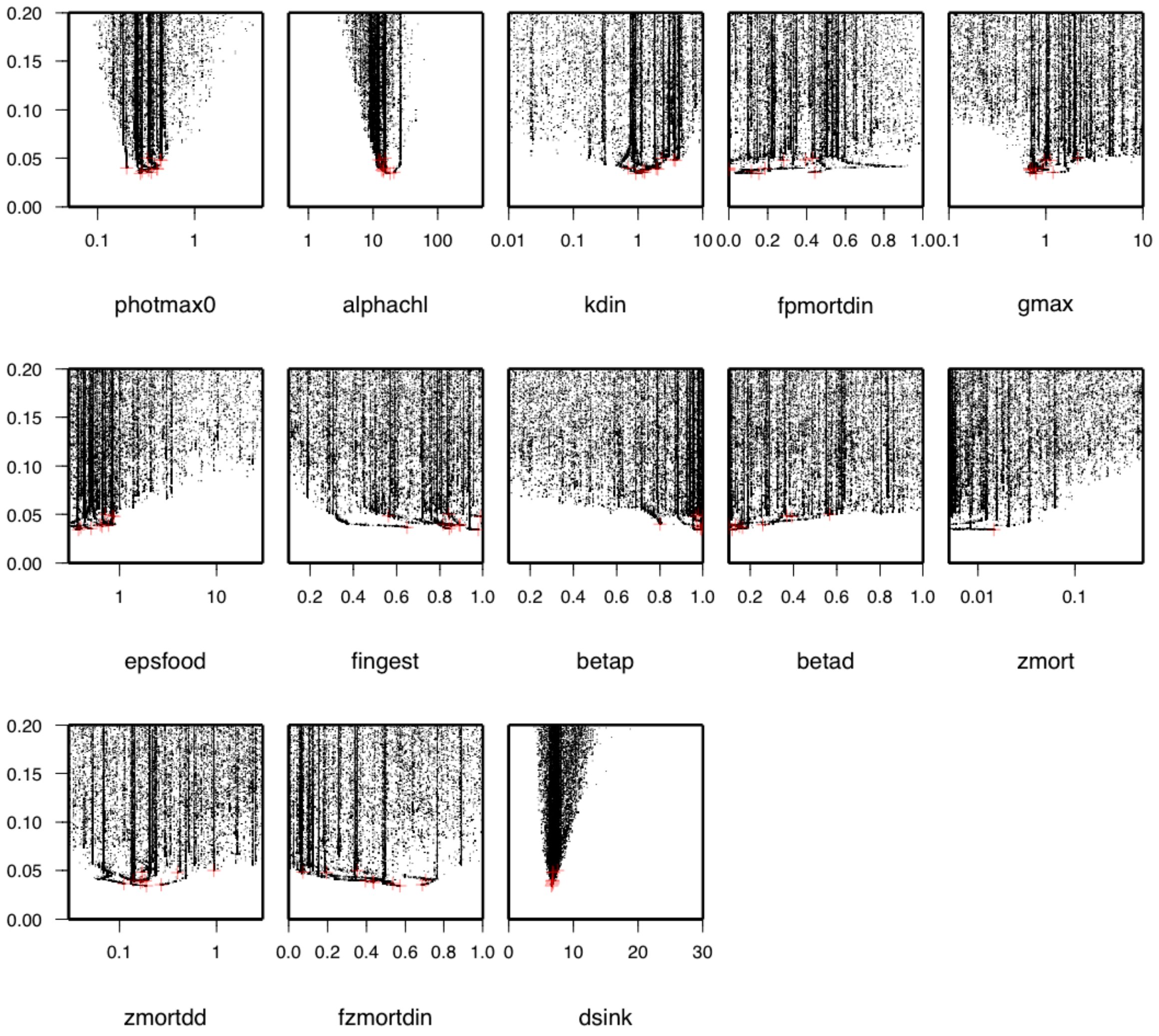

Figure 12: Cost function evaluations for Experiment OPT13CC-NPZF. Final results for 10 different optimizer initializations are shown by red crosses. See Table 5 for parameter units.

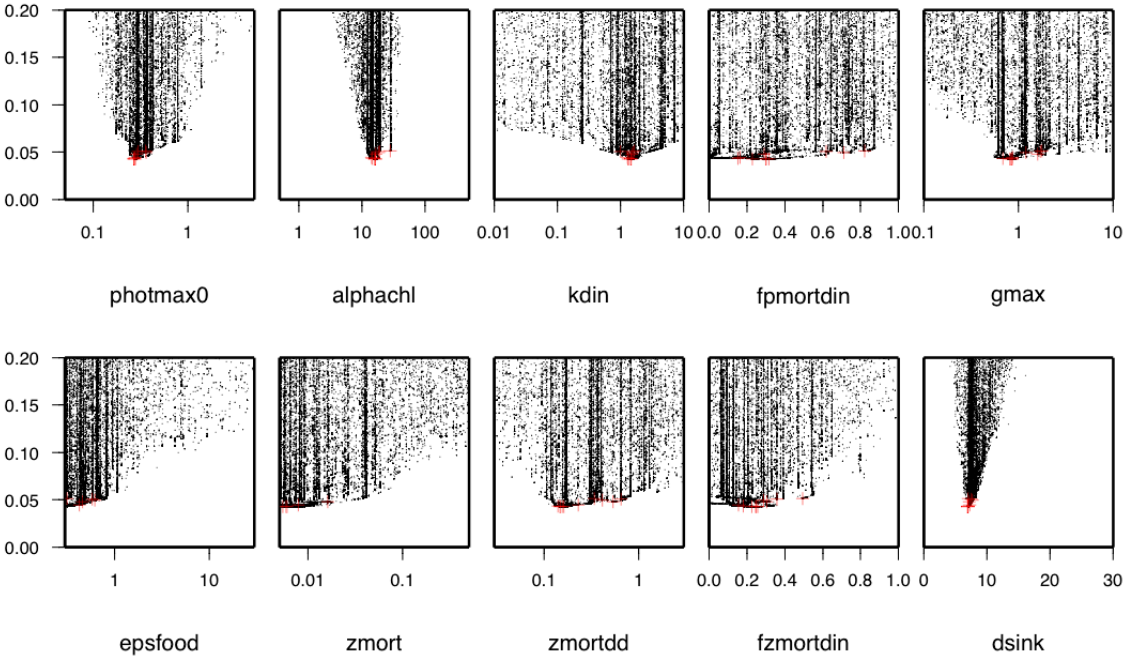

Figure 13: Cost function evaluations for Experiment OPT10CC-NPZF. Final results for 10 different optimizer initializations are shown by red crosses. See Table 5 for parameter units.

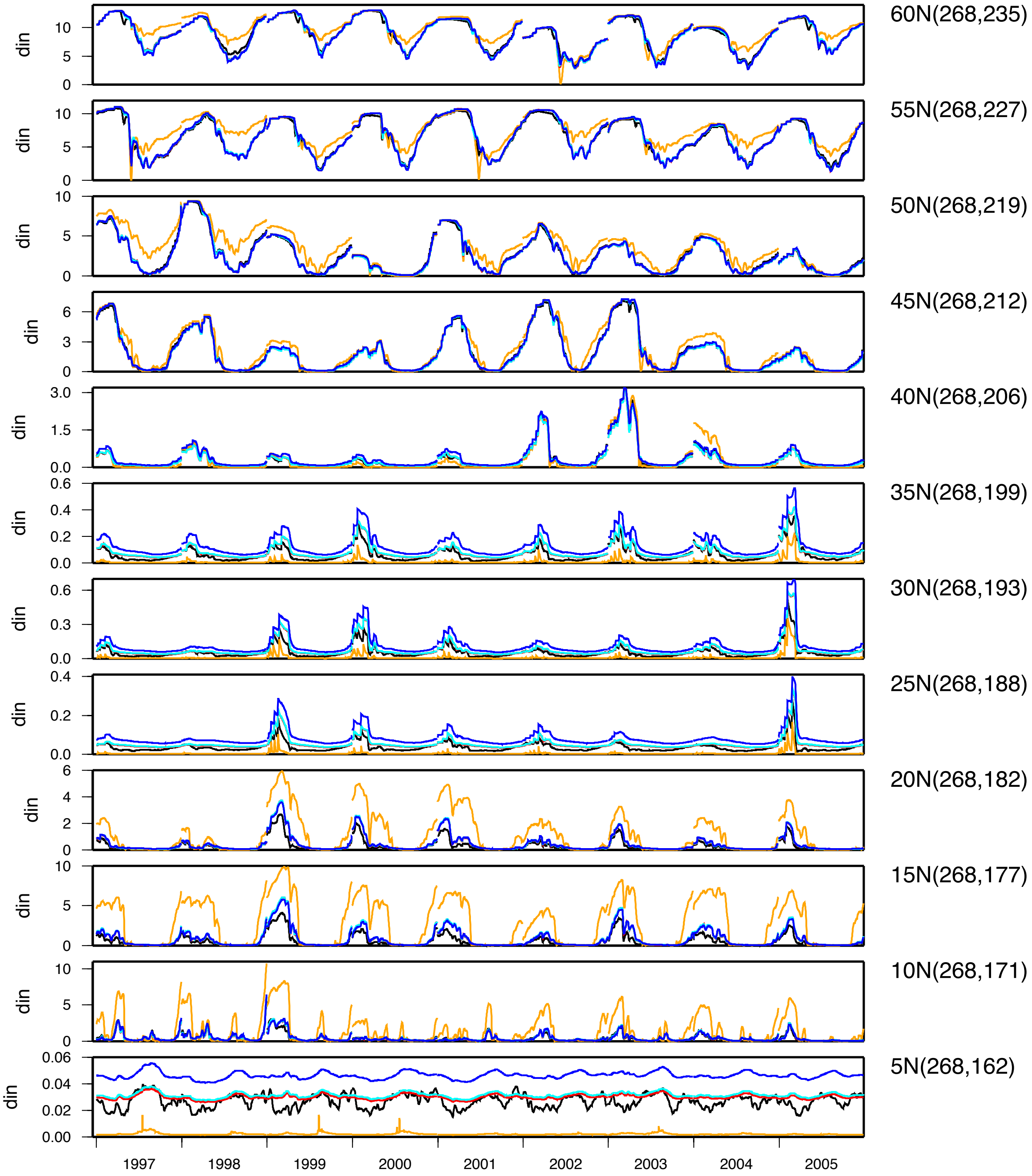

Figure 14: Surface level DIN (mmol N m^{-3}) from second year of 2 year simulations: MEDUSA (black); HadOCC with prior parameters (orange), OPT13CC-NPZ parameters (red), OPT13CC-NPZF parameters (cyan), OPT10CC-NPZF parameters (blue). Plots are for each calibration site on the 20W transect. The ORCA1 grid reference is shown in brackets.

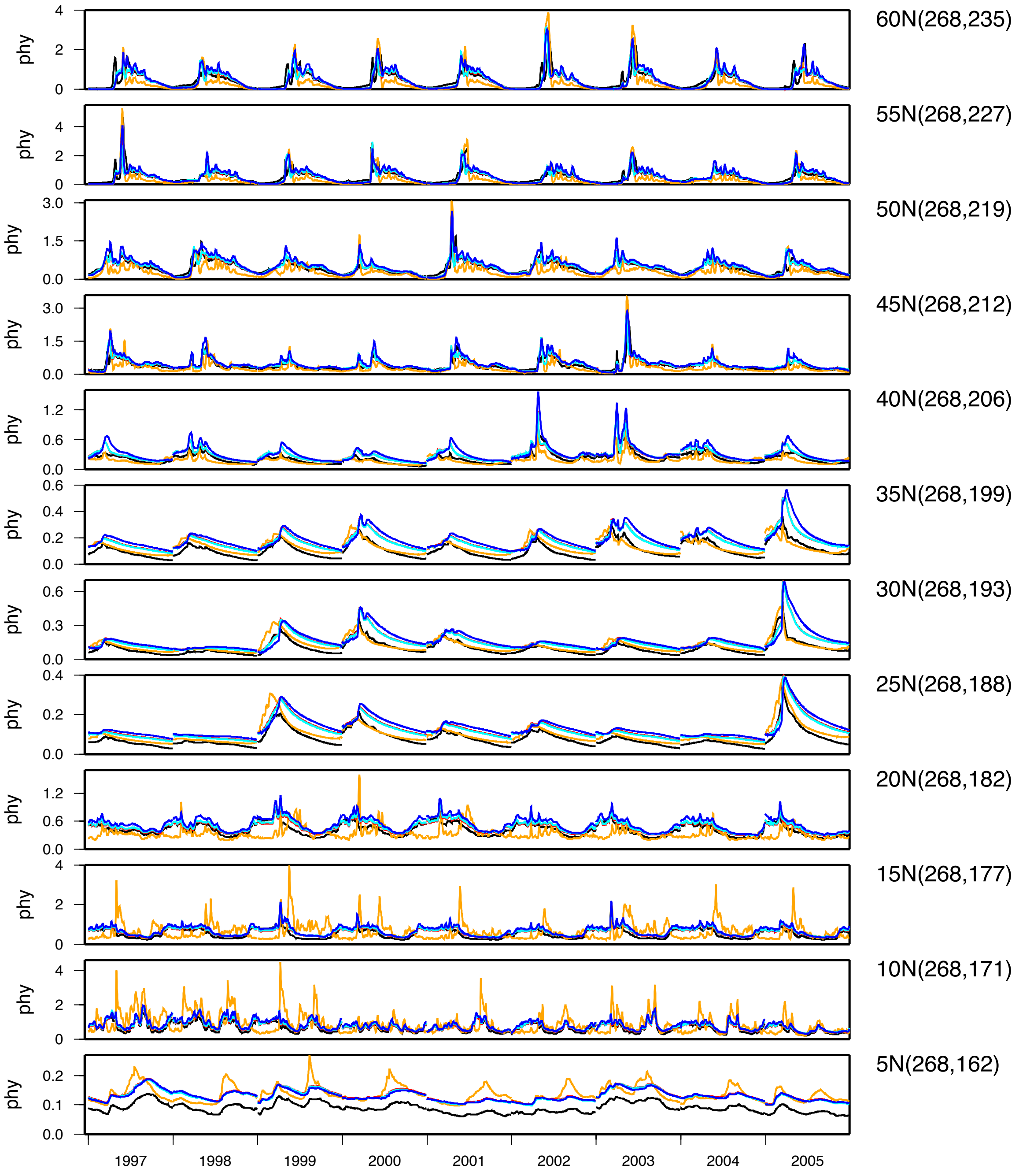

Figure 15: Surface level phytoplankton (mmol N m^{-3}) from second year of 2 year simulations: MEDUSA (black); HadOCC with prior parameters (orange), OPT13CC-NPZ parameters (red), OPT13CC-NPZF parameters (cyan), OPT10CC-NPZF parameters (blue). Plots are for each calibration site on the 20W transect. The ORCA1 grid reference is shown in brackets.

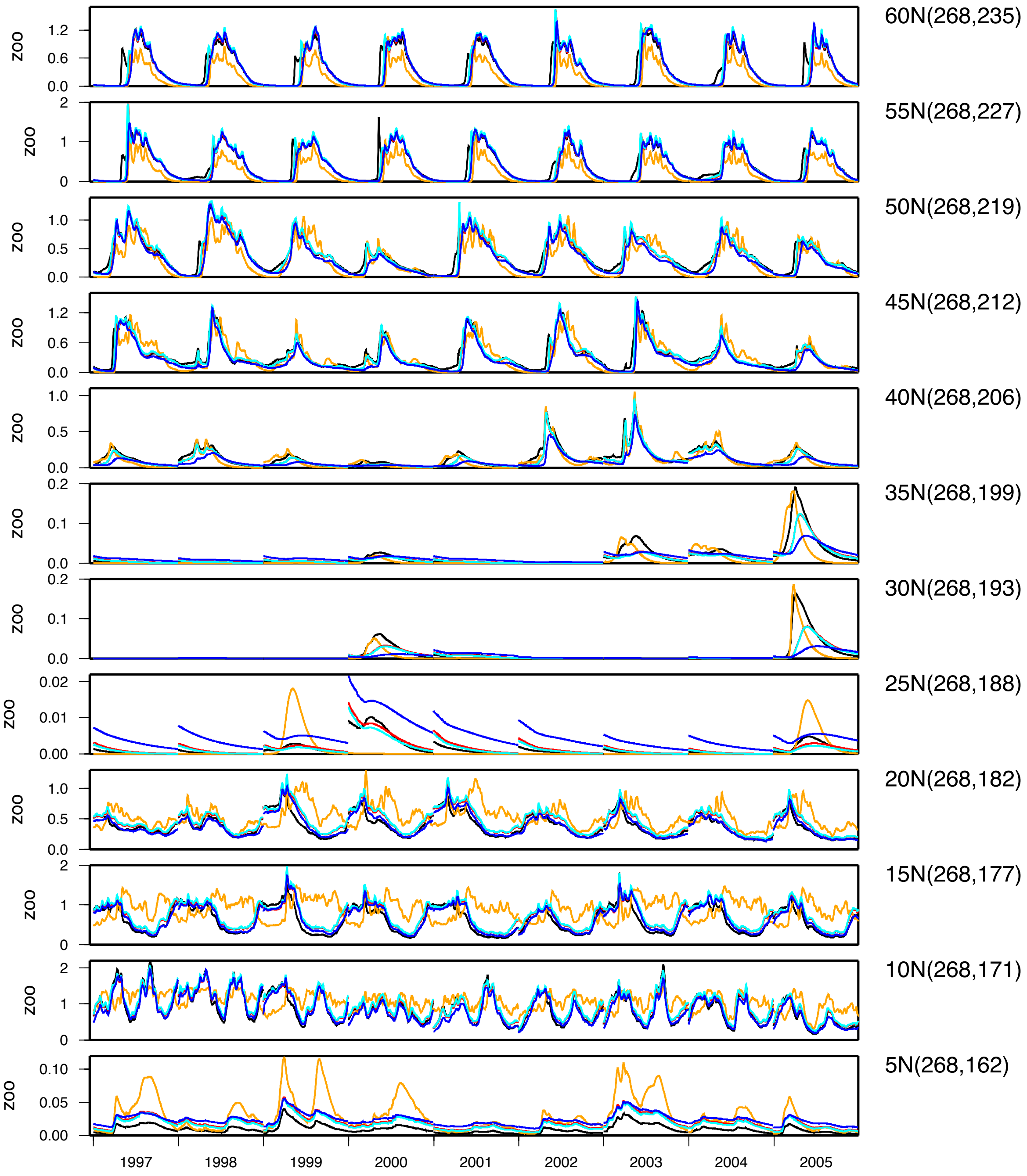

Figure 16: Surface level zooplankton (mmol N m^{-3}) from second year of 2 year simulations: MEDUSA (black); HadOCC with prior parameters (orange), OPT13CC-NPZ parameters (red), OPT13CC-NPZF parameters (cyan), OPT10CC-NPZF parameters (blue). Plots are for each calibration site on the 20W transect. The ORCA1 grid reference is shown in brackets.

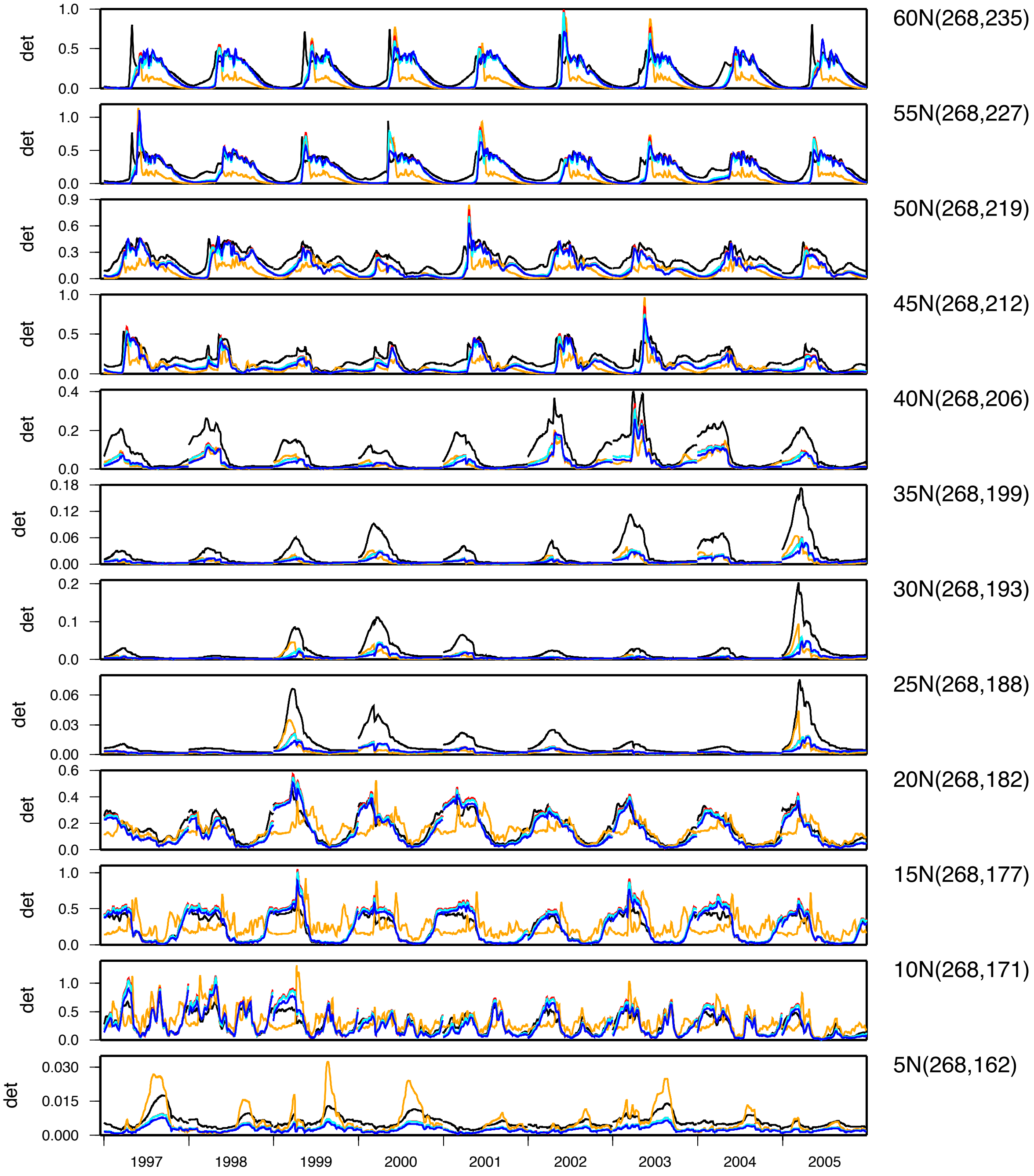

Figure 17: Surface level detritus (mmol N m^{-3}) from second year of 2 year simulations: MEDUSA (black); HadOCC with prior parameters (orange), OPT13CC-NPZ parameters (red), OPT13CC-NPZF parameters (cyan), OPT10CC-NPZF parameters (blue). Plots are for each calibration site on the 20W transect. The ORCA1 grid reference is shown in brackets.

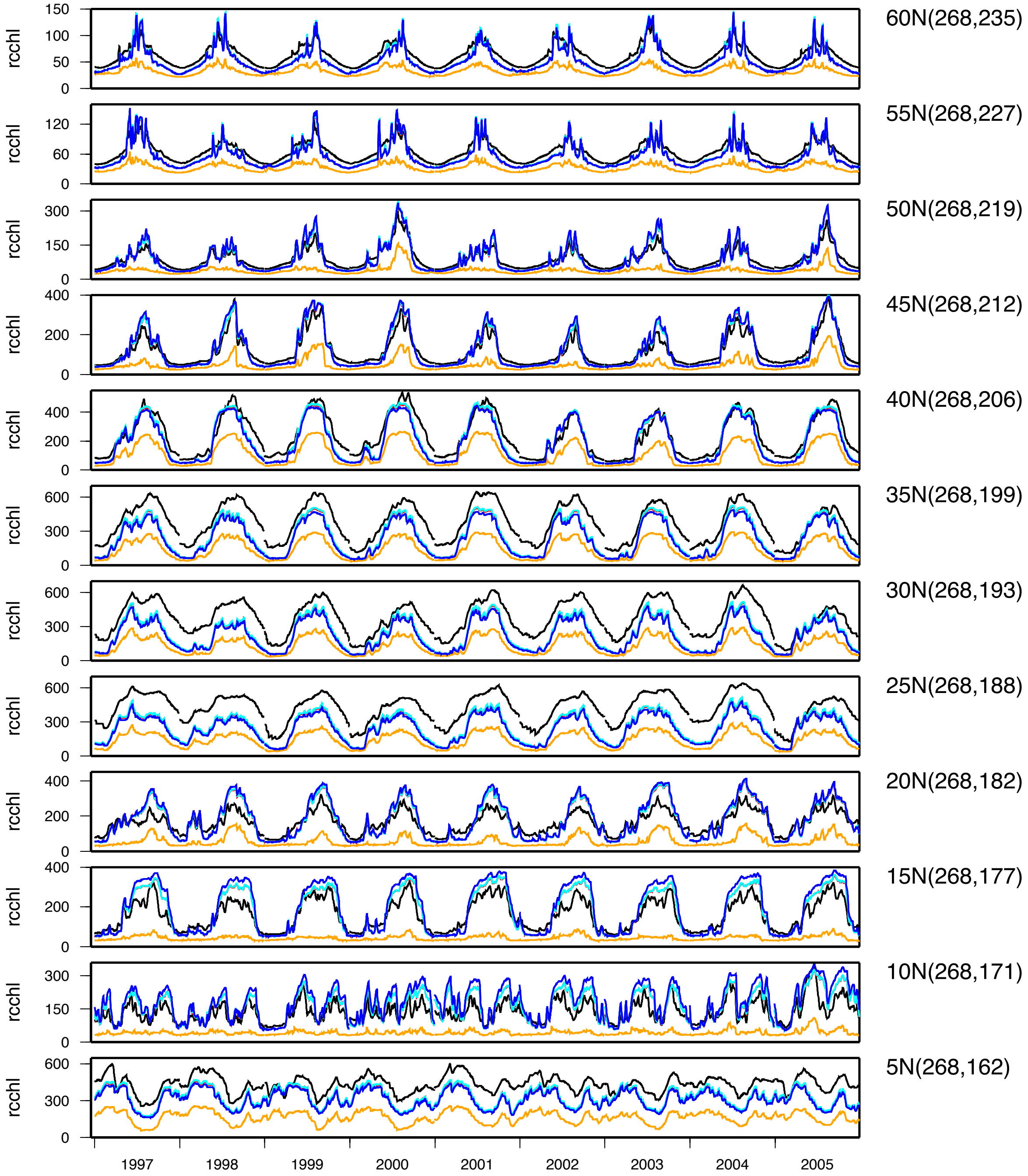

Figure 18: Surface level C:Chl ratio (gC gChl^{-1}) from second year of 2 year simulations: MEDUSA (black); HadOCC with prior parameters (orange), OPT13CC-NPZ parameters (red), OPT13CC-NPZF parameters (cyan), OPT10CC-NPZF parameters (blue). Plots are for each calibration site on the 20W transect. The ORCA1 grid reference is shown in brackets.

biogeochemical cycles: A comparison of results for iron-replete and iron-limited areas, and the SOIREE iron-enrichment experiment, Deep-Sea Res. I, 53, 333–366, 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.

Geider, R. J., MacIntyre, H.L. and T. M. Kana, T. M., 1997. Dynamic model of phytoplankton growth and acclimation: Responses of the balanced growth rate and the chlorophyll *a*:carbon ratio to light, nutrient-limitation and temperature. Mar. Ecol. Prog. Ser. *148*, 187-200.

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.

Hemmings, J. C. P. and Challenor, P. G., 2012. 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.

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

Kennedy, M. C. and O'Hagan, A. 2001. Bayesian calibration of computer models. J. R. Statist. Soc. B *63*, 425-464.

Krishnakumar, K. 1989. Micro-genetic algorithms for stationary and non-stationary function optimization, Proc. SPIE: Intelligent Control and Adaptive Systems *1196*, Philadelphia, PA, 289-296.

McDonald, C. P, Bennington, V., Urban, N. R. and McKinley, G. A. 1-D test-be dcalibration of a 3-D Lake Superior biogeochemical model. Ecol. Model. *225*, 115-126.

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.

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

Oxlade, R. H., 2012. Comparing multiple simulators using Bayesian emulators, Ph.D. thesis, Durham University. http://etheses.dur.ac.uk/4943/

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

Powell, M. J. D.: An efficient method for finding the minimum of a function of several variables without calculating derivatives, Comput. J., 7, 155–162, 1964.

Press, W. H., Flannery, B. P., Teukolsky, S. A., and Vetterling, W. T.: Numerical Recipes in C: the Art of Scientific Computing, Cambridge University Press, Cambridge, 1992.

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

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.

Yool, A., Popova, E. E., Anderson, T. R., 2011. 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.