Calibrating a global three-dimensional biogeochemical ocean model (MOPS-1.0)

Iris Kriest[1], Volkmar Sauerland[2], Samar Khatiwala[3], Anand Srivastav[2], and Andreas Oschlies[1]

[1]GEOMAR Helmholtz-Zentrum für Ozeanforschung Kiel, Düsternbrooker Weg 20, D-24105 Kiel, Germany
[2]Institut für Informatik, Christian-Albrechts-Universität zu Kiel, Christian-Albrechts-Platz 4, D-24098 Kiel, Germany
[3]Department of Earth Sciences, University of Oxford, South Parks Road, Oxford OX1 3AN, UK

Correspondence to: Iris Kriest (ikriest@geomar.de)

Abstract.

Global biogeochemical ocean models contain a variety of different biogeochemical components and often much simplified representations of complex dynamical interactions, which are described by many ($\approx 10 - \approx 100$) parameters. The values of many of these parameters are empirically difficult to constrain, due to the fact that in the models they represent processes for a range of different groups of organisms at the same time, while even for single species parameter values are often difficult to determine in situ. Therefore, these models are subject to a high level of parametric uncertainty. This may be of consequence for their skill with respect to accurately describing the relevant features of the present ocean, as well as their sensitivity to possible environmental changes.

We here present a framework for the calibration of global biogeochemical ocean models on short and long time scales. The framework combines an offline approach for transport of biogeochemical tracers with an Estimation of Distribution Algorithm (Covariance Matrix Adaption Evolution Strategy, CMA-ES). We explore the performance and capability of this framework by five different optimizations of six biogeochemical parameters of a global biogeochemical model, simulated over 3000 years. First, a twin experiment explores the feasibility of this approach. Four optimizations against a climatology of observations of annual mean dissolved nutrients and oxygen determine the extent, to which different setups of the optimization influence model fit and parameter estimates. Because the misfit function applied focuses on the large-scale distribution of inorganic biogeochemical tracers, parameters that act on large spatial and temporal scales are determined earliest, and with the least spread. Parameters more closely tied to surface biology, which act on shorter time scales, are more difficult to determine. In particular the search for optimum zooplankton parameters can benefit from a sound knowledge of maximum and minimum parameter values, leading to a more efficient optimization. It is encouraging that, although the misfit function does not contain any direct information about biogeochemical turnover, the optimized models nevertheless provide a better fit to observed global biogeochemical fluxes.

1 Introduction

Global ocean models that simulate biogeochemical interactions are subject to many uncertainties, among them those related to initial conditions, forcing, and parameterizations of physical and biological processes, as well as the adequacy of the chosen

model complexity with respect to the scientific problem under investigation. It is generally assumed that all these 'input' factors affect the simulation results in ways that may be different for different models, but a thorough understanding of how uncertainties in input map onto model output (residuals, i.e., deviations from the true state) is still lacking. Quantitative estimates of the effect of model uncertainty on model residuals are generally obtained from individual sensitivity studies, model intercomparison or model ensemble studies, where the spread of model results is regarded as a measure of model uncertainty. This procedure is, for example, followed in the assessment reports of the Intergovernmental Project of Climate Change (IPCC). The Ocean Carbon Model Intercomparison Project (OCMIP) applied a strict protocol regarding the description of biogeochemical processes to a suite of different ocean circulation models to show that the effect of uncertainties in the simulated circulation on biogeochemical tracer distributions and their residuals can be considerable (Orr et al., 2001; Najjar et al., 2007). However, the effect of uncertainties in the formulation of biogeochemical models on simulated biogeochemical tracers and fluxes can be of similar magnitude (Kriest et al., 2010) and is often difficult to disentangle from other sources of uncertainty (e.g., Cabre et al., 2015; Seferian et al., 2016). One reason for diverging results of global biogeochemical models can be related to the uncertainty with respect to biological constants and equations. In addition to often poorly constrained parameters, it is, so far, not even clear how complex a biogeochemical model should be (e.g. what state variables it should contain) in order to realistically reproduce observed global tracer distributions (Kriest et al., 2012). As a consequence, the diversity of biogeochemical models ranges from simple, "nutrient-only" models to far more complex ones, comprising different elemental cycles and biological components.

Uncertainties in biogeochemical model setup partly arise from sparse observations, particularly in the open ocean and during winter season in the high latitudes (Kriest et al., 2010). Further, the combined effects of shallow and deep biogeochemistry and ocean circulation introduce a variety of timescales, from minutes to millennia, hampering a complete and thorough investigation of the combined effects of the different process parameterizations. Finally, even quite simple biogeochemical models are often characterized by non-linear interactions, complicating the a posteriori analysis of model results. By performing a relatively "coarse sweep" of the multidimensional model parameter space, Kriest et al. (2010, 2012) illustrated the impact of different model complexities and parameter sets on simulated tracers and their fit to observations. This first attempt to systematically explore the impacts of biogeochemical parameter uncertainty in global models may well have missed optimal regions in parameter space, making it difficult to decide whether a model performs badly due to ill-chosen parameters, or due to an inappropriate model structure. The development of automatic optimization of global ocean biogeochemical models that is the goal of this study should enable a more thorough search for "best" parameters, and thus facilitate inter-model comparison.

An under-sampled ocean, together with a large variety of time and space scales and a high level of structural model complexity, poses a challenge for optimization, and for a full, and dense enough, scan of the parameter space on a global scale. Therefore, optimization of marine biogeochemical models has mostly been carried out in a local, 0- or 1-dimensional setting (e.g., Fasham and Evans, 1995; Athias et al., 2000; Rückelt et al., 2010; Ward et al., 2010). The variability of biogeochemical processes has been addressed by simultaneous optimization at different sites (and physical forcings) in the North Atlantic by Schartau and Oschlies (2003a, b). Given the high computational demands, and the sparsity of biogeochemical data on a global scale, attempts to address the indeterminacy of global simulations of ocean biogeochemistry via optimization have resorted to

rather simple biogeochemical systems (Kwon and Primeau, 2006, 2008) or to rather coarse physical model resolution (Tjiputra et al., 2007). To constrain parameters related to dissolved organic matter production and decay on short and long time scales, Letscher et al. (2015) alternated between a simplified biogeochemical system and a more complex model, which is limited in terms of spin-up time. Recent attempts begin to combine complex, local models and a detailed three-dimensional global environment for optimization (Hemmings et al., 2014). To our knowledge, however, the experiments presented here are the first one that, for a state-of-the-art global biogeochemical ocean model, carry out a parameter optimization that targets at parameters relevant for biogeochemical processes on both large and small scales in the full spatio-temporal domain.

In this paper we first test the global biogeochemical model optimization against synthetic data, derived from a previous model experiment with perturbed model parameters in so-called twin experiments. We then present four optimizations against a global, synoptic data set of observed phosphate, nitrate, and oxygen.

2 Methods

2.1 Biogeochemical ocean model

2.1.1 Circulation framework

For easy and generic coupling between different biogeochemical models and circulation fields, as well as fast and efficient computation we use the "Transport Matrix Method" (TMM), developed by Samar Khatiwala (Khatiwala, 2007), and available via Github (`https://github.com/samarkhatiwala/tmm`). This efficient "offline" method for ocean passive tracer transport represents advection and mixing in the form of transport matrices that have been calculated from an ocean circulation model simulation prior to the biogeochemical simulations performed here.

For optimization, we use the TMM with monthly mean transport matrices derived from a 2.8° global configuration of the MIT ocean model with 15 levels in the vertical (Marshall et al., 1997). Using this rather coarse spatial grid, a time step length of 1/2 day for tracer transport and 1/16 day for biogeochemical interactions, each biogeochemical model setup with seven tracers has been simulated for 3000 years, after which most of the tracers approach steady state.

2.1.2 Biogeochemical model

The biogeochemical model employed as representative of current state-of-the-art models is the same as presented by Kriest and Oschlies (2015, hereafter called MOPS - Model of Oceanic Pelagic Stoichiometry), and we only describe it briefly here. Based on phosphorus, it consists of seven tracers, namely phosphate, dissolved inorganic nitrogen (hereafter termed and compared to nitrate), phytoplankton, zooplankton, detritus, dissolved organic matter (DOM) and oxygen. For conversion between the different elements we apply a constant global stoichiometry of $R_{-\mathrm{O2:P}} = 170$ mmol O_2:mmol P for the ratio between O_2:P, and 16 mmol N:mmol P for the N:P ratio of particular and dissolved organic matter. The stoichiometry of aerobic and anaerobic remineralization is based on Paulmier et al. (2009). Remineralization of detritus and DOM is parameterized via a constant nominal remineralization rate, $r = 0.05$ [d^{-1}]. However, aerobic remineralization is restricted to regions with suffi-

cient oxygen. If oxygen declines, nitrate is used as electron acceptor, thereby mimicking denitrification. If both oxygen and nitrate are depleted, remineralization of organic matter is suppressed in the model. Aerobic and anaerobic remineralization are parameterized as saturation (Monod-type) curves that regulate the rates of these processes using either oxidant, as well as the inhibition of denitrification by oxygen. Thus, the actual remineralization rate may differ from r, depending on oxidant availability. Temperature dependent nitrogen fixation resupplies fixed nitrogen lost through denitrification via relaxation at the sea surface to the stoichiometric ratio of 16. Thus, while total phosphate inventory is conserved, oxygen and fixed nitrogen inventory may change during the course of the simulation, with the long-term, steady state inventory depending on physics and biogeochemistry (Kriest and Oschlies, 2015).

Sinking of detritus is simulated using a sinking speed increasing with depth $w = a\,z$ [d^{-1}]. Assuming constant remineralization rate r, equilibrium conditions and absence of horizontal or vertical advection, this would result in a particle flux profile defined by $F(z) \propto z^{-b}$, where $b = r/a$ (see also Kriest and Oschlies, 2008). For better comparison to observed particle flux profiles (e.g., Martin et al., 1987), in the following we express the sinking speed via the parameter $b = r/a$ (see Kriest and Oschlies, 2008). The model also includes burial of particulate organic phosphorus and nitrogen arriving at the sea floor, which is resupplied globally as phosphate and nitrate via river runoff (Kriest and Oschlies, 2013).

Simulating both surface (primary production, grazing, egestion and excretion by zooplankton) as well as deep (sinking and decay of organic matter) processes before the background of ocean circulation and seasonally varying forcing, the model thus encompasses processes that act on a variety of time scales, from the order of hours to days (surface) to months and years.

2.2 The optimization algorithm CMA-ES

2.2.1 Population-based search heuristics

The TMM as described above is fast enough to be used together with meta-heuristic methods for parameter optimization, such as Evolutionary Algorithms (EAs) or Estimation of Distribution Algorithms (EDAs). Although these methods require more function evaluations to converge to some local optimum than gradient-based methods, they are of advantage in complicated, irregular "search landscapes" with local optima (which might be far worse than the global optimum), or discontinuities.

The common goal of such population-based meta heuristics is to strike a good balance of both search properties, exploration (search for promising solutions in a wide area of the search space) and exploitation (search within small regions around good solutions to fast reach local optima). Classical evolutionary algorithms as depicted on the left of Fig. 1 mimic principles of natural evolution to pursue that goal. They use randomized procedures to select, combine, mutate and reinsert candidate solutions (individuals) from/into a given solution set (population). In each iteration, these mechanisms (red operations in Fig. 1) indirectly imply a probability distribution on the search space with respect to which individuals are likely to appear in the next "generation". The implied probability distribution changes in each generation, tending to increase the probabilities of good solutions and to decrease the probabilities of poor solutions due to the survival-of-the-fittest principle.

In contrast to classical EAs, estimation of distribution algorithms (sketched on the right of Fig. 1) use an explicit (parameterized) probability distribution from which candidate solutions are sampled, directly. In each iteration, the probability

distribution is also updated directly by utilizing good solutions of the current iteration. Good solutions of preceding iterations are (optionally) considered by involving preceding probability distributions into the update process using auxiliary variables. Evolutionary frameworks use operators (EAs) and probability distributions (EDAs) that are appropriate for the search space under consideration. For example, so called quantum inspired evolutionary algorithms (QiEA) have shown to be very suitable EDAs for binary problems (e.g. Kliemann et al., 2013; Patvardhan et al., 2015, 2016). QiEA versions for continuous problems have also been investigated in the literature (Babu et al., 2009).

Here we use a state-of-the-art EDA for optimization of (firstly) six parameters. Our task can be classified as a continuous optimization problem with bound-constraints, i.e. boundaries for the parameters. One appropriate EA/EDA tool is the Covariance Matrix Adaption Evolution Strategy (CMA-ES; Hansen and Ostermeier, 2001; Hansen, 2006), which has shown good performance with respect to quality and efficiency (in terms of function evaluations) in similar applications. Hansen et al. (2010) compare 31 algorithms on a test bed of 24 continuous benchmark functions presented in Hansen et al. (2009a), finding CMA-ES versions to perform well, particularly on multi-modal test functions. CMA-ES is invariant regarding both order preserving transformations of the objective function and rotations and translations of the search space. Invariances of a strategy justify generalizations of empirical results, which encouraged us to choose CMA-ES for our application.

We essentially follow the description of the $(\mu/\mu_w, \lambda)$-CMA-ES in Hansen (2016). In Subsubsection 2.2.2, we illustrate how the distribution is sampled and modified. For the sake of completeness, we present the guiding ideas behind the exact procedures in Subsubsections 2.2.3 - 2.2.6. The algorithm outline can be found in 2.3. This basic version does not consider bound constraints. We therefore use a penalty function based boundary handling (Hansen et al., 2009b) which we will briefly explain in Subsubsection 2.2.7

2.2.2 Normal distributions

In CMA-ES the distribution from which candidate solutions (BGC parameter vectors in our application) are sampled is a multi-variate normal-distribution. It generalizes the usual normal distribution, also known as Gaussian distribution, from $\mathbb{R}$ to the vector space $\mathbb{R}^n$ with arbitrary dimension n, given by the number of biogeochemical parameters to be estimated. The position and the shape of the one-dimensional normal distribution (more precisely, its density function) is uniquely defined by its mean and its variance, respectively.

A measure of "diversity" of a probability distribution is the so called (differential) entropy. For a given variance, the normal distribution has the maximum entropy amongst all distributions with the same variance (Cover and Thomas, 2006; Hansen, 2016). Entropy is used as an index of diversity, though it does not directly mean the same as diversity (Jost, 2006).

An EDA that works with Gaussian distributions is supposed to carefully update both defining distribution parameters mean and variance, in order to balance its exploration and exploitation ability. This update process is illustrated in Fig. 2. The left side shows a run of the CMA-ES algorithm on a uni-variate test function (a misuse to some degree, as CMA-ES is actually not suggested to be applied with problem dimensions less than 5). The test function has many local optima in which a gradient based search might get stuck. From the distributions (the blue density functions), we draw 10 samples per iteration

(some samples more than the suggested default number, which depends on the problem dimension, cf. Subsubsection 2.3.1). Each sample together with its function value is marked with a dot. The distribution is updated by involving the better half (CMA-ES default portion) of the samples (blue dots). Drawing more samples per iteration generally improves the exploration capability of the algorithm but requires correspondingly more function evaluations. We can observe that the mean of the distribution is attracted towards the good samples, then. Also, the distribution shape widens, after good samples had some distance to each other and/or some distance to the current mean. Vice versa, if all good samples are close to the mean, the shape will narrow, again. Now, the mean of the distribution is supposed to drift towards the global optimum and should then start to narrow more and more. This behavior is observed in iterations 16, 22 and 28. So, when necessary, the procedure is supposed to become less exploring but more exploiting.

Similarly to the definition of the uni-variate Gaussian distribution by mean and variance, a multi-variate normal-distribution can be uniquely identified by a mean vector $\overline{\boldsymbol{x}}$ and a positive definite matrix $\mathbf{C}$ of covariances, respectively, and is denoted by $\mathcal{N}(\overline{\boldsymbol{x}}, \mathbf{C})$. Again, the mean defines the center of the distribution while the covariance matrix defines its shape. The area of one standard deviation which is an interval $[x-\sigma, x+\sigma]$ in the one-dimensional case becomes an n-dimensional ellipsoid, now (cf. the ellipses on the right side of Fig. 2 for $n=2$). It can be shown that the principal axes of the ellipsoid correspond to $\mathbf{C}$'s eigenvalues and eigenvectors, respectively. More precisely, an eigenvector defines the orientation of a principal axis and the square root of the corresponding eigenvalue defines the length of that principal axis.

2.2.3 Sampling the distribution

Sampling a multi-variate normal distribution $\mathcal{N}(\overline{\boldsymbol{x}}, \mathbf{C})$ can be practically implemented using an eigendecomposition $\mathbf{C} = \mathbf{B}\mathbf{D}^2\mathbf{B}^{\mathrm{T}}$, where $\mathbf{D}^2$ is a diagonal matrix of eigenvalues of $\mathbf{C}$ and $\mathbf{B}$ is a matrix of corresponding orthonormal eigenvectors of $\mathbf{C}$. One sample $\boldsymbol{x} \in \mathbb{R}^n$ of $\mathcal{N}(\overline{\boldsymbol{x}}, \mathbf{C})$ can be realized by drawing n independent random numbers from the uni-variate standard normal distribution $\mathcal{N}(0,1)$ to be the components of a random vector $\boldsymbol{z} \in \mathbb{R}^n$ and setting $\boldsymbol{x} = \overline{\boldsymbol{x}} + \mathbf{B}\mathbf{D}\boldsymbol{z}$.

Note, that for our problem there are bound constraints on the parameters such that samples of a normal distribution might be infeasible, regardless of whether the distribution mean is feasible or not. However, a boundary handling procedure (see Subsubsection 2.2.7) will ensure that the optimization result of CMA-ES is feasible.

2.2.4 Updating the distribution: basic principle

Empirical (re)estimates $\overline{\boldsymbol{x}}_{\mathrm{emp}}$ and $\mathbf{C}_{\mathrm{emp}}$ of the distribution parameters can be calculated from a set $S = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_\lambda\}$ of λ samples, such that the expectation of $\overline{\boldsymbol{x}}_{\mathrm{emp}}$ is $\overline{\boldsymbol{x}}$ and the expectation of $\mathbf{C}_{\mathrm{emp}}$ is $\mathbf{C}$:

$$\overline{\boldsymbol{x}}_{\mathrm{emp}} = \frac{1}{\lambda}\sum_{i=1}^{\lambda} \boldsymbol{x}_i$$

$$\mathbf{C}_{\mathrm{emp}} = \frac{1}{\lambda - 1}\sum_{i=1}^{\lambda} (\boldsymbol{x}_i - \overline{\boldsymbol{x}}_{\mathrm{emp}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}}_{\mathrm{emp}})^{\mathrm{T}}.$$

Note, that each vector $\boldsymbol{v}$ is this work is a column vector and its transposed vector $\boldsymbol{v}^T$ is a row vector. The products under the sum in the second formula are therefore n-by-n matrices.

Clearly, the estimates become the more reliable the larger λ is. We may assume that the population S is increasingly ordered (ranked) with respect to the considered objective function $f : \mathbb{R}^n \longrightarrow \mathbb{R}$, that is $f(\boldsymbol{x}_1) \leq f(\boldsymbol{x}_2) \cdots \leq f(\boldsymbol{x}_\lambda)$. Now, by involving only the better half of $\mu = \lfloor \frac{\lambda}{2} \rfloor$ samples, their distribution estimate $\mathcal{N}(\overline{\boldsymbol{x}}_\mu, \mathbf{C}_\mu)$ with corresponding parameters $\overline{\boldsymbol{x}}_\mu$ and $\mathbf{C}_\mu$ will be biased towards reproducing that μ samples with higher probability than the other $\lambda - \mu$ samples. CMA-ES uses positive values $w_1 \geq w_2 \geq \cdots \geq w_\mu$ with $\sum_{i=1}^{n} w_i = 1$ to give solutions a rank dependent weight in the updating process of both, $\overline{\boldsymbol{x}}_\mu$ and $\mathbf{C}_\mu$. The exact CMA-ES formula for the w-values and information about its background is found in Subsubsection 2.3.1. The new mean is, thus, calculated as $\overline{\boldsymbol{x}}_\mu = \sum_{i=1}^{\mu} w_i \boldsymbol{x}_i$. A subtlety is the choice of the reference mean value used for estimating $\mathbf{C}_\mu$. Instead of the new empirical mean $\overline{\boldsymbol{x}}_\mu$, the mean $\overline{\boldsymbol{x}}$ of the former distribution is chosen and yields

$$\mathbf{C}_\mu = \sum_{i=1}^{\mu} w_i (\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^{\mathrm{T}}. \tag{1}$$

It has the effect that the new distribution is elongated into directions of descent (see iteration 2 in the right example of Fig. 2).

2.2.5 Updating the distribution: reliability with small populations

As mentioned above, reliable distribution estimates require a sufficiently large number of samples. But, for a competitive computational performance we must get along with a rather small number of samples. CMA-ES therefore involves the information of former populations by updating the covariance matrix $\mathbf{C}$ to be a (convex) combination of both the current $\mathbf{C}$ and its estimate $\mathbf{C}_\mu$, that is

$$\mathbf{C} \leftarrow (1 - c_\mu)\mathbf{C} + c_\mu \mathbf{C}_\mu. \tag{2}$$

Using this formula, it can be shown that 37% of the current matrix $\mathbf{C}$'s information dates back at least $\lfloor \frac{1}{c_\mu} \rfloor$ generations, that is, the choice of the smoothing factor c_μ decides about the backward time horizon of the update procedure. The smaller the factor c_μ in (2) is the more former samples contribute to the current distribution estimate, slowing down learning but being more reliable with less samples per iteration. For example, the experiments in this paper use $n = 6$ parameters and $\lambda = 10$ samples per generation. Using (2) to update C and the (compromise) c_μ value defined for CMA-ES (see Table 1), the samples of the last 23 iterations would contribute roughly 63% of the over all information in C.

Another feature that facilitates small population sizes λ is to calculate and update a vector $\boldsymbol{p}_c$ that represents iteration averaged changes of the distribution mean and to use $\boldsymbol{p}_c$ for a so called rank-one estimate $\mathbf{C_1} = \boldsymbol{p}_c \boldsymbol{p}_c^{\mathrm{T}}$ of the covariance matrix. The idea behind this approach is that, using $\mathbf{C}_\mu$, distribution elongations into directions of descent do not distinguish for the sign of the directions. The use of the vector $\boldsymbol{p}_c$ (called evolution path) mitigates this effect. Consecutive changes of the distribution mean into opposite directions would cancel out each other. Similar to the smoothing with factor c_μ in the update of $\mathbf{C}$, above, the update of $\boldsymbol{p}_c$ is done with a smoothing factor c_c. With a further smoothing factor c_1 for the rank-one estimate $\mathbf{C_1}$, the combined covariance matrix update reads

$$\mathbf{C} \leftarrow (1 - c_\mu - c_1)\mathbf{C} + c_\mu \mathbf{C}_\mu + c_1 \mathbf{C_1}.$$

While $\mathbf{C}_\mu$ efficiently involves information from the current population into the update process, $\mathbf{C_1}$ exploits correlations between generations. The former is important in large populations, the latter is particularly important in small populations.

2.2.6 Step size control

Finally, there is an additional explicit adaption of the over all scale (the step size) of the distribution by adapting a scaling factor σ, actually using $\mathcal{N}(\overline{\boldsymbol{x}}, \sigma^2\mathbf{C})$ instead of $\mathcal{N}(\overline{\boldsymbol{x}}, \mathbf{C})$. Similar to the evolution path $\boldsymbol{p}_c$ for the rank-one covariance matrix estimates above, the adaption of the scale σ involves an evolution path $\boldsymbol{p}_\sigma$ that mirrors cumulative changes of the mean. The difference between the update formulas of both evolution paths $\boldsymbol{p}_\sigma$ and $\boldsymbol{p}_c$ is that for $\boldsymbol{p}_\sigma$ each change is re-scaled (normalized) with respect to the isotropic normal distribution $\mathcal{N}(0, \mathbf{I})$. Since covariances are always re-estimated with respect to the mean of the former iteration (cf. equation (1)) the expected normalized change of the distribution mean per iteration is therefore the expected length of a sample of $\mathcal{N}(0, \mathbf{I})$, which is

$$\chi := \mathbb{E}\big(\|\mathcal{N}(0, \mathbf{I})\|\big) \approx \sqrt{n}(1 - \frac{1}{4n} + \frac{1}{21n^2}).$$

Now, a rather small length $\|\boldsymbol{p}_\sigma\|$ compared to χ indicates that consecutive normalized moves of the mean canceled each other out, meaning that the overall scale of the distribution should be reduced with σ. Vice versa, an evolution path $\boldsymbol{p}_\sigma$ longer than χ indicates consecutive distribution drifts into correlated directions which justifies a larger overall scale of the distribution.

2.2.7 Boundary handling

In order to consider boundary constraints we use the procedure proposed in Hansen et al. (2009b, Section IV B) for CMA-ES. It applies if the distribution mean runs out of bounds. In this case, the fitness of an infeasible sample $\boldsymbol{x}$ becomes the sum of the fitness of its closest feasible point $\boldsymbol{x}_{\mathrm{feas}}$ and a weighted quadratic penalty function of its distance $\|\boldsymbol{x} - \boldsymbol{x}_{\mathrm{feas}}\|$ to the feasible box (to $\boldsymbol{x}_{\mathrm{feas}}$). Feasible samples are not penalized, i.e., the penalty function is 0 within the feasible box. Thus, the minimum of the sum of the actual fitness function and the penalty function is taken inside the feasible box or on its boundary. The quadratic penalty function has coordinate-wise weights $\frac{\gamma_i}{\xi_i}$, where ξ_i scales the out of bounds distance in the i-th coordinate with regard to the shape of the current distribution. The γ_i are suitably initialized with the range of former (unpenalized) objective function values and is multiplied with a constant > 1 in every iteration in which $\overline{\boldsymbol{x}}_i$ is more than 3 standard deviations off its bounds.

In our implementation of CMA-ES, the feasible box we operate on is the unit cube $[0, 1]^n \subseteq \mathbb{R}^n$. The samples are then linearly transformed (encoded) with respect to the actual bound constraints before evaluating the objective (misfit) function.

2.3 Implementation of the optimization algorithm

2.3.1 Algorithm outline

The CMA-ES approach described in Subsection 2.2 allows for reliable covariance matrix estimates with a relatively small population size. The default population size of $\lambda = 4+3\log(n)$ individuals and all further operational constants are successively derived from the problem dimension n as outlined in Table 1.

Here, μ counts the good portion of individuals that are selected from the λ samples in each iteration and used to update the probability distribution. As mentioned in Subsubsection 2.2.4, sampled individuals are always sorted with respect to their function values ($f(\boldsymbol{x_1}) \leq \cdots \leq f(\boldsymbol{x_\lambda})$).

The μ recombination weights w_i sum up to 1 and are monotonically decreasing in order to give better selected samples a higher weight in the updating formulas. Hansen (2016) suggest to use the value μ_{eff} as a quality measure for the weights and states that

$$\mu_{\text{eff}} = \frac{\lambda}{4} \tag{3}$$

indicates a good choice. Indeed, equation (3) is approximately satisfied by the given weighting scheme. We can only briefly sketch the history behind the suggestion: With equal weights $\frac{1}{\mu}$ in the distribution update, all the best μ independent samples would count with the same influence. For this case it has been shown with an exemplary uni-modal function (the infinite-dimensional sphere function) that the setting $\mu = 0.27 \cdot \lambda$ is optimal in the sense that the "expected progress per sample" towards the global optimum is maximized (Hansen et al., 2015, Section 4.2.2)(cf. Beyer, 2001, Chapters 3.1.1 & 3.2.1.2). Hansen considers the value μ_{eff} to be a generalization of the number of selected independent samples that influence the distribution, consequently using the similar equation (3) for the case of rank dependent weights. Note that μ_{eff} takes its maximum μ with equal weights and its minimum 1, if all but one weight are zero. Actually, theoretically optimal non-equal weights and, thus, the optimal value for μ_{eff} are also known for the infinite-dimensional sphere function (Arnold, 2006, 3.2). These include non-zero weights for all λ samples and negative weights for the worse $\frac{\lambda}{2}$ samples (hence doubling the value of μ_{eff}). However, negative weights are not considered to be a robust enough practical choice.

Together with the problem dimension n, the generalized number of independent selected samples μ_{eff} appears in the calculation of the four smoothing constants $c_\sigma, c_c, c_\mu, c_1$ used in the update formulas of both the evolution paths and the covariance matrix. Their dependence on n and μ_{eff} have been derived empirically. The constant χ (cf. Subsubsection 2.2.6) is approximately the expected norm of the n-dimensional standard normal distribution $\mathcal{N}(0, \mathbf{I})$.

The algorithm details are summarized in Algorithm 1. It starts with the identity matrix $\mathbf{I}$ for the covariances, that is, with an isotropic distribution. Assuming the optimum solution to reside within the unit cube $[0,1]^n \subseteq \mathbb{R}^n$, the mean $\overline{\boldsymbol{x}}$ and the over all scale σ are initialized according to Hansen (2016). Actually, having bound constraints (cf. Subsubsection 2.2.7) we operate on the unit cube and shift and scale obtained samples into their real bounds before calculating their objective function values. New samples are drawn as described in Subsubsection 2.2.3. The $\boldsymbol{y}_k$ correspond to the $\boldsymbol{x}_k - \overline{\boldsymbol{x}}$ considered there, divided by the step size σ. The new $\overline{\boldsymbol{x}}$ is calculated according to $\overline{\boldsymbol{x}}_\mu$ in Subsubsection 2.2.4. Note that $\overline{\boldsymbol{y}}$ is the σ-adjusted move of the mean while $\overline{\boldsymbol{y}}^*$ adjusts the move of the mean with respect to the (isotropic) standard normal distribution. The evolution paths which cumulate the drifts of the distribution mean (adjusted with regard to the overall scale and with regard to isotropy, respectively) are updated using the corresponding smoothing factors. Here, the factors before $\overline{\boldsymbol{y}}$ and $\overline{\boldsymbol{y}}^*$ act as normalization constants (Hansen, 2016). Finally, the overall step size and the covariances are updated as described in Subsubsections 2.2.6 and 2.2.5, respectively. For the given weighs w the factor $\frac{c_\sigma}{1+c_\sigma}$ in the update formula of σ is equal to a more general formulation used, e.g., in the CMA-ES tutorial (Hansen, 2016). We stop either after the predefined number of iterations or if the current

Algorithm 1 The $(\mu/\mu_w, \lambda)$-CMA-ES

Initialization:

Set λ, μ, w, μ_{eff}, χ, c_σ, c_c, c_μ, c_1 according to Table 1

Set $\overline{\boldsymbol{x}} = (\frac{1}{2}, \ldots, \frac{1}{2})^{\mathrm{T}}$

Set $\boldsymbol{p}_\sigma = \boldsymbol{p}_c = 0,\ \mathbf{C} = \mathbf{B} = \mathbf{D} = \mathbf{I}$ and $\sigma = 0.5$

while stopping criterion*) is not met **do**

Sample probability distribution:

for $k = 1, \ldots, \lambda$ **do**

Sample $\boldsymbol{z}_k \in \mathbb{R}^n$ from $\mathcal{N}(0, \mathbf{I})$ by sampling its entries from $\mathcal{N}(0,1)$

Set $\boldsymbol{y}_k = \mathbf{B}\mathbf{D}\boldsymbol{z}_k$ and $\boldsymbol{x}_k = \overline{\boldsymbol{x}} + \sigma \boldsymbol{y}_k$

end for

Update probability distribution:

Update mean:

$\overline{\boldsymbol{x}} \leftarrow \sum_{k=1}^{\mu} w_k \boldsymbol{x}_k$

Set $\overline{\boldsymbol{y}} = \sum_{k=1}^{\mu} w_k \boldsymbol{y}_k$ and $\overline{\boldsymbol{y}}^* = \mathbf{B}\mathbf{D}^{-1}\mathbf{B}^{\mathrm{T}}\overline{\boldsymbol{y}}$

Update evolution paths:

$\boldsymbol{p}_\sigma \leftarrow (1 - c_\sigma)\boldsymbol{p}_\sigma + \sqrt{c_\sigma(2 - c_\sigma)\mu_{\text{eff}}}\ \overline{\boldsymbol{y}}^*$

$\boldsymbol{p}_c \leftarrow (1 - c_c)\boldsymbol{p}_c + \sqrt{c_c(2 - c_c)\mu_{\text{eff}}}\ \overline{\boldsymbol{y}}$

Update covariances and scaling:

$\sigma \leftarrow \sigma \cdot \exp\left(\frac{c_\sigma}{1+c_\sigma}\left(\frac{\|\boldsymbol{p}_\sigma\|}{\chi} - 1\right)\right)$

Set $\mathbf{C}_\mu = \sum_{k=1}^{\mu} w_k \boldsymbol{y}_k \boldsymbol{y}_k^{\mathrm{T}}$ and $\mathbf{C_1} = \boldsymbol{p}_c \boldsymbol{p}_c^{\mathrm{T}}$

$\mathbf{C} \leftarrow (1 - c_\mu - c_1)\mathbf{C} + c_1\mathbf{C_1} + c_\mu\mathbf{C}_\mu$

Determine $\mathbf{B}$ and $\mathbf{D}$ from eigendecomposition $\mathbf{C} = \mathbf{B}\mathbf{D}^2\mathbf{B}^{\mathrm{T}}$

end while

*) our stopping criterion is that either a predefined number of iterations is reached or the fitness distribution is flat (see text)

population shows a flat misfit distribution, i.e., if the fitness of the better 70% of the individuals deviate less than $\epsilon = 10^{-5}$ from the very best one.

2.3.2 Algorithm parallelization

Our current technical implementation of the parallel framework can be easily transferred to other EAs/EDAs. The iterative optimization process is carried out via a series of chain jobs, where short serial jobs (the actual optimizer) that update the
5 population of model evaluations ("individuals"; i.e. parameter sets for biogeochemistry) alternate with parallel jobs of function evaluations ("generations"), i.e. forward integrations of the coupled ocean model with different parameter sets. Parameters of the optimizer are population size λ and the termination criterion for convergence, additionally a maximum number of iterations.

As noted above, the framework presented here is set up such that a serial script `serial.job` calls the optimization routine (in our case CMA-ES), which computes a population of size = λ of parameter vectors, stored in ASCII files. The same script then calls a parallel script `parallel.job`, which starts λ model simulations. During these simulations, the parameter files are read, and a spinup is carried out for each individual setup. The individual model runs then output the misfit function to specified files. When all jobs are finished, script `parallel.job` invokes script `serial.job` again, etc.. Thus, communication between both alternating steps (creation of parameter vectors and computation of resulting misfit function) is carried out by these parameter and misfit files. In addition, file `nIter.txt` keeps track of the progress of optimization, and provides the information which generation is to be computed; it also contains the runtime parameters for the optimizer, CMA-ES. See the information in supplement for more details on how this setup works, and how to specify biogeochemical and optimizer parameters used, e.g., in the work presented here.

2.4 Misfit function

As a first approach to optimization, we have calculated the root-mean-square error RMSE between simulated and observed (or twin) annual mean phosphate, nitrate, and oxygen concentrations on a global scale, weighted by the volume V_i of each individual grid box, expressed as fraction of total ocean volume, V_T. To sum the three different components of the misfit function we have to divide them by some typical value. Here we use the global mean concentration of observed tracers. The resulting misfit function J thus reads:

$$J = \sum_{j=1}^{3} \frac{1}{\overline{o_j}} \sqrt{\sum_{i=1}^{N} (m_{i,j} - o_{i,j})^2 \frac{V_i}{V_\mathrm{T}}} \tag{4}$$

for the annual mean concentrations of three tracers phosphate ($j = 1$), nitrate ($j = 2$) and oxygen ($j = 3$), at $N = 52749$ locations (model grid boxes) of the model domain. $\overline{o_j}$ is the global average observed (or twin) concentration of the respective tracer. $m_{i,j}$ and $o_{i,j}$ are model and observations (or twin results), respectively. By weighting the model mismatch with volume, we put some emphasis on the deep ocean, down-weighting deviations in surface grid boxes relative to those of deep boxes. Thus, our misfit function serves more as a long time-scale geochemical estimator, in contrast to a function that focuses on (rather fast) turnover in the surface layer.

2.5 Parameters to be estimated

Although the model contains more than 20 parameters (even more, if we consider the empirically derived parameters for benthic burial, nitrogen fixation, denitrification and air-sea gas exchange; see Kriest and Oschlies, 2013, 2015), for this first approach we only consider six parameters for optimization. As a stringent test for the framework we chose parameters that encompass a large range of time and space scales, reflect different trophic levels and dependencies between internal (interactions between compartments) and external (dependence on light) factors. We aimed to avoid simultaneous optimization of

parameters that are obviously related to each other, such as maximum growth rates and half-saturation constants, or sinking speed and remineralization rate.

Four parameters are more relevant for biological interactions at the sea surface. Phytoplankton growth is controlled by the half-saturation for light (I_c, in W m^{-2}) and phosphate (K_{PHY}, in mmol P m^{-3}). For optimization of zooplankton parameters we chose its maximum grazing rate (μ_{ZOO}, in d^{-1}) and quadratic mortality rate (κ_{ZOO}, in (mmol P m^{-3})$^{-1}$ d^{-1}). Two parameters are of importance for the transport and decay of particulate organic matter to/in the deep ocean, namely the ratio of oxygen consumption to phosphate release during aerobic remineralization ($R_{-O2:P}$, mmol O_2:mmol P), and the parameter for vertical increase of sinking speed of organic matter, a (d^{-1}). Note that as stated above, in the following, and during optimization, we express this last parameter through $b = r/a$, with r held constant at $r = 0.05$ d^{-1}.

For each parameter we initially chose a rather wide range of possible parameter values (Table 2). The lower value of $R_{-O2:P}$ was set to 150 mmol O_2:mmol P (Anderson, 1995), while its upper value is at the upper end of observed values (Boulahdid and Minster, 1989), and closer to value used in previous model studies (Paulmier et al., 2009). b is allowed to vary between low values observed mainly in oxygen minimum zones (Van Mooy et al., 2002), and twice the global open ocean composite derived by Martin et al. (1987); its range is slightly larger than the range applied in previous modeling studies (Kwon and Primeau, 2006; Kriest and Oschlies, 2008; Kriest et al., 2012), or the range of b determined from in situ observations (e.g., Martin et al., 1987; Buesseler et al., 2007). It agrees with the range of b derived from indirect estimates of b (Henson et al., 2012; Marsay et al., 2015).

Ranges for parameters of parameters related to surface processes were more difficult to assign. Due to the highly aggregated form of the organic biological components in the model these parameters are supposed to reflect a variety of processes such as species shift and adaptation (e.g., half-saturation constants for nitrate uptake may vary over several orders of magnitude; see Collos et al., 2005). We therefore initially assigned very wide boundaries for I_c, K_{PHY}, μ_{ZOO} and κ_{ZOO}, which allow the optimization to pick parameters that virtually may shut down certain biological fluxes and processes. The choice of these wide boundaries, its consequences for optimization and model performance and the effects of narrower boundaries will be examined and discussed below.

2.6 Setup and performance of optimization

Using the combined framework described above, i.e. TMM+MOPS+CMA-ES, we carried out five different, full optimizations, with the aim to determine the four parameters related to surface biology and two parameters more closely tied to deep biogeochemistry mentioned above. The experiments differ with respect to the observations used for the misfit function (model output, climatologies of observations), population size λ of CMA-ES (10 or 20 individuals per generation), parameter boundaries, and the sampling strategy of CMA-ES. They are explained in detail below.

2.6.1 Twin experiment

First we tested the ability of CMA-ES to recover known parameters of a model simulation that applied the same biogeochemical parameters as MOPS-RemHigh of Kriest and Oschlies (2015, setup "base", i.e., with a particle flux described by $b = 0.858$, or

$a = 0.058275$, and a high affinity of oxic and suboxic remineralization to oxidants). This is done by optimization against its simulated annual average phosphate, nitrate and oxygen of year 3000. We refer to this experiment as "TWIN". TWIN applies rather wide boundaries for all parameters (see table 2), and a population size for CMA-ES of $\lambda = 10$, which was deemed sufficient for six parameters, given the default configuration of the CMA-ES (see above).

2.6.2 Optimizations against observed tracers

Four further optimizations were carried out against observations of annual mean phosphate, nitrate, and oxygen (Garcia et al., 2006a, b), gridded onto the model geometry. These are referred to as OBS-WIDE, OBS-WIDE-20, OBS-NARR and OBS-NARR-R. To investigate the robustness of CMA-ES with respect to different setups of the algorithm itself, the experiments differ in the upper and lower boundaries of the search space for zooplankton parameters, the population size λ of CMA-ES and its sampling strategy. This is done in a stepwise fashion.

Experiment "OBS-WIDE" differs from TWIN only with respect to the observations that enter the misfit function. In OBS-WIDE we encountered an unlikely (with respect to biological tracer concentrations) solution, pointing towards a potential local minimum in the misfit function. We therefore set up two experiments to investigate strategies to improve the performance of CMA-ES with respect to more plausible solutions. The experiments both increase the search density in the parameter space with respect to OBS-WIDE. In experiment "OBS-WIDE-20" search density is increased by doubling the population size of CMA-ES to $\lambda = 20$. Otherwise, its setup is the same as OBS-WIDE. In experiment "OBS-NARR" we keep $\lambda = 10$ of OBS-WIDE, but restrict the boundaries for zooplankton parameters to $\pm 50\%$ of the value of the reference run of MOPS.

Because optimization OBS-NARR showed the best results with respect to misfit function, biogeochemical fluxes and optimization performance (see below; tables 3 and 4), in experiment "OBS-NARR-R" we finally evaluate the robustness of optimization OBS-NARR by repeating this optimization with a different random selection of the parameter values from the distribution calculated by CMA-ES.

2.6.3 Performance

The internal termination criterion of CMA-ES was reached after 95, 173, 182 and 140 generations for OBS-WIDE, OBS-WIDE-20, OBS-NARR and OBS-NARR-R, respectively. For the twin experiment, we restricted the maximum number of generations to 200, at which TWIN had approached the target parameters, the misfit declined to < 0.0004 (i.e., on average less that $0.2‰$ of global mean tracer concentrations; see Eqn. 4) and fitness variance declined to $< 10^{-9}$. As presented above, in each "generation" we computed 10 (20) different "individuals" (model simulations over 3000 years) in parallel. One simulation of each generation on average took ≈ 1.25 hours, on 40 (80) nodes of Intel Xeon IvyBridge or Intel Xeon Haswell at the North-German Supercomputing Alliance (HLRN). We note that tests on either hardware (two iterations of the coupled code, started from generation 80 and 160 of experiment TWIN) did not reveal any differences in the estimated fitness. The CMA-ES - which, due to its very short runtime, is not parallelized - was always computed on one core of Intel Xeon IvyBridge.

3 Results

3.1 Twin experiment (TWIN)

The optimization starts with a wide range of potential parameters (see Fig. 3), with individual parameters sometimes even exceeding the prescribed boundaries. This results in high maximum and minimum misfit (Fig. 4), and this high variability is maintained over about 10-20 generations. The trajectory of transient average parameter values and their variance depend strongly on the parameter itself: while the two parameters associated with rather long time scales namely the stoichiometric ratio $R_{-\mathrm{O2:P}}$ and exponent b describing particle sinking, approach their target values quite early (about generation 20-40), parameters associated with surface biogeochemistry stay far away from their target value for $\approx$ 80 generations (I_{c}, K_{PHY}, κ_{ZOO}), or oscillate around it (μ_{ZOO}). After $\approx$ 160 generations, most of the parameters reached their target value, the exception being the half-saturation constant of phytoplankton for phosphate uptake, K_{PHY} (Table 3). This parameter still shows considerable variability at the end of the optimization (generation 200), although by that time is it quite close to the - rather low - target value.

The misfit function, its variance and the parameter variance do not decrease monotonically throughout the optimization trajectory. In particular, after an initial decline over ca. 60 generations, parameter and misfit variance increase again. Further increases in variance can be seen around generation 100, and at the end, when the algorithm widens its search area again, probably in search for an optimal K_{PHY}. It seems encouraging that the algorithm does not get stuck in a local minimum, but, at the expense of deterioration of the misfit, continues to search for an even better parameter set.

The largest fraction of the misfit function is related to oxygen, followed by the misfit to nitrate, and then phosphate. The dominance of oxygen and nitrate is not surprising, as these tracers are not conservative; i.e., their global inventory might change due to air-sea gas exchange, denitrification and nitrogen fixation (see also Kriest and Oschlies, 2015), so that the model may not only err with respect to the spatial distribution of these tracers, but also with respect to their global mean concentration.

In Fig. 5 we finally exploit the shape of the misfit function, shown on a color scale for each two pairs of parameters. As can be seen from misfit plotted against $R_{-\mathrm{O2:P}}$ and b (upper right corner), these two parameters are quite well constrained, with a very well defined minimum around the target value. All other parameters show more or less elongated search "canyons". Much of the algorithm search starts away from the target value; however, the algorithm finally manages to approach the target value even when the search path is not straight, but curved in the two-dimensional projections of the parameter space. Further, even when the algorithm exceeds the target value (e.g., for the maximum growth rate of zooplankton, μ_{ZOO}; lower right corner), despite of the already low misfit function the algorithm finally returns to the somewhat lower value (compare also to Fig. 3, lower left panel).

Summarizing, CMA-ES seems capable to deal even with our irregular search landscape, when iterated for a long enough time and with a sufficiently large population size. Some problem remains with regards to the half-saturation constant of phytoplankton for phosphate uptake: zooming into the scatter plot presented in Fig. 5 reveals that for this parameter the search landscape becomes quite uninformative (Fig. 6), with similar misfits around $\pm 2\%$ of its last value. Thus, a low misfit can be achieved within over a wide range of this parameter.

One reason for this low sensitivity of the misfit function to K_{PHY} may be found in the fact, that in the twin, against which the model is optimized, only very few (1%) phosphate values are at or below the target value of $K_{\mathrm{PHY}} = 0.03125$ mmol P m^{-3}. Therefore, besides the dominance of oxygen in the misfit function (Fig. 4) the misfit function is further dominated by phosphate concentrations outside the oligotrophic surface regions, rendering it quite insensitive to changes in the half-saturation constant at low values. In addition, a closer look at the misfit topography (Fig. 5) points towards a potential correlation of μ_{ZOO} and κ_{ZOO}, which may complicate the algorithm's search for an optimum set of parameters, thereby slowing down its convergence.

3.2 Optimization against observed nutrients and oxygen distributions

3.2.1 Wide boundary constraints for zooplankton (OBS-WIDE, OBS-WIDE-20)

When optimizing the model against observed concentrations with exactly the same setup as for experiment TWIN, optimization OBS-WIDE reaches the internal termination criterion of the CMA-ES at generation 95. Instead of declining exponentially towards zero, the misfit only declines from an average initial value of ≈ 0.8 to 0.477 (Fig. 7, Table 3), i.e. only slightly less than the misfit of the reference run (0.529). Also, the variance of misfit, as well as that of the parameters show a more or less gradual decline, without any intermittent increase (see supplement). Another notable difference to TWIN is the higher contribution of phosphate to the misfit function (Fig. 7).

Some parameters diverge strongly from those of there reference run. In particular, the phytoplankton's half-saturation constant for light, I_{c}, increases strongly up to its upper boundary (Fig. 8; Table 3; see also supplement for a plot of topography of the misfit function). However, the stronger light-limitation of phytoplankton growth is counteracted by a strong decrease in zooplankton growth rate, μ_{ZOO}, and a strong increase in its quadratic mortality rate, κ_{ZOO}. As a consequence, average and maximum zooplankton concentrations are $< 25\%$ and $< 50\%$ of that of the reference run in the surface layer (Fig. 9), while phytoplankton is strongly increased, when compared to the reference run. Most likely because the zooplankton-detritus pathway is nearly shut off, DOM concentrations are strongly increased. The reorganization of the pelagic food web in this optimized model scenario is reflected in the global annual biogeochemical fluxes: primary production is enhanced by almost 14%, but loss through grazing is reduced to about 1/3 of that of the reference run (Table 4). As a consequence, the largest fraction of recycling is through remineralization of detritus and DOM ($> 95\%$ of annual production), and only 4% through zooplankton excretion, while in the reference run zooplankton recycles almost 15% of annual production. Due to the reduced particle sinking speed shallow (130 m) and deep (2030 m) particle flux are reduced, as is benthic burial. While some of the simulated fluxes are within the observed estimates, too low zooplankton concentration, as well as resulting low zooplankton grazing are far outside observed estimates (see Table 4).

Therefore, although optimization OBS-WIDE against observations has decreased the misfit to observations to $\approx 90\%$ of that of the (subjectively tuned) reference run, the outcome is not overly satisfying with respect to the optimized parameters and the resulting dynamical behavior of the model. Obviously, the very wide boundary constraints we chose for the zooplankton parameters led to a solution where zooplankton is almost dead - a statistically optimal but biologically meaningless solution.

To examine if this optimization became trapped in a local minimum, in experiment OBS-WIDE-20 we increased the population size of CMA-ES from $\lambda = 10$ to $\lambda = 20$. Due to a larger population, in this optimization the variability of fitness (Fig. 10) and parameter values (Fig. 11) is maintained over a longer period, again, as for optimization TWIN, with intermittent increases of variance during the course of the optimization. Most importantly, using the setup of OBS-WIDE-20 the optimization finds very different parameters for many of the biogeochemical components:

$R_{-\mathrm{O2:P}}$ is now closer to the a priori value of 170, while optimal b has increased considerably to $b = 1.34$ (Table 3). The largest difference to both the reference run as well as optimization OBS-WIDE occurs for the four biogeochemical parameters that are more closely tied to surface processes: I_c decreases to less than 50% of its a priori value, while K_PHY is at its upper boundary of $0.5\,\mathrm{mmol\,P\,m^{-3}}$. Encouragingly, zooplankton parameters are now such that zooplankton is viable (Fig 9). Its maximum growth rate is very close to the a priori value of $2\,\mathrm{d^{-1}}$. Its mortality rate is still quite high; however, because of its high growth rate zooplankton plays a considerable role in the pelagic nitrogen budget , with global fluxes much closer to the observed ones than for optimization OBS-WIDE (Table 4). The topography of the - rather dense - scan of the parameter space of OBS-WIDE-20 (Fig. 12) points towards a potential correlation between K_PHY, μ_ZOO and κ_ZOO. In this projection, low misfit values occur along a concomitant increase of K_PHY with either μ_ZOO or κ_ZOO. This is also reflected in the high level of parametric uncertainty, as revealed by a large range of parameter values in the vicinity of the optimum (Table 3).

Summarizing, using a larger population size and thus a denser scan of the parameter space (see Fig. 12), CMA-ES has found a better solution, with respect to the misfit function (see Table 3) as well as a closer fit to biogeochemical fluxes and more plausible biological patterns.

3.2.2 Narrow boundary constraints for zooplankton (OBS-NARR and OBS-NARR-R)

Optimizations with a population size of $\lambda = 20$, as for OBS-WIDE-20, are computationally quite expensive, especially when iterated over a large number of generations (Table 3). Via the quite wide boundary constraints for zooplankton parameters, we have assumed to have almost no knowledge about zooplankton. In the following two sensitivity experiments we examine the impact of this assumption on optimization performance, by restricting zooplankton parameters to a narrower range. These experiments are again carried out with a population size of $\lambda = 10$.

To enforce live zooplankton, we restricted the range of zooplankton parameters to $\pm 50\%$ of their reference value. This results indeed in a solution with organic tracer concentrations close to that of the reference run or OBS-WIDE-20 (Fig. 9). After 182 generations, the algorithm terminates with a misfit of 0.45 (Fig. 13), i.e. better than experiment OBS-WIDE, but the same as for optimization OBS-WIDE-20 (Table 3). As in TWIN and OBS-WIDE-20, misfit variance shows intermittent increases, and the contribution of nitrate to the misfit function dominates over that of phosphate. Likewise, resulting optimal parameter values are quite close to those of OBS-WIDE-20 (Table 3). Thus, OBS-NARROW is more similar to OBS-WIDE-20 than to OBS-WIDE, demonstrating the importance of good a priori knowledge about parameter values.

As for OBS-WIDE-20, the quadratic mortality of zooplankton, κ_ZOO and the half-saturation constant of phosphate uptake for phytoplankton, K_PHY show a strong increase; the latter up to its upper prescribed boundary, which may be interpreted as an attempt of the algorithm to force the model towards higher surface nutrient concentrations in the subtropical gyres. A reduced

half-saturation constant for light, on the other hand, counteracts the grazing pressure exerted by zooplankton, particularly in the high latitudes. Most likely because of increased detritus production by zooplankton - and thus increased export from the surface layer (Table 4) - particle flux to the deep ocean is reduced by an increase in b, i.e. relatively slow particle sinking speed.

A closer look at the topography of the misfit function shows that the misfit is quite insensitive to changes in some parameters (Fig. 15; see supplement for a detailed plot of misfit topography around $\pm 2\%$ of the optimal parameters). While again the parameters $R_{-\mathrm{O2:P}}$ and b, that tend to exert an influence on large temporal and spatial scales, are quite well constrained, many of the surface-related parameters, that act on smaller time scales, such as K_{PHY}, show a wide scatter across the parameter space (see also Table 3), with very little differences in the misfit function.

However, variations in parameters after ≈40 generations do not strongly improve the model fit to observations (Figures 13 and 14). The rather constant misfit after generation 40 is quite surprising, given that some parameters still show some significant excursions after that time, indicating that - as already shown before - the misfit function is quite uninformative about these parameters. This insensitivity of inorganic tracers is also illustrated in Fig. 16, which shows the deviation of vertically integrated tracers from observations, plotted for individuals of three different generations of OBS-NARR (see also blue vertical lines in Fig. 14) The parameters of these individuals differ mainly with respect to their combination of K_{PHY} and κ_{ZOO}. While the reference run applies very low $K_{\mathrm{PHY}} = 0.03125$ mmol P m^{-3} and moderate $\kappa_{\mathrm{ZOO}} = 3.2$ (mmol P m^{-3})$^{-1}$ d^{-1}, individuals of the optimization are characterized by medium (generation 61) to high (generations 110 and 182) K_{PHY}, and moderate (generations 61 and 110) and high (generation 182) κ_{ZOO} (see also blue vertical lines in Fig. 14). All individuals differ from the reference run; yet the difference among them is almost not visible in the simulated tracer distributions. Thus, annual mean tracer concentrations on a global scale do not seem to suffice in constraining some of the parameters related to the very dynamic biological turnover at the sea surface, leading to a large parametric uncertainty (Table 3), possibly amplified by correlation among these three parameters.

Except for deep particle fluxes, all biogeochemical fluxes are increased compared to the reference run or experiment OBS-WIDE, but similar to that of OBS-WIDE-20 (Table 4). Therefore, although the misfit function so far only optimized towards inorganic constituents, the optimized model with narrow zooplankton parameter boundaries shows a much better fit to observed global fluxes to primary production, zooplankton grazing, shallow and deep particle flux, and benthic burial. The seemingly better dynamical biogeochemical behavior of this model setup gives some confidence that the model's fit to inorganic tracers is not improved at the cost of any other tracer.

Repeating optimization OBS-NARR with a different random selection of parameters from the parameter distribution in each generation (OBS-NARR-R) yields the same, or very similar, best values for most of the parameters (see Table 2), the exception being the two zooplankton parameters, μ_{ZOO} and κ_{ZOO}. These two parameters of OBS-NARR-R are 7% (μ_{ZOO}) and 16% (κ_{ZOO}) lower than in OBS-NARR; however, the misfit of both optimizations is the same (0.45). The low sensitivity of the misfit function to zooplankton parameters is mirrored in similar nutrient and oxygen distributions (see supplement) and almost identical biogeochemical fluxes (see Table 4).

4 Discussion

4.1 Computational performance

Our results suggest that the CMA-ES optimization algorithm performs well, particularly for the twin experiment, even though the parameters to be estimated involve diverse temporal and spatial scales. CMA-ES manages to set up curved search paths in parameter space, and therefore is capable to approach an optimum within a rather complex topography of the misfit function. Its sometimes elongated and/or curved shape resembles many of those resulting from earlier 1D (Athias et al., 2000; Schartau et al., 2001; Schartau and Oschlies, 2003a; Ward, 2009) or 3D (Kwon and Primeau, 2006, 2008) optimizations of marine biogeochemical models. However, when imposing wide boundary constraints for zooplankton parameters, OBS-WIDE becomes trapped in a local minimum; only with a larger population size or narrower parameter boundaries we find a solution that results in realistic concentrations and fluxes of all components. Clearly, the number of experiments conducted here is too small to make statistically significant statements about the optimizers' exploration capability with respect to the population size. But similar to other population based heuristics, examinations with multimodal test functions have given evidence that larger populations increase CMA-ES' chance to find good local optima (or even a global optimum; Hansen and Kern, 2004). It remains to be investigated, whether different configurations of the CMA-ES, or a different optimization algorithm, e.g., gradient-based methods or evolutionary algorithms, perform better or worse with respect to the number of model evaluations required, or their ability to avoid local minima (see also Athias et al., 2000). However, there is some indication that genetic algorithms perform better with respect to a rough topography of the misfit function, when compared to a variational adjoint method, with otherwise equally good fit to marine biogeochemical observations (Ward et al., 2010).

As the computational effort remains a challenge in parameter optimization of global ocean BGC models, further possibilities to accelerate model evaluations within the optimization process are desirable. Surrogate-assisted approaches use meta-models to approximate model evaluations within optimization (Priess et al., 2013). They are becoming practice within evolutionary frameworks coping with computational expensive model functions (Jin, 2011). It should be worth considering surrogate approaches with CMA-ES as investigated in Kern et al. (2006), Auger et al. (2013) and Loshchilov et al. (2012). A general approach with EA and EDA frameworks is to prematurely abort the fitness calculation after detecting that the corresponding individual will not be better than the worst member of the current population. We can benefit from such short-cut fitness computation if the optimizers' implementation supports asynchronous communication. An example for this approach is dealt with in Kliemann et al. (2013). There, aborting fitness calculations reduces the computational effort by orders of magnitude, since the considered combinatorial problem is of minimax-type. However, short-cut fitness computation concerning ocean models requires a more elaborated method and is not expected to reach similar savings.

4.2 Misfit function and parameter identifiability

In our study we chose annual means of dissolved nutrients and oxygen on a rather coarse spatial grid as a measure for model skill. By doing so, we avoid problems associated with time lags (e.g., in phytoplankton blooms, which would result in time lags of nutrient depletion) or meso- and submesoscale spatial structures (see, e.g., Wallhead et al., 2006), obviously at the cost

of precisely resolving parameters related to the biological system in surface layers. Possibly as a consequence of this particular misfit function, the parameters that could be fitted best are parameters that are mostly influential in determining the nutrient or oxygen distribution on large spatial and temporal scales, such as the stoichiometric ratio between oxygen and phosphorus, $R_{-\mathrm{O2:P}}$, or the parameter that determines particle sinking speed, b (see also Kriest et al., 2012). Our model optimizations against observations so far confirm a stoichiometry of $R_{-\mathrm{O2:P}} \approx 170$ mmol O_2:mmol P, in agreement with observational estimates (Takahashi et al., 1985; Anderson and Sarmiento, 1994), but suggest an increase of b towards ≈ 1.3. The latter is to some extent in agreement with results obtained by Kwon and Primeau (2006, 2008), who found an optimal b of 1, when fitting a simple global model against observed inorganic tracers. It should be kept in mind, however, that the b obtained in our study represents not only particle sinking speed, but also accounts for the effect of numerical diffusion in our rather coarse vertical grid (Kriest and Oschlies, 2011). Accordingly, the "true" b can be regarded as being about 10-20% smaller than obtained by our study. Also, as has been shown earlier (Kriest and Oschlies, 2013), the lower boundary condition simulated by benthic exchange can be very important for the ability of phosphate and oxygen to constrain particle sinking; therefore, the results obtained in our study should be regarded as specific to this particular biogeochemical model.

Our optimizations against observations with wide and narrow boundaries for zooplankton parameters produced two solutions with quite similar misfit, but with very different biological parameters, and consequently different fluxes and concentrations of organic components in the surface layers. Using wide boundary constraints for zooplankton parameters resulted in a solution where zooplankton is almost extinct, while phytoplankton and DOM concentration are far too high. Solutions of optimizations with unrealistic parameter values or concentrations for zooplankton have been observed earlier (Schartau et al., 2001; Ward et al., 2010), and point towards a necessity to better constrain this compartment. Increasing the population size λ of CMA-ES in optimization OBS-WIDE-20 could cure this problem, but at the cost of a high computational demand. Restricting the range of zooplankton parameters resulted in a better fit to nutrient and oxygen; more importantly, concentrations and fluxes in the latter solution are much more realistic, confirming in the latter parameter set. This illustrates the potential benefit of a sound a priori knowledge of parameter ranges, both in terms of biogeochemical and computational performance.

Another possibility to avoid undesired effects like nearly extinct zooplankton is to introduce further criteria that take account of this issue. A technically easy approach would be to add further objective terms to the misfit function. But facing complex model interactions, it can become difficult to find suitable weights for the different terms in order to force solutions to become a desired compromise of objectives. An alternative is to deal with more than one objective function, say $f_1, f_2, \ldots, f_k$. For example, we can define the deviation of zooplankton mass from observed values as a second objective. Now, two solutions $x \neq y$ are said to be incomparable if $f_i(x) > f_i(y)$ but $f_j(x) < f_j(y)$ for some $i \neq j$. Multi-objective optimization algorithms aim to find (a limited number of) good incomparable solutions, from which the user can make a final choice that is a good compromise in his/her opinion. The topic of multi-objective optimization is intensively regarded with EAs (Deb, 2001) and EDAs (Hauschild and Pelikan, 2011), including CMA-ES (Igel et al., 2007).

Nevertheless, even for the more realistic optimizations OBS-WIDE-20, OBS-NARR and OBS-NARR-R we find similar misfits for a rather wide range of some phyto- and zooplankton parameters, pointing towards an indeterminacy of these parameters when using the current misfit function. While it cannot be ruled out that this arises from a correlation among these parameters,

even simpler biogeochemical models with less degrees of freedom might be difficult to constrain from nutrient data alone: problems were also encountered by Kwon and Primeau (2006), when optimizing b, DOP production and its decay rate against phosphate on a global scale. They found that phosphate data alone were not sufficient to resolve parameters associated with DOP, but several equally good fits could be obtained with different sets of parameters. It remains to be investigated whether this is related to the lack of seasonal data, or to phosphate concentration being weakly dependent on dissolved or particular organic matter concentration. Subsequent studies with different misfit functions, that for example resolve monthly changes, target at the representation of surface nutrients (e.g., by using a weighted, relative misfit; Kriest et al., 2010) or add additional tracers to the misfit function (e.g., combining chlorophyll derived from remote sensing with nitrate observations; see also Tjiputra et al., 2007) will reveal the effect of the assumptions made for the misfit function with respect to constraining these parameters.

4.3 Future directions

Even the use of observations more closely related to surface biology may not resolve the problem of indeterminacy, as shown by Ward et al. (2010) in optimizations of two different, 0D-biogeochemical models. As in earlier, 0D and 3D studies (e.g., Friedrichs, 2001; Schartau et al., 2001; Kwon and Primeau, 2006, 2008), they found almost identical misfits for a wide range of parameters, an indication that these models are underdetermined, particularly when attempting to estimate more than about 10 parameters. In our study we have chosen to tune a rather moderate number of six parameters, but already noted some difficulty in constraining two of these. A potential solution could be to fix certain parameters to prior values, and thereby decrease the dimension of the parameter space to be estimated. However, as pointed out by Ward et al. (2010), this may lead to an underestimate of model uncertainty, and therefore not be the ultimate solution to this problem. Future studies will address these problems by testing different combinations of parameters, in conjunction with different misfit functions.

The above mentioned problems may even increase if we move towards more sparsely sampled, biased, or noisy data. So far, for the twin experiment as well as for the optimization against observations we assume perfect data coverage. However, sparse data sets (as usually available from cruises or time series stations) as well as the influence of noise have been shown to be very influential for the ability of an optimization to recover results from 0D (Friedrichs, 2001; Schartau et al., 2001; Löptien and Dietze, 2015) and 3D (Tjiputra et al., 2007) twin experiments. The presence of noise or measurement errors should be reflected in the termination criterion for optimization; this will, for some parameters, influence the estimates optimum values (see Fig. 8 of Schartau et al., 2016). Future studies will have to address to what extent noise will affect the 3D optimizations presented here and how this parameter uncertainty will map onto model fluxes, or even transient scenarios.

While we found a decrease of the twin experiment's misfit to almost zero, the misfit of the optimization against observations remained relatively high (on average, about 15% of global mean tracer concentrations). Potential reasons for this are an inappropriate biogeochemical model structure, wrong choice of parameters to be optimized, or flaws in the physical model. For example, it is well known that coarse resolution models do not resolve physical processes of the Equatorial Pacific current system (Dietze and Loeptien, 2013), which may result in an attempt of the optimization to "cure" deficient physics by changing biogeochemical parameters. This feature might also explain some of the sensitivities - or lack of - found by Kwon and Primeau

(2006). Solutions to this potential flaw could be to exclude regions from the misfit, that are known to be not well represented by the physical model, or to weigh biogeochemical misfits by the model's fit to observations of physical data.

To summarize, any global model study that aims to inversely determine parameters of a global biogeochemical ocean model in an attempt to find the model setup "best" suited for a particular application (and circulation), has to consider five tasks: (1) investigate model solutions on the appropriate (depending on tunable parameters) time scales, possibly including long, millennial simulations; (2) address the potential of local minima (depending on the topography of the misfit function); (3) investigate different parameter combinations and boundaries, including the misfit function's sensitivity to them; (4) disentangle the effects of physical and biogeochemical model on model-data misfit; and (5) investigate the effect of misfit function, including data distribution and availability on model assessment. This last point also includes decisions about weights applied to different data sets, or for a particular form of misfit function, which may be very influential for the optimal parameter choice (Evans, 2003). It also depends on the desired application of the model, and the scientific question it is supposed to address.

5 Conclusions

We have presented a framework for the optimization of global biogeochemical ocean models, that combines an offline approach for transport of biogeochemical tracers with an Estimation of Distribution Algorithm (Covariance Matrix Adaption Evolution Strategy, CMA-ES). A twin experiment revealed a good performance of this algorithm with respect to recovering six parameters, that are associated with various time and space scales. Optimizations against observations of annual mean nutrients and oxygen could reduce the misfit of the model to some extent; however, even for the "best" model solution the remaining misfit is still $\approx 15\%$ of global mean tracer concentrations, which might be related to inappropriate physics. Tests with different circulation (which is easy to exchange with the current framework) will provide more insight into the impact of physical forcing on the ability of the biogeochemical model to fit the observations.

Encouragingly, parameter sets associated with the lowest misfit to dissolved inorganic tracers also show the best fit to global mean tracer fluxes not considered during optimization. This increases our confidence in the method presented here. Some parameter estimates are associated with a rather high level of uncertainty. Incorporating different or additional data sets, that more closely relate to the parameters to be optimized, can help to improve estimates for these parameters. Likewise, observations that provide information about the upper and lower bounds of biological parameters - such as zooplankton grazing and mortality rates - will provide a good guidance for future optimization studies, and lower their computational demand.

6 Code availability

The source code of MOPS coupled to TMM, as well as the optimization framework are available as supplement. The most recent TMM source code, forcing, etc. are available under
`https://github.com/samarkhatiwala/tmm`.

Appendix A: Source code

As research questions may diverge strongly (and therefore, also the different user groups, hardware, biogeochemical models and circulations), we aimed to construct a tool that is as generic and universally applicable as possible, with a high level of portability among different architectures. The model-optimization framework of TMM comprises new subroutines for data assimilation and misfit function evaluation, as well as monitor routines to facilitate run-time checks of model state, and a more generic coupling interface for biogeochemistry. It can thus easily be applied within an optimization framework. While we here focus on the coarse resolution model, we note that the generic structure of the TMM framework allows the user to easily switch between transport matrices, once these are available. Likewise, coupling different biogeochemical models to the framework only requires editing of a (few) interface subroutines. Finally, in principle it should be possible to exchange the optimization algorithm by any other algorithm, that requires only model misfit as input, and provides a set of parameter files as output.

A1 MOPS-2.0 biogeochemical subroutines

Besides the stand-alone, forward integration of a global biogeochemical model two additional tasks are required for optimization: computation and output of misfit, and input of trial sets of parameters passed to the model by the optimizer. In the following, files relevant for input of parameter vectors and computation of misfit that have been added or changed (with respect to MOPS-1.0; see Kriest and Oschlies, 2015) are shown in bold face. An overview of the model structure and layout, with emphasis on those parts that affect computation of biogeochemical fluxes and tracers, optimization and parameter handling is given in the supplement.

As noted in Kriest and Oschlies (2015), the code consists of two files with outer routines, that connect to the main driver code `tmm_main.c`, and inner routines that contain the local biogeochemical sources and sinks, and define the biogeochemical parameters. These routines communicate via common blocks in header files.

(1) `external_forcing_mops_biogeochem.c` is the first interface between MOPS and the TMM. Besides in- and output of files and runtime parameters it determines from runtime options whether a parameter file should be read, as well as its name. It assembles model equivalents for the misfit function and passes it to the main driver code, `tmm_main.c`. It calls the following subroutines:

`(1.2) mops_biogeochem_ini.F`: interface between `(1)` and `(1.2.1)`. It calls

(1.2.1) `BGC_INI.F` assigns biogeochemical parameters. The routine distinguishes between parameters that stay fixed, and derived parameters that depend on parameters read during runtime. For example, the stoichiometric ratio O_2:P determines the stoichiometry for nitrate loss during denitrification (Paulmier et al., 2009). Thus, if the former changes, the latter will have to be recalculated. The routine is called every time after a new parameter vector has been read.

(1.3) `mops_biogeochem_set_params.F` maps vector of parameters read by `(1)` to symbolic names used by MOPS. Each call to `(1.3)` is followed by a call to `(1.2)` and `(1.2.1)`, to recalculate dependent parameters.

(1.1) mops_biogeochem_copy_data.F: interface between (1) and (1.2) and (1.4).

(1.4) mops_biogeochem_model.F: interface between (1) and (1.4.1). It calls

(1.4.1) BGC_MODEL.F calculates biogeochemical sources and sinks. It now also assigns state variables to arrays that will be passed to the misfit function.

(1.5) mops_biogeochem_diagnostics.F: interface (for diagnostic output) between (1) and (1.4.1).

(1.6) mops_biogeochem_misfit.F: interface for misfit computation between (1) and (1.4.1).

(2) tmm_misfit.c initializes and carries out misfit computation. Writes misfit to either binary or ASCII files. It communicates with the biogeochemical model in (1) via (b).

Communication between the different modules is carried out mainly via several header files:

(a) mops_biogeochem.h introduces subroutines to (1).

(b) mops_biogeochem_misfit_data.h passes information related to misfit computation between (1) and (2).

(c) BGC_PARAMS.h passes biogeochemical parameters and profiles of tracers between all different modules called by (1).

(d) BGC_DIAGNOSTICS.h passes diagnostic variables from (1.4.1) to (1.5).

(e) BGC_MISFIT.h passes misfit variables from (1.4.1) to (1.6).

(f) BGC_CONTROL.h passes time step and geometry between (1.2) and (1.2.1), (1.4) and (1.4.1),

(g) **tmm_external_forcing.h** introduces subroutines in (1) to tmm_main.c.

(h) **tmm_misfit.h** introduces subroutines in (2) to tmm_main.c.

Finally, one may want to prevent computation of a simulation if during spinup some parameter values or concentrations lead to erroneous (e.g., negative) tracer concentrations. Routine tmm_monitor.c may serve as a module to monitor state variables, or other model properties (not used in the current setup presented here).

A2 Optimization

As noted above, the framework presented here is set up such that a serial script serial.job calls the optimization routine (in our case CMA-ES), which computes a population of size = λ of parameter vectors, stored in ASCII files. The same script then calls a parallel script parallel.job, which starts λ model simulations. During these simulations, the parameter files are read, and a spinup is carried out for each individual setup. The individual model runs then output the misfit function to specified files. When all jobs are finished, script parallel.job invokes script serial.job again, etc.. Thus, communication between both alternating steps (creation of parameter vectors and computation of resulting misfit function) is carried out

by these parameter and misfit files. In addition, file `nIter.txt` keeps track of the progress of optimization, and provides the information which generation is to be computed; it also contains the runtime parameters for the optimizer, CMA-ES. See information in supplement for more details on how this setup works, and how to specify biogeochemical and optimizer parameters used e.g., in the work presented here.

Acknowledgements. This work is a contribution to the DFG-supported project SFB754 and to the research platforms of the DFG cluster of excellence The Future Ocean. We thank Nikolaus Hansen for support on and open access to the CMA-ES code. Parallel supercomputing resources have been provided by the North-German Supercomputing Alliance (HLRN). The authors wish to acknowledge use of the Ferret program of NOAA's Pacific Marine Environmental Laboratory for analysis and graphics in this paper. We thank Momme Butenschön and an anonymous reviewer for their very constructive and helpful comments.

References

Anderson, L.: On the hydrogen and oxygen content of marine phytoplankton, Deep-Sea Res. Pt.I, 42, 1675–1680, 1995.

Anderson, L. and Sarmiento, J.: Redfield ratios of remineralization determined by nutrient data analysis, Global Biogeochem. Cy., 8, 65–80, 1994.

Arnold, D. V.: Weighted multirecombination evolution strategies., Theoretical Computer Science, 361, 18–37, 2006.

Athias, V., Mazzega, P., and Jeandel, C.: Selecting a global optimization method to estimate the oceanic particle cycling rate constants, J. Mar. Res., 58, 675–707, 2000.

Auger, A., Brockhoff, D., and Hansen, N.: Benchmarking the local metamodel CMA-ES on the noiseless BBOB'2013 test bed, in: Genetic and Evolutionary Computation Conference, GECCO 2013, Amsterdam, The Netherlands, July 6-10, 2013, Companion Material Proceedings, pp. 1225–1232, 2013.

Babu, G. S. S., Das, D. B., and Patvardhan, C.: Solution of real-parameter optimization problems using novel Quantum Evolutionary Algorithm with applications in power dispatch, in: Proceedings of the IEEE Congress on Evolutionary Computation, CEC 2009, Trondheim, Norway, 18-21 May, 2009, pp. 1920–1927, 2009.

Beyer, H.-G.: The theory of evolution strategies., Springer, Berlin, 2001.

Boulahdid, M. and Minster, J.-F.: Oxygen consumption and nutrient regeneration ratios along isopycnal horizons in the Pacific Ocean, Mar. Chem., 26, 133–153, 1989.

Buesseler, K., Lamborg, C., Boyd, P., Lam, P., Trull, T., Bidigare, R., Bishop, J., Casciotti, K., Dehairs, F., Elskens, M., Honda, M., Karl, D., Siegel, D., Silver, M., Steinberg, D., Valdes, J., Mooy, B. V., and Wilson, S.: Revisiting carbon flux through the ocean's twilight zone, Science, 316, 567–570, 2007.

Cabre, A., Marinov, I., Bernadello, R., and Bianchi, D.: Oxygen minimum zones in the tropical Pacific across CMIP5 models: mean state differences and climate change trends, Biogeosciences, 12, 5429–5454, doi:10.5194/bg-12-5429-2015, www.biogeosciences.net/12/5429/2015/, 2015.

Carr, M.-E., Friedrichs, M., Schmeltz, M., Aitac, M., Antoine, D., Arrigo, K., Asanuma, I., Aumont, O., Barber, R., Behrenfeld, M., Bidigare, R., Buitenhuis, E., Campbell, J., Ciotti, A., Dierssen, H., Dowell, M., Dunne, J., Esaias, W., Gentili, B., Gregg, W., , Groom, S., Hoepffner, N., Ishizaka, J., Kameda, T., Quere, C. L., Lohrenz, S., Marra, J., lino, F. M., Moore, K., Morel, A., Reddy, T., J.Ryan, Scardi, M., T.Smyth, Turpie, K., Tilstone, G., Waters, K., and Yamanaka, Y.: A comparison of global estimates of marine primary production from ocean color, Deep-Sea Res. Pt. II, 53, 741–770, doi:10.1016/j.dsr2.2006.01.028, 2006.

Collos, Y., Vaquer, A., and Souchou, P.: Acclimation of nitrate uptake by phytoplankton to high substrate levels, jp, 41, 466–478, f4a, 2005.

Cover, T. M. and Thomas, J. A.: Elements of Information Theory, John Wiley & Sons, Hoboken, NJ, 2006.

Deb, K.: Multi-objective optimization using evolutionary algorithms. John Wiley & Sons, Hoboken, NJ, 2001.

Dietze, H. and Loeptien, U.: Revisiting "nutrient trapping" in global coupled biogeochemical ocean circulation models, Global Biogeochem. Cy., 27, 265—284, doi:10.1002/gbc.20029, 2013.

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

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

Fasham, M. and Evans, G.: 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.

Friedrichs, M. A. M.: A data assimilative marine ecosystem model of the central equatorial Pacific: Numerical twin experiments, Jour. Mar. Res., 59, 859–894, 2001.

Garcia, H. E., Locarnini, R. A., Boyer, T. P., and Antonov, J. I.: World Ocean Atlas 2005, Vol. 4: Nutrients (phosphate, nitrate, silicate), in: NOAA Atlas NESDIS 64, edited by Levitus, S., U.S. Government Printing Office, Wash.,D.C., http://iridl.ldeo.columbia.edu/SOURCES/.NOAA/.NODC/.WOA05/, 2006a.

Garcia, H. E., Locarnini, R. A., Boyer, T. P., and Antonov, J. I.: World Ocean Atlas 2005, Vol. 3: Dissolved Oxygen, Apparent Oxygen Utilization, and Oxygen Saturation, in: NOAA Atlas NESDIS 63, edited by Levitus, S., U.S. Government Printing Office, Wash.,D.C., http://iridl.ldeo.columbia.edu/SOURCES/.NOAA/.NODC/.WOA05/, 2006b.

Hansen, N.: The CMA evolution strategy: a comparing review, in: Towards a new evolutionary computation. Advances on estimation of distribution algorithms, edited by Lozano, J. A., Larranaga, P., Inza, I., and Bengoetxea, E., pp. 75–102, Springer, 2006.

Hansen, N.: The CMA Evolution Strategy: a tutorial, arXiv:1604.00772v1, 2016.

Hansen, N. and Kern, S.: Evaluating the CMA Evolution Strategy on Multimodal Test Functions, in: Parallel Problem Solving from Nature PPSN VIII, edited by Yao, X. et al., vol. 3242 of *LNCS*, pp. 282–291, Springer, 2004.

Hansen, N. and Ostermeier, A.: Completely Derandomized Self-Adaptation in Evolution Strategies, Evolutionary Computation, 9, 159–195, 2001.

Hansen, N., Finck, S., Ros, R., and Auger, A.: Real-Parameter Black-Box Optimization Benchmarking 2009: Noiseless Function Definitions, Tech. Rep. inria-00362633, INRIA, Le Chesnay, France, https://hal.inria.fr/inria-00362633/document, 2009a.

Hansen, N., Niederberger, A. S. P., Guzzella, L., and Koumoutsakos, P.: A Method for Handling Uncertainty in Evolutionary Optimization with an Application to Feedback Control of Combustion, IEEE Transactions on Evolutionary Computation, 13, 180–197, 2009b.

Hansen, N., Auger, A., Ros, R., Finck, S., and Posík, P.: Comparing results of 31 algorithms from the black-box optimization benchmarking BBOB-2009, in: Proceedings of the Genetic and Evolutionary Computation Conference, GECCO 2010, Portland, Oregon, USA, July 7-11, 2010, Companion Material, pp. 1689–1696, 2010.

Hansen, N., Arnold, D. V., and Auger, A.: Evolution Strategies, in: Springer Handbook of Computational Intelligence, pp. 871–898, Springer, Berlin, 2015.

Hauschild, M. and Pelikan, M.: An Introduction and Survey of Estimation of Distribution Algorithms, http://medal-lab.org/files/2011004_rev1.pdf, 2011.

Hemmings, J. C. P., Challenor, P., and Yool, A.: Mechanistic site-based emulation of a global ocean biogeochemical model for parametric analysis and calibration, Geosci. Model Dev. Disc., 7, 6327–6411, doi:10.5194/gmdd-7-6327-2014, www.geosci-model-dev-discuss.net/7/6327/2014/, 2014.

Henson, S. A., Sanders, R., and Madsen, E.: Global patterns in efficiency of particulate organic carbon export and transfer to the deep ocean, Global Biogeochem. Cy., 26, GB1028, 2012.

Honjo, S., Manganini, S. J., Krishfield, R. A., and Francois, R.: Particulate organic carbon fluxes to the ocean interior and factors controlling the biological pump: A synthesis of global sediment trap programs since 1983, Prog. Oceanogr., 76, 217–285, doi:10.1016/j.pocean.2007.11.003, 2008.

Igel, C., Hansen, N., and Roth, S.: Covariance Matrix Adaptation for Multi-objective Optimization, Evolutionary Computation, 15, 1–28, 2007.

Jin, Y.: Surrogate-assisted evolutionary computation: Recent advances and future challenges, Swarm and Evolutionary Computation, 1, 61–70, 2011.

Jost, L.: Entropy and Diversity, OIKOS, 113, 363–375, 2006.

Kern, S., Hansen, N., and Koumoutsakos, P.: Local Meta-models for Optimization Using Evolution Strategies, in: Parallel Problem Solving from Nature - PPSN IX, 9th International Conference, Reykjavik, Iceland, September 9-13, 2006, Procedings, pp. 939–948, 2006.

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

Kliemann, L., Kliemann, O., Patvardhan, C., Sauerland, V., and Srivastav, A.: A New QEA Computing Near-Optimal Low-Discrepancy Colorings in the Hypergraph of Arithmetic Progressions, in: Proceedings of the 12th International Symposium on Experimental and Efficient Algorithms, Rome, Italy, June 2013 (SEA 2013), edited by Bonifaci, V., Demetrescu, C., and Marchetti-Spaccamela, A., no. 7933 in Lecture Notes in Computer Science, pp. 67–78, 2013.

Kriest, I. and Oschlies, A.: On the treatment of particulate organic matter sinking in large-scale models of marine biogeochemical cycles, Biogeosciences, 5, 55–72, http://www.biogeosciences.net/5/55/2008/, 2008.

Kriest, I. and Oschlies, A.: Numerical effects on organic matter sedimentation and remineralization in biogeochemical ocean models, Ocean Modell., 39, 275–283, doi:10.1016/j.ocemod.2011.05.001, 2011.

Kriest, I. and Oschlies, A.: Swept under the carpet: The effect of organic matter burial in global biogeochemical ocean models, Biogeosciences Disc., 10, 10 859–10 911, doi:10.5194/bgd-10-10859-2013, www.biogeosciences-discuss.net/10/10859/2013/, 2013.

Kriest, I. and Oschlies, A.: MOPS-1.0: towards a model for the regulation of the global oceanic nitrogen budget by marine biogeochemical processes, Geoscientific Model Development, 8, 2929–2957, doi:10.5194/gmd-8-2929-2015, www.geosci-model-dev.net/8/2929/2015/, 2015.

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

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

Kwon, E. Y. and Primeau, F.: Optimization and sensitivity study of a biogeochemistry ocean model using an implicit solver and in situ phosphate data, Global Biogeochem. Cy., 20, doi:10.1029/2005GB002631, 2006.

Kwon, E. Y. and Primeau, F.: Optimization and sensitivity of a global biogeochemistry ocean model using combined in situ DIC, alkalinity, and phosphate data, J. Geophys. Res., 113, doi:10.1029/2007JC004520, 2008.

Letscher, R., Moore, J. K., Teng, Y.-C., and Primeau, F.: Variable C : N : P stoichiometry of dissolved organic matter cycling in the Community Earth System Model, Biogeosciences, 12, 209–221, doi:10.5194/bg-12-209-2015, www.biogeosciences.net/12/209/2015/, 2015.

Löptien, U. and Dietze, H.: Constraining parameters in marine pelagic ecosystem models – is it actually feasible with typical observations of standing stocks?, Ocean Sci., 11, 573–590, doi:10.5194/os-11-573-2015, www.ocean-sci.net/11/573/2015/, 2015.

Loshchilov, I., Schoenauer, M., and Sebag, M.: Self-adaptive surrogate-assisted covariance matrix adaptation evolution strategy, in: Genetic and Evolutionary Computation Conference, GECCO 2012, Philadelphia, PA, USA, July 7-11, 2012, pp. 321–328, 2012.

Lutz, M., Caldeira, K., Dunbar, R., and Behrenfeld, M. J.: Seasonal rhythms of net primary production and particulate organic carbon flux to depth describe biological pump efficiency in the global ocean, J. Geophys. Res., 113, C10011, doi:10.1029/2007JC003706, 2007.

Marsay, C. M., Sanders, R., Henson, S., Pabortsava, K., Achterberg, E., and Lampitt, R.: Attenuation of sinking particulate organic carbon flux through the mesopelagic ocean, Proc. Nat. Acad. Sci., 112, 1089–1094, 2015.

Marshall, J., Adcroft, A., Hill, C., Perelman, L., and Heisey, C.: A finite-volume, incompressible Navier-Stokes model for studies of the ocean on parallel computers, J. Geophys. Res., 102, 5733–5752, 1997.

Martin, J. H., Knauer, G. A., Karl, D. M., and Broenkow, W. W.: VERTEX: carbon cycling in the Northeast Pacific, Deep-Sea Res., 34, 267–285, 1987.

Najjar, R. G., Jin, X., Louanchi, F., Aumont, O., Caldeira, K., Doney, S. C., Dutay, J.-C., Follows, M., Gruber, N., Joos, F., Lindsay, K., Maier-Reimer, E., Matear, R., Matsumoto, K., Monfray, P., Mouchet, A., Orr, J. C., Plattner, G.-K., Sarmiento, J. L., Schlitzer, R., Slater, R. D., Weirig, M.-F., Yamanaka, Y., and Yool, A.: Impact of circulation on export production, dissolved organic matter and dissolved oxygen in the ocean: Results from Phase II of the Ocean Carbon-cycle Model Intercomparison Project (OCMIP-2), Global Biogeochem. Cy., 21, doi:10.1029/2006GB002857, 2007.

Orr, J., Maier-Reimer, E., Mikolajewicz, U., Monfray, P., Sarmiento, J., Toggweiler, J., Taylor, N., Palmer, J., Gruber, N., Sabine, C., Le Quere, C., Key, R., and Boutin, J.: Estimates of anthropogenic carbon uptake from four three-dimensional global ocean models, Global Biogeochem. Cy., 15, 43–60, doi:10.1029/2000GB001273, 2001.

Patvardhan, C., Bansal, S., and Srivastav, A.: Quantum-Inspired Evolutionary Algorithm for difficult knapsack problems, Memetic Computing, 7, 135–155, 2015.

Patvardhan, C., Bansal, S., and Srivastav, A.: Parallel improved quantum inspired evolutionary algorithm to solve large size Quadratic Knapsack Problems, Swarm and Evolutionary Computation, 26, 175–190, 2016.

Paulmier, A., Kriest, I., and Oschlies, A.: Stoichiometries of remineralisation and denitrification in global biogeochemical ocean model, Biogeosciences, 6, 923–935, www.biogeosciences-discuss.net/6/923/2009, 2009.

Priess, M., Koziel, S., and Slawig, T.: Marine ecosystem model calibration with real data using enhanced surrogate-based optimization, Journal of Computational Science, 4, 423–437, doi:10.1016/j.jocs.2013.04.001, 2013.

Rückelt, J., Sauerland, V., Slawig, T., Srivastav, A., Ward, B., and Patvardhan, C.: Parameter Optimization and Uncertainty Analysis in a Model of Oceanic CO_2 Uptake Using a Hybrid Algorithm and Algorithmic Differentiation, Nonlinear Analysis: Real World Applications, 11, 3993–4009, 2010.

Schartau, M. and Oschlies, A.: Simultaneous data-based optimization of a 1D-ecosystem model at three locations in the North Atlantic: Part II – Standing stocks and nitrogen fluxes, J. Mar. Res., 61, 795–821, 2003a.

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, 2003b.

Schartau, M., Oschlies, A., and Willebrand, J.: Parameter estimates of a zero-dimensional ecosystem model applying the adjoint method, Deep-Sea Res. Pt. II, 48, 1796–1800, f1a, 2001.

Schartau, M., Wallhead, P., Hemmings, J., Löptien, U., Kriest, I., Krishna, S., Ward, B., Slawig, T., and Oschlies, A.: Reviews and syntheses: Parameter identification in marine planktonic ecosystem modelling, Biogeosciences Disc., doi:10.5194/bg-2016-242, 2016.

Schmoker, C., Hernandez-Leon, S., and Calbet, A.: Microzooplankton grazing in the oceans: impacts, data variability, knowledge gaps and future directions, Jour. Plank. Res., 35, 691–706, doi:10.1093/plankt/fbt023, 2013.

Seferian, R., Gehlen, M., Bopp, L., Resplandy, L., Orr, J., Marti, O., Dunne, J. P., Christian, J., Doney, S., Ilyina, T., Lindsay, K., Halloran, P., Heinze, C., Segschneider, J., Tjiputra, J., Aumont, O., and Romanou, A.: Inconsistent strategies to spin up models in CMIP5: implications for ocean biogeochemical model performance assessment, Geosci. Model Dev., 9, 1827—1851, doi:10.5194/gmd-9-1827-2016, www.geosci-model-dev.net/9/1827/2016/, f0d, 2016.

Takahashi, T., Broecker, W., and Langer, S.: Redfield ratio based on chemical data from isopycnal surfaces, J. Geophys. Res., 90, 6907–6924, 1985.

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

Van Mooy, B., Keil, R., and Devol, A.: Impact of suboxia on sinking particulate organic carbon: Enhanced carbon flux and preferential degradation of amino acids via denitrificiation, Geochim. Cosmochim. Ac., 66, 457–465, f5, 2002.

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

Wallmann, K.: Phosphorus imbalance in the global ocean?, Global Biogeochem. Cy., 24, doi:10.1029/2009GB003643, 2010.

Ward, B.: Marine Ecosystem Model Analysis Using Data Assimilation, Ph.D. thesis, Univ. Southampton, School of Ocean and Earth Science, 2009.

Ward, B., Friedrichs, M. A. M., Anderson, T., and Oschlies, A.: Parameter optimisation techniques and the problem of underdetermination in marine biogeochemical models, Jour. Mar. Systems, 81, 34–43, doi:10.1016/j.jmarsys.2009.12.005, 2010.

Table 1. Operational constants of the CMA-ES algorithm (cf. Initialization in Algorithm 1).

Selection and recombination	Step size control	Covariance matrix adaption
$\lambda = 4 + \lfloor 3 \log n \rfloor$	$\chi = \sqrt{n}\left(1 - \frac{1}{4n} + \frac{1}{21n^2}\right)$	$c_c = \frac{4+\mu_{\text{eff}}/n}{n+4+2\mu_{\text{eff}}/n}$
$\mu = \lfloor \frac{\lambda}{2} \rfloor$	$c_\sigma = \frac{\mu_{\text{eff}}+2}{n+\mu_{\text{eff}}+5}$	$c_\mu = \min\left(1 - c1, 2\frac{\mu_{\text{eff}}+1/\mu_{\text{eff}}-2}{(n+2)^2+\mu_{\text{eff}}}\right)$
$w_i = \frac{\log(\mu+0.5)-\log(i)}{\sum_{j=1}^{\mu}\log(\mu+0.5)-\log(j)}$		$c_1 = \frac{2}{(n+1.3)^2+\mu_{\text{eff}}}$
$\mu_{\text{eff}} = \frac{\left(\sum_{i=1}^{\mu} w_i\right)^2}{\sum_{i=1}^{\mu} w_i^2} = \frac{1}{\sum_{i=1}^{\mu} w_i^2}$		

Table 2. Experimental setup of optimization. "low" and "high" indicate boundary constraints of the optimizations, respectively.

Name	$R_{-\text{O2:P}}$		I_{c}		K_{PHY}		μ_{ZOO}		κ_{ZOO}		b[§]	
	low	high	low	high	low	high	low	high	low	high	low	high
TWIN	150	200	4.0	48	0.0001	0.5	0.1	4.0	0.0	10.0	0.4	1.8
OBS-WIDE	150	200	4.0	48	0.0001	0.5	0.1	4.0	0.0	10.0	0.4	1.8
OBS-WIDE-20	150	200	4.0	48	0.0001	0.5	0.1	4.0	0.0	10.0	0.4	1.8
OBS-NARR	150	200	4.0	48	0.0001	0.5	1.0	3.0	1.6	4.8	0.4	1.8
OBS-NARR-R	150	200	4.0	48	0.0001	0.5	1.0	3.0	1.6	4.8	0.4	1.8

[§] Note that from b (the optimized parameter) in the model we calculate the rate of vertical increase in sinking speed a, always assuming nominal detrital remineralization of $r = 0.05\ \text{d}^{-1}$. The resulting values for a are: 0.058275 (Target (Twin)), 0.125 (high) and 0.027778 (low).

Table 3. Optimization results (evaluations, i.e. number of individuals, λ, times number of generations, N), best model misfit M_{opt}, optimum parameters and their uncertainties. For each model and parameter, the first line gives the optimum parameter, followed by $p_{\min}$ and maximum $p_{\max}$ of all individuals, for which the misfit M_i is $(M_i - M_{opt})/M_{opt} \leq 0.001$. The third line additionally present in parentheses the percent of individuals for which this criterion holds, as well as the range of optimum parameters as percent of the average parameter of the last generation. We also give misfit and parameters of the reference run, against which the twin experiment was optimized.

Experiment	$\lambda \times N$	M_{opt}	$R_{-\mathrm{O2:P}}$	I_{c}	K_{PHY}	μ_{ZOO}	κ_{ZOO}	b
Reference	1	0.529	170.0	24.0	0.03125	2.0	3.2	0.858
TWIN	2000	0.0003	170.0	24.0	0.034	2.0	3.20	0.858
			170	24	0.033-0.035	2.0	3.19-3.20	0.858
	(< 1)		(< 1)	(< 1)	(5)	(< 1)	(< 1)	(< 1)
OBS-WIDE	950	0.477	179.5	48.0	0.12	0.28	6.15	1.10
			176-182	46-49	0.09-0.13	0.24-0.32	4.79-3.37	1.08-1.12
	(31)		(3)	(6)	(32)	(28)	(26)	(4)
OBS-WIDE-20	3460	0.450	167.7	9.9	0.5	2.05	5.83	1.34
			165-171	9.6-10.8	0.39-0.57	2.00-2.52	5.37-10.0	1.31-1.37
	(64)		(3)	(12)	(34)	(25)	(79)	(5)
OBS-NARR	1820	0.450	167.0	9.7	0.5	1.89	4.57	1.34
			165-170	9.0-10.3	0.39-0.53	1.57-2.02	2.95-4.66	1.30-1.36
	(39)		(3)	(14)	(28)	(23)	(37)	(4)
OBS-NARR-R	1400	0.450	166.7	9.6	0.5	1.76	3.82	1.34
			165-169	8.7-10.1	0.44-0.54	1.57-1.79	2.77-3.90	1.31-1.36
	(50)		(2)	(14)	(19)	(13)	(30)	(3)

Table 4. Global annual fluxes of primary production (PP), grazing (GRAZ), aerobic and anaerobic remineralization of detritus and DOM to nutrients (REM), excretion by zooplankton (EXCR) export production (F120, flux through 120 m), flux through 2030 m (F2030), and benthic burial (BUR), in Pg N y^{-1}, for the reference experiment, OBS-WIDE, OBS-WIDE-20 and OBS-NARR (two repeated experiment with different configurations of CMA-ES). We also show some globally derived, observed estimates. Conversion between different elements was carried out via N:P=16, and C:P=122.

Experiment	PP	GRAZ	REM	EXCR	F120	F2030	BUR
Reference	5.44	3.52	4.72	0.80	0.92	0.11	0.05
OBS-WIDE	6.20	1.24	5.94	0.25	0.81	0.06	0.02
OBS-WIDE-20	7.45	4.68	6.66	1.00	1.10	0.06	0.02
OBS-NARR	7.52	4.74	6.65	1.10	1.10	0.06	0.02
OBS-NARR-R	7.58	4.77	6.65	1.19	1.10	0.06	0.02
Observed[§]	7.68-8.09	4.79, 5.71	-	-	0.29-1.53	0.03-0.07	0.02

[§] Observed fluxes are from Carr et al. (2006, primary production), Honjo et al. (2008, particle flux), Lutz et al. (2007, particle flux), Dunne et al. (2007, particle flux), Schmoker et al. (2013, primary production, zooplankton grazing excluding/including mesozooplankton grazing) and Wallmann (2010, burial; without shelf and slope region).

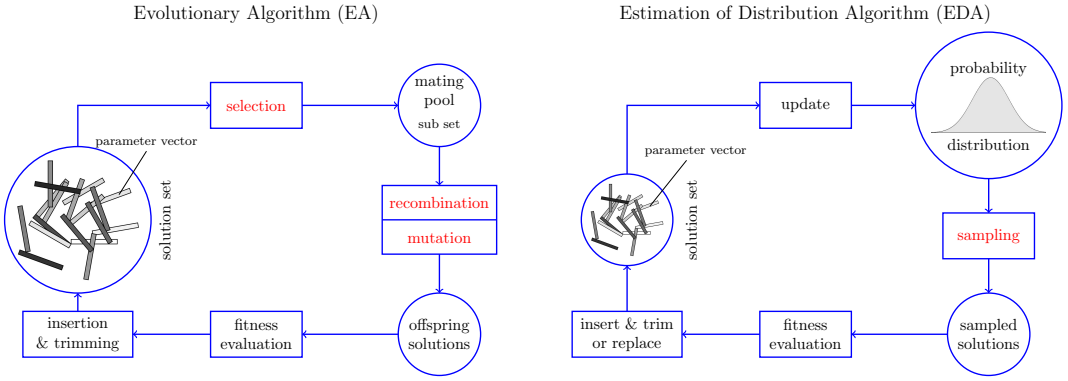

Figure 1. A general EA (left) and EDA (right) schematic. Cycles represent sets of solutions (vectors of BGC parameters in our case) or an explicit probability distribution from which new solutions can be drawn. Rectangle symbols depict operations. Operations displayed in red font depend on random decisions. EA: A set of candidate solutions (population) is iteratively updated. In each generation, candidate solutions compete to form a mating pool which is realized by a random selection operator. Offspring solutions are produced by recombining mates and/or introducing some mutation. Finally, there is a fitness based insertion back into the population, which is usually trimmed to a predefined population size. The random operators selection, recombination and mutation imply an implicit probability distribution on the search space with respect to which solutions are likely to appear in the next generation. EDA: Candidate solutions of the current iteration's population are used to update an explicit probability distribution such that the likelihood to sample good solutions increases. New candidate solutions are directly sampled from the probability distribution. Usually, the realization of the probability distribution update ensures that information of former solutions fades out slowly, resisting for several iterations. Therefore, the population may be smaller as an EA population and even be replaced with the entire set of new samples, which is the case for the CMA-ES algorithm we use.

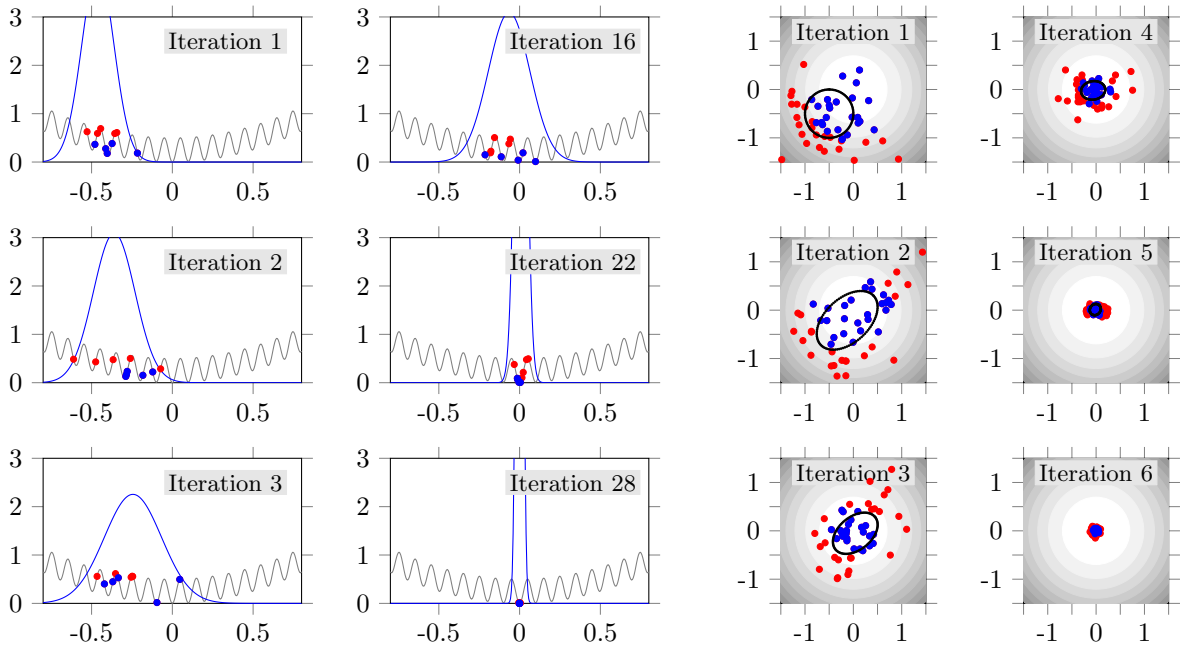

Figure 2. Iterations of the CMA-ES applied to test functions. Left: A uni-variate Griewank-type function f (grey curve). In each iteration we draw $\lambda = 10$ samples from the normal distribution (blue curve). For each sample x_i, the pair $(x_i, f(x_i))$ is marked with a dot. The $\mu = \frac{\lambda}{2} = 5$ better samples (blue dots) are involved into the normal distribution update for the next iteration. Right: Two-dimensional sphere function. Here, samples are marked with dots while function values are indicated by the grey levels in the counter plots; the i-th grey level represents the range $\left[\frac{i-1}{2}, \frac{i}{2}\right)$. More samples (50) than necessary are used to update the distribution, which is indicated by its standard deviation ellipse (black), here. Distributions tend to elongate into directions of descent (iteration 2). For the convex example function the algorithm converges after few iterations.

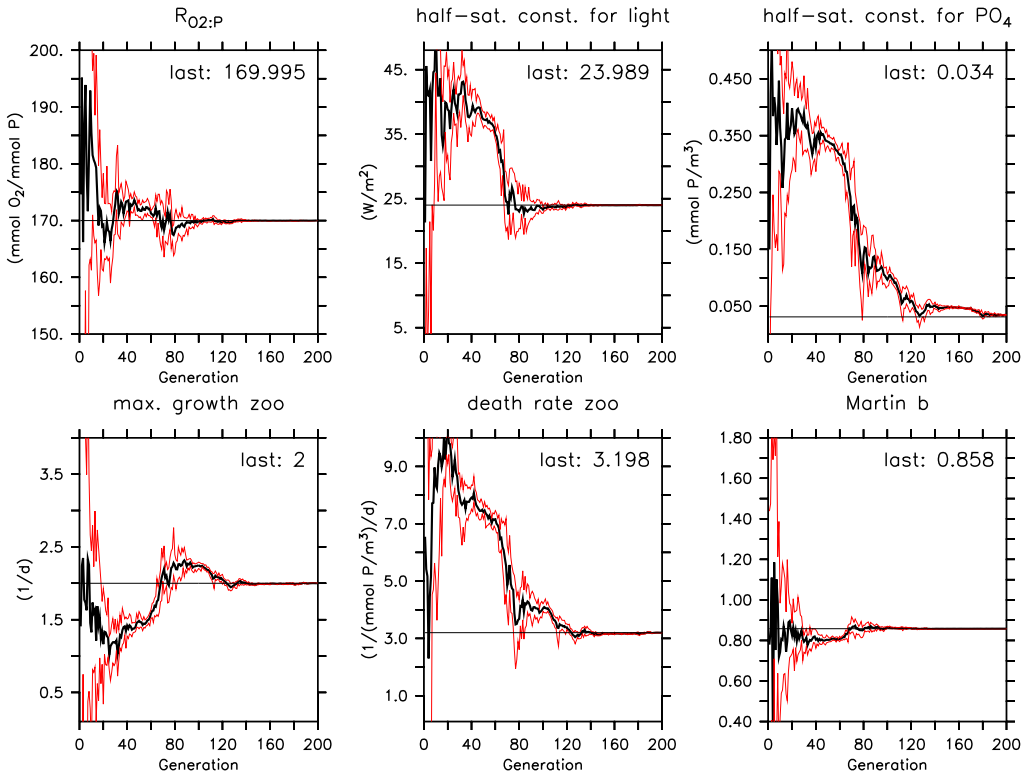

Figure 3. Optimization trajectory for six parameters of the twin experiment. Thick black line shows average parameter of all ten individuals of a generation. Red lines indicate their maximum and minimum parameter value. Horizontal black lines indicate the target parameter. Note that we restrict the y-axis to maximum and minimum boundary.

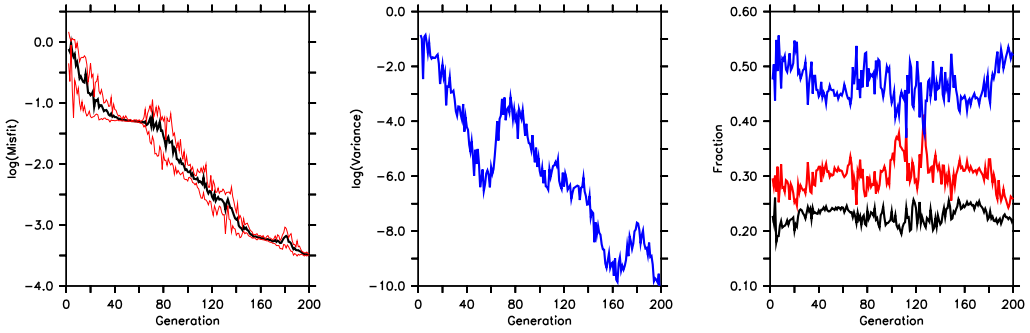

Figure 4. Model misfit, its variance, calculated from individuals of each population (both transformed logarithmically by $\log_{10}$) and components of the twin experiment. Left panel: Thick black line shows average misfit of all ten individuals of a generation. Red lines indicate maximum and minimum misfit. Mid panel: Variance of misfit. Right panel: contribution of each component of the misfit Function. Blue: oxygen. Red: nitrate. Black: phosphate.

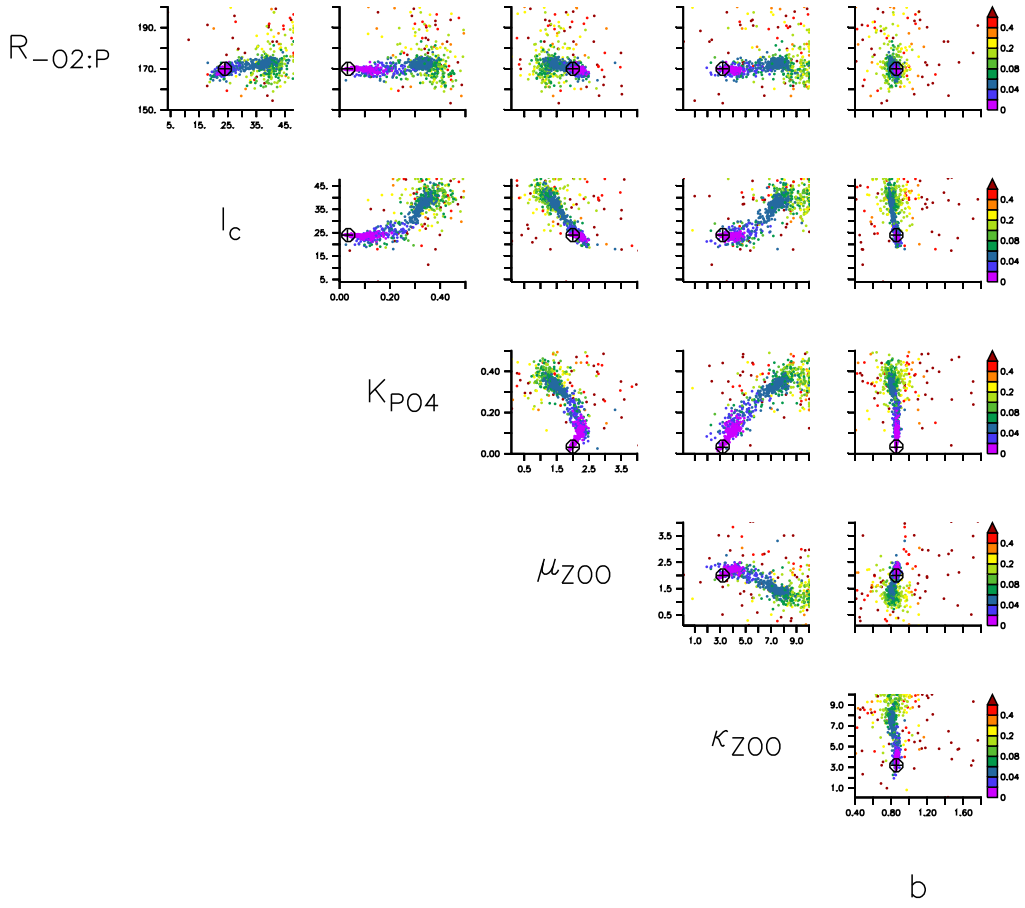

Figure 5. Model misfit, plotted for each pair of parameter combinations of the twin experiment. Color indicates misfit (see color bars on the right). A cross indicates the target value, i.e. the value of the reference experiment. A circle indicates the parameter of one individual of the last generation. Note that for better visibility we restrict the parameter range to its boundaries (see Table 2).

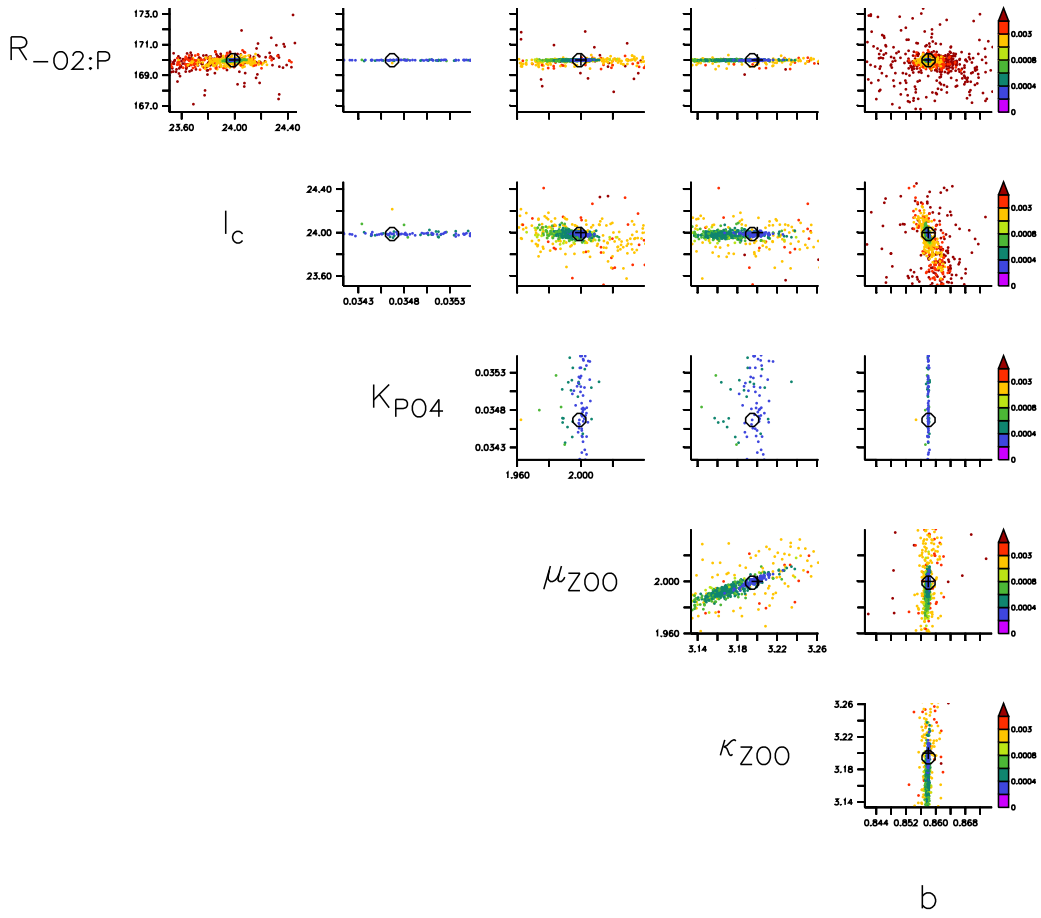

Figure 6. As Fig. 5, but only plotted for a region $\pm 2\%$ around the average parameter value of the last generation, regardless of generation and associated misfit. Note that these parameters can have occurred early in the optimization, and even be associated with a large misfit (that would arise from at least one of the other parameters causing a large misfit). Note that the color scale is different than in Fig. 5.

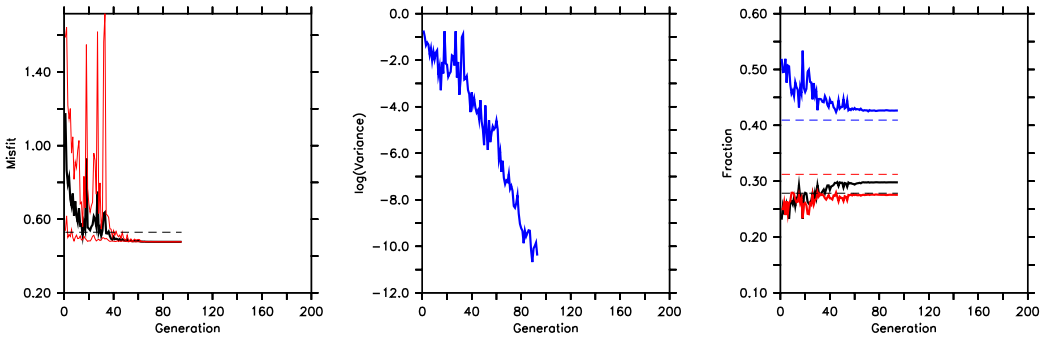

Figure 7. As Fig. 4, but for optimization OBS-WIDE. Note that in the left plot, we now show the raw value of the misfit function (not log transformed). The optimization finished at generation 95.

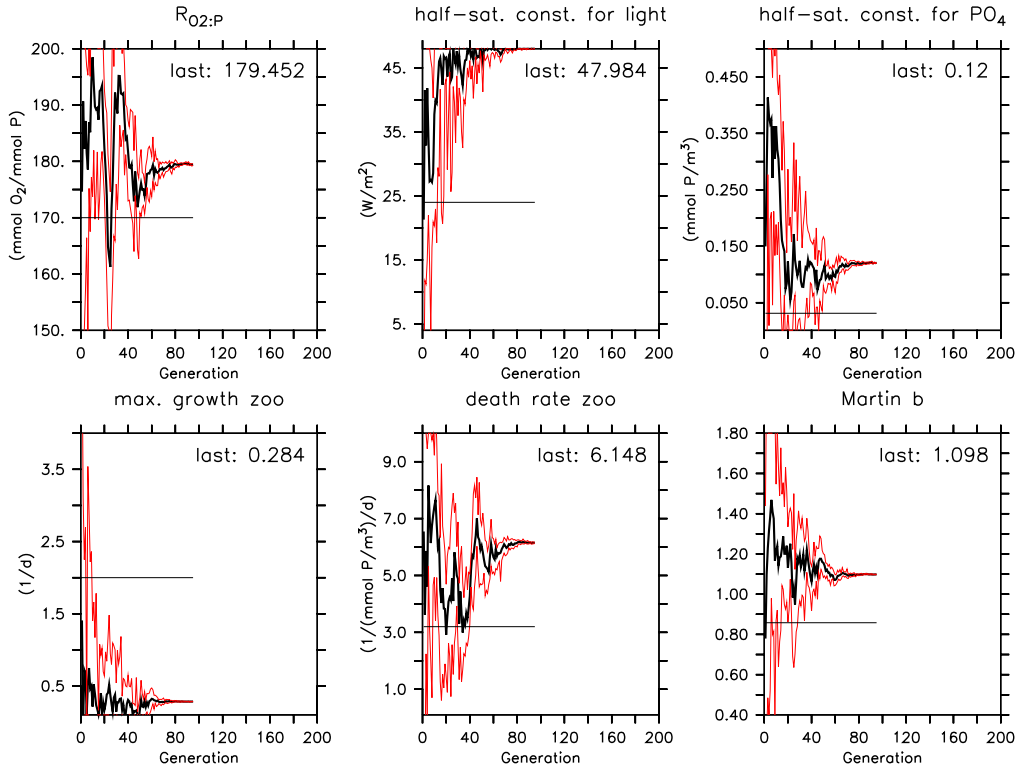

Figure 8. As Fig. 3, but for optimization OBS-WIDE. The optimization finished at generation 95.

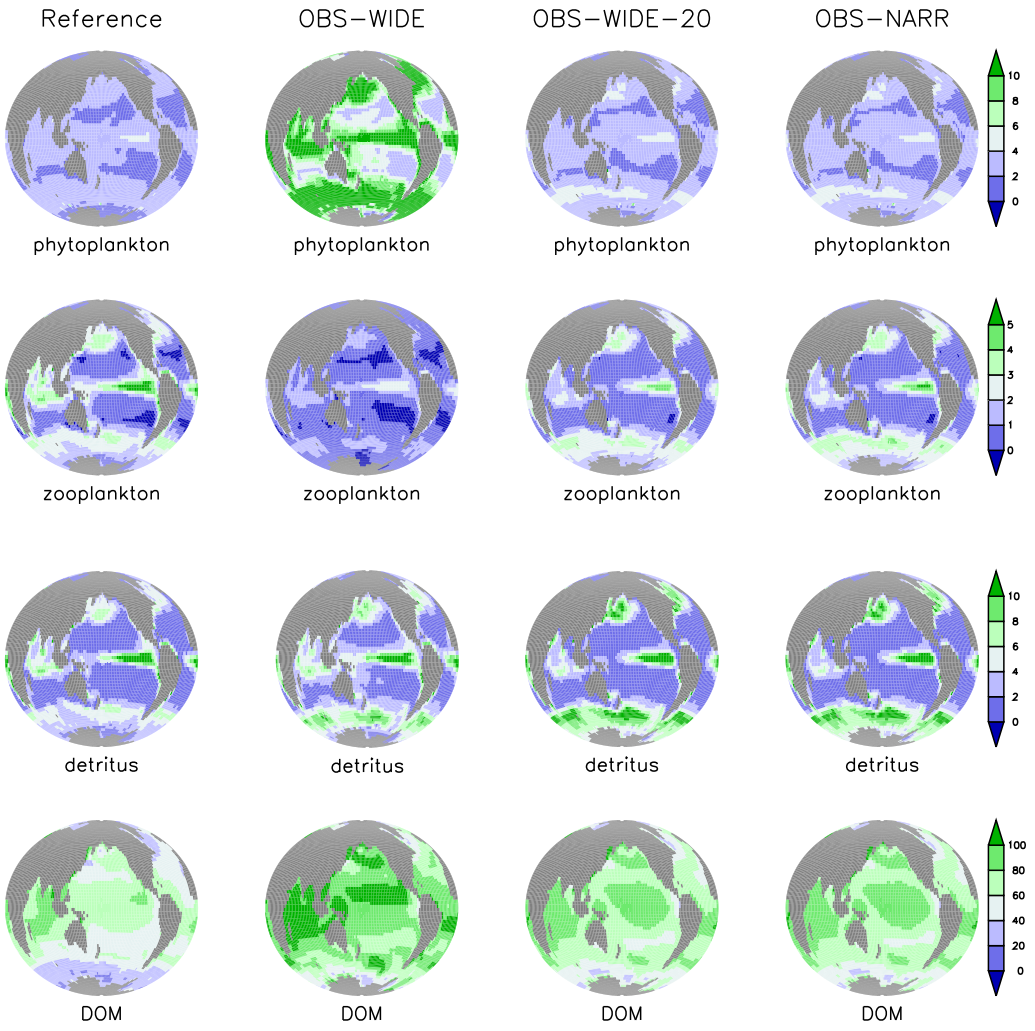

Figure 9. Surface (first) layer concentrations (in mmol C m^{-3}, converted via a C:P ratio of 122) for phytoplankton, zooplankton, detritus and DOM for the reference run, optimizations OBS-WIDE, OBS-WIDE-20 and OBS-NARR.

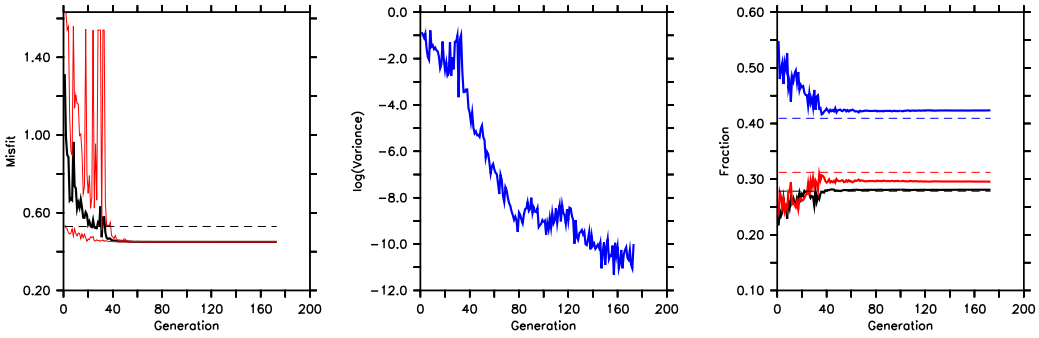

Figure 10. As Fig. 7, but for optimization OBS-WIDE-20. The optimization finished at generation 173.

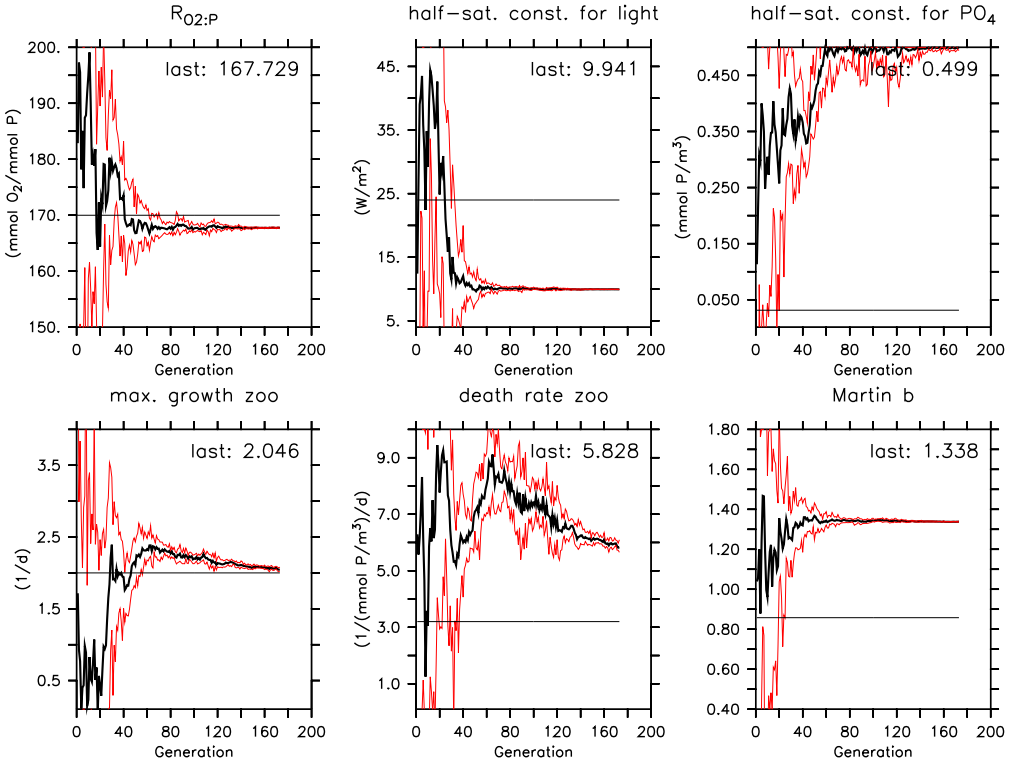

Figure 11. As Fig. 8, but for optimization OBS-WIDE-20. The optimization finished at generation 173.

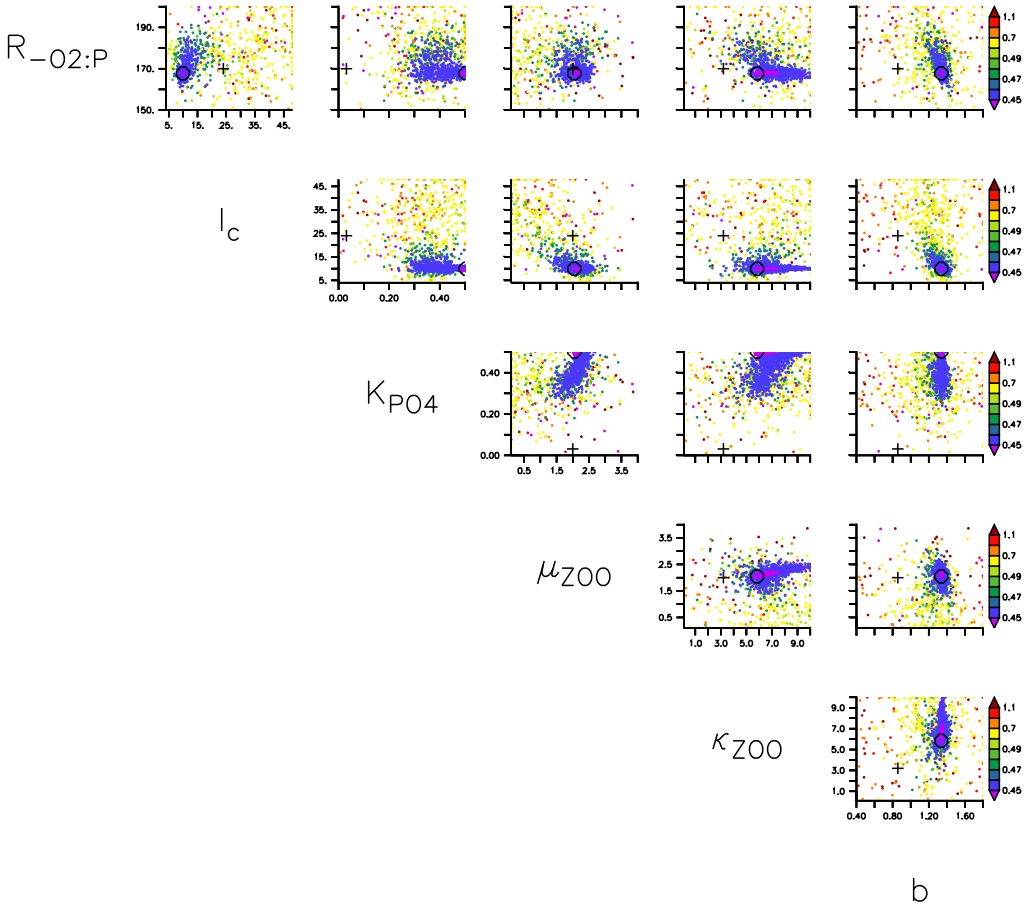

Figure 12. As Fig. 5, but for optimization OBS-WIDE-20. Note that the color scale differs from that of Fig. 5.

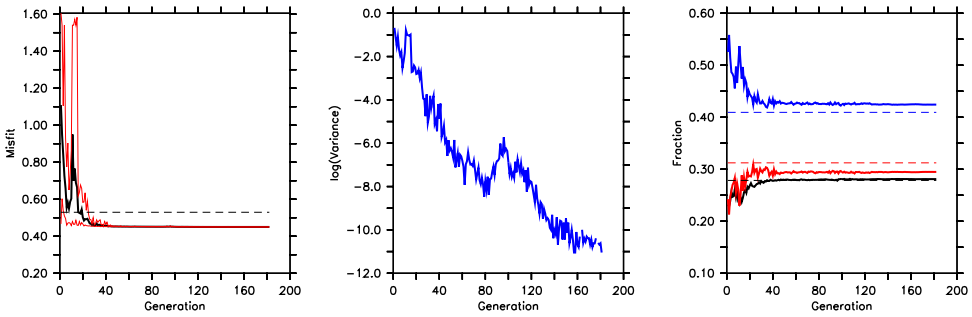

Figure 13. As Fig. 10, but for optimization OBS-NARR. The optimization finished at generation 182.

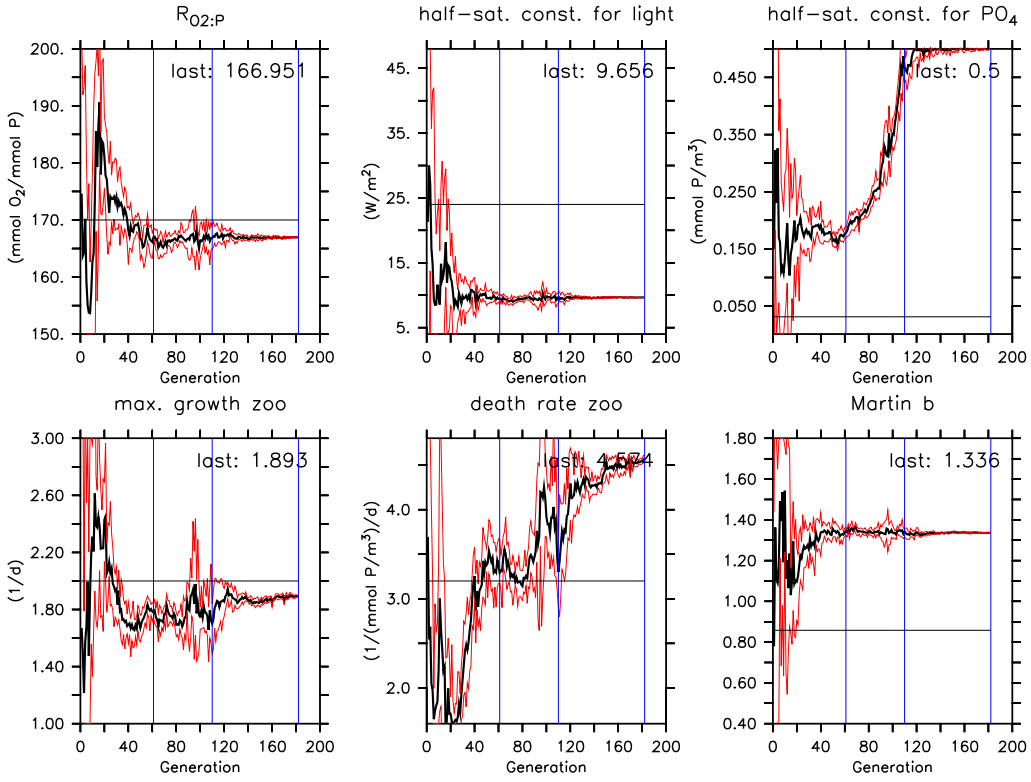

Figure 14. As Fig. 11, but for optimization OBS-NARR. The optimization finished at generation 182. Vertical blue lines indicate generation, for which we also present deviations from observation of vertically integrated nutrients and oxygen from in Fig. 16.

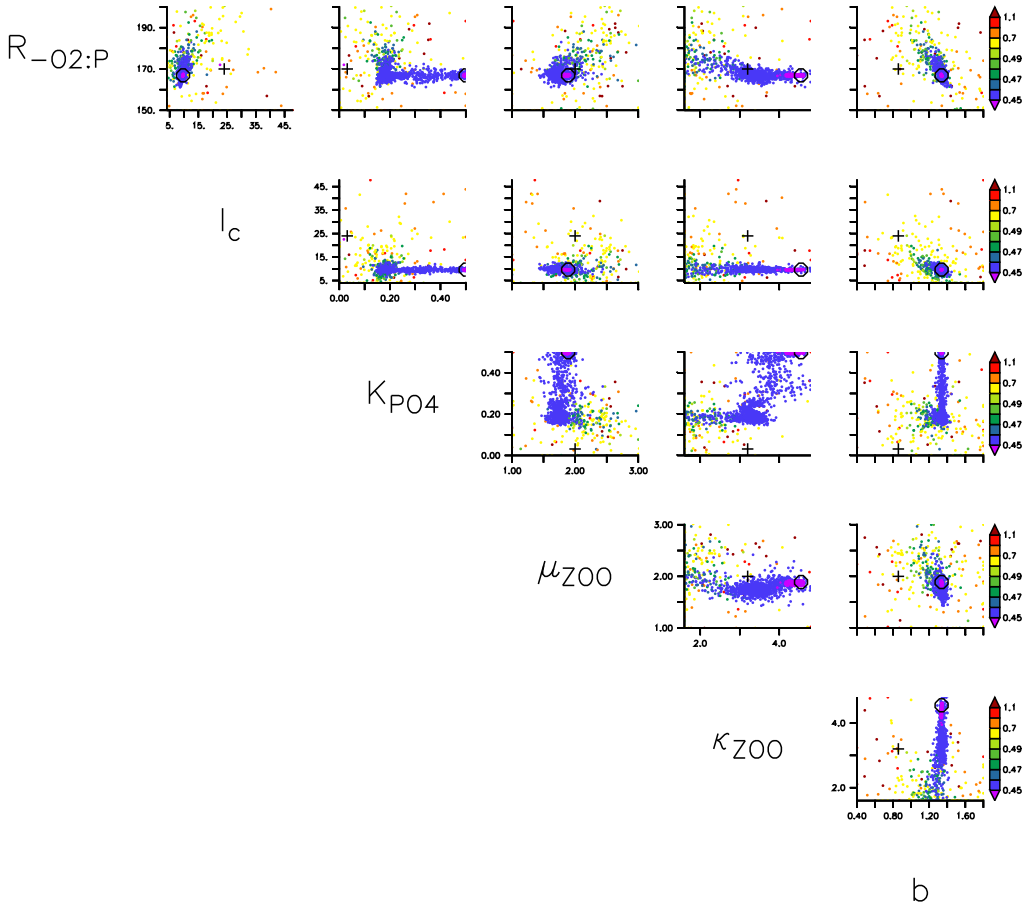

Figure 15. As Fig. 12, but for OBS-NARR.

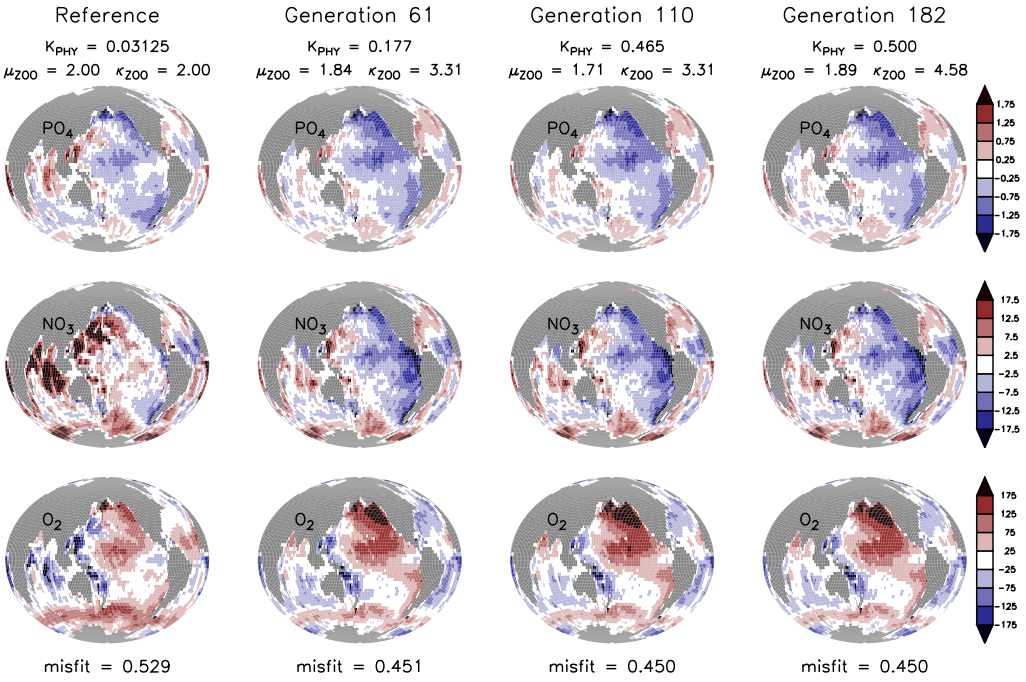

Figure 16. Model deviations from observations of vertically integrated phosphate (top), nitrate (middle) and oxygen (bottom) for the reference run, and three generations (61, 110, 182) of OBS-NARR. See blue lines in Fig. 14 for parameter values in this generation. For each generation, we chose the best (with respect to misfit) individual for plotting. Misfit is 0.451, 0.450 and 0.450 for generation 61, 110 and 182, respectively.