Geosci. Model Dev., 5, 471–498, 2012
www.geosci-model-dev.net/5/471/2012/
doi:10.5194/gmd-5-471-2012

Geoscientific Model Development

Addressing the impact of environmental uncertainty in plankton model calibration with a dedicated software system: the Marine Model Optimization Testbed (MarMOT 1.1 alpha)

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

National Oceanography Centre, European Way, Southampton SO14 3ZH, UK

Correspondence to: J. C. P. Hemmings (j.hemmings@noc.ac.uk)

Received: 16 June 2011 – Published in Geosci. Model Dev. Discuss.: 19 August 2011
Revised: 16 March 2012 – Accepted: 29 March 2012 – Published: 20 April 2012

Abstract. A wide variety of different plankton system models have been coupled with ocean circulation models, with the aim of understanding and predicting aspects of environmental change. However, an ability to make reliable inferences about real-world processes from the model behaviour demands a quantitative understanding of model error that remains elusive. Assessment of coupled model output is inhibited by relatively limited observing system coverage of biogeochemical components. Any direct assessment of the plankton model is further inhibited by uncertainty in the physical state. Furthermore, comparative evaluation of plankton models on the basis of their design is inhibited by the sensitivity of their dynamics to many adjustable parameters. Parameter uncertainty has been widely addressed by calibrating models at data-rich ocean sites. However, relatively little attention has been given to quantifying uncertainty in the physical fields required by the plankton models at these sites, and tendencies in the biogeochemical properties due to the effects of horizontal processes are often neglected.

Here we use model twin experiments, in which synthetic data are assimilated to estimate a system's known "true" parameters, to investigate the impact of error in a plankton model's environmental input data. The experiments are supported by a new software tool, the Marine Model Optimization Testbed, designed for rigorous analysis of plankton models in a multi-site 1-D framework. Simulated errors are derived from statistical characterizations of the mixed layer depth, the horizontal flux divergence tendencies of the biogeochemical tracers and the initial state. Plausible patterns of uncertainty in these data are shown to produce strong temporal and spatial variability in the expected simulation error variance over an annual cycle, indicating variation in the significance attributable to individual model-data differences. An inverse scheme using ensemble-based estimates of the simulation error variance to allow for this environment error performs well compared with weighting schemes used in previous calibration studies, giving improved estimates of the known parameters. The efficacy of the new scheme in real-world applications will depend on the quality of statistical characterizations of the input data. Practical approaches towards developing reliable characterizations are discussed.

1 Introduction

Ocean biogeochemical general circulation models (OBGCMs) have a key contribution to make to the goal of understanding biogeochemical cycles at global and regional scales. These models are highly simplified "mechanistic" models of a generic plankton ecosystem, coupled with 3-dimensional ocean circulation models that provide the physical environment to which the plankton models respond. Reliable plankton models are needed to make inferences about the potential role of the marine biota in environmental change. However, the contrast between the complexity of biological systems and the limited data available to empirically constrain model structure and parameter values has led to a wide range of different representations of the marine plankton system. Each model is one of a still wider set of competing hypotheses concerning the dominant mechanisms that control the biological response to change

in the physical and chemical environment. The level of complexity that can be justified in these models, given the available biogeochemical data, has been a subject of some debate (Anderson, 2005; Le Quéré, 2006). To resolve this we must be able to comparatively evaluate models on the basis of their structure and process formulations. Behaviour of plankton models in OBGCMs is sensitive to the details of the physical dynamics (Sinha et al., 2010). Dependence on a particular physical model in comparative assessments of model designs should therefore be avoided if future biogeochemical simulations are to benefit from improved representations of the physical environment.

Direct comparison of plankton models on the basis of their design is inhibited by parameter uncertainty: behaviour of each model depends on many adjustable parameters that are poorly known or difficult to quantify. Although some of these values can be determined experimentally under controlled conditions, the corresponding values in nature are generally highly variable in space and time or across taxa. Fasham and Evans (1995) and Matear (1995) started to address this problem by fitting plankton models to observations from time-series sites in the temperate North Atlantic and subarctic Pacific respectively, using non-linear data assimilation techniques to seek optimal parameter sets. Matear (1995) investigated 3 different ecosystem configurations with 3, 4 and 7 nitrogen compartments and concluded that the data from the study site were insufficient to justify either of the more complex models over the simple nitrate-phytoplankton-zooplankton model. Dadou et al. (2004) compared 3 alternative configurations, spanning a similar range of complexity, at an oligotrophic study site in the eastern North Atlantic and were not able to objectively discriminate between the designs on the basis of their misfit results.

To test models' predictive ability it is necessary to examine their misfit with respect to unassimilated data as in the more recent model inter-comparison experiments of Friedrichs et al. (2006, 2007). In an experiment with 12 models (Friedrichs et al., 2007), data from Arabian Sea and Equatorial Pacific sites were used and models calibrated at one site were cross-validated at the other. Here, the more complex models with multiple plankton functional groups tended to perform better, provided that only a small number of parameters were optimized, suggesting greater portability and predictive skill associated with model design.

The results obtained from all of these optimization experiments are dependent on the external inputs to the plankton model. Friedrichs et al. (2006) examined the impact of uncertainty in the physical forcing and demonstrated that likely errors in the physical forcing data can have a major impact on biogeochemical simulations, causing a calibration process to yield inappropriate parameter values. One approach to solving this problem is to improve the physical forcing. Joint assimilation of physical and biogeochemical data, as advocated by Friedrichs et al. (2006), seems likely to be beneficial. However, the inadequacy of data coverage combined with the sensitivity of plankton models to their forcing data inevitably makes the problem persistent, motivating a formal treatment of uncertainty.

The uncertainty introduced by horizontal processes poses a further problem for 1-D studies that has yet to be satisfactorily addressed. Flux divergences associated with mesoscale eddy activity are particularly problematical in this respect. The issue does not arise explicitly when calibrating a model to simulate a climatological annual cycle (Matear, 1995; Hurtt and Armstrong, 1996, 1999; Spitz et al., 1998, 2001; Schartau and Oschlies, 2003; Dadou et al., 2004; Losa et al., 2004). In these cases, mesoscale and inter-annual variability are both interpreted as noise superimposed on the average annual cycle. Alternatively, mesoscale variability can be treated as noise superimposed on spatially averaged plankton concentrations. On this basis, Hemmings et al. (2003, 2004) treated all satellite chlorophyll data within either 150 km or 100 km as equally representative of the calibration site. A problem with both approaches is that averaging tends to smooth out features such as blooms, in effect changing the apparent response of the system that we are attempting to model.

Simulating the dynamics for specific years at specific locations seems preferable, particularly if we want plankton models that will benefit from increased resolution in general circulation models, but it requires more supporting data. Year-specific forcing can be derived from in situ observations (Fasham and Evans, 1995; Schartau et al., 2001; Fasham et al., 2006), from a 1-D physical model with appropriate meteorological forcing (Prunet et al., 1996a,b; Faugeras et al., 2003, 2004; Kettle, 2009), from a 3-D circulation model (Fennel et al., 2001; Schartau et al., 2001) or from a combination of in situ and 3-D model data (Friedrichs et al., 2006, 2007). However, the local forcing is only relevant when local effects are dominant. The presence of strong mean flows in some regions, together with the ubiquity of mesoscale patchiness associated with fronts and eddies means that such dominance cannot generally be assumed. Friedrichs et al. (2007) determined that horizontal advective divergence of nutrients could have first order effects on the biogeochemistry at the Equatorial Pacific site and introduced an additional source/sink term computed from a 1/3° coupled biological-physical model to account for these, while acknowledging the issue of unknown error in the 3-D model.

Other approaches to the horizontal flux divergence problem have been applied with some success to specific data sets. Fasham et al. (1999) used data from a 3 week North Atlantic spring bloom survey that followed a drogued buoy, deployed within an anti-cyclonic eddy, to minimize contamination of the biological dynamics by non-local effects. In a calibration exercise using data from the SOIREE iron fertilization experiment, Fasham et al. (2006) parameterized diffusive flux divergence effects using a mixing rate based on the dilution of a passive tracer added to the iron enriched water. A novel "variable lag" fitting technique introduced

by Wallhead et al. (2006) allows for phase differences associated with mesoscale patchiness. Survey data from a relatively wide area could thereby be combined without explicitly resolving mesoscale processes yet avoiding the risk of smoothing out temporal variability.

It is clear that a thorough investigation of the impact of uncertainty in all factors that contribute to uncertainty in plankton model simulations is a high priority. The associated data management issues, in combination with the need to perform a wide range of computationally expensive model analyses involving many different simulations has been a factor inhibiting rapid progress in this area. The MarMOT software system has been developed as a generic tool applicable to different plankton models with the aim of removing this barrier.

In Sect. 2, existing model calibration schemes are reviewed and a new scheme is proposed that includes an explicit treatment of environmental uncertainty. Section 3 describes an evaluation of the scheme in idealized model-twin experiments where the true system is known, exploiting key features of the MarMOT system. The challenges of applying the scheme to real-world data and the wider role of MarMOT in plankton model analysis are discussed in Sect. 4 and a summary is presented in Sect. 5.

2 Cost function design

In inverse analyses of plankton ecosystem models, parameter optimization is generally performed by minimization of a cost function. Maximum likelihood methods have also been employed (Hurtt and Armstrong, 1996, 1999), in which an optimizer is applied to the problem of maximizing a function describing the likelihood of the parameter values conditional on the observation set. The two techniques are essentially equivalent and give point estimates of the model parameters. Alternatively, in a fully Bayesian scheme, the likelihood is multiplied by prior probability distributions for the parameters to estimate their complete posterior distributions (Harmon and Challenor , 1997) or combined distributions for the parameters and the system state (Dowd and Meyer, 2003).

2.1 Generic cost function

The use of cost functions as metrics for summarizing the overall performance of simulations against multivariate observational data sets is discussed by Stow et al. (2009). The MarMOT system supports a generic cost function for multiple variable types and multiple simulation cases of the form

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

$$N = \sum_{k=1}^{C} \sum_{j=1}^{m} \sum_{i=1}^{n_k} p_{ijk} \quad (2)$$

where C is the number of cases, 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. 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.

Model-data differences may be calculated in transformed variable space. log or square root transformations are sometimes used, in which case x is replaced by $\log_{10} x$ or $\sqrt{x}$, respectively and y is likewise replaced by $\log_{10} y$ or $\sqrt{y}$. Log transformations emphasize relative error and are appropriate for variables that tend to exhibit log-normal distributions. However, in ecological analyses it is often unclear whether absolute or relative errors should be considered. Square root transformations have been applied as a compromise in some studies for this reason (Fasham and Evans, 1995; Evans, 1999; Dadou et al., 2004; Fasham et al., 2006).

The basic cost function described here could be extended to introduce additional constraints. In particular, parameter penalty terms are often included to inhibit excessive deviation of parameters from their prior expected values. Such terms allow subjective prior information about the parameters to be included which can be particularly valuable in the analysis of under-determined systems. Although MarMOT does not presently support penalty terms in the cost function, parameter bounds can be imposed independently of the cost function using optimizer features described in Appendix C. An alternative approach is to reduce the size of the adjustable parameter set to one that can be adequately constrained by the available data (Friedrichs et al., 2007). While parameter constraints are generally useful, omitting them can reveal useful information about deficiencies in model design if the data constraints cause parameters to take values outside their expected ranges.

2.2 Weighting of model-data differences

The weight given to individual model-data misfits in a particular cost function or likelihood function is fundamental to the effectiveness of data assimilation for controlling model parameter values. As discussed by Evans (2003), a wide variety of different approaches have been used in the literature, having a potentially major impact on parameter estimates and the resultant estimates of key biogeochemical quantities from the calibrated model simulations.

Unweighted misfits have been used (Fasham and Evans, 1995) or sometimes weights have been used in a subjective way to give more influence to observations that are felt to be more reliable or more important to fit (Fasham et al., 1999, 2006). The square root transform used by Fasham and Evans

(1995), Evans (1999) and Fasham et al. (2006), while not weighting individual misfits explicitly, has the effect of giving more influence to misfits occurring when values of model and data are low. This is a compromise between treatment of absolute and relative errors; absolute errors might be considered more important in the context of estimating total element fluxes, whereas relative errors might be favoured by arguments based on representing ecological structure (Evans, 1999). Hurtt and Armstrong (1996); Fasham et al. (1999); Hurtt and Armstrong (1999) scaled model-data differences relative to the model values at the observation points, giving equal weight to equal relative departures.

More typically, some characteristic scale is determined for each assimilated data type, designed to reflect its variability relative to other data types over the whole data set. Weights w_j (Eq. 1) are chosen to be inversely proportional to the mean of all observations of the same type (Spitz et al., 2001), the square of the mean (Kuroda and Kishi, 2004) or their variance (Friedrichs et al., 2006, 2007; Kettle, 2009; Ward et al., 2010). Friedrichs et al. (2007) and Ward et al. (2010) found it necessary to introduce a subjective up-weighting of misfit to primary production observations due to the high variability of these data. Evans (2003) suggested that if focusing on the cycle of a particular element it may be desirable to give the same weight to all misfits for that element, regardless of the form in which it occurs. Dadou et al. (2004) therefore used a single scaling factor for all nitrogen variables, based on the maximum observed nitrate, and used intuitive arguments to determine relative scaling factors for primary production and particle fluxes based on the maximum observed values of other relevant properties.

In general, characteristic scales are used because of the absence of information required to properly estimate error variances. In some studies though, the variable-specific weight is presented as the reciprocal of an assumed or estimated observation error variance (Prunet et al., 1996a,b; Fennel et al., 2001; Faugeras et al., 2003, 2004); for a particular variable, either absolute or relative error variances are taken to be constant. Schartau et al. (2001) used a combination of constant absolute and relative error variance estimates for chlorophyll and primary production data. Finally, seasonally varying observation error variance estimates have been used in inverse modelling of the annual cycle (Matear, 1995), while Hemmings et al. (2003, 2004) estimated error variances specific to individual chlorophyll observations from spatial variances in satellite data.

There are some other weighting considerations that are unrelated to the expected error variances. Cases are common in the literature where different numbers of observations are available for different data types. They are generally treated in one of two ways: in some studies, misfits for different variables are weighted by the reciprocal of the number of observations of each type (Hurtt and Armstrong, 1999; Schartau et al., 2001; Schartau and Oschlies, 2003; Faugeras et al., 2003, 2004; Hemmings et al., 2003, 2004; Friedrichs et al., 2007; Kettle, 2009; Ward et al., 2010), while in others, no such weighting is applied (Matear, 1995; Prunet et al., 1996a; Hurtt and Armstrong, 1996; Spitz et al., 2001; Friedrichs, 2002; Dadou et al., 2004; Kuroda and Kishi, 2004; Friedrichs et al., 2006). The choice is significant: Fasham and Evans (1995) performed experiments with and without a weighting factor that increased the influence of the small number of zooplankton observations in their data set, obtaining two different optimal parameter sets for which simulated primary production differed by a factor of about 2.

Explicit weighting to balance the contributions of different data types is objectively justifiable if error correlations are much greater between variables of the same type than between different data types. However, this cannot generally be assumed and Evans (2003) argues that, while such balancing has the advantage of emphasizing scarce but important measurements, it may not be desirable in a formal procedure. As discussed by Evans (2003), we can expect simulation errors arising from model error or external factors to introduce both serial correlations and correlations between variables via the model dynamics. This could be allowed for by the use of a non-diagonal covariance matrix in the cost function formulation. However, the issue has not been addressed in previous studies and a full treatment is not presently supported in MarMOT.

Another issue arises when optimizing over multiple sites. Schartau and Oschlies (2003) optimized parameters for three Atlantic sites simultaneously and found with their initial weighting scheme that observations at a particular site had a much greater influence than those at the other sites. This was a consequence of order-of-magnitude variations in property concentrations between sites. The problem was countered by introducing a weight based on variables' mean values at each site, an approach also adopted by Friedrichs et al. (2007) in simultaneous optimizations for sites in the Arabian Sea and Equatorial Pacific. No site-specific weighting was used in the two-site calibration of Hurtt and Armstrong (1999) or the multi-site calibrations of Hemmings et al. (2003, 2004).

When the objective is to achieve a particular compromise between sites or between variables that is dictated by an application of the model then some subjective weighting can be justified. However, when it is to make inferences about the model such weighting is undesirable. Furthermore, it is possible that improved normalization of model-data misfits could reduce the need for it.

2.3 An uncertainty-based weighting scheme

A formal weighting scheme is developed here, with explicit consideration given to the different sources of error contributing to the model-data misfit. Misfit arises from a combination of error in the observations and error in the simulation. Error in the observations arises from both measurement error and error of representativeness. The latter is error due to small-scale variability or, more specifically, the mismatch

between the volume of water sampled and the minimum scale resolved by the simulation. It includes error due to small-scale variations in both space and time. Error in the simulation is the result of model error, attributable to deficiencies in the model, and environment error, attributable to error in its environmental inputs (forcing data and boundary conditions). For a model with optimizable parameters, model error can be treated as the sum of parameter error and structural error components. The structural error is the residual error for the true parameter set (assuming such a set exists conceptually). It is the error associated with the model design and includes error attributable to values of any fixed model parameters.

If we assume that all errors are additive and independent, the simulated and observed values of variable j at observation point i at site k can be expressed as

$$x_{ijk} = x_{ijk\mathrm{T}} + \epsilon_{ijk\mathrm{ENV}} + \epsilon_{ijk\mathrm{P}} + \epsilon_{ijk\mathrm{S}} \quad (3)$$

$$y_{ijk} = x_{ijk\mathrm{T}} + \epsilon_{ijk\mathrm{OBS}} \quad (4)$$

where $x_{ijk\mathrm{T}}$ is the true value (i.e. that for a perfect simulation) and $\epsilon_{ijk\mathrm{ENV}}$, $\epsilon_{ijk\mathrm{P}}$, $\epsilon_{ijk\mathrm{S}}$ and $\epsilon_{ijk\mathrm{OBS}}$ are the environment error, parameter error, structural error and observation error, respectively. Observation error here is the sum of measurement error and representativeness error. The model data difference or residual is then:

$$x_{ijk} - y_{ijk} = \epsilon_{ijk\mathrm{ENV}} + \epsilon_{ijk\mathrm{P}} + \epsilon_{ijk\mathrm{S}} - \epsilon_{ijk\mathrm{OBS}}. \quad (5)$$

While it is unreasonable to assume that the simulation error sources are truly independent, the interpretation here is useful if they are in some sense separable. For the purposes of this study, mean errors are assumed to be zero. The lack of any explicit treatment of bias is consistent with previous studies. However, it is acknowledged as a potential limitation. A further assumption is that errors are normally distributed.

The appropriate normalization variance (reciprocal of w_{ijk}) for the cost function depends on the objective. If it is to evaluate the goodness-of-fit of a given simulation, then all simulation errors are fixed. $\epsilon_{ijk\mathrm{OBS}}$ is treated as a random variable and the expected variance of the residual for a perfect simulation is the observation error variance $\sigma^2_{ijk\mathrm{OBS}}$. However, if the aim is to evaluate the plankton model itself then we must take into account environmental uncertainty. For a model with a prescribed parameter set $\epsilon_{ijk\mathrm{P}}$ and $\epsilon_{ijk\mathrm{S}}$ are fixed, while both $\epsilon_{ijk\mathrm{OBS}}$ and $\epsilon_{ijk\mathrm{ENV}}$ should be treated as random variables. The residual for a perfect model (obtained by setting the fixed errors in Eq. (5) to zero) has an expected variance equal to the sum of the observation and environment error variances $\sigma^2_{ijk\mathrm{OBS}} + \sigma^2_{ijk\mathrm{ENV}}$. The corresponding cost function for quantifying model goodness-of-fit is:

$$J = \frac{1}{N} \sum_{ijk} \frac{(x_{ijk} - y_{ijk})^2}{\sigma^2_{ijk\mathrm{OBS}} + \sigma^2_{ijk\mathrm{ENV}}}. \quad (6)$$

The significance of each individual model-data misfit is reduced to take into account the effect of uncertainty in the model's environmental inputs. The expected value of J for a perfect model is 1; if the model-data difference is no larger on average than might be expected as a result of observation error and environment error then there is no evidence for model error so the data give us no cause to reject the model.

If the aim is to estimate model parameters we need also to take into account structural uncertainty so $\epsilon_{ijk\mathrm{S}}$ becomes a random variable. The cost function for evaluating a particular parameter set is:

$$J = \frac{1}{N} \sum_{ijk} \frac{(x_{ijk} - y_{ijk})^2}{\sigma^2_{ijk\mathrm{OBS}} + \sigma^2_{ijk\mathrm{ENV}} + \sigma^2_{ijk\mathrm{S}}}. \quad (7)$$

If the model-data difference is no larger than might be expected as a result of observation error, environment error and structural error then there is no evidence for parameter error so no cause to reject the parameter set.

In ecosystem models, parameters typically do not correspond to well defined physical constants and the hypothetical "true" parameter set is likely to be model-specific. In such cases the distinction between parameter error and structural error becomes unclear. In addition, the problem of estimating the structural error and its varying contribution between data points is less tractable than that of estimating the environment error. For pragmatic reasons, we might therefore permit parameter values to compensate for structural error and ignore the structural error term. The free parameters would then be adjusted to minimize Eq. (6).

A value for $\sigma_{ijk\mathrm{OBS}}$ can in principle be derived from repeat observations, if available. An appropriate value for $\sigma_{ijk\mathrm{ENV}}$ can be obtained from ensemble integrations of the model with input data representative of the probability distributions of the forcing variables and boundary conditions. The uncertainty in these external fields is propagated to the model-estimated properties via the simulation and $\sigma_{ijk\mathrm{ENV}}$ is then determined from the resulting probability distribution at each data point. The method relies on a good characterization of uncertainty in forcing data and boundary conditions, requiring a thorough analysis of relevant satellite and in situ data available for the site and its surroundings. Local modelling studies, including data assimilating hindcasts, might provide additional information. In a calibration exercise, parameter error is non-zero and the issue of separability of parameter error and environment error arises. This is addressed by taking into account parameter uncertainty in the derivation of $\sigma_{ijk\mathrm{ENV}}$.

3 Twin experiments

The potential of the proposed calibration method is examined by way of identical twin experiments in which the true parameter values are known. We focus on the design of the cost function. The more general calibration problem normally includes a parameter selection phase guided by a sensitivity

analysis to determine which parameters can be independently constrained and/or which parameters are likely to impact on model outputs of particular scientific interest. The selection phase is outside the scope of the present study.

In the twin experiments, synthetic observations are generated from a model with a particular parameter set, taken to represent the true system. The same model with 5 free parameters is then optimized to fit these data in an attempt to recover the original "true" parameter values. Results for the proposed method are compared with those obtained using established weighting schemes.

A set of three experiments is performed with different weighting schemes, one using a characteristic scale for each variable, another based on the known observation error statistics and a third using these in combination with expected simulation error variances. The first two schemes are representative of established schemes described in Sect. 2.2.

In Experiment 1, we consider observation error variance estimates based on the inherent variability in the data set and, following Friedrichs et al. (2006), set the uncertainty to 25 % of the standard deviation s for all observations of the same type at the same site. So for variable j at site k the weights in the MarMOT cost function (Eq. 1) are:

$$w_{ijk} = \frac{16}{s_{jk}^2}. \tag{8}$$

In Experiment 2, the known observation error statistics are used, so:

$$w_{ijk} = \frac{1}{\sigma_{j\mathrm{OBS}}^2}. \tag{9}$$

Model-data differences are calculated in transformed space, corresponding to that in which the observation errors are generated. In Experiment 3, the weights are derived following the new method proposed in Sect. 2.3, using environmental simulation error variance estimates $s_{ijk\mathrm{ENV}}^2$. Structural error is zero by definition so the cost function formulation is that given in Eq. (6). Weights for individual misfits are of the form:

$$w_{ijk} = \frac{1}{s_{ijk\mathrm{OBS}}^2 + s_{ijk\mathrm{ENV}}^2}. \tag{10}$$

All model-data differences are calculated in square root space. The expected observation error in square root space $s_{ijk\mathrm{OBS}}$ is derived from $\sigma_{j\mathrm{OBS}}$ according to the observation type.

3.1 Method

3.1.1 Experimental design

The first step is to create a statistical characterization of the environmental inputs representing a given scenario with reasonably realistic patterns of uncertainty. One realization of this synthetic environment represents the true environment and the corresponding simulation is used to generate the observation set. A second realization is treated as the best available estimate of the true environment and used to drive trial simulations with varying parameter vectors in the optimization experiments. This realization of the environmental data is referred to as the optimization environment. To examine the robustness of the results with respect to environment error, the set of optimization experiments is repeated for different realizations. Estimates of the environment error variances $\sigma_{ijk\mathrm{ENV}}^2$ are determined from ensemble realizations using the same synthetic environment model, so they reflect the impact of known uncertainty in the environmental inputs. In a real-world experiment, the reliability of the data assimilation results will depend on how well the uncertainty is characterized.

The plankton model is a version of the HadOCC (Hadley Centre Ocean Carbon Cycle) model, based on the model of Palmer and Totterdell (2001), in which organic carbon fluxes are controlled by a 4 compartment nitrogen cycle. The state variables are dissolved inorganic nitrogen (DIN), phytoplankton, zooplankton and detritus. A full description of the model's nitrogen cycle is given in Appendix D. The parameters to be optimized are specified in Table 1.

The chosen scenario is based on an annual cycle at three sites with 1-D simulations being driven by environmental input data from a global ocean biogeochemical general circulation model; the NEMO (Nucleus for European Modelling of the Ocean) model coupled with the plankton model MEDUSA (Model for Ecosystem Dynamics, carbon Utilisation, Sequestration and Acidification) is used to provide local physical forcing data and statistics for the horizontal flux divergence tendencies of the biogeochemical properties. The plankton model providing the divergence tendencies is therefore different from the plankton model being analyzed in the 1-D simulations. This is acceptable in the context of the twin experiments: there is no requirement for the lateral forcing to be consistent with the model being calibrated, provided that it is representative of likely conditions in real-world experiments. For model assessment in a real-world scenario, the relevant divergence tendencies are estimates of those that would be obtained by running the same model in a hypothetical perfect 3-D physical simulation.

The environmental input data from NEMO-MEDUSA are derived from 5 day mean fields from a 3-D simulation at 1/4° resolution with 64 vertical levels, referred to as ORCA025-N201 (Popova et al., 2010). The run was undertaken at the National Oceanography Centre as part of the DRAKKAR collaboration (Barnier, 2006) with model integration being performed on HECToR, the UK National Supercomputing Centre facility. The selected sites are at 31° N 64° W, 47° N 20° W and 59° N 19° W corresponding to the BATS, NABE and OWS-INDIA sites used by Schartau and Oschlies (2003).

Table 1. Free parameter space.

Parameter	Unit	Symbol	Minimum	Maximum	Transformation	Local Search
Initial slope of photosynthesis-PAR curve	$\mathrm{mg\,C\,(mg\,Chl)^{-1}}$	α_{surf}	0.5	50	log	unbounded
Half-saturation conc. for nutrient uptake	$\mathrm{mmol\,N\,m^{-3}}$	k_{N}	0.01	1	log	unbounded
Maximum grazing rate	$\mathrm{d^{-1}}$	g_{max}	0.1	10	log	unbounded
Zooplankton density-dependent mortality	$\mathrm{d^{-1}(mmol\,N\,m^{-3})^{-1}}$	m_2	0.03	3	log	unbounded
Detrital sinking velocity	$\mathrm{m\,d^{-1}}$	w_{D}	0	100	none	bounded

3.1.2 1-D simulations

All of the 1-D simulations were performed using the Marine Model Optimization Testbed (Fig. 1). Following the testbed concept of Friedrichs et al. (2006, 2007), MarMOT provides a common physical and computational environment in which different plankton ecosystem models can be calibrated and compared. It is designed to support computationally intensive experiments in which model integrations are performed many times with different input data. MarMOT does not include a 1-D physical model. All physical forcing is instead provided by external input data. Plankton model responses to a wide range of different physical environments can be examined by providing different instances of the forcing data derived from models or observations or a combination of both. Further details of the system design are given in Appendix A. The features described here relate to each individual simulation case.

The equation for the evolution of a biogeochemical tracer concentration C_i in a MarMOT simulation is:

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

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 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 (in the absence of horizontal gradients). SMS_i is the source-minus-sink term from the selected plankton model which is a function of the state vector $\boldsymbol{C}$ and a forcing vector $\boldsymbol{F}$. w_i is also provided by the plankton model.

The last two terms define the boundary condition: p_i is a perturbation term driven by an applied perturbation $p^\star$, which may be stochastic, and 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}. The sum of these terms can be interpreted to represent missing tendencies due to horizontal flux divergence or they might be used to introduce or correct for errors in the simulation.

In addition to the tendency terms in Eq. (11), rapid mixing of the upper mixed layer can be parameterized by complete homogenization of tracers above an externally specified mixed layer depth at each time step. Optional partial mixing of the model level spanning the specified depth is included.

Forcing data for the model can be periodic, representing a repeating annual cycle, or year specific. The standard forcing variables for a 1-D plankton model simulation determine the light availability at the sea surface and the transport of passive tracers in the water column. In MarMOT, they comprise the downwelling solar radiation incident on the sea surface, either as a daily mean or a point-in-time estimate, the mixed layer depth, the depth-dependent turbulent diffusion coefficient K_ρ and the vertical velocity w_{p}. Additional model-specific forcing variables are also catered for.

In perturbed simulations, the perturbation for an individual tracer can be independent of the concentration C_i or it can be applied to log-transformed or square root-transformed concentration so that p_i becomes a function of concentration. In either case, the applied perturbation $p_i^\star$ is given by the sum of a prescribed perturbation μ_i^{pert} and a stochastic term. The latter is modelled as a first order auto-regressive process such that the perturbation at time step n is

$$p_i^\star = \mu_i^{\mathrm{pert}} + q_n \quad (12)$$

where:

$$q_n = a q_{n-1} + \epsilon_n. \quad (13)$$

The value a is determined from the auto-correlation coefficient for q at a time lag of 24 h as specified by a fixed simulation parameter. ϵ_n is a normally distributed random variable with zero mean. Its standard deviation is set to give an expected $p_i^\star$ standard deviation matching that prescribed by external data σ_i^{pert} in cases where the process is stationary. The actual perturbation process can be non-stationary: μ_i^{pert} and σ_i^{pert} are handled as forcing variables and both can

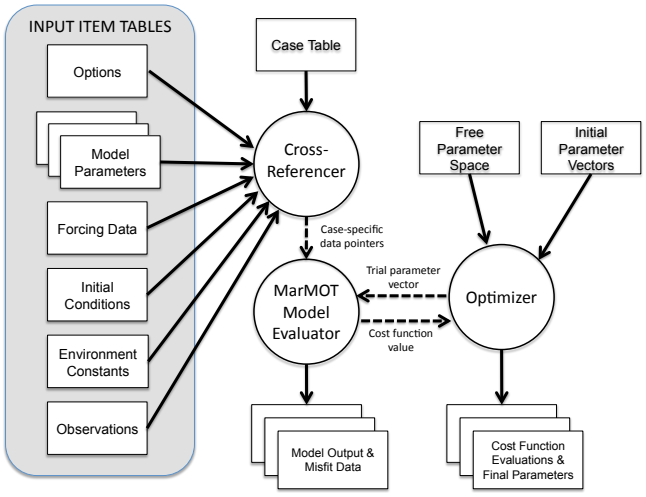

Fig. 1. Simplified schematic of the MarMOT system, showing the main system components and data flows. Data flows shown by dotted lines are purely internal.

be time- and depth-dependent. ϵ_n covaries at all depths and is scaled according to the local value of σ_i^{pert}. Any negative post-perturbation tracer concentrations are set to zero.

Each relaxation rate r_i is handled as a forcing variable, as is the reference concentration C_i^{ref} for each tracer. Any of these variables can vary independently in time and/or depth if required. A further option allows relaxation to be restricted to grid points above or below the mixed layer depth, the euphotic zone depth (1 % light level) or the greater of the two.

A number of different 1-D simulations were performed at each site in connection with the twin experiments. An overview is given in Table 2. Simulation Group A provides a synthetic climatology used to create an ensemble of initial states. Simulation Group B is an ensemble simulation giving state estimates of the form $x_{ijk\text{T}} + \epsilon_{ijk\text{ENV}}$, used to estimate the environment error variances $s^2_{ijk\text{ENV}}$. The Group B variances are for a known system; they do not take into account parameter uncertainty but serve to illustrate the impact of environmental uncertainty on the simulation. Simulation Group C gives state estimates $x_{ijk\text{T}} + \epsilon_{ijk\text{ENV}} + \epsilon_{ijk\text{P}}$. These are used to calculate parameter-independent environment error variance estimates for optimization Experiment 3. Simulation Group D provides the true system state for the true environment $x_{ijk\text{T}}$. This state is used in generating synthetic observations $x_{ijk\text{T}} + \epsilon_{ijk\text{OBS}}$ with observation error of known variance $\sigma^2_{j\text{OBS}}$. These observations are then assimilated in a set of optimization experiments comprising Simulation Group E. The observed variables are DIN, particulate organic nitrogen (PON), phytoplankton chlorophyll and primary production. For the purposes of this experiment, PON is defined as the sum of the organic nitrogen tracers (phytoplankton, zooplankton and detritus).

3.1.3 Statistical characterization of the synthetic environment

Estimates of the mixed layer depth, horizontal advection tendencies and the initial state at each site are treated as uncertain and represented by input ensembles. The methods for ensemble generation are described below. The potential impact of uncertainty in solar radiation, vertical velocity and interior vertical diffusion is not explored: the "true" values of these variables are used throughout. The true monthly means for the horizontal advection tendencies are also used, the uncertainty in these tendencies being restricted to their shorter time-scale anomalies. The vertical diffusion coefficient is set to zero so that only numerical diffusion occurs below the mixed layer.

For the mixed layer depth, the level of uncertainty is based on the assumption that time-varying mixed layer depth statistics for a 1° square area are known. Mixed layer depth at a given time is described by a log-normal distribution with mean and variance determined from the distribution of turbocline depths over all ORCA025 grid points within a 1° square area centred on each site location, using data from the scenario year (2005). The turbocline depth at each grid point is taken to be an equally likely representation of the depth of the mixed layer at the site. Mixed layer depth values are generated at 5 day intervals with no temporal inter-dependency and linearly interpolated between these times. The characteristics of the mixed layer depth input ensemble are summarized in Fig. 2.

It is assumed, for the purposes of quantifying uncertainty, that data for the horizontal advection tendencies are available from a model-based climatology in the form of depth-dependent monthly mean and standard deviation estimates and that these statistics are not strongly parameter dependent. In a real-world experiment it would be important to ensure that the statistics were consistent with the model being analyzed. Here, this is not required and advection tendency statistics derived from the ORCA025-N201 output are used. Inter-annual variability in the 3-D simulation provides separate realizations of the circulation, the statistical properties of which are taken be representative of uncertainty in our knowledge of the true circulation affecting conditions in the scenario year. The 3-D model resolution is eddy-permitting, so the advective flux divergences can be expected to represent some limited eddy diffusion effects.

All perturbations are applied in transformed tracer space so are concentration dependent. A square root transformation was chosen for all tracers at all sites, giving a rate of change for tracer concentration C_i of

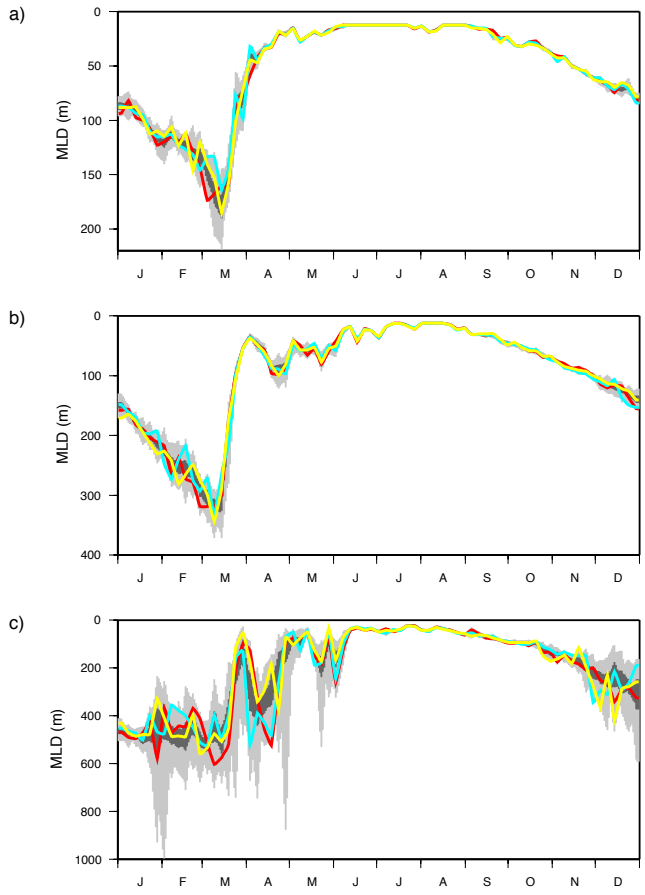

Fig. 2. Illustration of 100 member mixed layer depth ensemble at **(a)** BATS, **(b)** NABE and **(c)** OWS-INDIA sites, showing full ranges (light grey), inter-quartile ranges (dark grey) and three example members (coloured).

$$p_i = 2\sqrt{C_i}\, p_i^\star \tag{14}$$

in response to a perturbation $p_i^\star$ applied to $\sqrt{C_i}$. The choice of tracer transformation was a compromise supported by a Box and Cox (1964) analysis in which a maximum likelihood method is used to determine the optimum variance-stabilizing transformation from those available in MarMOT (log, square root or none). The applied perturbation was derived from the advection tendency of the transformed tracer as determined from the 5 day mean concentration and velocity fields output by the 3-D model, so

$$p_i^\star = -\boldsymbol{u}_\mathrm{h} \cdot \nabla_\mathrm{h} \sqrt{C_i} \tag{15}$$

where the subscript h denotes vectors in the horizontal plane and $\boldsymbol{u}_\mathrm{h}$ is the current velocity. This is calculated for all times and depth levels over 15 yr of the 3-D simulation (1991–2005) and binned by month to obtain statistics μ_i^{pert} and σ_i^{pert} for one annual cycle. The resulting tracer perturbation input fields for each site are shown in Figs. 3 and 4. Different realizations of the perturbation rate anomaly, consistent with σ_i^{pert}, are generated internally from different input seed values. A 24 h auto-correlation coefficient of 0.5 is used for all simulations.

There is clearly strong correlation between state variables in the mean advection tendencies represented in Fig. 3. Correlation structure arising from the plankton dynamics would likewise be expected in any anomalies, although the present MarMOT system only generates perturbation rate anomalies for different variables independently. Functionality to introduce correlation structure on the basis of input statistics would be a useful extension.

For the initial state, it is assumed that multi-variate monthly climatological statistics are available for all tracers at depths of 5, 10, 20, 40, 60, 80, 100, 150, 200, 250, 300, 500, 750 and 1000 m. A synthetic climatology is created for each site from a 15 yr HadOCC integration to the start of the scenario year with the true parameter set (Simulation Group A). Excessive model drift due to absent horizontal processes is avoided by relaxing the DIN tracer towards climatology at all depths below the combined mixed layer and euphotic zone, with a 60 day relaxation time scale ($r = 0.0167\,\mathrm{d}^{-1}$). The reference concentrations for relaxation are given by local annual mean nitrate profiles from the World Ocean Atlas (Garcia et al., 2010) and the 15 yr integrations are initialized from a steady state annual cycle obtained from repeat integrations of the first year. Monthly statistics from the resulting climatology are used to construct a probability model for randomly generating system states as needed, preserving vertical covariances and covariances between tracers as characterized by the first 5 principal components of the anomalies. These explain 76 %, 62 % and 74 % of the variance at BATS, NABE and OWS-INDIA sites, respectively. A multi-variate state representative of December or January is selected with equal probability to initialize simulations at the start of the calendar year. The main characteristics of the initial state input ensemble are summarized in Fig. 5.

3.1.4 Environment error for a known system

Given a statistical characterization of the input data, the expected environment error in the simulation is dependent on the plankton model and its parameter values. Estimates of the environment error fields for a known system, specifically the HadOCC model with default parameters (Table D1), are given by a 100 member ensemble simulation at each site (Simulation Group B). A square root transformation is applied to each observed variable, on the basis of a Box-Cox analysis (Box and Cox, 1964), to stabilize the ensemble variance. The ensemble standard deviation for each transformed variable gives an estimate of its expected r.m.s. environment error. Estimates are shown in Fig. 6 as a function of depth and time.

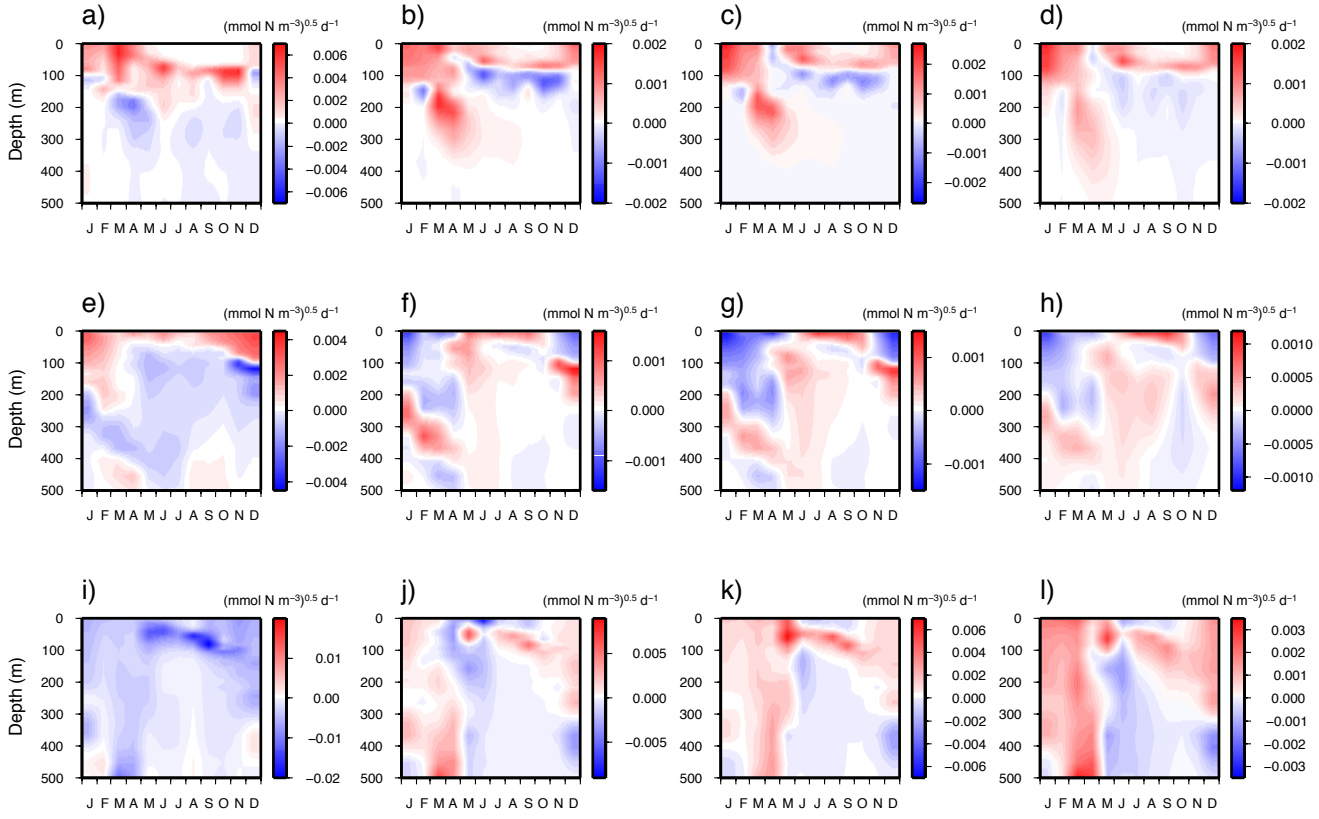

Fig. 3. Perturbation rate mean μ_i^{pert} for transformed state variables. For the BATS site: **(a)** DIN ($\sqrt{N}$), **(b)** phytoplankton ($\sqrt{P}$), **(c)** zooplankton ($\sqrt{Z}$) and **(d)** detritus ($\sqrt{D}$). **(e–h)** Same variables at the NABE site. **(i–l)** Same variables at the OWS-INDIA site.

There are particular patterns in Fig. 6 that are directly linked with uncertainty in mixed layer depth (Fig. 2) during seasonal deepening of the boundary layer. At NABE and OWS-INDIA, clear bands of high standard deviation in transformed PON and chlorophyll are evident in the region of the maximum mixing depth from late summer onwards. These are also seen at BATS and NABE from January to March where corresponding bands are present in the DIN plots. In contrast, at OWS-INDIA where there is much greater variability in mixed layer depth over the ensemble, there are no obvious peaks in the depth distributions of the ensemble standard deviation over the winter period. At OWS-INDIA particularly high ensemble variance occurs in transformed DIN as the mixed layer deepens in the autumn. This extends throughout the boundary layer and appears to be the result of high variability in the advective DIN tendency (see Fig. 4), much of which is above the mixed layer depth. Variability in DIN flux divergence is similarly high at BATS at this time but below the mixed layer depth, contributing to a sub-surface band of high simulation variance in DIN from spring through to the end of the year. Other high variance patterns in late spring and early summer appear to be associated with the biological response to spring shoaling of the mixed layer. These are symptomatic of more complex interactions between the variance in the input ensemble and the biological dynamics.

Another important point with respect to the transformed DIN ensemble standard deviation is its strong increase over the year at OWS-INDIA. Here, the ensemble variance is much higher over the full depth range at the end of December than at the beginning of the year. The situation is similar at BATS, although less obvious. In contrast, the DIN pattern at NABE is much more suggestive of a repeatable annual cycle. The presence of the net growth in error variance at BATS at depths down to 400 m, well below the ensemble maximum in the mixed layer depth, suggests that much of the error growth is due exclusively to the variance in the advective tendencies. At OWS-INDIA where the winter mixing is deeper and the ensemble variance in mixed layer depth is greater, it is likely to be the result of strong interaction between the effects of the variances in the advective tendencies and the mixed layer depth.

At both BATS and OWS-INDIA, the error growth may be due in part to deficiencies in the statistical representation of the advective tendencies. This should be further investigated with a view to possible refinement of the boundary condition.

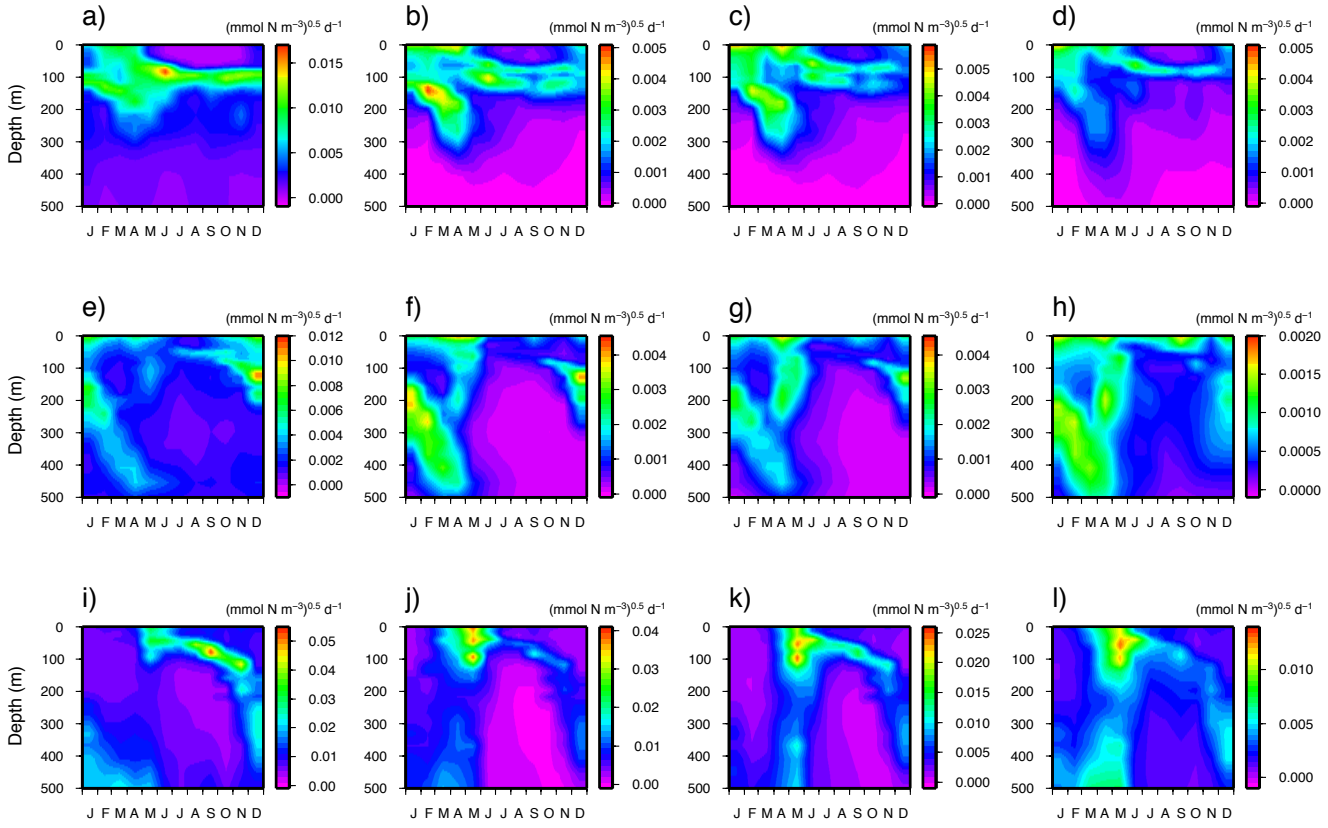

Fig. 4. Perturbation rate standard deviation σ_i^{pert} for transformed state variables. For the BATS site: **(a)** DIN ($\sqrt{N}$), **(b)** phytoplankton ($\sqrt{P}$), **(c)** zooplankton ($\sqrt{Z}$) and **(d)** detritus ($\sqrt{D}$). **(e–h)** Same variables at the NABE site. **(i–l)** Same variables at the OWS-INDIA site.

In particular, the use of a square root transformation in the tendency calculation (Eq. 15) may not be appropriate over all times and depths at which it is applied. Preliminary analyses of the 3-D biogeochemical simulation suggest that the tendencies might be better represented using a variable power law transformation that adapts to time and depth variations in their probability distributions.

3.1.5 Parameter-independent environment error

The proposed cost function formulation for parameter optimization (Eq. 6) is based on the assumption that the environment error and parameter error are independent. If this were truly the case, then the environment error variances determined for our known system could be applied to the minimization problem. However, in practice we expect dependencies to exist. The problem is alleviated to an extent by considering parameter uncertainty in the derivation of the environment error variance.

An estimate of $s^2_{ijk\text{ENV}}$ that is not dependent on a particular parameter set was determined by pooling variances calculated for 100 different parameter vectors in the 5-dimensional parameter space described by the bounds given in Table 1. The choice of bounds is arbitrary and solely for the purposes of evaluating the cost function designs. The vectors were chosen 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 pairs of sample points. For each parameter vector, the error variance estimates were determined using 100 realizations of the environment, requiring 10 000 simulations at each site in Simulation Group C.

The parameter-independent field estimates from the 10 000 member ensemble are shown in Fig. 7. The differences between the error standard deviation patterns shown in Figs. 6 and 7 give an indication of the effect of parameter uncertainty. While the patterns are broadly similar, it is clear that many of the details are sensitive to the parameter values, suggesting that the use of parameter-specific environment error estimates in the cost function could be beneficial. This option would be computationally more expensive and is not explored in the optimization experiments presented here.

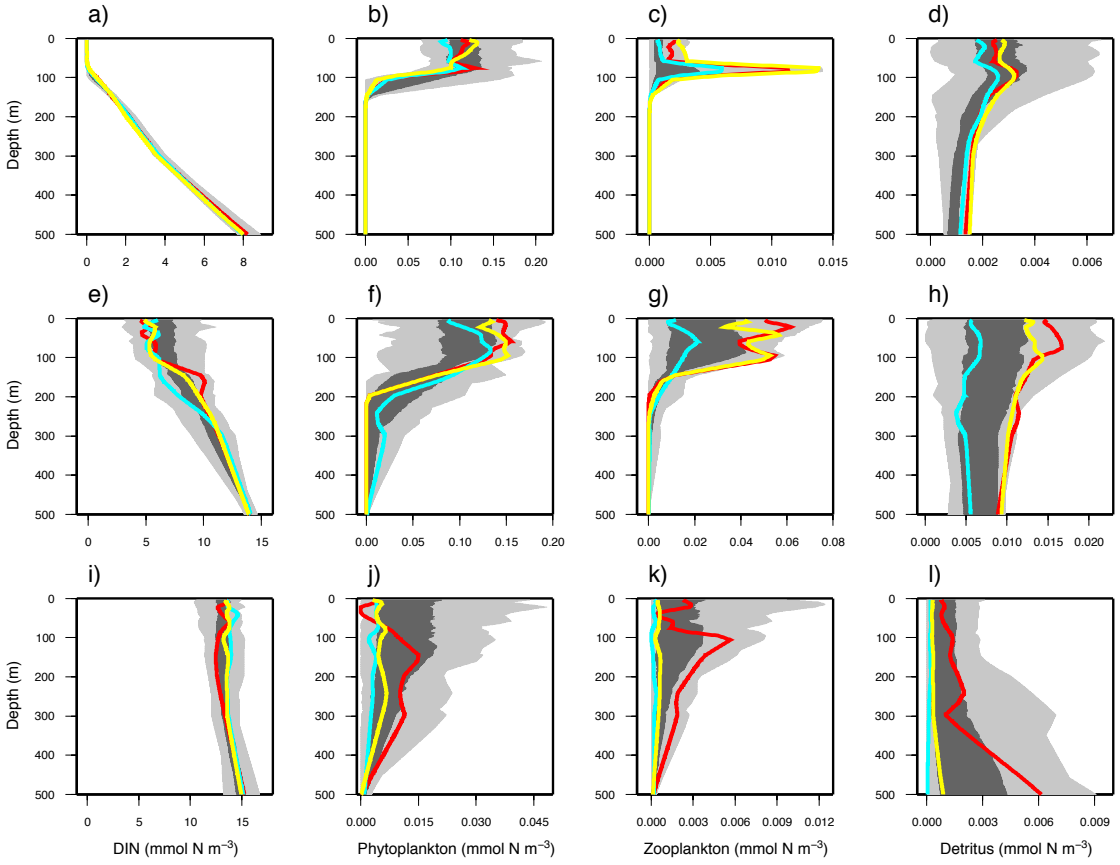

Fig. 5. Illustration of the 100 member initial state ensemble. For the BATS site: **(a)** DIN (N), **(b)** phytoplankton (P), **(c)** zooplankton (Z) and **(d)** detritus (D). **(e–h)** Same variables at the NABE site. **(i–l)** Same variables at the OWS-INDIA site. Full ranges (light grey), inter-quartile ranges (dark grey) and three example members (coloured) are shown.

3.1.6 Synthetic observations

The observations for the scenario year are generated from a simulation with the true environment (Simulation Group D) by sampling the output and adding observation errors. The resulting observation data set comprises monthly DIN and PON concentrations at depths of 10, 30, 50, 100, 200, 300 and 500 m, monthly chlorophyll concentrations and primary production fluxes at 10, 30, 50, 100 and 200 m and upper mixed layer chlorophyll concentrations at 5 day intervals. Plausible errors are applied to square root or log transformed values as specified in Table 3. For the log-transformed biological variables (chlorophyll, PON and primary production) the error standard deviations are derived from nominal relative errors by averaging positive and negative errors in log space. The actual relative errors are shown in brackets.

In Experiment 3, all model-data differences are calculated in square root space. The observation error for the log-transformed variables is specified in $\log_{10}$ space and the expected error in square root space depends on the untransformed observation value $\Omega = y_{ijk}^2$ according to

$$s_{ijk\mathrm{OBS}} = \ln(10)\frac{\sqrt{\Omega}}{2}\sigma_{j\mathrm{OBS}}. \quad (16)$$

For DIN:

$$s_{ijk\mathrm{OBS}} = \sigma_{j\mathrm{OBS}}. \quad (17)$$

While the presence of significant correlation structure in the simulation error between sample points is acknowledged, no allowance is made for covariances in the cost function weighting. The adverse effects of this limitation are reduced by removing duplicate simulation values that occur at multiple sampling depths within the upper mixed layer. Where this occurs, all mixed layer observations below 10 m are excluded.

3.1.7 Parameter optimization

The MarMOT optimizer is well suited to non-linear problems in multi-dimensional parameter space: it includes a genetic algorithm for identifying promising areas of a bounded parameter space and a non-gradient direction set algorithm for

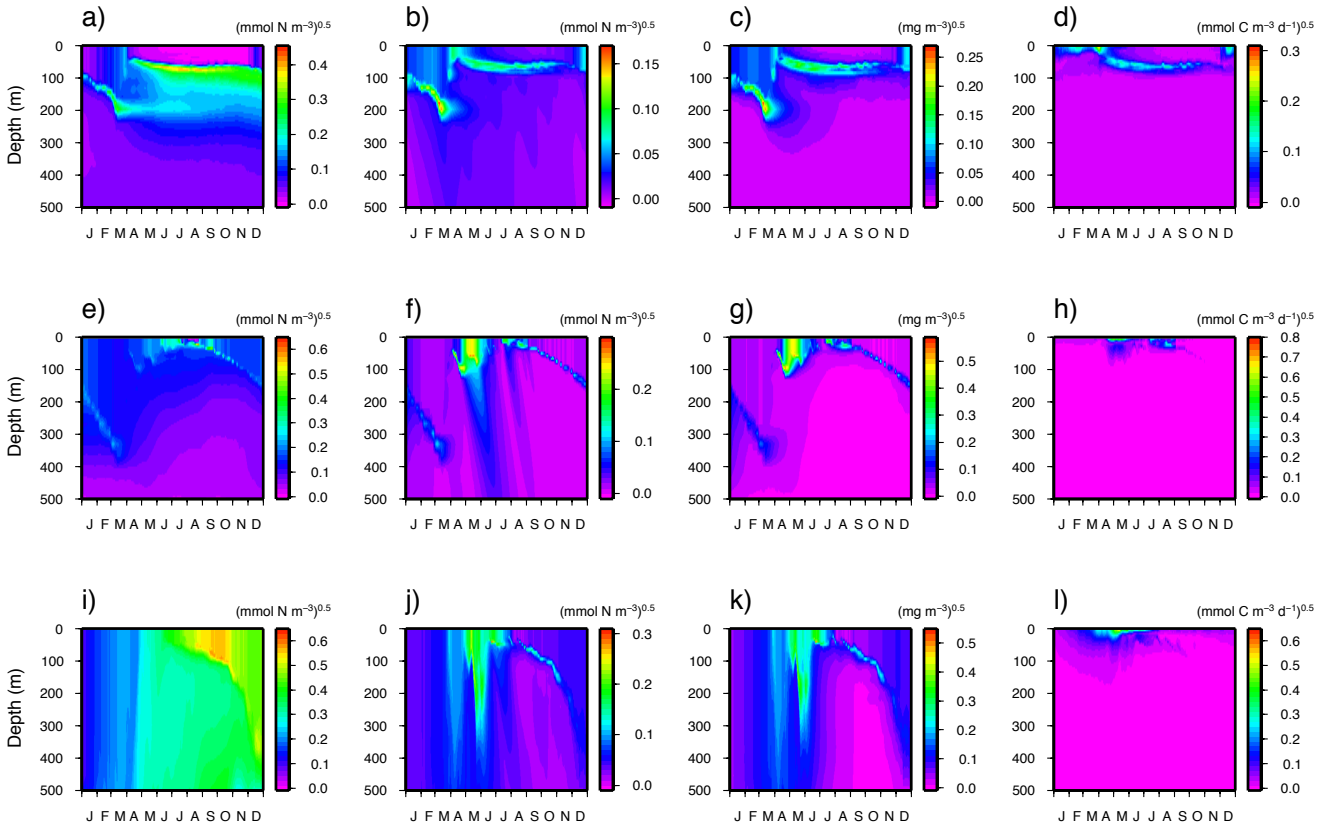

Fig. 6. Ensemble standard deviation of square-root transformed variables from Simulation Group B: estimated environment error for the HadOCC model with the default parameter set. For the BATS site: **(a)** DIN ($\sqrt{N}$), **(b)** PON ($\sqrt{P+Z+D}$), **(c)** Chlorophyll ($\sqrt{12.01(\frac{\theta_P}{\theta_{chl}})P}$) and **(d)** primary production ($\sqrt{\bar{\mu}_P \theta_P P}$). **(e–h)** Same variables at the NABE site. **(i–l)** Same variables at the OWS-INDIA site.

bounded or unbounded local minimization. The two algorithms can be used in combination or independently. The genetic algorithm is a global method in the sense that it is able to locate multiple minima in the cost function. 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. Local algorithms can be applied to global problems by performing repeated searches from different initial points in parameter space to increase the likelihood of locating the global minimum.

The genetic algorithm provided 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), Weber et al. (2005) and Kettle (2009) and by 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 in Appendix C. 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 (Fasham and Evans, 1995; Fasham et al., 1999; Evans, 1999; Hemmings et al., 2003, 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. On any convergence in the parameter vector population,

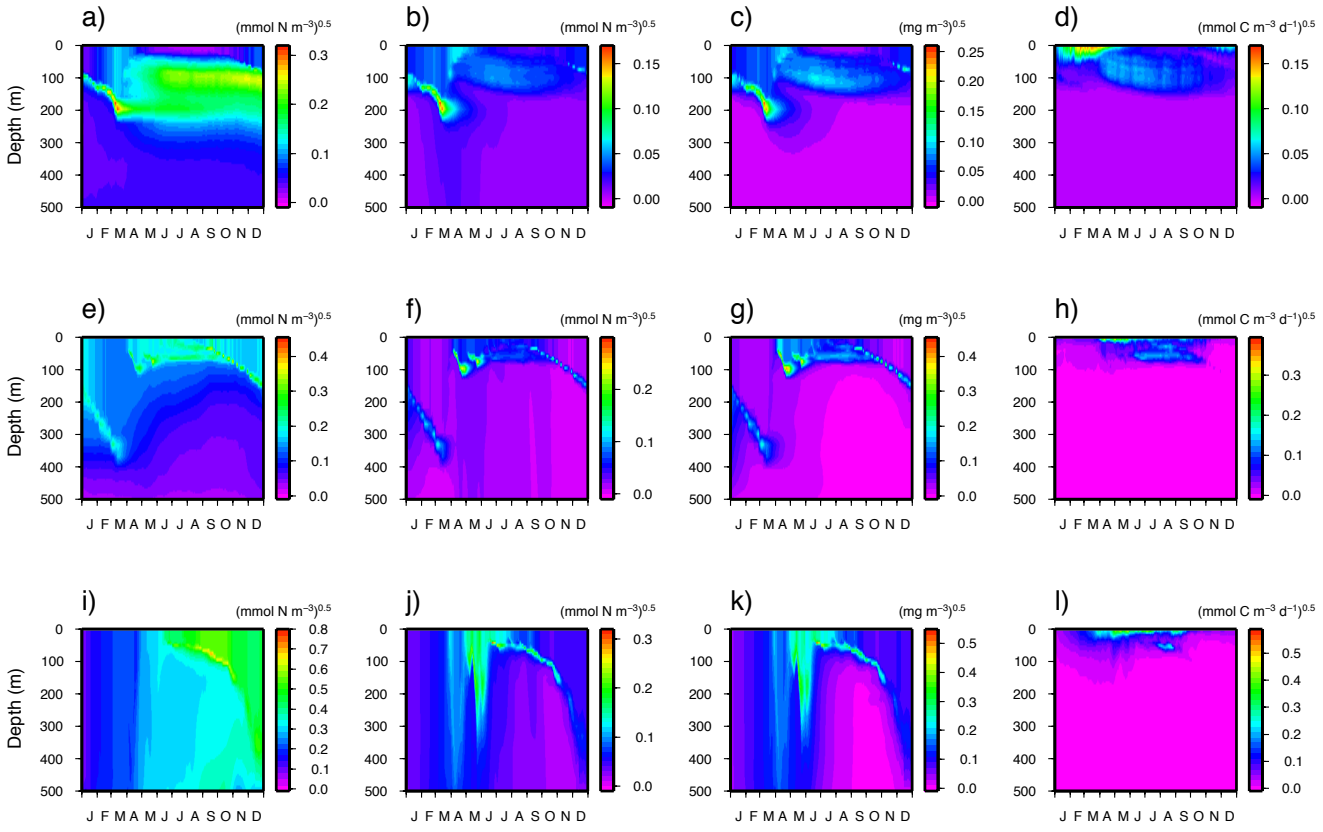

Fig. 7. Ensemble standard deviation of square-root transformed variables from Simulation Group C: estimated environment error $s_{ijk\mathrm{ENV}}$ applicable to the HadOCC free parameter space defined by the μGA optimizer bounds. For the BATS site: **(a)** DIN ($\sqrt{N}$), **(b)** PON ($\sqrt{P+Z+D}$), **(c)** Chlorophyll ($\sqrt{12.01(\frac{\theta_\mathrm{P}}{\theta_\mathrm{chl}})P}$) and **(d)** primary production ($\sqrt{\bar{\mu}_\mathrm{P}\theta_\mathrm{P}P}$). **(e–h)** Same variables at the NABE site. **(i–l)** Same variables at the OWS-INDIA site.

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 uniform cross-over between bit strings at a probability of 0.5. Bounds are required for the μGA but were removed for the local search to avoid enforcing artificial constraints when locating minima close to the boundaries of the parameter space. Log transformations were used to prevent parameters taking negative values. An exception was made for the detrital sinking velocity for which bounds were retained in the local search to avoid potential problems with numerical instability. Details of the free parameter space are summarized in Table 1. Within the μGA, each parameter was represented by 8 bits giving 256 possible values prior to refinement by the local searches.

The population size for the μGA was 5, 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. The direction set algorithm was applied to each unique parameter vector in the final population 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 5 alternative designs, choosing those with the largest minimum Euclidean distances from a sample of 500 randomly generated hypercubes.

A single set of three optimization experiments is referred to as Simulation Group E. Simulation Group E was repeated for 10 different realizations of the optimization environment. Because the mixed layer depth varies between different realizations of the environment error, a slightly different observation set is used for each set of experiments. A further set of three optimization experiments was performed using the true environment.

Table 2. Overview of 1-D plankton model simulations.

Simulation group id	Product(s)	Time period	Simulations at each site	Model parameters	Initial state	Forcing (ORCA025)	Boundary condition
A	initial state statistics	1990–2004	1	true parameter vector	1990 repeat cycle	on-site data	DIN relaxation to climatology
B	expected environment error for true system	2005	100 member environment ensemble	true parameter vector	initial state ensemble (100 members)	on-site solar rad., w_p MLD ensemble (100 members)	perturbation ensemble (100 members)
C	estimated parameter-independent environment error	2005	100 member environment ensemble $\times$ 100 param. vectors	sample from parameter space (100 vectors)	initial state ensemble (100 members)	on-site solar rad., w_p MLD ensemble (100 members)	perturbation ensemble (100 members)
D	observation set	2005	1 (true environment)	true parameter vector	1 initial state realization	on-site solar rad., w_p 1 MLD realization	1 perturbation realization
E	optimal parameter vectors (Expts. 1–3)	2005	1 optimization environment, trial parameter vectors	free parameter space	1 initial state realization	on-site solar rad.,w_p 1 MLD realization,	1 perturbation realization

Table 3. Observation errors.

Observation type	HadOCC equivalent	Transformation	Error std. dev.	Relative error
DIN	N	sqrt	0.05 (mmol N m^{-3})$^{0.5}$	variable
PON	$P+Z+D$	log	0.239 $\log_{10}$ units	50 % (−42 %, +73 %)
Surface chlorophyll	$12.01\frac{\theta_\mathrm{P}}{\theta_\mathrm{chl}}P$	log	0.159 $\log_{10}$ units	35 % (−31 %,+44 %)
Sub-surface chlorophyll	$12.01\frac{\theta_\mathrm{P}}{\theta_\mathrm{chl}}P$	log	0.088 $\log_{10}$ units	20 % (−18 %,+22 %)
Primary production	$\bar{\mu_\mathrm{P}}\theta_\mathrm{P}P$	log	0.184 $\log_{10}$ units	40 % (−35 %,+53 %)

3.2 Results

3.2.1 Cost function minimization

Results of the cost function minimization procedure in each optimization experiment are shown in Table 4, together with the cost function value for the true parameter vector $J(\boldsymbol{P}_\mathrm{true})$. The initial minima and maxima show the range of the cost J over a super population of 25 parameter vectors, comprising the 5 distinct initializations of the μGA population. The final cost range is that for the 5 output parameter vectors, each being the lowest cost vector for one μGA initialization after local minimization. Final cost ranges are small indicating low sensitivity to the details of optimizer initialization.

In Experiment 3, the final cost values $J(\boldsymbol{P}_\mathrm{opt})$ and the true parameter costs $J(\boldsymbol{P}_\mathrm{true})$ both tend to be close to unity in the presence of environmental error. The costs for the other optimization experiments are consistently larger, indicating that the level of uncertainty present is greater than that allowed for in the cost function design. If the true parameter vector were not known a priori, there would be a risk of such high costs leading to rejection of the true hypothesis. Cost function values are particularly large in Experiment 2. This is a consequence of relative errors in organic tracer concentrations that are much larger than the small observation errors associated with small concentrations. The effect can be attributed to our simple treatment of observation error, which inevitably underestimates expected error as the observed concentration tends to zero. A more sophisticated treatment would be to represent the error as a sum of absolute and relative terms as done by Schartau et al. (2001). Where the true environment is used, Experiment 2 gives cost values close to unity ($J(\boldsymbol{P}_\mathrm{opt}) = 1.2$ and $J(\boldsymbol{P}_\mathrm{true}) = 1.2$) since the weighting used is consistent with the uncertainty present. In

Table 4. Cost minimization.

Optimization Experiment	Environment	Initial Cost Minimum	Initial Cost Maximum	Final Cost $J(\boldsymbol{P}_{\text{opt}})$	Final Cost Range	True Parameter Cost $J(\boldsymbol{P}_{\text{true}})$	Cost Difference $J(\boldsymbol{P}_{\text{opt}}) - J(\boldsymbol{P}_{\text{true}})$
1	1	6.5	68	3.9	0.0002	4.5	−0.5
1	2	7.5	80	5.1	7×10^{-5}	6.2	−1.1
1	3	6.8	73	5.1	0.0005	8.0	−3.0
1	4	5.6	64	3.6	0.02	6.4	−2.9
1	5	10.2	95	8.3	0.004	20.6	−12.3
1	6	7.0	79	4.9	7×10^{-5}	8.5	−3.6
1	7	6.5	85	4.6	0.0002	5.7	−1.0
1	8	11.0	102	9.1	0.01	9.7	−0.6
1	9	14.2	91	12.1	0.009	14.4	−2.3
1	10	8.5	81	6.3	0.1	6.8	−0.5
	MEAN	8.4	82	6.3	0.02	9.1	−2.8
1	TRUE	5.3	63	2.6	1×10^{-5}	2.9	−0.3
2	1	27.9	99	23.6	0.005	23.7	−0.1
2	2	40.5	108	37.2	0.008	37.6	−0.4
2	3	29.6	98	26.3	0.4	27.2	−0.9
2	4	17.8	89	12.6	0.02	14.5	−1.9
2	5	28.8	112	25.5	0.6	27.7	−2.2
2	6	20.6	85	15.9	0.01	16.2	−0.3
2	7	40.1	108	32.4	0.02	34.9	−2.4
2	8	36.1	98	33.2	0.003	33.5	−0.4
2	9	49.9	115	48	0.1	49.8	−1.8
2	10	24.9	95	20.1	0.0008	20.4	−0.2
	MEAN	31.6	101	27.5	0.1	28.5	−1.1
2	TRUE	5.6	78	1.2	1×10^{-5}	1.2	0.0
3	1	1.71	15.1	1.04	2×10^{-5}	1.07	−0.03
3	2	1.90	14.9	1.12	2×10^{-5}	1.26	−0.14
3	3	1.68	14.4	1.01	1×10^{-5}	1.03	−0.02
3	4	1.62	12.7	0.95	0.0002	1.32	−0.38
3	5	2.09	16.3	1.40	2×10^{-5}	1.64	−0.24
3	6	1.76	13.7	1.01	2×10^{-5}	1.07	−0.07
3	7	1.82	12.7	1.13	1×10^{-5}	1.17	−0.05
3	8	2.06	14.3	1.30	0.0007	1.34	−0.04
3	9	2.14	15.0	1.57	0.0002	2.02	−0.45
3	10	1.94	13.0	1.31	0.0009	1.33	−0.02
	MEAN	1.87	14.2	1.18	0.0002	1.33	−0.14
3	TRUE	1.43	13.5	0.53	3×10^{-6}	0.54	0.00

contrast, the corresponding Experiment 3 results show much lower costs ($J(\boldsymbol{P}_{\text{opt}}) = 0.53$, $J(\boldsymbol{P}_{\text{true}}) = 0.54$).

The final costs are always less than $J(\boldsymbol{P}_{\text{true}})$ except when the true environment is used, indicating some degree of over-fitting. This is expected where the cost function is distorted by error in the observations or environmental inputs but should be reduced by an effective weighting scheme. The cost differences $J(\boldsymbol{P}_{\text{opt}}) - J(\boldsymbol{P}_{\text{true}})$ suggest that over-fitting is worst in Experiment 1, with a mean cost difference of −2.8 in the presence of non-zero environment error compared with −1.1 and −0.14 in Experiments 2 and 3, respectively. The Experiment 1 mean cost difference is a factor of 20 greater than that for Experiment 3. This contrasts with factors of about 4 and 6 for the initial cost minima and maxima respectively, so is not simply due to a parameter-independent scaling of the cost function. It should also be noted that for 5 out of 10 environment error realizations, the Experiment 1 cost function is greater at the location of the true parameter vector than the cost function minimum found prior to any application of the optimizer. This is a clear indication of a high over-fitting risk not seen in the Experiment 2 or 3 results.

3.2.2 Parameter recovery

The final parameter values obtained in each experiment for each input environment are shown in Fig. 8. All distinct

Table 5. Posterior parameter errors.

Parameter	True	Unit	R.M.S. Error			Bias		
	Value		Expt. 1	Expt. 2	Expt. 3	Expt. 1	Expt. 2	Expt. 3
α_{surf}	5.56	$\text{mg C (mg Chl)}^{-1}\ (\text{E m}^{-2})^{-1}$	1.80 (32 %)	0.78 (14 %)	0.48 (8.7 %)	−1.62 (−29 %)	+0.11 (2 %)	−0.41 (−7 %)
k_{N}	0.1	mmol N m^{-3}	0.056 (56 %)	0.045 (45 %)	0.021 (20 %)	−0.016 (−16 %)	−0.002 (−2 %)	+0.001 (1 %)
g_{max}	0.8	d^{-1}	0.55 (68 %)	0.48 (60 %)	0.39 (48 %)	+0.37 (45 %)	+0.24 (30 %)	+0.18 (23 %)
m_2	0.3	$\text{d}^{-1}(\text{mmol N m}^{-3})^{-1}$	0.59 (195 %)	0.41 (136 %)	0.31 (103 %)	+0.45 (149 %)	+0.20 (66 %)	+0.16 (52 %)
w_{D}	10	m d^{-1}	12.1 (121 %)	1.8 (18 %)	1.4 (14 %)	+7.1 (71 %)	−0.3 (−3 %)	+0.9 (9 %)

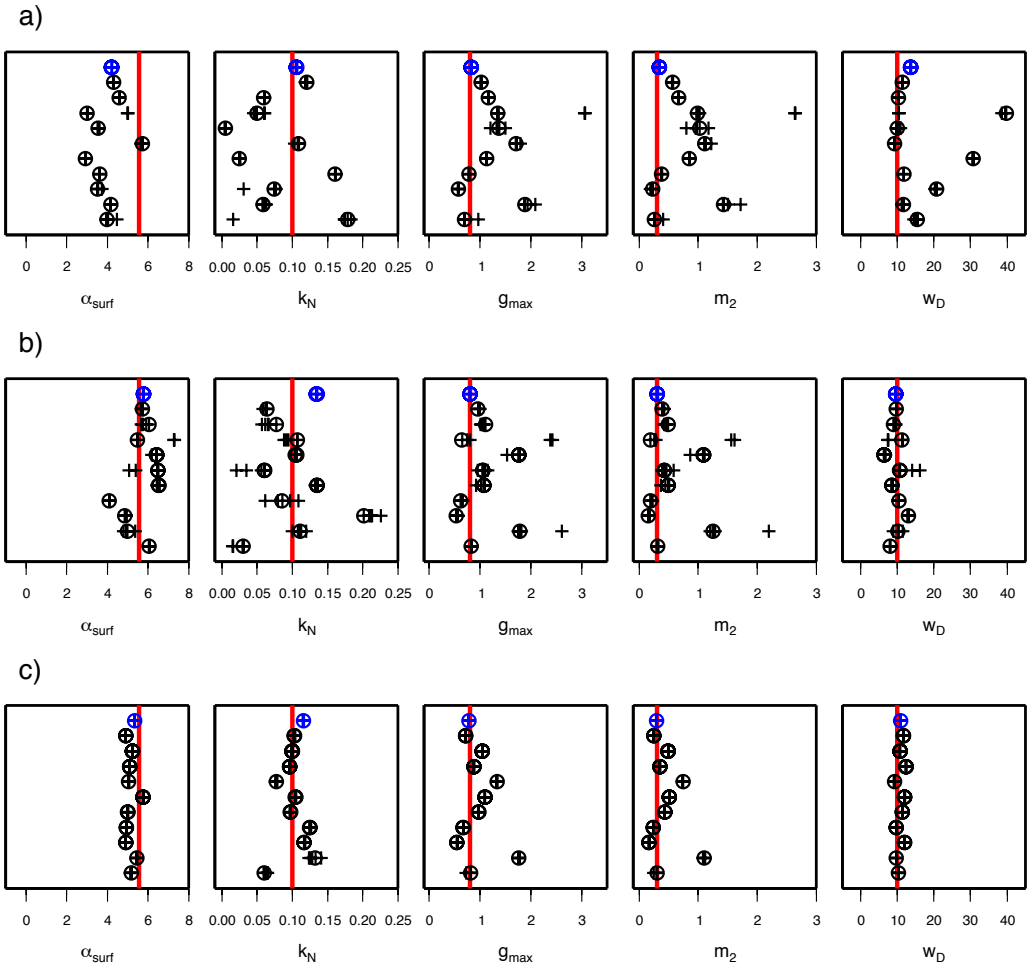

Fig. 8. Parameter recovery results for **(a)** Experiment 1, **(b)** Experiment 2 and **(c)** Experiment 3. Red lines represent the true values for each parameter. Crosses in each row show optimizer output parameter values for the true environment (blue) and for each of the 10 realizations of the optimization environment (black). One cross is shown for each distinct parameter value obtained with 5 different optimizer initialization cases. Optimal values are circled. Crosses not highlighted thus are values associated with higher cost function values.

values are shown for each of the 5 optimizer initialization cases but only the optimal values (those associated with the minimum costs) are highlighted. Table 5 gives summary statistics for the parameter recovery errors over all non-zero realizations of the environment error. Where multiple initialization of the optimizer produced more than one final parameter vector with the same cost (to 6 significant digits), parameter values are first averaged to give a single value for each environment error realization.

Parameter recovery is generally improved in Experiment 3, where both error sources are accounted for. There is also less sensitivity in final parameter values to the initial μGA population The Experiment 2 design, where the weights are based on observation error, performs better than the characteristic scale weighting used in Experiment 1,

particularly with regard to parameter biases. In Experiment 1, the initial P-E slope α_{surf} is estimated low for all but 1 case of the environment error and has a strong negative bias (−29 %). There are also some very high estimates of the sinking velocity parameter w_{D}, leading to a 71 % bias. Furthermore, the r.m.s. errors in the final parameter values show the expected error to be consistently higher than for the other two experiments. In Experiment 2, although the r.m.s. errors are consistently higher than those for Experiment 3, the biases are smaller for 2 of the 5 parameters suggesting some room for improvement in the environment error weighting.

Closer inspection of the values for the parameters $g_{\max}$ and m_2 from all experiments shows them to be highly correlated. This is perhaps unsurprising considering their role in the model dynamics, since the maximum grazing rate $g_{\max}$ impacts directly on nitrogen transfer into the zooplankton pool and the density-dependent mortality m_2 impacts directly on transfer out. It is thus possible to compensate for excessively high values of one parameter by high values of the other, keeping zooplankton nitrogen stable. This leads to a positive bias in both parameters. High values do increase the throughput of nitrogen from phytoplankton food to DIN potentially impacting on chlorophyll and DIN observations but this effect is attenuated by recycling which fuels more phytoplankton growth. Nevertheless, other features of the system make some observational constraint possible. It is notable that the cost function design in Experiment 3 appears more robust in the face of this correlation tendency between parameters than either of the other designs.

3.2.3 Impact of parameter error

To examine the implication of the parameter recovery errors for model estimates of key carbon fluxes, simulations were run with each of the 10 optimal parameter vectors using the true environment. Table 6 gives error statistics, over this 10 member ensemble, for the annual mean primary production integrated over the water column at each site. Corresponding estimates of the export flux of sinking particles to the ocean interior are given in Table 7. The export is represented by the downward flux of particulate carbon at a site-dependent reference depth z_{ref}. The flux is $w_{\mathrm{D}}\theta_{\mathrm{D}}D(z_{\mathrm{ref}})$, that is the product of the sinking velocity, carbon:nitrogen ratio and nitrogen concentration of the detritus. z_{ref} is set at 250, 400 and 100 m for BATS, NABE and OWS-INDIA respectively, just below the maximum depth of winter mixing for all ensemble members.

In all experiments, the sinking particle flux r.m.s. errors and biases are consistent across sites and strongly reflect the statistics for the sinking rate parameter w_{D}. While particle flux is also affected by error in the detritus concentration $D(z_{\mathrm{ref}})$, such errors are not consistent over the year so have a relatively small impact on the annual mean.

The r.m.s. errors in both primary production and sinking particle flux are lowest for the Experiment 3 parameter vectors and highest for the Experiment 1 parameter vectors at all sites. In contrast, the biases are generally smallest for Experiment 2, rather than Experiment 3, with the sinking particle flux biases being less than half those given by the Experiment 3 parameter vectors. This underlines the need for further refinements to the new weighting scheme, despite its improved performance generally over both established schemes. The characteristic scale weighting used in the Experiment 1 cost function leads to r.m.s. errors in primary production due to environment error of 14–20 %. The corresponding errors with the new method are reduced by a factor of about 3 at each site. The sinking particle flux errors when the characteristic scale weighting is used are more serious at 122–128 %. These are reduced by an order of magnitude in Experiment 3. The twin experiment configuration is of course idealistic. It may not be possible to achieve such improvements in real-world experiments, where characterization of uncertainty is a much more difficult problem. Nevertheless, the poor performance of the widely used characteristic scale method in the presence of a fairly modest amount of synthetic environment error, combined with error in the observation data set, should be seen as a strong motivation for developing reliable statistical characterizations for both sources of uncertainty.

4 Discussion

4.1 Uncertainty in model calibration

The new cost function weighting scheme tested here clearly has the potential to perform well against existing schemes in the presence of environment error. In particular, it is seen that the traditional schemes are prone to over-fitting in the presence of environment error, leading to relatively poor parameter recovery. Over-fitting occurs as the optimizer attempts to adjust parameter values to compensate for the environment error. A key feature of the new scheme is that the relative importance of individual misfits is reduced at data points where the impact of environmental uncertainty in the solution is expected to be high. The posterior solution is less constrained by the data in these areas but the overall constraints are more appropriate, reflecting our knowledge of the uncertainty introduced into the system and the system response. The risk that a large environment error value will have a detrimental impact of the calibration is thereby reduced. The results obtained with the new scheme provide strong evidence that the more appropriate weighting can reduce the problem of overfitting.

The possibility of further improvements should be investigated by refining the scheme to use parameter-dependent simulation error variances. Ideally, simulation error variances would be computed for all trial parameter vectors in an optimization experiment, but the computational cost of this solution is high. A less expensive alternative would be

Table 6. Error in annual mean primary production.

	Production (mmol C m^{-2} d^{-1})					
	R.M.S. Error			Bias		
Optimization Experiment	BATS	NABE	INDIA	BATS	NABE	INDIA
1	1.7 (14 %)	8.5 (20 %)	4.4 (20 %)	−1.3 (−11 %)	−7.3 (−18 %)	−3.6 (−16 %)
2	1.0 (8 %)	3.7 (9 %)	2.6 (12 %)	+0.3 (2 %)	+0.6 (1 %)	+0.7 (3 %)
3	0.5 (4 %)	2.4 (6 %)	1.5 (7 %)	−0.3 (−2 %)	−1.7 (−4 %)	−0.7 (−3 %)

Table 7. Error in annual mean sinking particle flux.

	Particle Flux at Reference Depth (mmol C m^{-2} d^{-1})					
	R.M.S. Error			Bias		
Optimization Experiment	BATS (250 m)	NABE (400 m)	INDIA (1000 m)	BATS (250 m)	NABE (400 m)	INDIA (1000 m)
1	0.154 (125 %)	0.879 (122 %)	0.825 (128 %)	+0.089 (73 %)	+0.512 (71 %)	+0.48 (74 %)
2	0.022 (18 %)	0.128 (18 %)	0.116 (18 %)	−0.004 (−3 %)	−0.024 (−3 %)	−0.021 (−3 %)
3	0.016 (13 %)	0.093 (13 %)	0.086 (13 %)	+0.010 (8 %)	+0.056 (8 %)	+0.051 (8 %)

to use a sample of simulation error variances calculated for different points in the parameter space, as in the analysis of our Simulation Group C, selecting the nearest neighbour for each trial parameter vector. A further refinement likely to be beneficial is the inclusion of simulation error covariances in the cost function weighting scheme.

In a real-world context, obtaining reliable statistical characterizations of the required environmental input data will be a major challenge. These are required for all plankton model assessments, with or without parameter optimization. To constrain the probability distributions for these inputs we must make use of a much wider range of supporting data than is traditionally used when comparing biogeochemical model outputs with observations.

Background climatological statistics for physical forcing can be based on analyses of 3-D physical simulations. These should ideally be eddy-resolving. Furthermore, it is important that they are evaluated against observational climatologies so that information on biases can be included. Available satellite and in situ observations contemporary with the biogeochemical evaluation data can be used to further constrain the physical forcing statistics. Assimilative physical model output might also be used, although these data may be less reliable than output from free-running simulations if used to infer relationships between observed and unobserved variables. The details will depend on the performance of the balancing schemes used to preserve physical laws in the assimilation process.

Successful application of the horizontal flux divergence scheme depends on obtaining good estimates for the perturbation rate statistics for each tracer. The biogeochemical flux divergence tendencies required for model assessment are those for the trial model in a perfect 3-D physical simulation. Thus they do not exist in reality and cannot be derived directly from observations. Furthermore, they are inevitably parameter-dependent. For these reasons we are forced to rely on broad-based statistics derived from biogeochemical simulations. Multiple 3-D simulations should be analyzed to explore sensitivity to model structure and parameters with the aim of developing climatological statistics that are reasonably robust to model differences. This would allow consistent unbiased boundary conditions to be applied to any trial model configuration. The model-based background statistics should be further constrained by observations giving information about the contemporary physical environment. In situ current data can be used, if available. Otherwise, surface geostrophic current estimates derived from satellite altimetry might be used. Evidence of physical gradients from satellite sea-surface temperature or ocean colour measurements is also relevant since horizontal flux divergence is likely to be increased in frontal regions, especially if there is evidence of a cross-frontal velocity component. These types of

information can be used to modify the climatological probability distributions.

In common with the flux divergence boundary condition, the initial conditions in a model assessment could be chosen to be consistent with a spin up of the trial model in a perfect physical simulation. While the idea is conceptually appealing, a reliable characterization of this hypothetical system state is likely to be elusive. A more practical alternative is to use an estimate of the real-world state, explicitly restricting any inferences about the model to its behaviour over relatively short time scales. The state estimate would be based on observational data where possible.

In the absence of observations, initial conditions for 1-D simulations are often determined by a steady state analysis based on a repeating annual cycle. The same approach might be taken in an ensemble simulation, provided that error growth associated with uncertainty in the forcing data and boundary conditions does not prevent achievement of a statistical steady state. Here, the boundary condition should be based on estimates of real-world flux divergence tendencies. A suitable scheme is described in Appendix B2. The scheme relies on a climatological reference state. For unobserved state variables this would need to be primarily model-based with an appropriately high level of uncertainty.

For some state variables, relevant measurements exist but the relationship between model variables and the real-world observations is uncertain due to a combination of observing system limitations and simplifying assumptions made in model design. In such cases, the observational data can be use to partially constrain model-based estimates. For example, chlorophyll measurements can be used to constrain phytoplankton nitrogen subject to the uncertainty introduced by an unknown nitrogen:chlorophyll ratio. PON measurements might be used to constrain the combined phytoplankton, zooplankton and detritus variables in the HadOCC model. However, they are affected by plankton avoidance of sampling bottles so will tend to under-represent zooplankton. They could therefore be used as an upper bound estimate for the sum of phytoplankton and detrital nitrogen or a lower bound estimate for the total organic nitrogen. A similar argument was used by Fasham and Evans (1995) to compare PON observations with values derived from simulated phytoplankton, bacteria, detritus and zooplankton concentrations.

4.2 Role of the MarMOT facility

We have focussed on the application of MarMOT to model calibration but the system is also designed to be a generic tool for model assessment and inter-comparison. The aim is to provide a facility for evaluating plankton models independently from a particular host model describing the physical circulation.

Model inter-comparison may be performed separately from model assessment or models may be comparatively assessed with reference to observational data. In the first case, MarMOT provides a flexible environment for comparing the responses of alternative model designs to many different instances of their input data. Such comparisons will lead to an improved understanding of the relationships between models and the implications of different design decisions. Effective calibration will allow models to be comparatively assessed, with reference to independent observations, on the basis of their design.

The large number of adjustable parameters in most plankton models makes the inverse problem particularly challenging. Sensitivity analyses are often used as a basis for reducing the size of the adjustable parameter vector prior to formal optimization. The size and dimensionality of the input spaces involved typically limit the effectiveness of Monte Carlo methods. However, output from ensemble integrations performed in MarMOT can be used to build fast statistical emulators (O'Hagan, 2006) with which coverage can be achieved more efficiently.

Other applications include the comparison of plankton models at the level of individual processes and the provision of 1-D state estimates for specific locations of interest. Comparison at the process level is achieved by holding individual tracer concentrations constant or by fully prescribing their variation using external input fields. In addition, the scope of model inter-comparison studies can be reduced to focus on the biogeochemical interactions by applying a common photosynthesis sub-model, selecting from a number of photosynthesis light-limitation options (Appendix B1). 1-D state estimates with uncertainty measures can be determined on the basis of one or more plankton models.

MarMOT development is on-going. The software will be adapted to address some of the specific issues identified in this study, including cost function support for parameter-dependent simulation error variances and covariances. In addition, the system is being extended to support models of varying biogeochemical complexity with the aim of establishing a valuable community resource for plankton model evaluation in global and regional applications. At present, MarMOT is not generally available and queries regarding accessibility of the code should be addressed to the corresponding author.

5 Summary

Plankton models cannot easily be assessed against biogeochemical data because they are reliant on external drivers, typically provided by a physical simulation. Skill metrics are more readily derived for a coupled system (Stow et al., 2009) but these metrics provide only indirect information about the performance of the plankton model since the biogeochemical error fields are the combined result of errors in the plankton model and errors introduced by inaccurate physics. Inferences about the plankton model must be made against this background of environmental uncertainty.

The potential impact of environmental uncertainty on model calibration has been investigated here in an idealized experimental framework. It has been demonstrated that a modest amount of error introduced into a plankton model simulation via the model's external drivers can have an important detrimental effect on calibration results obtained using established cost function weighting schemes. A new weighting scheme that includes a formal treatment of simulation error due to error in the external drivers has been evaluated in the same experimental framework with promising results.

The scheme's effectiveness relies on good quality statistical characterizations of the plankton model's uncertain environmental input data to drive ensemble 1-D simulations from which environment error variances are determined. Its successful transition to a real-world situation will be challenging, requiring a major effort in uncertainty quantification for multiple drivers at each calibration site to be used. An approach to using model results in combination with the available observational data has been outlined here. Many of the required variables will inevitably remain poorly constrained due to the non-availability of suitable data and the levels of uncertainty assigned to these variables may initially be rather subjective. Nevertheless, an explicit treatment of uncertainty is likely to be beneficial in reducing the problem of overfitting and can be refined to produce more robust results as new measurements become available.

If a sound treatment of uncertainty in plankton model parameters and their environmental input data can be achieved then the plankton sub-models within more comprehensive environmental models can be assessed independently as hypotheses concerning the dominant biogeochemical processes they are designed to represent. The MarMOT software provides a flexible facility that can be readily adapted to support the developments in data assimilation and uncertainty analysis that will be needed and to ensure their applicability to a wide range of candidate models for improving our ability to understand and predict environmental change.

Appendix A

MarMOT design concepts

Figure 1 gives an overview of the MarMOT system in terms of its main components and the data flows between them. Simulations are controlled by data selected from a number of input tables, referred to as "item tables", each containing one or more instances of a particular input item. Different instances of each item are combined according to entries in a further input table: the "case table". Each case table entry defines a simulation case determined by a specific combination of input data and identified by a site name (or number) and an ensemble member name (or number).

A particular case table defines a set of simulations for one or more ensemble members at one or more sites. The set of ensemble members may vary between sites if required. Ensemble configurations for multiple sites can involve site-specific information (e.g. water depth), ensemble-member specific information (e.g. plankton model identifier in a multi-model comparison experiment), information specific to the combination of site and ensemble member (e.g. forcing data) and independent information (e.g. simulation time period). A cross-referencer links the appropriate item instances to the case table, determining the required data for each item either from an explicit reference or from the context implied by the site and/or the ensemble member. Free model parameters can be optimized over all cases in a given case table, so it is straight-forward to set up multi-site calibration experiments. Multi-member calibrations are likewise possible.

The core of the system is the MarMOT Model Evaluator (MME) that performs plankton ecosystem model runs according to the specifications in the case table. It calculates a cost function value dependent on the misfit between simulation variables and a set of observations or other reference values provided as an additional case-dependent input item. It can also provide a range of different output tables that are selected or de-selected according to user requirements.

The MME is implemented as a specific application within a system called the Generic Function Analyzer (GFAn). GFAn provides a cross-referencer for input selection and an optimizer for cost function minimization over the model parameter space. It also provides a generic data management framework that adapts to the requirements of the MME application to provide a MarMOT-specific user interface. GFAn is essentially an analysis engine with a well-defined application interface that makes all of its functionality available to any compatible application. The full functionality of both GFAn and the MME can likewise be applied to any plankton ecosystem model for which the basic input requirements are supported. This layered approach ensures the widest possible applicability of on-going improvements to the functionality of both GFAn and the MME. The GFAn code and MME user interface are written in C and the plankton model interface is in Fortran.

A1 Data management

An integrated data management system is essential for efficiently handling the diverse data requirements of different experiments. GFAn handles 3 different kinds of input data item used in MarMOT: parameter set items, gridded domain items and non-gridded domain items. Each instance of a parameter set item consists of a number of individually named values, such as plankton model parameters. There is one parameter set item for each supported plankton model, containing one or more instances of the model's parameter set. Further parameter set items provide model-independent

information. Gridded-domain items consist of one or more data arrays defined on a common regular grid with axes corresponding to one or more dimensions of the simulation domain. These are used to define the vertical grid, initial conditions, boundary conditions and forcing data. Non-gridded domain items are one or more vectors of values co-located at arbitrary points on the model domain axes. Observations or other reference data for comparison with the simulation output are input in this form.

An important design consideration is the need for the system to support complex experiments while at the same time being easily configurable for simple experiments. Individual items are optional wherever possible. Forcing data can be supplied in a number of different ways: as full-depth time-varying fields or as data fixed in space or in time or simply as environmental constants. Boundary condition data are treated likewise. Time-varying fields can be provided at any regular interval. The interval need not necessarily be the same for all variables: different forcing variables can be distributed arbitrarily among a number of different input item tables, typically one for each user-defined grid. Forcing data interpolated to the model time step is available in the simulation output.

GFAn provides multi-case support in the form of a flexible cross-referencing algorithm that determines the required data for each simulation case. This is done either by context or explicitly by using alphanumeric key variables to identify particular instances of each item. The data instances selected for each case by the cross-referencer are indicated in the log file. For each item having multiple instances, the cross-referencing method is determined by the presence or absence of an item key in the input item table. Items without keys are to be referenced contextually and their instances are identified by one or more variables referred to as case variables. In MarMOT, there are 2 case variables: site and ensemble member. Input data can be associated with a particular site or a particular ensemble member or both. Both case variables are used to identify particular simulations in the input case table and in any output tables produced.

MarMOT is configured by providing a set of input tables and optionally produces a set of output tables, in addition to the cost function value. Each table is contained in an ASCII file. For each input item, a table is expected with one entry for each instance of the data. For domain items, this table contains metadata describing the structure of the data and the actual data values are extracted from a separate table. Alternatively, for gridded data items, data can be extracted automatically from one or more NetCDF data sets (Rew and Davis, 1990) to populate a user-defined grid. A case table is needed for any experiment involving more than one simulation. Further input tables are required for setting up optimization experiments and output variable selection where applicable. Finally, an "experiment control table" is used for assigning experiment-specific file names to all other input and output tables. The experiment control table can specify one or more experiments to be run, each either with or without parameter optimization. Batches of experiments are run without the overhead of re-loading resident data. Comprehensive, customizable log output provides a record of the experimental configurations. An example of the input and output for a simple experiment is given in the Supplement.

A2 Plankton model interface

The MarMOT Model Evaluator handles a superset of prognostic and diagnostic variables and the necessary information is transferred between the MME data area and the active plankton model at each time step, allowing plankton models to be implemented with minimal changes to their native variables and code. Each model must provide a specific set of Fortran subroutines to perform basic functions such as defining the model parameter names for use in the data management system, setting fixed model variables (e.g. model grid, time step) and providing MarMOT with source-minus-sink tendencies. Generic socket subroutines on the MME side of the interface are responsible for calling the appropriate model-specific subroutines, according to the model selected for the current simulation.

MarMOT maintains two sets of tracers: primary tracers and derived tracers. The concentration of each derived tracer is determined by the concentration of one or more primary tracers and zero or more ratios describing the composition of particular ecosystem components. Derived tracers such as total nitrogen or total carbon are made available for diagnostic purposes only, while other derived tracers can be prognostic variables.

The initial conditions required for a simulation are model-dependent. For a given plankton model the initial state is defined by profiles for each applicable primary tracer and any composition ratios that will vary dynamically. Where tracers are linked by composition ratios, whether variable or fixed, there are alternative sets of prognostic variables and those used within the model may be different from those initialized. MarMOT uses nitrogen variables as the primary tracers for all organic components. Forcing data requirements are also model-dependent. Each model indicates to the MME what forcing data it requires and the MME selects the information from the input data available. Only data relevant to the currently selected model appear in the simulation output.

Three plankton models are currently supported: the 4 compartment nitrogen model of Oschlies and Garçon (1999), a version of the Hadley Centre Ocean Carbon Cycle model developed by Palmer and Totterdell (2001) and the MEDUSA model of Yool et al. (2011). The first two models are of the NPZD class, representing the nitrogen cycle in terms of fluxes between dissolved inorganic nitrogen (DIN), phytoplankton, zooplankton and detritus. The HadOCC model also includes a carbonate system in the form of additional tracers for total dissolved inorganic carbon and alkalinity. MEDUSA is a slightly more complex model that includes

two types of phytoplankton and two types of zooplankton. A wider range of models, of varying complexity, will be supported in future versions.

Appendix B

MarMOT simulation features

Two important features of the MME not detailed in the main text are the provision of different options for the light limitation of phytoplankton photosynthesis and support for the parameterization of real-world horizontal flux divergences.

B1 Photosynthesis options

Different parameterizations can be applied independently for the attenuation of photosynthetically available radiation (PAR) in the water column, the chlorophyll-specific absorption of light energy by the phytoplankton and the photosynthetic response.

The PAR attenuation coefficient can be modelled as a linear function of pigment concentration G provided by the plankton model

$$K_{\mathrm{dPAR}} = k_{\mathrm{water}} + k_{\mathrm{pig}} G \tag{B1}$$

where k_{water} is the attenuation due to water, k_{pig} is the attenuation due to pigment. Although widely used, this formulation ignores the effect of changes in the spectral distribution of the energy in the PAR waveband on the attenuation coefficient as the light quality changes with depth. An alternative option is available that accounts for these changes: an empirical approximation to the 61 wave-band model of Morel (1988), developed by Anderson (Anderson, 1993) for use in OBGCMs. Light penetration is based on a 3 layer model of the attenuation coefficient K_{dPAR}, as a function of a depth-invariant pigment concentration. The three optical layers are divided by layer boundaries at 5 m and 23 m. K_{dPAR} is determined from the local pigment concentration at each depth level. Where the depth level boundaries for the current simulation do not coincide with optical layer boundaries, K_{dPAR} is depth averaged within levels.

The K_{dPAR} profile from the attenuation model can optionally be adjusted, following Oschlies and Garçon (1999), to allow for the geometric effect of the sun's zenith angle on the path length between the surface and a given depth. The correction factor is based only on the direct path effect, tending to bias K_{dPAR} high. However, a compensating bias is introduced by basing the factor on the zenith angle at noon, when path length is at its daily minimum. The true effect of zenith angle on the attenuation coefficient is strongly wavelength dependent and decreases with depth (Zheng et al., 2002). The depth dependency is not currently modelled.

Chlorophyll-specific light absorption by phytoplankton varies with depth, due to changes in spectral distribution. This directly affects the initial slope of the photosynthesis-PAR curve. In many plankton models, this effect is ignored and a constant value is used for the initial slope. This option is supported in MarMOT, together with an alternative option to use the spectrally-averaged chlorophyll absorption model of Anderson (1993). Like the attenuation coefficient model, this is based on an empirical approximation to a 61 waveband model (Morel, 1988, 1991).

Three alternative parameterizations are provided for the light limitation of photosynthesis: two for calculating the daily mean photosynthetic rate over each simulation level and one for calculating a point-in-time rate for each level that allows the diel cycle to be resolved explicitly when high resolution forcing data are available. The available parameterizations for daily mean photosynthesis are those of Evans and Parslow (1985) and Platt et al. (1990). These are based on triangular and sinusoidal representations of the diel cycle respectively and use different formulations of the photosynthesis-PAR curve. The point-in-time rate is calculated using the same photosynthesis-PAR curve as Evans and Parslow (1985).

B2 Horizontal flux divergence

Parameterization of horizontal flux divergence by perturbation of the local state as described in Sect. 3 is consistent with the aim of emulating the behaviour of a plankton model in a 3-D system for the purposes of assessing model skill. MarMOT also supports a modified parameterization for use in state estimation when independent information, in the form of a prior estimate of the real-world state, is available. The prior state C_i^{ref} would typically be a high uncertainty estimate based on climatology that could potentially be improved upon by a plankton model's response to local forcing data.

The parameterization is designed to represent uncertain real-world flux divergence tendencies. It combines stochastic perturbations with a relaxation tendency towards the reference state. Perturbations, constrained where possible by observed current velocities and property gradients, represent the effect of lateral processes moving the system trajectory away from that of a locally forced system. The relaxation term ensures that as information is lost, the solution tends towards the prior state estimate C_i^{ref}. The prior is effectively assimilated during integration and the magnitude of the relaxation tendency is balanced against that of the perturbation to ensure consistency with the expected change due to flux divergence.

In practice, the maximum relaxation rate is constrained by the perturbation standard deviation σ_i^{pert}. At each time step, a new perturbation-limited relaxation rate

$$r_i' = \min\left(\frac{R_i}{|C_i^{\mathrm{ref}} - C_i|}, r_i^{\mathrm{ext}}\right) \tag{B2}$$

is determined for each tracer i, where R_i is a maximum permitted magnitude for the rate of change in concentration due to relaxation and r_i^{ext} is the input relaxation rate. The degree of limitation is determined by a relaxation control factor ψ such that

$$R_i = \psi \sigma_i^{\text{pert}} \tag{B3}$$

so ψ controls the significance of the relaxation change, relative to the random perturbations. Alternatively, if σ_i^{pert} is defined in transformed variable space

$$R_i = C_i \psi \sigma_i^{\text{pert}} \tag{B4}$$

or

$$R_i = 2\sqrt{C_i}\psi \sigma_i^{\text{pert}} \tag{B5}$$

for log and square root transformations, respectively. The optimal value for ψ rate depends on the relative quality of the two different estimates of the local state provided by the model and the prior.

The maximum permitted relaxation rate is determined separately for each tracer. However, it is desirable to use the same relaxation rate for all tracers to preserve relationships between different tracers in the prior state estimate. At each time step, a universal relaxation rate

$$r_i = \min_i(r_i') \tag{B6}$$

is therefore applied to all relaxed tracers.

Appendix C

Implementation of bounded parameter optimization

The micro-genetic algorithm employed by the MarMOT optimizer is designed to work with a bounded parameter space, while Powell's direction set algorithm treats the parameter space as infinite. To support bounded minimizations with the direction set algorithm, transformations can be applied to any parameter value P, in original or log space, to provide an unbounded value $P^\star$ for the optimizer

$$P^\star = \left\{ \begin{array}{l} \frac{P-P_{\text{mid}}}{P-P_{\text{lower}}} \ , \ P < P_{\text{mid}} \\ \frac{P-P_{\text{mid}}}{P_{\text{upper}}-P} \ , \ P > P_{\text{mid}} \end{array} \right\} \tag{C1}$$

$$P_{\text{mid}} = \frac{1}{2}(P_{\text{lower}} + P_{\text{upper}}) \tag{C2}$$

where P_{lower} and P_{upper} are the required bounds in the original finite parameter space. The magnitude of $P^\star$ tends to infinity as P approaches either bound, so any point $P^\star$ in the infinite space seen by the optimizer maps to a value P where $P_{\text{lower}} < P < P_{\text{upper}}$. The behaviour of the search algorithm with respect to the original parameter space is affected as a consequence of the modified cost function $J'(P^\star) = J(P)$ presented to the optimizer. Transformations are dimension specific, so bounded and unbounded parameters can be optimized simultaneously.

The parameter transformation is based on that introduced by Fasham et al. (1999) for the same purpose. In that study, a parameter penalty term was also included in the cost function formulation to weight against large deviations of the transformed parameters from their prescribed prior values. In MarMOT, prior parameter information is provided purely in terms of allowable ranges so that the value of the cost function $J(P)$ is unaffected by the parameter values, except via the simulation.

Appendix D

HadOCC nitrogen cycle simulation

The HadOCC model described here is a modified version of the model of Palmer and Totterdell (2001) incorporating a number of subsequent developments (Totterdell, personal communication, 2005). The nitrogen tracers are phytoplankton P, zooplankton Z, detritus D and dissolved inorganic nitrogen N. The main differences from the original version are the introduction of a variable carbon:chlorophyll ratio and changes to the pathways of material originating from grazing and mortality. In addition, spectrally-averaged photosynthesis is parameterized using the Anderson (1993) approximations (see Appendix B1). There is no temperature limitation of photosynthesis and DIN limitation is applied to the photosynthesis-PAR curve maximum, rather than the light-limited photosynthetic rate, reducing its effect at low light levels. A different parameterization of depth variation in the detrital remineralization rate is used and a number of the parameters common to both model versions are assigned different values. Process parameterizations and source-minus-sink terms are defined below. Refer to Table D1 for parameter values.

Photosynthesis: daily mean biomass-specific growth rate $\bar{\mu_{\text{P}}}$ is calculated for each model level using the integral approximation of Platt et al. (1990). The photosynthesis-PAR response at depth z and time t is

$$\mu_{\text{P}}(z,t) = P_{\max}\left[1 - \exp\left(-\frac{\alpha_{\text{chl}}(z)E_{\text{d}}(z,t)}{\theta_{\text{chl}}P_{\max}}\right)\right] \tag{D1}$$

where the maximum nutrient-limited photosynthetic rate is given by:

$$P_{\max} = V_{\max}\frac{N}{N+k_{\text{N}}}. \tag{D2}$$

Table D1. HadOCC model parameters.

Parameter	Symbol	Value
Minimum C:Chl ratio	$\theta_{\min}$	$20\,\mathrm{g\,C\,(g\,Chl)^{-1}}$
Maximum C:Chl ratio	$\theta_{\max}$	$200\,\mathrm{g\,C\,(g\,Chl)^{-1}}$
C:N ratio for phytoplankton	θ_{P}	6.625
C:N ratio for zooplankton	θ_{Z}	5.625
C:N ratio for detritus	θ_{D}	7.5
Maximum photosynthetic rate	$V_{\max}$	$2\,\mathrm{d^{-1}}$
Initial slope of photosynthesis-PAR curve	α_{surf}	$5.56\,\mathrm{mg\,C\,(mg\,Chl)^{-1}\,(E\,m^{-2})^{-1}}$
Half-saturation conc. for nutrient uptake	k_{N}	$0.1\,\mathrm{mmol\,N\,m^{-3}}$
Phytoplankton density-dependent mortality	m_{o}	$0.05\,\mathrm{d^{-1}(mmol\,N\,m^{-3})^{-1}}$
Phytoplankton specific respiration	η	$0.05\,\mathrm{d^{-1}}$
Maximum grazing rate	$g_{\max}$	$0.8\,\mathrm{d^{-1}}$
Half-saturation conc. for grazing	k_{F}	$0.5\,\mathrm{mmol\,N\,m^{-3}}$
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
Zooplankton specific mortality	m_1	$0.05\,\mathrm{d^{-1}}$
Zooplankton density-dependent mortality	m_2	$0.3\,\mathrm{d^{-1}(mmol\,N\,m^{-3})^{-1}}$
Detrital sinking velocity	w_{D}	$10\,\mathrm{m\,d^{-1}}$
Parameters derived from C:N ratios (above):		
Biomass-equivalent:N ratio for phytoplankton	B_{P}	1
Biomass-equivalent:N ratio for zooplankton	B_{Z}	0.87
Biomass-equivalent:N ratio for detritus	B_{D}	1.11

The carbon:chlorophyll ratio is given by the balanced growth photo-acclimation model of Geider et al. (1997):

$$\theta_{\mathrm{chl}} = \min\left(\sqrt{\theta_{\min}\frac{\alpha_{\mathrm{chl}}E_{\mathrm{d}}}{\mu_{\mathrm{P}}(\theta_{\mathrm{chl}})}}, \theta_{\max}\right). \tag{D3}$$

Downwelling PAR E_{d} is determined by the light attenuation coefficient model of Anderson (1993), without the direct path adjustment of Oschlies and Garçon (1999). A ratio of chlorophyll to total pigment concentration of 0.8 is assumed and $E_{\mathrm{d}}(0,t)$ is taken to be 43 % of total downwelling solar radiation at the sea surface. The chlorophyll-specific initial slope α_{chl} is determined from model parameter α_{surf} using the Anderson (1993) chlorophyll light absorption model.

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:

$$h = \frac{B_{\mathrm{Z}}Z}{F_{\mathrm{tot}}}g_{\max}\frac{F^2}{F^2+K_{\mathrm{F}}^2}; \tag{D4}$$

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

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

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

Detrital remineralization: $\lambda = 0.1\,\mathrm{d^{-1}}$ for $z < 100\,\mathrm{m}$, otherwise $\lambda = \frac{8.58}{z}\,\mathrm{d^{-1}}$.

Nitrogen equations:

$$\mathrm{SMS_P} = \bar{\mu_{\mathrm{P}}}P - M_{\mathrm{P}} - \eta P - G_{\mathrm{P}} \tag{D5}$$

$$\mathrm{SMS_Z} = \phi_{\mathrm{I}}(\beta_{\mathrm{P}}G_{\mathrm{P}} + \beta_{\mathrm{D}}G_{\mathrm{D}}) - M_{\mathrm{Z}} \tag{D6}$$

$$\begin{aligned}\mathrm{SMS_D} = {} & \frac{\theta_{\mathrm{P}}}{\theta_{\mathrm{D}}}(0.99M_{\mathrm{P}}) + \frac{\theta_{\mathrm{Z}}}{\theta_{\mathrm{D}}}(0.33M_{\mathrm{Z}}) \\ & + \frac{\theta_{\mathrm{P}}}{\theta_{\mathrm{D}}}a_{\mathrm{PD}}G_{\mathrm{P}} + (a_{\mathrm{DD}}-1)G_{\mathrm{D}} - \lambda D\end{aligned} \tag{D7}$$

$$\begin{aligned}\mathrm{SMS_N} = {} & \left\{0.01 + \left(1 - \frac{\theta_{\mathrm{P}}}{\theta_{\mathrm{D}}}\right)0.99\right\}M_{\mathrm{P}} + \eta P \\ & + \left\{0.67 + \left(1 - \frac{\theta_{\mathrm{Z}}}{\theta_{\mathrm{D}}}\right)0.33\right\}M_{\mathrm{Z}} \\ & + 0.1(1-\phi_{\mathrm{I}})(G_{\mathrm{P}} + G_{\mathrm{D}}) + \left(1 - \frac{\theta_{\mathrm{P}}}{\theta_{\mathrm{D}}}\right)a_{\mathrm{PD}}G_{\mathrm{P}} \\ & + \lambda D - \bar{\mu_{\mathrm{P}}}P\end{aligned} \tag{D8}$$

where $a_{\mathrm{PD}} = 0.9(1-\phi_{\mathrm{I}}) + (1-\beta_{\mathrm{P}})\phi_{\mathrm{I}}$ and $a_{\mathrm{DD}} = 0.9(1-\phi_{\mathrm{I}}) + (1-\beta_{\mathrm{D}})\phi_{\mathrm{I}}$. 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.

Numerical configuration: the vertical grid has 63 levels with 35 levels in the top 1000 m. These 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, corresponding to those of the ORCA025 model. Levels spanning the mixed layer depth are partially mixed. The advection scheme is an upstream differencing scheme. The time step is 1 h.

Supplementary material related to this article is available online at: http://www.geosci-model-dev.net/5/471/2012/gmd-5-471-2012-supplement.zip.

Acknowledgements. The MarMOT system has been developed with funding from the UK Natural Environment Research Council. We would like to thank Andreas Oschlies for access to code used in the vertical transport schemes and to Rachel Oxlade and Peter Craig for help with system testing and valuable feedback. Thanks are due also to the UK Met Office for access to the HadOCC model code and to Philip Wallhead and 2 anonymous reviewers for their valuable comments on the original manuscript. This work was supported by NERC via the National Centre for Earth Observation (a NERC collaborative centre).

Edited by: A. Ridgwell

References

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

Anderson, T. R.: A spectrally averaged model of light penetration and photosynthesis, Limnol. Oceanogr., 38, 1403–1419, 1993.

Anderson, T. R.: Plankton functional type modelling: running before we can walk?, J. Plankton Res., 27, 1073–1081, 2005.

Barnier B., Madec, G., Penduff, T., Molines, J.-M., Treguier, A.-M., Le Sommer, J., Beckmann, A., Biastoch, A., Böning, C., Dengg, J., Derval, C., Durand, E., Gulev, S., Remy, E., Talandier, C., Theetten, S., Maltrud, M., McClean, J., and de Cuevas, B.: Impact of partial steps and momentum advection schemes in a global ocean circulation model at eddy permitting resolution, Ocean Dynam., 56, 543–567, doi:10.1007/s10236-006-0082-1, 2006.

Box, G. E. P. and Cox, D. R.: An analysis of transformations, J. Roy. Stat. Soc. B Met., 26, 211–252, 1964.

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, NJ, 61–80, 1973.

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.

Dowd, M. and Meyer, R.: A Bayesian approach to the ecosystem inverse problem, Ecol. Model., 168, 39–55, 2003.

Evans, G. T.: The role of local models and data sets in the Joint Global Ocean Flux Study, Deep-Sea Res. I, 46, 1369–1389, 1999.

Evans, G. T.: Defining misfit between biogeochemical models and data sets, J. Marine Syst., 40–41, 49–54, 2003.

Evans, G. T. and Parslow, J. S.: A model of annual plankton cycles, Biol. Oceanogr., 3, 327–347, 1985.

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

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

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

Faugeras, B., Lévy, M., Mémery, L., Verron, J., Blum, J. and Charpentier, I.: Can biogeochemical fluxes be recovered from nitrate and chlorophyll data? A case study assimilating data in the Northwestern Mediterranean Sea at the JGOFS-DYFAMED station, J. Marine Syst., 40–41, 99–125, 2003.

Faugeras, B., Bernard, O., Sciandra, A., and Lévy, M.: A mechanistic modelling and data assimilation approach to estimate the carbon/chlorophyll and carbon/nitrogen ratios in a coupled hydrodynamical-biological model, Nonlin. Processes Geophys., 11, 515–533, doi:10.5194/npg-11-515-2004, 2004.

Fennel, K., Losch, M., Schröter, J., and Wenzel, M.: Testing a marine ecosystem model: sensitivity analysis and parameter optimization, J. Marine Syst., 28, 45–63, 2001.

Friedrichs, M. A. M.: Assimilation of JGOFS EqPac and SeaWiFS data into a marine ecosystem model of the central equatorial Pacific Ocean, Deep-Sea Res. II, 49, 289–319, 2002.

Friedrichs, M. A. M., Hood, R. R., and Wiggert J. D.: Ecosystem model complexity versus physical forcing: Quantification of their relative impact with assimilated Arabian Sea data, Deep-Sea Res. II, 53, 576–600, 2006.

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

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

Geider, R. J., MacIntyre, H. L., and Kana, T. M.: 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, 1997.

Harmon, R. and Challenor, P.: A Markov chain Monte Carlo method for estimation and assimilation into models, Ecol. Model., 101, 41–59, 1997.

Hemmings, J. C. P., Srokosz, M. A., Challenor, P., and Fasham, M. J. R.: Assimilating satellite ocean-colour observations into oceanic ecosystem models, Philos. T. Roy. Soc. A, 361, 33–39, 2003.

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.

Hurtt, G. C. and Armstrong, R. A.: A pelagic ecosystem model calibrated with BATS data, Deep-Sea Res. II, 43, 653–683, 1996.

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

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

Kettle, H.: Using satellite-derived backscattering coefficients in addition to chlorophyll data to constrain a simple marine biogeochemical model, Biogeosciences, 6, 1591–1601, doi:10.5194/bg-6-1591-2009, 2009.

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

Kuroda, H. and Kishi, M. J.: A data assimilation technique applied to estimate parameters for the NEMURO marine ecosystem model, Ecol. Model., 172, 69–85, 2004.

Lafore, J. P., Stein, J., Asencio, N., Bougeault, P., Ducrocq, V., Duron, J., Fischer, C., Héreil, P., Mascart, P., Masson, V., Pinty, J. P., Redelsperger, J. L., Richard, E., and Vilà-Guerau de Arellano, J.: The Meso-NH Atmospheric Simulation System. Part I: adiabatic formulation and control simulations, Ann. Geophys., 16, 90–109, doi:10.1007/s00585-997-0090-6, 1998.

Le Quéré, C.: Reply to horizons article "Plankton functional type modelling: running before we can walk" Anderson (2005): I. Abrupt changes in marine ecosystems?, J. Plankton Res., 28, 871–872, 2006.

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

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

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.

Morel, A.: Optical modelling of the upper ocean in relation to its biogenous matter content (case 1 waters), J. Geophys. Res., 93, 10749–10768, 1988.

Morel, A.: Light and marine photosynthesis: A spectral model with geochemical and climatological implications, Prog. Oceanogr., 26, 263–306, 1991.

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

Oschlies, A. and Garçon, V.: An eddy-permitting coupled physical-biological model of the North Atlantic, 1. Sensitivity to advection numerics and mixed layer physics, Global Biogeochem. Cy., 13, 135–160, 1999.

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

Popova, E. E., Yool, A., Coward, A. C., Aksenov, Y. K., Alderson, S. G., de Cuevas, B. A., and Anderson, T. R.: Control of primary production in the Arctic by nutrients and light: insights from a high resolution ocean general circulation model, Biogeosciences, 7, 3569–3591, doi:10.5194/bg-7-3569-2010, 2010.

Platt, T., Sathyendranath, S., and Ravindran, P.: Primary production by phytoplankton: Analytic solutions for daily rates per unit area of water surface, P. Roy. Soc. Lond. B Bio., 241, 101–111, 1990.

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.

Prunet, P., Minster, J. F., Echevin, V., and Dadou, I.: Assimilation of surface data in a one-dimensional physical-biogeochemical model of the surface ocean, 2. Adjusting a simple trophic model to chlorophyll, temperature, nitrate, and pCO_2 data, Global Biogeochem. Cy., 10, 139–158, 1996a.

Prunet, P., Minster, J. F., Ruiz-Pino, D., and Dadou, I.: Assimilation of surface data in a one-dimensional physical-biogeochemical model of the surface ocean, 1. Method and preliminary results, Global Biogeochem. Cy., 10, 111–138, 1996b.

Rew, R. and Davis, G.: NetCDF – an interface for scientific data access, IEEE Comput. Graph., 10, 76–82, 1990.

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.

Schartau, M., Oschlies, A., and Willebrand, J.: Parameter estimates of a zero-dimensional ecosystem model applying the adjoint method, Deep-Sea Res. II, 48, 1769–1800, 2001.

Sinha, B., Buitenhuis, E. T., Le Quéré, C., and Anderson, T. R.: Comparison of the emergent behaviour of a complex ecosystem model in two ocean general circulation models, Prog. Oceanogr., 84, 204–224, 2010.

Spitz, Y. H., Moisan, J. R., Abbott, M. R., and Richman, J. G.: Data assimilation and a pelagic ecosystem model: parameterization using time series observations, J. Marine Syst., 16, 51–68, 1998.

Spitz, Y. H., Moisan, J. R., and Abbott, M. R.: Configuring an ecosystem model using data from the Bermuda Atlantic Time Series (BATS), Deep-Sea Res. II, 48, 1733–1768, 2001.

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

Wallhead, P. J., Martin, A. P., Srokosz, M. A., and Fasham, M. J. R.: Accounting for unresolved spatial variability in marine ecosystems using time lags, J. Mar. Res., 64, 881–914, 2006.

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.

Weber, L., Volker, C., Schartau, M., and Wolf-Gladrow, D. A.: Modeling the speciation and biogeochemistry of iron at the Bermuda Atlantic Time-series Study site, Global Biogeochem. Cy., 19, GB1019, doi:10.1029/2004GB002340, 2005.

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

Zheng, X., Dickey, T., and Chang, G.: Variability of the downwelling diffuse attenuation coefficient with consideration of inelastic scattering, Appl. Optics, 41, 6477–6488, 2002.