JAMES | Journal of Advances in Modeling Earth Systems

AGU100 ADVANCING EARTH AND SPACE SCIENCE

RESEARCH ARTICLE
10.1029/2018MS001510

Key Points:
- Calibrating parameterized models is important for model intercomparison
- Multiobjective optimization helps revealing dependencies between parameter values and model skills

Supporting Information:
- Data Set S1
- Data Set S2

Correspondence to:
V. Sauerland,
sauerland@math.uni-kiel.de

Citation:
Sauerland, V., Kriest, I., Oschlies, A., & Srivastav, A. (2019). Multiobjective calibration of a global biogeochemical ocean model against nutrients, oxygen, and oxygen minimum zones. *Journal of Advances in Modeling Earth Systems*, *11*, 1285–1308. https://doi.org/10.1029/2018MS001510

Received 23 SEP 2018
Accepted 24 MAR 2019
Accepted article online 1 APR 2019
Published online 8 MAY 2019

Multiobjective Calibration of a Global Biogeochemical Ocean Model Against Nutrients, Oxygen, and Oxygen Minimum Zones

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

[1]Department of Mathematics, Kiel University, Kiel, Germany, [2]GEOMAR Helmholtz-Zentrum für Ozeanforschung Kiel, Kiel, Germany

Abstract Global biogeochemical ocean models rely on many parameters, which govern the interaction between individual components, and their response to the physical environment. They are often assessed/calibrated against quasi-synoptic data sets of dissolved inorganic tracers. However, a good fit to one observation might not necessarily imply a good match to another. We investigate whether two different metrics—the root-mean-square error to nutrients and oxygen and a metric measuring the overlap between simulated and observed oxygen minimum zones (OMZs)—help to constrain a global biogeochemical model in different aspects of performance. Three global model optimizations are carried out. Two single-objective optimizations target the root-mean-square metric and a sum of both metrics, respectively. We then present and explore multiobjective optimization, which results in a set of compromise solutions. Our results suggest that optimal parameters for denitrification and nitrogen fixation differ when applying different metrics. Optimization against observed OMZs leads to parameters that enhance fixed nitrogen cycling; this causes too low nitrate concentrations and a too high global pelagic denitrification rate. Optimization against nutrient and oxygen concentrations leads to different parameters and a lower global fixed nitrogen turnover; this results in a worse fit to OMZs. Multiobjective optimization resolves this antagonistic effect and provides an ensemble of parameter sets, which help to address different research questions. We finally discuss how systematic model calibration can help to improve models used for projecting climate change and its effect on fisheries and climate gas emissions.

1. Introduction

Global biogeochemical ocean models are not only applied to estimate the ocean biota's impact on climate-relevant gases such as atmospheric carbon dioxide and nitrous oxide (Ciais et al., 2013) but also to assess its impact on other, societal relevant processes. For example, they are used to investigate the influence of a changing climate on the potential expansion of oxygen minimum zones (OMZs, e.g., Cocco et al., 2013; Stramma et al., 2012), which might be of consequence for marine fisheries (Stramma et al., 2012), on ocean acidification, which affects the spatial distribution of coral reefs and other calcaerous organisms (e.g., Bopp et al., 2013), and on biogeochemical interactions and regime shifts (e.g., Laufkötter et al., 2013, 2016; Segschneider & Bendtsen, 2013), which, in turn, might affect higher trophic levels or the storage of carbon in the deep ocean. Thus, any biogeochemical model has to represent a plethora of different processes and organisms in an appropriate way. This includes not only a “correct” present mean state but also the right sensitivities to transient forcing.

However, a correct representation of one variable (e.g., global nutrient distribution) does not necessarily imply that a model correctly represents all inherent processes and tracers. For example, a model that correctly represents oxygen in equatorial upwelling regions might do so at the cost of nitrate concentrations (Moore & Doney, 2007) or benthic exchange processes (Kriest & Oschlies, 2013). So far, optimization of global biogeochemical ocean models against observed nutrient and oxygen concentrations, which were combined into one single metric, has shown that this single-objective optimization can help to improve simulated global fluxes such as primary and export production, which are of relevance for the simulation of air-sea gas exchange (Kriest, 2017; Kriest et al., 2017). However, to date it has not been shown that this also holds for other diagnostics, such as ocean oxygen content or the volume and location of OMZs, which might be anticorrelated to nutrient concentrations (as indicated by the experiments carried out by; Moore

& Doney, 2007). Further, combining different tracers and diagnostics into one single metric usually requires some form of normalization and, possibly, weighting, all of which can strongly influence the outcome of optimization (Evans, 2003).

Given these complications, and the ongoing application of global models to investigate various processes of societal relevance, it might be worthwhile to develop a tool that provides model solutions calibrated against different objectives (model-data misfit measures). From these ensembles of model solutions we can then pick a posteriori those solutions (or parameterizations), which are suitable to be applied to a specific research question or scenario.

One such tool is multiobjective optimization (MO). Rather than seeking a single compromise parameter set by optimizing a weighted sum of misfit measures, MO yields a collection of compromise solutions with respect to the different misfit measures of interest. Ideally, the collection of compromise solutions contains (nearly) optimal solutions for each single objective as “extremes” and further well-distributed compromises in between. The latter solutions are of special interest if we face diverging model calibration results with respect to different types of real-world observations. For example, dealing with the most common case of two objectives, it is often observed that compromise solutions obtained by MO follow the so-called *Pareto principle*, meaning that we only need to sacrifice 20% of the best value with respect to one objective in order to reach 80% of the other objectives' optimum.

MO approaches gained most attention in economy and operations research but have also been applied for geoscientific problems, mainly and long-established in the field of hydrology (see, e.g., Efstratiadis & Koutsoyiannis, 2010; Mostafaie et al., 2018; Yapo et al., 1998). Recent contributions also deal with emulators/surrogates of earth system models in order to apply MO in the face of different prediction tasks (see, e.g., Langenbrunner & Neelin, 2017a, 2017b; Price et al., 2009).

Here, we parallelize and apply a state-of-the art estimation of distribution algorithm for MO (Igel et al., 2007) in order to calculate the required set of good compromise parameter sets at once. While exemplarily demonstrating the approach with our OMZ test case, MO can be applied to any situation where single-objective model calibrations tend to be in conflict, either with respect to different observation types or with respect to observations from different sites. MO will help to obtain more universally reliable model parameters, especially when cross validation of single-objective parameter optimization results fails due to multiple sources of uncertainty. More background about the issues that we aim to facilitate with MO, for example, simultaneous data assimilation with respect to different ocean sites or cross validation of model calibration results, can be found, for example, in sections 6.1 and 7.2 of the overview article by Schartau et al. (2017).

We first investigate if calibration against concentrations of nutrient and oxygen of a coupled global biogeochemical model (i.e., the model's root-mean-square error, RMSE) improves the model's fit to OMZs, that is, whether the two objectives—fit to global nutrients and oxygen and fit to OMZs—are correlated. This initial analysis examines a posteriori the optimization trajectory, of a calibration presented in Kriest (2017). We then present and investigate an optimization where the fit to OMZs is added to the fit to nutrient and oxygen concentrations. In case of uncorrelated objectives, this will result in a compromise solution of the optimization. We contrast these two solutions with those of a third optimization that applies MO and thus targets both objectives (RMSE and OMZ) simultaneously, yet independently.

The paper is structured as follows: After a brief presentation of the biogeochemical model, its coupling to the off-line circulation framework and to the optimization framework, we introduce the MO applied here and describe the metric for OMZs (our second objective), which is based on the metric developed by Cabre et al. (2015). We then present and discuss the results of the three different optimizations against different misfit functions. In our analysis we focus on the correlation between both objectives and on their ability to constrain the model parameters. We finally discuss the impact of optimization and parameter values on global biogeochemical model fluxes (as an independent estimate of model fit), as well as its relation to different criteria for OMZs, which might be relevant for higher trophic levels such as fish.

2. Models, Experiments, and Optimizations

2.1. Circulation and Physical Transport

All model simulations apply the transport matrix (TM) method (github.com/samarkhatiwala/tmm; Khatiwala, 2007) for tracer transport, with monthly mean TMs, wind, temperature, and salinity (for air-sea gas exchange) derived from a 2.8° global configuration of the Massachusetts Institute of Technology General Circulation Model (MITgcm), with 15 levels in the vertical, as described in Marshall et al. (1997) and Dutkiewicz et al. (2005). The circulation model was forced with climatological annual cycles of wind, heat, and freshwater fluxes and subject to a weak restoring of surface temperature and salinity to observations. Its configuration is similar to that applied in the Ocean Carbon-Cycle Model Intercomparison Project (Orr et al., 2000), which has been assessed against observations of temperature, salinity, and mixed layer depth (Doney et al., 2004), CFCs (Dutay et al., 2002; Matsumoto et al., 2004), and radiocarbon (Graven et al., 2012; Matsumoto et al., 2004). Overall, its performance is comparable to other global models.

Using this efficient off-line approach for describing the transport of the tracers of the biogeochemical model (see below), a time step length of (1/2) day for tracer transport and (1/16) day for biogeochemical interactions, simulation of 3,000 years requires about 0.5–1.5 hr on four nodes (24 core Intel Xeon Ivybridge) at a High Performance Computing Centre (www.hlrn.de). After 3,000 years most tracers have approached steady state (see also; Kriest & Oschlies, 2015, for long time trends of MOPS simulated in a different circulation), and the transient of the misfit function becomes very small. The last year is used for model analysis and evaluation of the misfit function.

2.2. Biogeochemical Model Structure

The biogeochemical model applied in this study is of intermediate complexity and represents the pelagic cycle of phosphorus, nitrogen, and oxygen with seven components, which are coupled via constant stoichiometric ratios (Model of Oceanic Pelagic Stoichiometry, MOPS; Kriest & Oschlies, 2015). In MOPS phosphate, phytoplankton, zooplankton, dissolved organic phosphorus, and detritus are calculated in units of millimoles of phosphorus per cubic meter. All organic components are characterized by a constant N:P stoichiometry of $d = 16$. Oxygen is coupled to the P cycle with a constant stoichiometry given by $R_{\text{-O2:P}}$. Aerobic remineralization of organic matter follows a saturation curve, with half-saturation constant K_{O2}. It ceases when oxygen declines, depending on K_{O2}; at the same time, denitrification takes over, as long as nitrate is available above a defined threshold, DIN_{min}. Like the oxic process, suboxic remineralization follows a saturation curve for oxidant nitrate, with half-saturation constant K_{DIN}. MOPS does not explicitly resolve the different oxidation states of inorganic nitrogen (nitrite, N_2O, and ammonium) but assumes immediate coupling of the different processes involved in nitrate reduction, the end-product being dinitrogen (see also; Kriest & Oschlies, 2015; Paulmier et al., 2009). On long time scales (Kriest & Oschlies, 2015), loss of fixed nitrogen is balanced by a simple parameterization of nitrogen fixation by cyanobacteria, which relaxes the surface layer nitrate-to-phosphate ratio to d with a time constant, μ^*_{NFix}.

Detritus sinks with a vertically increasing sinking speed: $w = az$. Assuming a constant degradation rate r, in equilibrium this would result in a particle flux curve given by $F(z) \propto z^{-b}$, with $b = r/a$. For better comparison with values of b derived from observations (e.g., Buesseler et al., 2007; Martin et al., 1987; Van Mooy et al., 2002) a is here expressed in terms of b (assuming constant, nominal $r = 0.05\ \text{day}^{-1}$). A fraction of detritus deposited at the seafloor (at the bottom of the deepest vertical box) is buried instantaneously in some hypothetical sediment. The fraction buried depends on the deposition rate onto the sediment. Nonburied detritus is resuspended into the last box of the water column, where it is treated as regular detritus. The phosphorus budget is closed on an annual time scale through resupply via river runoff. More details about the biogeochemical model and parameters and their effects on model behaviour can be found in Kriest and Oschlies (2013, 2015).

2.3. Parameter Calibrations

Kriest and Oschlies (2015) presented a “hand-tuned,” a priori setup of MOPS, which we refer to as MOPS$^{\text{r}}$ (see also Table 1). The first systematic calibration of this model was presented by Kriest et al. (2017), who optimized six parameters against the RMSE (equation (1)) to observed annual mean phosphate, nitrate, and oxygen. Four of the parameters were related to the dynamic phosphorus turnover by plankton in the surface layer. The remaining two parameters were related to particle flux (parameter b) and the stoichiometric oxygen demand of remineralization ($R_{-\text{O2:P}}$). The six optimal parameters led to a better agreement of simulated biogeochemical fluxes to observations of primary and export production, zooplankton grazing, particle flux at 2,000 m, and benthic burial (Kriest et al., 2017).

In a follow-up calibration (Kriest, 2017) kept the four optimal parameters related to phytoplankton and zooplankton dynamics of the calibration by Kriest et al. (2017) constant. Instead, four parameters related to remineralization and nitrogen fixation were optimized against the same data set and misfit, together

Table 1
Experimental Design and Optimization Results

Experiment:	MOPS^{r}	MOPS^{o}		$\text{MOPS}^{\text{o+}}$		MOPS^{mo}|RMSE		MOPS^{mo}|OMZ	
Reference:	(Kriest & Oschlies, 2015)	(Kriest, 2017)		This study		This study		This study	
Calibration:	—	Single-objective		Single-objective		Multiobjective			
Cost function:	—	J_{RMSE} (equation (1))		$J_{\text{RMSE}} + J_{\text{OMZ}}$ (equation (4))		J_{RMSE} (equation (1))		J_{OMZ} (equation (3))	
Parameters									
	Θ	Θ^*	$R_\Theta(\Omega)$	Θ^*	$R_\Theta(\Omega)$	Θ^*	$R_\Theta(\Omega)$	Θ^*	$R_\Theta(\Omega)$
b	0.86	1.39	[1.33–1.45]	1.39	[1.32–1.48]	1.40	[1.34–1.45]	1.38	[1.21–1.46]
$R_{-\text{O2:P}}$	170	172	[166–178]	177	[170–187]	176	[166–183]	185	[175–200]
μ_{NFix}	2.00	1.19	[1.08–1.84]	1.99	[1.68–2.64]	1.98	[1.47–3.00]	2.99	[1.88–3.00]
DIN_{min}	4.00	15.80	[12.18–16.00]	7.59	[6.95–16.00]	10.32	[3.93–16.00]	3.40	[1.78–14.16]
K_{DIN}	8.00	31.97	[17.18–32.00]	31.61	[15.80–32.00]	25.97	[12.15–32.00]	17.58	[5.57–28.51]
K_{O2}	2.00	1.00	[1.00–8.48]	1.00	[1.00–12.10]	1.00	[1.00–12.87]	1.00	[1.00–5.89]
Optimization performance									
$M(\Omega)$	—	705		1,051		746		969	
$\lambda \times M$	1	1,190		1,530		3,400			
J^*	0.529	**0.439**		**1.130**		—			
J^*_{RMSE}	0.529	**0.439**		0.444		**0.439**		0.474	
J^*_{OMZ}	0.791	0.704		0.686		0.701		**0.673**	

Note. Minimum misfit function J^*, optimal parameters Θ^*, and their uncertainties R_Θ. To determine parameter uncertainty, we selected a group Ω of individuals defined by a misfit $J_i\,:\,J_i/J^* - 1 \leq 0.01$. For each parameter the first column gives the optimal parameter Θ^*, and the second column presents the parameter range in Ω. For the single-objective optimizations we present the respective optimal misfit J^* and the value for equations (1) and (3) for this individual. For MOPS^{mo} we present the values for the two outmost individuals of the Pareto front, that is, those best with respect to either misfit function and the corresponding misfit of the other objective. All analyses are restricted to individuals whose parameters lie within the prescribed boundaries. Values of unoptimized MOPS^{r} are also shown for comparison. The smallest misfit values (for the respective metrics) that we have found in all our experiments are shown in bold. MOPS = Model of Oceanic Pelagic Stoichiometry; OMZ = oxygen minimum zone; RMSE = root-mean-square error.

with the stoichiometric ratio for oxygen demand of remineralization, $R_{-\text{O2:P}}$ and b. In the following, this latter, single-objective optimization against nutrients and oxygen carried out by Kriest (2017) is referred to as MOPS^{o} (see also Table 1). The selection of parameters optimized in MOPS^{o} was motivated by the large uncertainty regarding extent and expansion of OMZs in models (Cabre et al., 2015; Cocco et al., 2013) and by the fact that very little knowledge exists about parameters (or parameterizations) that might affect nitrate and oxygen in subsurface layers. Therefore, upper and lower boundaries of parameters to be optimized were set to a rather wide range (see Kriest, 2017).

Given the large number of parameters required even by simple models such as MOPS, it would be appealing to calibrate many more, if not all, parameters of the biogeochemical model. However, because of possible correlations between the parameters, it is very likely that not all of them can be determined; this depends on the cost function applied and the ability of available data to constrain the parameters (Matear & Holloway, 1995; Ward et al., 2010). Especially noninformative ("flat") cost functions, or those with a rough topography, might cause the optimization to become trapped in local minima or result in biologically meaningless parameters (e.g., Kriest et al., 2017; Schartau et al., 2001; Ward et al., 2010). To avoid these complications, and because our cost functions do not include direct information about planktonic interactions, in this study we therefore restrict optimizations again to the six parameters calibrated by Kriest (2017).

2.4. Objectives (Data Components of the Metric)

2.4.1. Objective 1: RMSE to Nutrients and Oxygen

As in Kriest et al. (2017) the standard misfit to observations J is defined as the RMSE between simulated and observed annual mean phosphate, nitrate, and oxygen concentrations of spatially interpolated data provided in Garcia et al. (2006b) and Garcia et al. (2006a), mapped onto the three-dimensional model geometry. Although regridding the observations onto the coarser model geometry removes some of the variability, this method is computationally more efficient in an optimization framework. Also, a sensitivity study with a similar coupled model showed that accounting for the variance inherent in the observational data and arising from regridding did not have a large influence on the misfit function (Kriest et al., 2010). In our model experiments we did not carry out any weighting for different spatial densities of observations.

Deviations between model and observations are weighted by the volume of each individual grid box, V_i, expressed as fraction of total (model) ocean volume, V_T. The resulting sum of weighted deviations is then normalized by the volume-weighted, global mean concentration of the respective observed tracer:

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

$j = 1, 2, 3$ indicates the tracer type and $i = 1, \dots, N$ are the model locations for $N = 52,749$ model grid boxes. $\overline{o_j}$ is the global average observed concentration of the respective tracer. $m_{i,j}$ and $o_{i,j}$ are model and observations, respectively. By weighting each individual misfit with volume, J_{RMSE} serves more as a long time scale geochemical estimator, in contrast to a misfit function that, for example, focuses on (rather fast) turnover in the surface layer, or resolves the seasonal cycle.

2.4.2. Objective 2: Fit to OMZs

The model's fit to observed OMZs, C, is calculated following Cabre et al. (2015): Let $V_i^{\mathrm{m}}(c)$ be the fractional volume of a model box i defined by an oxygen concentration of $O_2 < c$ and $V_i^{\mathrm{o}}(c)$ the observed volume. Then, with N being the total number of all model grid boxes, $V^{\mathrm{m}}(c) = \sum_{i=1}^{N} V_i^{\mathrm{m}}(c)$ and $V^{\mathrm{o}}(c) = \sum_{i=1}^{N} V_i^{\mathrm{o}}(c)$ are the global volume of suboxic water for model and observations, respectively.

Further, let $V_i^{\cup}(c)$ be the fractional volume of a box where either model *or* observation have an oxygen concentration of $O_2 < c$. Then $V^{\cup}(c) = \sum_{i=1}^{N} V_i^{\cup}(c)$ is the global volume of water where either model or observation is suboxic with respect to the given criterion c (i.e., the union of all suboxic water masses). Let further $V_i^{\cap}(c)$ be the fractional volume of a box where both model *and* observation have an oxygen concentration of $O_2 < c$. Then $V^{\cap}(c) = \sum_{i=1}^{N} V_i^{\cap}(c)$ is the global volume of water where both model and observation are suboxic (the intersection volume). The degree of overlap C is then given by the union divided by the intersection:

$$C = \frac{V^{\cap}(c)}{V^{\cup}(c)} = \frac{V^{\cap}(c)}{V^{\mathrm{m}}(c) + V^{\mathrm{o}}(c) - V^{\cap}(c)} \tag{2}$$

Obviously, this metric varies between 0 (model and observations do not overlap at all or $V^{\cap}(c) = 0$) and 1 (perfect match; $V^{\cap}(c) = V^{\cup}(c)$). The advantage of the metric is that even if the total simulated OMZ volume matches perfectly that of observations ($V^{\mathrm{m}}(c) = V^{\mathrm{o}}(c)$), but if simulated OMZs are located in the wrong place (think of all simulated OMZ waters being around Antarctica), C will be 0. Likewise, if the entire simulated ocean was suboxic according to criterion c ($V^{\mathrm{m}}(c) = 1$), the intersection would equal the observed volume ($C = V^{\cap}(c) = V^{\mathrm{o}}(c)$) and thus be very small (unless we choose a rather large criterion c).

For evaluation of the cost function for optimization we then define

$$J_{\mathrm{OMZ}} = 1 - C, \tag{3}$$

which varies between 0 (optimal fit) and 1 (no overlap between observed and simulated OMZs). The new term J^{OMZ} is of the same order of magnitude as J_{RMSE}: For example, in the reference model MOPS$^{\mathrm{r}}$ $J_{\mathrm{RMSE}} = 0.53$ and $J_{\mathrm{OMZ}} = 0.79$ for an OMZ criterion of $c = 50$ mmol/m^3 (see Table 1).

2.5. Experimental Setup

In addition to the uncalibrated model simulation MOPS$^{\mathrm{r}}$ (Kriest & Oschlies, 2015), and the single-objective calibration MOPS$^{\mathrm{o}}$ presented by (see also section 2.2 ; Kriest, 2017), we here present two new optimizations, which account in different ways to the model's fit to OMZs (see also Table 1).

Optimization MOPS$^{\mathrm{o+}}$ is based on a single metric that is the sum of two components of misfit, namely, the RMSE to nutrients and oxygen (equation (1)) and the model's match to observed OMZ (equation (3)):

$$J = J_{\mathrm{OMZ}} + J_{\mathrm{RMSE}} \tag{4}$$

Because J_{RMSE} varies typically between ≈0.4 and 1.7 (see above and; Kriest, 2017) and J_{OMZ} between 0 and 1, the two components are of equal order of magnitude. This new misfit function J thus pursues a single compromise solution.

A second experiment MOPS$^{\mathrm{mo}}$ treats the two objectives J_{OMZ} and J_{RMSE} separately by applying MO, as described in detail in section 2.7 below. In short, this new approach evaluates a trade-off between the two different components by a so-called Pareto front, which includes parameter sets that fit either objective to a varying degree. Among those sets are also individuals that form a compromise (an equally weighted solution with regard to both objectives), similar to the best solution of optimization MOPS$^{\mathrm{o+}}$.

2.6. Single-Objective Optimization

In order to optimize a single-objective function $f : A \to \mathbb{R}, A \subseteq \mathbb{R}^n$, we use the $(\mu/\mu_w, \lambda)$-CMA-ES (Hansen, 2006, 2016) and our parallel implementation described in Kriest et al. (2017). Note, that in our case, n is the number of free model parameters, A defines the reliable (feasible) parameter domain, and the objective function f is composed of a model simulation and the associated misfit metric, that is, J_{RMSE}, J_{OMZ}, or J.

In a nutshell, the $(\mu/\mu_w, \lambda)$-CMA-ES works as follows: A population of λ candidate solutions (= λ different sets of n parameters) is sampled from a multidimensional normal distribution, $\mathcal{N}(\bar{x}, \mathrm{C})$, defined by a mean vector $\bar{x}$ and a matrix of covariances C (similar to the definition of a univariate normal distribution $\mathcal{N}(\bar{x}, \sigma)$ by its mean $\bar{x}$ and its standard deviation σ). Following λ model simulations with these parameters, and evaluation of λ different misfit values, a new normal distribution is empirically reestimated from the better half of $\mu = \frac{\lambda}{2}$ samples, and the new probability distribution is used for an elaborated and smooth update of the former distribution, which in turn is sampled in the next iteration. The procedure is supposed to efficiently move the normal distribution toward good regions of the searchspace and to converge to a single solution x^*, that is, to $\mathcal{N}(x^*, 0)$, where, ideally, x^* is a (local) optimum of the given misfit function J in the n-dimensional parameter space.

Using single-objective optimization, we can address different targets by minimizing a (weighted sum of) certain model-data misfit(s). This approach is realized by minimizing the misfit metric J, introduced above. Optionally, we might impose constraints on certain model-data misfits. However, it appears difficult to properly join different model-data misfits into a single-objective function or to impose a proper bound on the model-data misfit to some set of observations while minimizing the misfit to another set of observations. Decisions about weights applied to different data sets, or a particular form of misfit function, may be very influential for the optimal parameter choice (Evans, 2003).

2.7. MO

A single-objective optimization converges to a single parameter set, only. Facing diverging model calibration results with respect to different target observations, we expect enhanced model assessment by calculating a couple of well-distributed compromise parametrizations at once. That task is what MO methods are designated for: MO provides a collection of parameter sets such that the user obtains a nearly optimal parameter set with respect to each target but also has the option to choose a best compromise parametrization in his/her opinion. Furthermore, the obtained collection of parameter sets might allow more insights about how specific model skills are influenced by the values of specific parameters.

Multiple, say, k, targets are addressed by corresponding objective functions $f_1, \dots, f_k$. In this situation a solution (parameter set) $x \in \mathbb{R}^n$ is said to *dominate* another $y \in \mathbb{R}^n$ if $f_i(x) \le f_i(y)$ for all $i \in \{1, \dots, k\}$ and $f_j(x) < f_j(y)$ for some $j \in \{1, \dots, k\}$. We write $x \prec y$ to express that x dominates y. On the contrary, two solutions $x \neq y$ are said to be *incomparable* if there are $i \neq j$ with $f_i(x) < f_i(y)$ but $f_j(x) > f_j(y)$. Ideally, provided a set $X \subseteq \mathbb{R}^n$ of allowed solutions, we would like to know the subset of all solutions that are not dominated by other ones, that is, the entire set of "best compromises"

$$\mathrm{ndom}(X) := \{x \in X \,|\, \texttt{There is no} \;\; y \in X \;\; \texttt{with} \;\; y \prec x\}.$$

The elements of the set ndom(X) are called *Pareto optimal* and the image of ndom(X) in the k-dimensional object space $\mathbb{R}^k$ is called *Pareto frontier*. Figure 1 illustrates the situation for $k = 2$ objectives (the case we deal with in this paper). The complement of ndom(X) is the set dom(X) $= X \setminus$ ndom(X) of all dominated elements in X. Clearly, like it is hard to find global optima subject to complex single-objective functions, it is even harder to find some (or all) *Pareto optimal* solutions in the multiobjective case. Therefore, the aim is to find a good *Pareto approximate set*, that is, a set of solutions, which are (in objective space)

1. incomparable,
2. close to the Pareto frontier, and
3. well distributed.

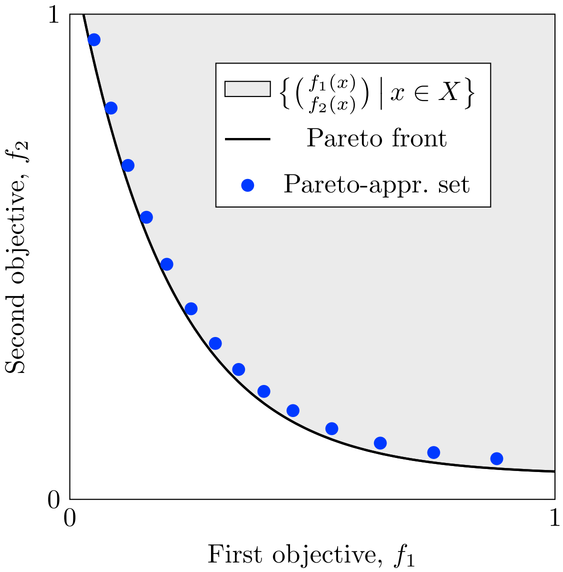

Figure 1. A set $X \subseteq \mathbb{R}^n$ of feasible parameters, mapped to the objective space (given $k = 2$ objectives $f_1, f_2 : X \to \mathbb{R}$), its Pareto front (the image of ndom(X)) and the image of a Pareto approximate set.

An example that satisfies all desired properties is given as the blue dots in Figure 1. Population-based algorithms like evolutionary algorithms ("EA"; Deb, 2009) and estimation of distribution algorithms ("EDA"; Hauschild & Pelikan, 2011) have shown to be quite suitable for this aim.

A popular approach to accomplish items 1–3 is to order solutions by a hierarchical ranking criterion combining the so-called *level of nondominance* criterion (mainly addressing items 1 and 2) with a secondary ranking criterion like the *crowding distance* (Deb et al., 2002) or the *contributing hyper volume* (Emmerich et al., 2005; mainly addressing item 3). We will apply a multiobjective variant of the CMA-ES, which uses the combined level of nondominance and contributing hyper volume criteria for ranking. We introduce the ranking procedures in Appendices A1 and A2 and outline the multiobjective CMA-ES in Appendix A3.

Our parallel implementation of the MO-CMA-ES is available on GitHub (https://github.com/vsauerland/calibrate2O). A permanent version of the code we used for the experiments of this paper is archived in a public Zenodo repository (Sauerland, 2018). For compilation, usage, and further notes, we refer to the README file contained in the repository.

3. Results

The optimizations converge after $M = 119$ (MOPS^{o}) to $M = 170$ (MOPS^{mo}) generations (Table 1). Therefore, MO is computationally more

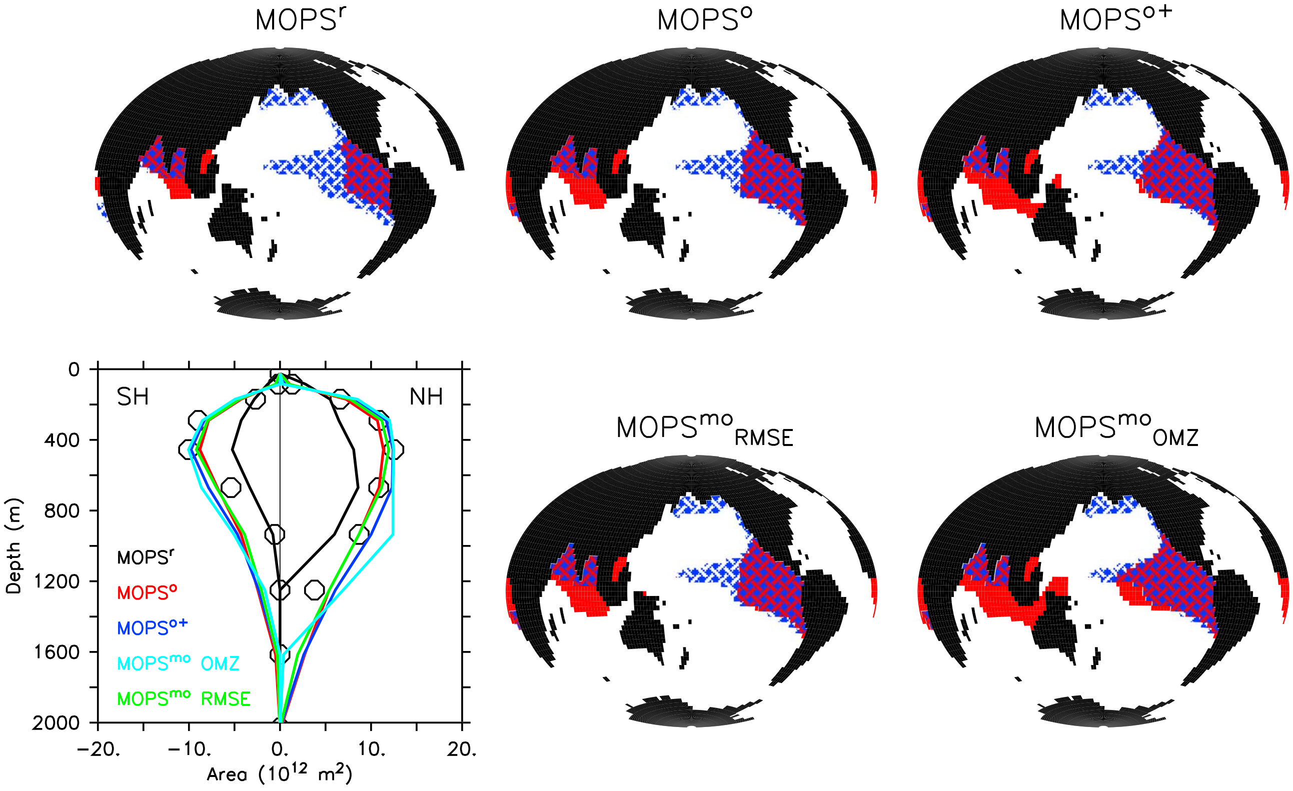

Figure 2. Extent of region with $O_2 \leq 50$ mmol/m^3 at 400 m of observations (blue hatched area) and model (red), for model simulation MOPS^{r}, MOPS^{o}, MOPS^{o+}, and two individuals of MOPS^{mo}, with best fit to RMSE (lower mid panel) and best fit to OMZ (lower right panel). The lower left panel shows the OMZ area in the eastern tropical Pacific (east of 140°W, between 20°S and 20°S) of observations (circles), MOPS^{r} (black lines), MOPS^{o} (red lines), MOPS^{o+} (blue lines), and MOPS^{mo}. For MOPS^{mo} the best individual with respect to J_{RMSE} is shown with green lines and the best with respect to J_{OMZ} with light blue lines. MOPS = Model of Oceanic Pelagic Stoichiometry; OMZ = oxygen minimum zone; RMSE = root-mean-square error.

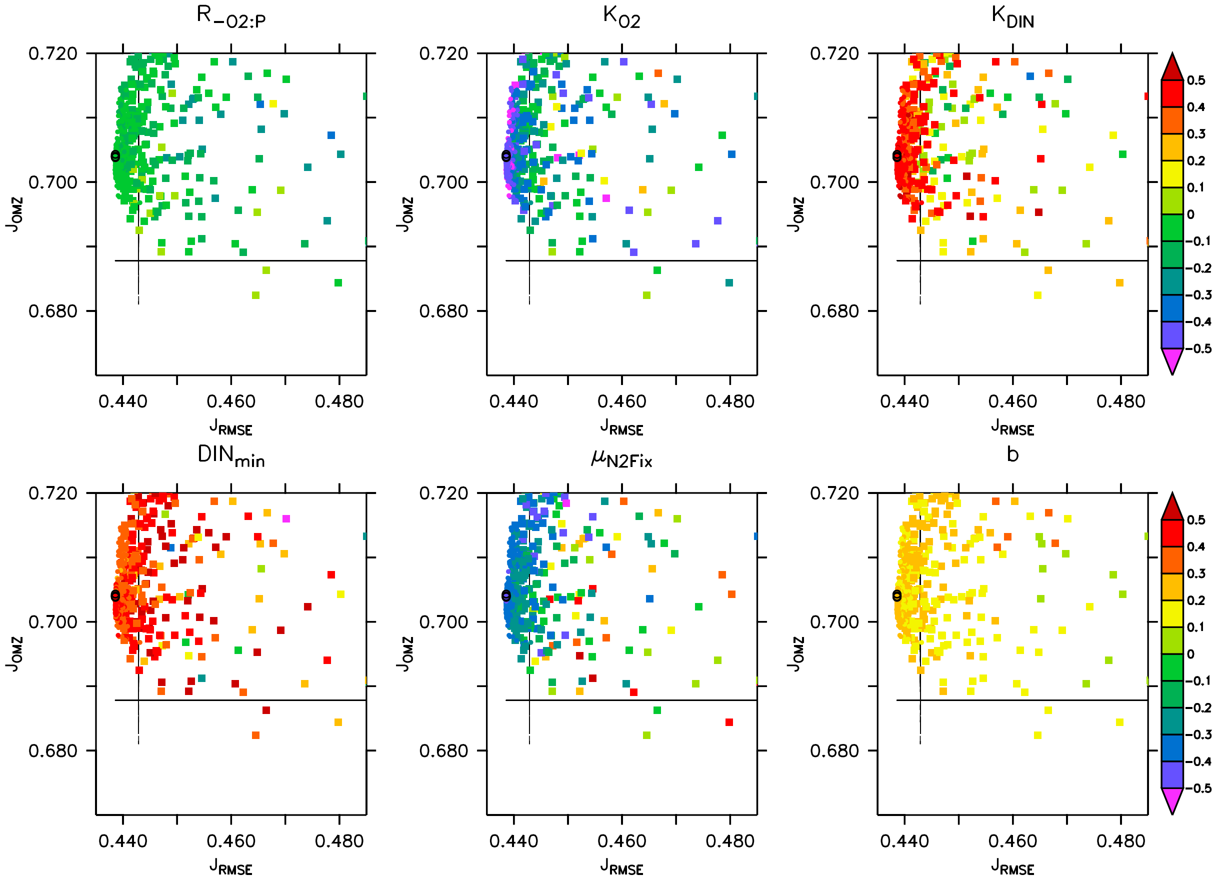

Figure 3. Projection of parameter values (normalized, color scale) onto model misfits, in the vicinity of the optimal solution of optimization MOPS^{o}. Parameter values have been scaled by $(\theta - \theta_c)/(\theta_u - \theta_l)$, where θ is the parameter value, θ_u and θ_l are upper and lower boundary constraints, and $\theta_c = (\theta_u + \theta_l)/2$ is the center of the allowed interval. The first 50 generations are plotted as small squares, and the remaining parameters as circles. Lined circles denote the parameters of the last generation. Horizontal and vertical lines denote the minimum value of each objective +1%. OMZ = oxygen minimum zone; RMSE = root-mean-square error.

expensive than single-objective optimization, because of a slower convergence and a larger population size of $\lambda = 20$ (MOPS^{mo}) instead of $\lambda = 10$ (MOPS^{o} and MOPS^{o+}).

The generally rather slow rate of convergence is partly due to the type of the misfit function, which is quite insensitive to some parameters. For example, a large number of individuals of MOPS^{o} and MOPS^{o+} almost equally good (within 1% of the optimal fit; see Table 1, $M(\Omega)$) with respect to the misfit function. Some parameters relevant for OMZs (such as K_{DIN} and DIN_{min}) are not well constrained by equation (1), as indicated by their large variation in the vicinity of the optimum. This is because OMZs occupy only a small fraction of the global ocean volume, but equation (1) is influenced strongly by the deep waters of the global ocean.

However, adding the fit to OMZs as in MOPS^{o+} does not narrow the uncertainty range of the parameters but results in a wider range of “good” parameters: While for MOPS^{o} almost 60% of the parameters result in a misfit (with respect to equation (1)) that is almost as good as the optimal model, the fraction increases to almost 70% for MOPS^{o+}, which optimizes against equation (4), the combined fit to both objectives. In contrast, MOPS^{mo} only 22% and 29% of all individual model simulations are in the vicinity of the respective optimal fit; this lower percentage is due partly to the large number of model simulations ($\lambda \times M = 3,400$). Another reason is that single-objective optimization converges to a single solution, while MO explores the parameter space along the Pareto front. Thus, the search for a simultaneous good fit to both objectives dealing with 20 compromise solutions results in a quite high additional computational demand. In the following sections we will see that this additional computational burden is justified in terms of model performance and flexibility.

Table 2
Parameters and Cost Functions for Model Simulations With Minimum J_{RMSE} and J_{OMZ} for the Entire Optimization Trajectory of the Three Optimizations MOPS^o MOPS^{o+} and MOPS^{mo}

Experiment	MOPS^o		MOPS^{o+}		MOPS^{mo}	
	J^{min}_{RMSE}	J^{min}_{OMZ}	J^{min}_{RMSE}	J^{min}_{OMZ}	J^{min}_{RMSE}	J^{min}_{OMZ}
b	1.39	1.19	1.41	1.26	1.40	1.34
$R_{-O2:P}$	172	168	171	182	173	187
μ_{NFix}	1.20	3.00	1.76	1.94	1.56	3.00
DIN$_{min}$	15.8	7.1	10.7	1.0	13.2	2.3
K_{DIN}	32.0	20.9	27.0	32.0	25.8	15.5
K_{O2}	1.00	3.18	1.00	1.36	1.00	1.00
J_{RMSE}	0.439	0.502	0.440	0.487	0.439	0.474
J_{OMZ}	0.704	0.681	0.704	0.679	0.701	0.673

Note. All analyses are restricted to individuals whose parameters lie within the prescribed boundaries.

3.1. Does Calibration Only against Nutrient and Oxygen Concentrations Improve OMZs? (MOPS^o)

In a perfect model an improvement in one objective would also improve the other, and therefore, a good fit to observed nutrients and oxygen would coincide with a good match to OMZs. As coupled models contain many simplifications on the physical and biological side, this is likely not the case. Instead, an improvement in one metric might result in deterioration of a second one. To investigate the potential trade-off between J_{OMZ} and J_{RMSE}, we here first analyze the output of all 1,190 individuals (simulations to near steady state) of the optimization by Kriest (2017), a single-objective optimization only against J_{RMSE}, with respect to their fit to OMZs.

The final, best (with respect to J_{RMSE}) solution of MOPS^o leads to a slight extension of the OMZ at 400-m depth (Figure 2) and improves the fit to OMZs (C of equation (2)) by about 50%, relative to the uncalibrated model MOPS^r, resulting in $J_{OMZ} = 0.704$ (Table 1). However, this solution does not perform best with respect to OMZs (Figure 3). Instead, the best fit to observed OMZs of all 1,190 model simulations of the optimization trajectory is $J_{OMZ} = 0.681$ (Table 2). This solution is achieved with different parameters than for optimal J_{RMSE}, namely, a lower affinity of aerobic remineralization to oxygen (K_{O2}), a higher affinity of denitrification to nitrate (K_{DIN}), and a lower threshold for the onset of denitrification (Table 2).

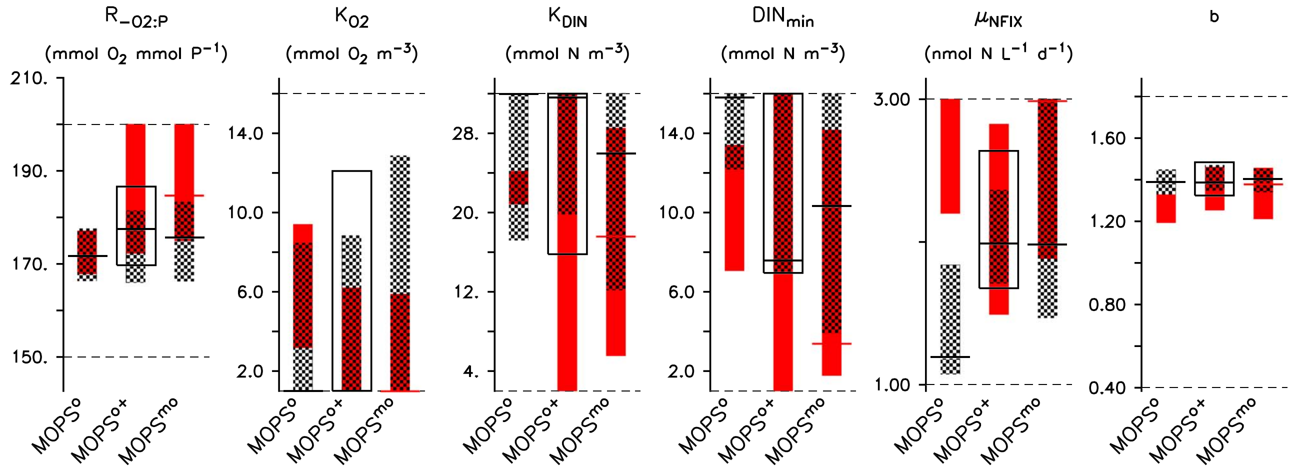

Figure 4. Parameter ranges of all individuals of MOPS^o, MOPS^{o+}, and MOPS^{mo} whose misfit is not worse that 1% of the best individual with respect to J_{RMSE} (hatched black bars) and J_{OMZ} (red bars). Horizontal thick lines denote the final misfit J of the optimization; for MOPS^{mo} these are those individuals of the last generation that are optimal with respect to J_{RMSE} (black) and J_{OMZ} (red). For MOPS^{o+} the open rectangle displays the range of parameters with respect the combined misfit, J. Only those parameters inside the prescribed boundaries (see Kriest, 2017, here indicated by horizontal dashed lines) are considered in the analysis. MOPS = Model of Oceanic Pelagic Stoichiometry.

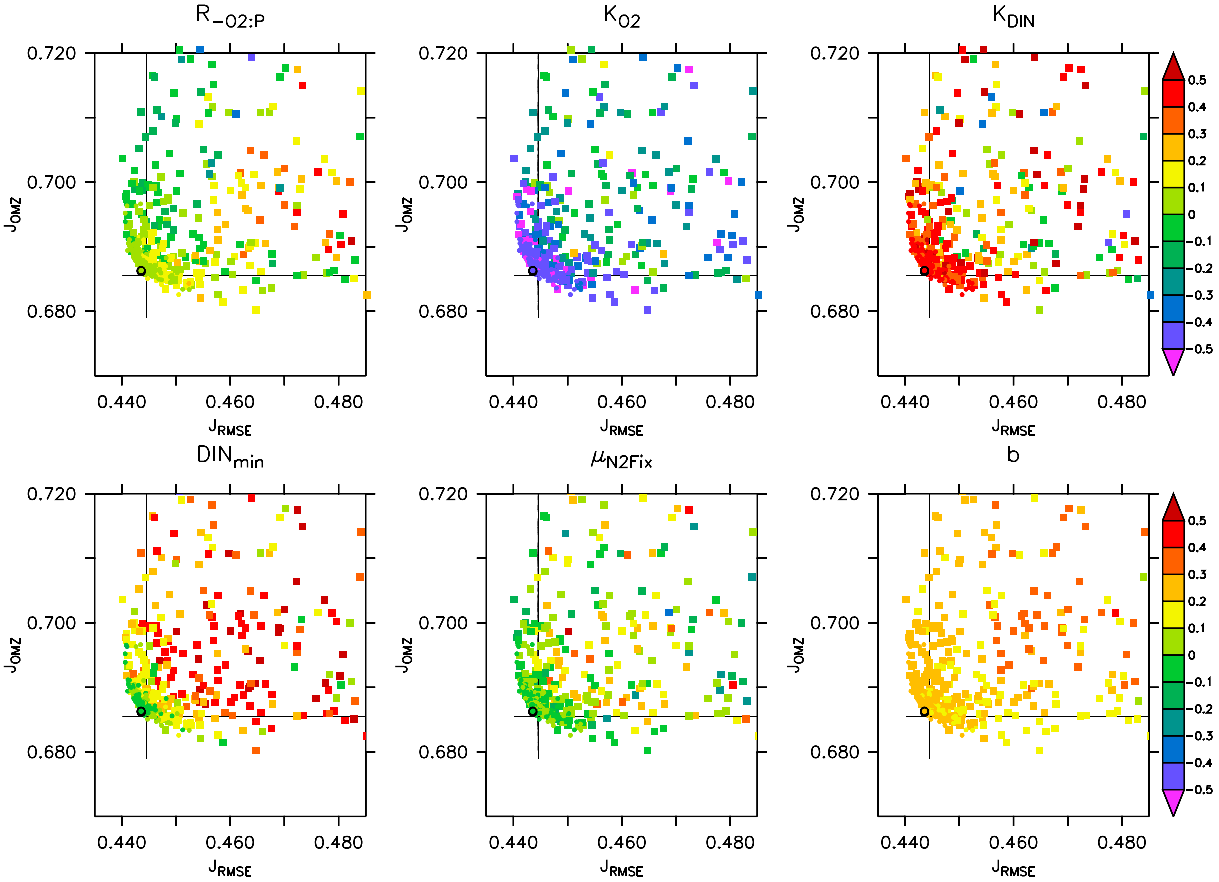

Figure 5. As Figure 3 but for optimization MOPS^{o+}.

Together with a maximum nitrogen fixation rate μ_{NFix} that is more than twice as high as for best J_{RMSE}, this indicates an enhanced fixed nitrogen turnover. Also, a lower value for b indicates that a longer particle flux length scale is of advantage to achieve a good match to OMZs. The improvement with respect to OMZs comes at the cost of a larger misfit to dissolved tracer concentrations, resulting in $J_{\mathrm{RMSE}} = 0.502$. Thus, the a posteriori analysis of the simulations from single-objective optimization clearly illustrates a trade-off between these two diagnostics, and the need for different parameter sets, that are required for a good match to either tracer concentrations or OMZs.

Yet a quite wide range of parameters accommodates a good fit to observed OMZs. This holds particularly for the affinity of remineralization to oxygen (K_{O2}), the threshold for the onset of denitrification ($\mathrm{DIN}_{\mathrm{min}}$), and maximum nitrogen fixation rate (μ_{NFix}), as illustrated in Figure 4. For the latter parameter, values optimal with respect to J_{OMZ} (red areas in Figure 4) diverge strongly from those for J_{RMSE} (black hatched areas in Figure 4). Thus, even in the presence of a large indeterminacy of parameters relevant for OMZs, a good representation of these areas is likely associated with a high fixed nitrogen turnover.

Unfortunately, the area of the parameter space that might be beneficial for J_{OMZ} is only sparsely explored by the optimization algorithm, due to the applied misfit function. This is observed by the colors in Figure 3, which represent normalized parameter values in terms of relative deviations from the center of the corresponding boundary constraints. This fact, the above-mentioned trade-off between the two objectives, and the different optimal parameter sets indicate that we might need to explicitly include the fit to the OMZs in the cost function of the optimization.

3.2. Combining Two Objectives Into One Single Metric (MOPS^{o+})

Optimization against equation (4) ($J = J_{\mathrm{RMSE}} + J_{\mathrm{OMZ}}$) results in an extension of the OMZ in the eastern equatorial Pacific at 400-m depth but at the cost on an overestimate of OMZ extent in the Indian Ocean (Figure 2). The parameter space for both J_{RMSE} and J_{OMZ} is now exhaustively explored in the vicinity of the

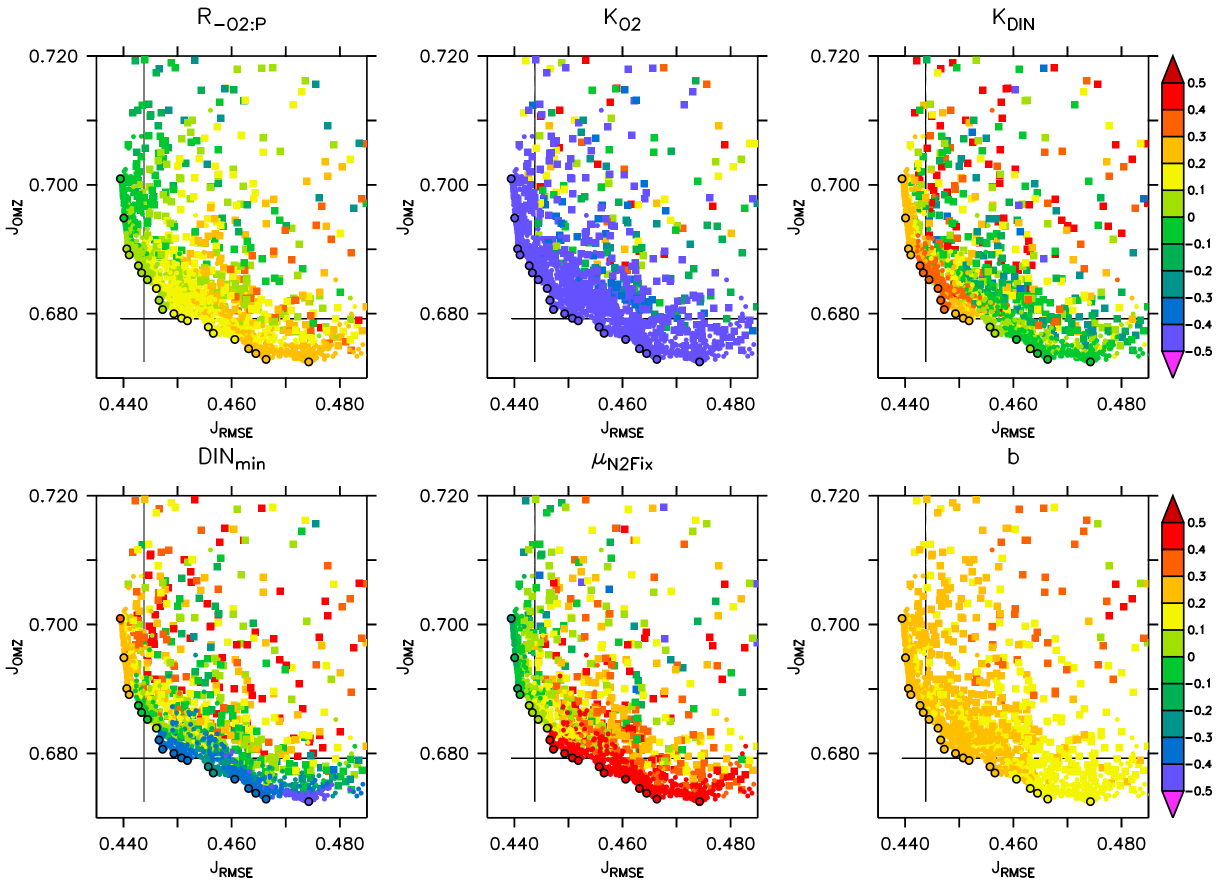

Figure 6. As Figure 3 but for optimization MOPS^{mo}.

optimum, and a good model solution with respect to RMSE is now closer to the optimal value for the OMZ (Figure 5).

Optimization MOPS^{o+} results in a maximum nitrogen fixation rate μ_{NFix} that is almost twice as high as that of MOPS^{o} (Table 1) and is associated with a reduced threshold for the onset of denitrification. This tendency of enhanced nitrogen cycling agrees with the results obtained from the a posteriori analysis of MOPS^{o} (see above). However, again a very wide range of parameters related to the oxidant affinity allows a good fit to observed OMZs (Figure 4). Because this optimization targets a compromise solution, the two ranges of maximum nitrogen fixation rate, which are optimal for the two objectives, now overlap, instead of being distinctly different (as in MOPS^{o}; Figure 4).

As for MOPS^{o} the best fit to observed OMZs of all model simulations of MOPS^{o+} is achieved with a sightly lower affinity of aerobic remineralization to oxygen and a lower threshold for the onset of denitrification (Table 2), together with a slightly increased maximum nitrogen fixation rate. Also, a smaller *b* points toward the necessity of faster particle sinking (deeper flux penetration) for a good representation of OMZs. The divergence between parameters optimal for the different objectives is generally less than for MOPS^{o}, likely because the metric applied here attempts to fit both objectives at the same time.

Therefore, because both MOPS^{o} and MOPS^{o+} point toward different sets of parameters to satisfy different objectives, it seems worthwhile to apply MO, which treats both objectives to some extent independently.

3.3. MO Against Nutrients and Oxygen Concentration and OMZs (MOPS^{mo})

The extension of the OMZ in the eastern equatorial Pacific at 400-m depth improves considerably, when selecting the individual of the Pareto front that matches best with respect to J_{OMZ} (Figure 2). However, this comes at the cost of an overestimate of OMZ extent in the Indian Ocean. Globally, the combination of both effects leads to an only small improvement in J_{OMZ}.

MO of MOPS against both metrics J_{RMSE} and J_{OMZ} does not result in one single, best solution but in a selection of 20 different parameter sets, in which one objective can only be improved at the cost of the other objective. Within this group, the individuals are quite uniform with respect to the same, low value of K_{O2} (Figure 6 and Table 1). The fit to J_{OMZ} improves with larger $R_{-O2:\mathrm{P}} = 185$ and a lower b, as was already evident in the a posteriori analysis of MOPS$^{\mathrm{o}}$ and MOPS$^{\mathrm{o+}}$. The MO clearly indicates distinct sets of parameters related to nitrogen cycling: For a low misfit with respect to J_{OMZ} a very low affinity to nitrate (K_{DIN}) and threshold for the onset of denitrification ($\mathrm{DIN}_{\mathrm{min}}$) are necessary, together with a high rate of nitrogen fixation (μ_{NFix}). In contrast, a good fit to J_{RMSE} requires a high nitrate affinity and threshold and a reduced nitrogen fixation rate. This result qualitatively agrees with the a posteriori analysis of the two optimizations presented above.

The optimal parameters of the Pareto front are of a slightly narrower range than the parameters that perform best throughout the entire (explorative) optimization trajectory (Table 2) but the tendency of faster particle sinking and enhanced nitrogen cycling as an improvement for OMZs that we found for MOPS$^{\mathrm{o}}$ and MOPS$^{\mathrm{o+}}$ remains.

To summarize, optimization leads to an improvement of OMZ representation. All three optimizations indicate a trade-off between the two objectives and require different parameter sets for the best representation of tracer concentrations or OMZs. While the former are better simulated with b representing slow particle sinking and fixed nitrogen turnover, the latter benefit from slightly enhanced particle sinking a more rapid nitrogen cycling in OMZs. This change might be of consequence for simulated different global nitrogen inventory or OMZ volume, which can be of importance for model dynamics and higher trophic levels.

4. Discussion

4.1. Robustness of Calibration

For our model calibrations we apply variants of the covariance matrix adaption evolution strategy, CMA-ES, which has proven successful with regard to many benchmark problems (Hansen et al., 2010). CMA-ES requires rather few model simulations compared to other population-based search heuristics but more model simulations to converge to some local optimum than gradient-based methods. However, it is a suitable choice in the face of complicated, irregular "search landscapes" with local optima (which might be far worse than the global optimum), or discontinuities.

There is also evidence that population-based search heuristics, especially evolution strategies like CMA-ES, are comparably robust with respect to measurement errors (e.g., and the first references therein; Jeballa et al., 2011). Sauerland et al. (2018) examine methods to calculate lower bounds on the best attainable model-data misfit with respect to noise disturbed data.

Concerning MO and the concept of Pareto optimality, population-based algorithms have proven to be the best choice. For example, many MO studies in the field of hydrology applied the genetic algorithm NSGA-II (Deb et al., 2002). The CMA-ES variant we apply has shown to perform well in comparison to NSGA-II (Igel et al., 2007).

4.2. MO as Tool for Global Model Calibration

Our experiments show that calibration of a global biogeochemical model against "only" nutrient and oxygen distributions in MOPS$^{\mathrm{o}}$ helps to improve the model with respect to the representation of OMZs; yet the parameter space is not exhaustively explored and other parameter sets may achieve an even better fit to the observed OMZs. Adding several objectives into one single metric (as in MOPS$^{\mathrm{o+}}$) results in compromise solutions and may, to a certain extent, be sufficient. In this experiment the individual components of the combined misfit were of equal magnitude—this does not have to be the case for other tracers or diagnostics. Then it is necessary to assign weights to the different terms of the misfit function. Weighting might be very influential for the resulting parameters and fluxes (Evans, 2003), and any decision about weights will add a somewhat subjective component to the procedure, best parameter set, and model dynamics, as indicated by the very different global nitrogen turnover of MOPS$^{\mathrm{mo}}$|RMSE and MOPS$^{\mathrm{mo}}$|OMZ. Thus, a model that has been calibrated successfully against one objective might not perform optimally for other tasks. One way out of this dilemma could be multiple, single-objective optimizations against different misfit functions and/or data. However, it is likely undesirable to calibrate a global model against very specific objectives without preserving at least a reasonable fit against observed dissolved inorganic tracers, which are at the heart of global biogeochemical models, and are available from quality-controlled, and very dense data sets.

Considering the drawbacks of weighting mentioned above, MO provides a tool that enables us to calibrate global models with a large flexibility regarding later applications of the model, in particular when objectives are not clear a priori.

The experiment with MO has shown good performance with respect to the exploration of the parameter space. It accomplishes our intention to find a collection of compromise solutions that is well distributed in objective space, where the unique compromise which provides the lowest misfit with respect to observed distributions of nutrients and oxygen (first objective) is almost as good as the corresponding single-objective calibration result. The subsequent analysis of the "best" individuals provides a wealth of information and offers a high flexibility for model and parameter choice (when these are used, e.g., for further application such as projections). The parameter set to be picked depends very much on the research question. For single-objective, those solutions that are close to J (e.g., within 1% of the optimum; Table 1) can be deemed equally good. For MO, in the current example it depends if we are more interested in a good representation of nutrient and oxygen concentrations—then we would pick any of the parameter sets providing nearly optimal J_{RMSE}. If we are interested in the representation of OMZ, we would pick a parameter set with nearly optimal J_{OMZ}. In any case, MO assures that the other objective is not entirely neglected—the benefit compared to two single-objective optimization, whenever the objectives are uncorrelated. The most versatile model configuration might be represented by a parameter set that is close to the "knee" of the Pareto front, that is, the region where one objective's gain is in balance with the other objective's loss (Figure 6), as such configuration is likely to imply the most universal applicability of a model and might help to narrow suitable parameter ranges. Indeed, the multiobjective approach can help model developers to setup particularly new models in a more versatile way, rendering them applicable to a wider range of research questions. This could be done by, for example, providing the model code together with the range of parameters (at the Pareto front from MO).

Model assessment can be facilitated by examining the final collection of compromise solutions obtained by MO: if certain parameter values show a clean monotonic tendency while traversing the final Pareto approximate set from one objective's best solution to another "extreme" solution, we gain more evidence about dependencies between processes and the parameters than by considering two single-objective model calibration results. This is the case, for example, for our parameters $R_{-\mathrm{O2:P}}$ and μ_{NFix}, as indicated by the corresponding color gradients in Figure 6.

Dealing with two objectives, the chosen population size of $\lambda_{\mathrm{MO}} = 20$ normal distributions seems sufficient for a good MO exploration of the parameter space. Note that the single-objective CMA-ES deals with only one distribution but (by default for six parameters) samples a population of $\lambda = 10$ individuals. Further, the two extreme compromise solutions of the Pareto approximate set are as good as the corresponding solutions obtained by two single-objective optimizations. Our choice of λ_{MO} is in accordance with a rule of thumb suggested in Deb (2009, Chapter 8.8). It defines a threshold value for the maximum portion of nondominated solutions in a randomly generated start population. According to this rule, the population size grows exponentially with the number of objectives, but for two objectives a threshold between 30% and 20% implies a minimum population size between 10 and 20 individuals. Summarizing, although of higher computational cost (because of a larger population size), MO can be a very helpful and flexible tool for model calibration, with many, so far unexplored future applications.

4.3. OMZs and the Pelagic Nitrogen Cycle

As noted earlier (Kriest, 2017) optimization of global models against J_{RMSE} results in a much better fit to observed global biogeochemical fluxes as well, as evident from enhanced global primary production, reduced particle flux at 2,000-m depth, lower global pelagic denitrification and a strong reduction in benthic burial (see also Figure 7). Many of these changes are likely related to the reduction in particle sinking (higher b), as well as a lower affinity of denitrification to nitrate (Table 1). When using J_{OMZ} as objective, as in MOPS$^{\mathrm{o+}}$ or MOPS$^{\mathrm{mo}}$|OMZ, the model again overestimates pelagic denitrification. The overestimate is most severe in the solution of MOPS$^{\mathrm{mo}}$|OMZ that is optimal for J_{OMZ} (Figure 7) and disappears when selecting the individual of the Pareto front that is optimal for J_{RMSE}, as in MOPS$^{\mathrm{mo}}$|RMSE. This response of the model to changes in the parameters is most likely due to the antagonistic effects of the two oxidants oxygen and nitrate and the parameters relevant for these processes. A high affinity of denitrification to nitrate (as mediated through low K_{DIN} and $\mathrm{DIN}_{\mathrm{min}}$) "protects" oxygen from becoming depleted in regions of weak

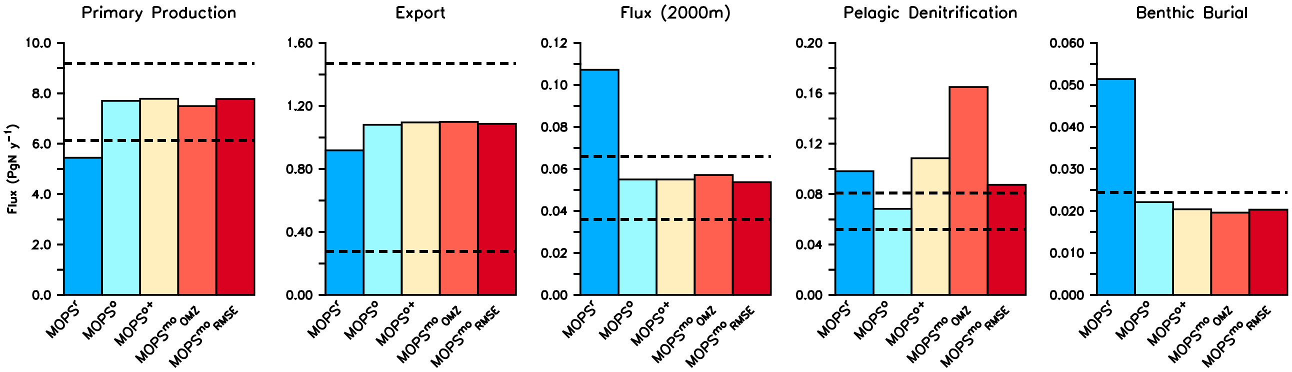

Figure 7. Global biogeochemical fluxes of primary and export productions, organic particle flux at 2,000-m depth, pelagic denitrification and benthic burial, simulated with MOPS^r, MOPS^o, MOPS^{o+}, and MOPS$^{mo}_*$. For the latter optimization we give the flux of the individual with best root-mean-square error and the individual with best oxygen minimum zone. Horizontal dashed lines denote observed estimates. For global observed primary production we refer to a range of 40–60 Gt C/year (Carr et al., 2006). The ranges for particle flux and export production have been determined from Honjo et al. (2008), Lutz et al. (2007), and Dunne et al. (2007). Observed global pelagic denitrification has been compiled from various sources, as listed in Table 2 of Kriest and Oschlies (2015). The estimate for observed benthic burial is based on Wallmann (2010). MOPS = Model of Oceanic Pelagic Stoichiometry.

ventilation but at the cost of a strong decline in nitrate (see Figure 8). This causes an increase in the model's misfit to nitrate, which is reflected in J_{RMSE}.

This latter result is in agreement with results obtained by Moore and Doney (2007). In their global model experiments Moore et al. had to impose a high threshold of 32 mmol/m^3 nitrate for denitrification, to prevent the model from simulating unrealistically low nitrate concentrations in the eastern equatorial Pacific and excessive fixed N loss. This is equivalent to the optimization against J_{RMSE}, which targets at very high K_{DIN} and $\mathrm{DIN}_{\mathrm{min}}$. The better fit to J_{RMSE} comes at the cost of a worse representation of OMZs, which are better represented with low K_{DIN} and $\mathrm{DIN}_{\mathrm{min}}$. Therefore, even model MOPS, that explicitly resolves the dependency

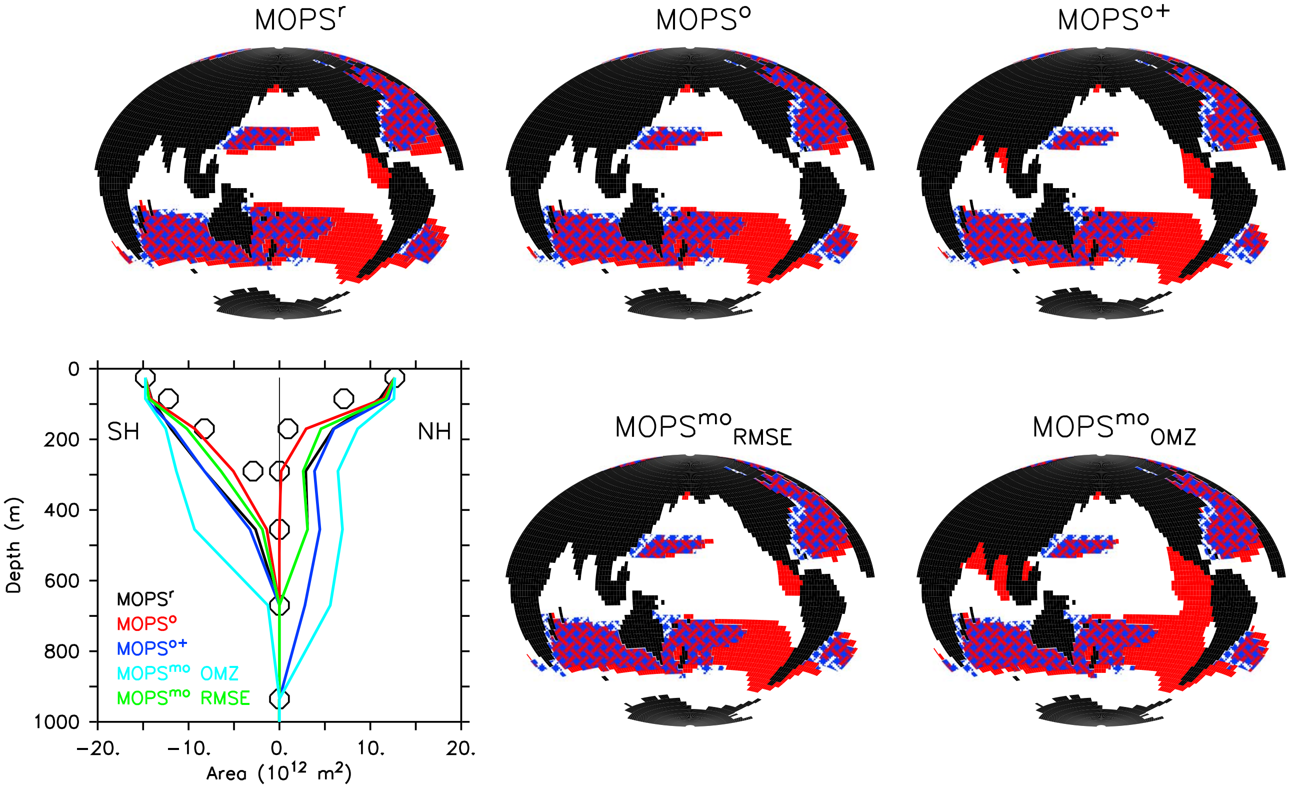

Figure 8. As Figure 2 but for nitrate and a criterion of $NO_3 \leq 20$ mmol/m^3 at 400 m. Note that axes in the lower left panel differ from that of Figure 2.

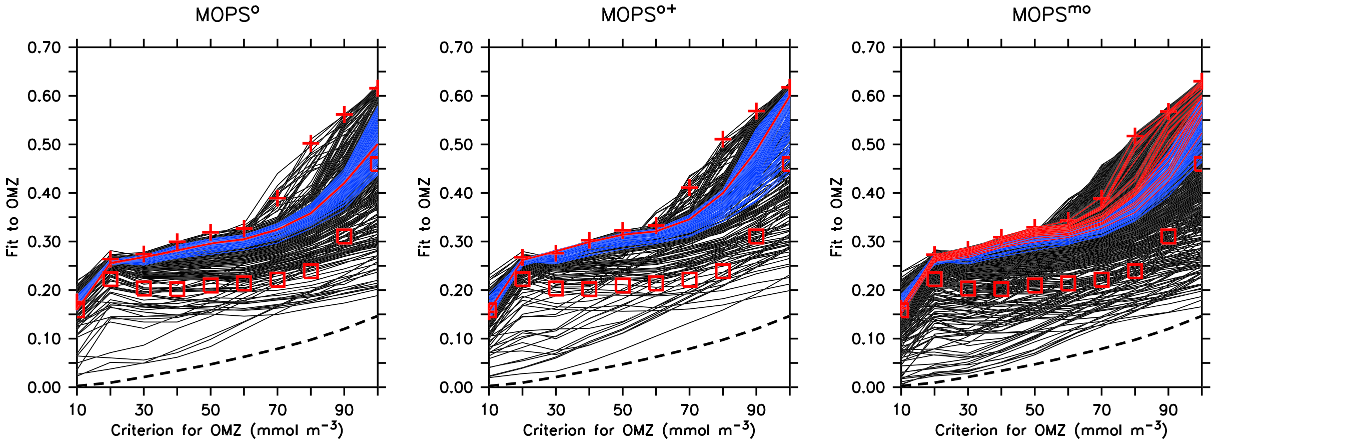

Figure 9. Fit to OMZ, for model solutions of optimization trajectory (lines) of MOPS^o (left), MOPS^{o+} (middle), and MOPS^{mo}. The plots are similar to Figure 7a of Cabre et al. (2015), but for the global ocean: The *x* axis shows the criterion *c* for OMZ definition ($O_2 < c$), and the *y* axis the degree of overlap between model and observations (equation (2)). Black lines: model solutions with $J_{\mathrm{RMSE}} \quad / J^*_{\mathrm{RMSE}} \quad - 1 > 0.01$. Blue lines: model solutions with $J_{\mathrm{RMSE}} \quad / J^*_{\mathrm{RMSE}} - 1 \leq 0.01$. Red line: individuals of the last generations of the optimization. Red pluses: best individual with respect to J_{OMZ} of all simulations. Red squares: MOPS^r. Only individuals with parameters inside the prescribed boundaries are considered in the analysis. Bold dashed lines indicate (hypothetical) solution for an entirely suboxic ocean. MOPS = Model of Oceanic Pelagic Stoichiometry.

of remineralization on organic carbon and oxidant supply cannot circumvent or resolve the discrepancy of too large demands for oxidants in the weakly ventilated OMZs.

A likely candidate for the better simulation of these regions is a more advanced circulation model, which allows for a better representation of physical processes. However, we note that this comes at much higher computational costs, complicating the calibration of the biogeochemical module (if not rendering it impossible). A possible solution to this dilemma could be to optimize biogeochemical model parameters in a coarse resolution model, whose properties resemble those of the higher resolved one and then apply the optimized parameters to the latter. This approach is similar to the surrogate-based optimization proposed by Priess et al. (2013), which alternates between so-called high-fidelity and low-fidelity models. Future research will show if, and to what extent, we can transfer parameters calibrated in different circulations among the different setups. Another solution might be the application of model emulators or metamodeling techniques on a high-resolution or more complex global model. While the TM method approach applied here can be classified with physical emulators, statistical emulators based on uncertainty propagation (e.g., utilizing the latin hypercube approach; Battaglia & Joos, 2018; Urban & Fricker, 2010) and combinations of both,

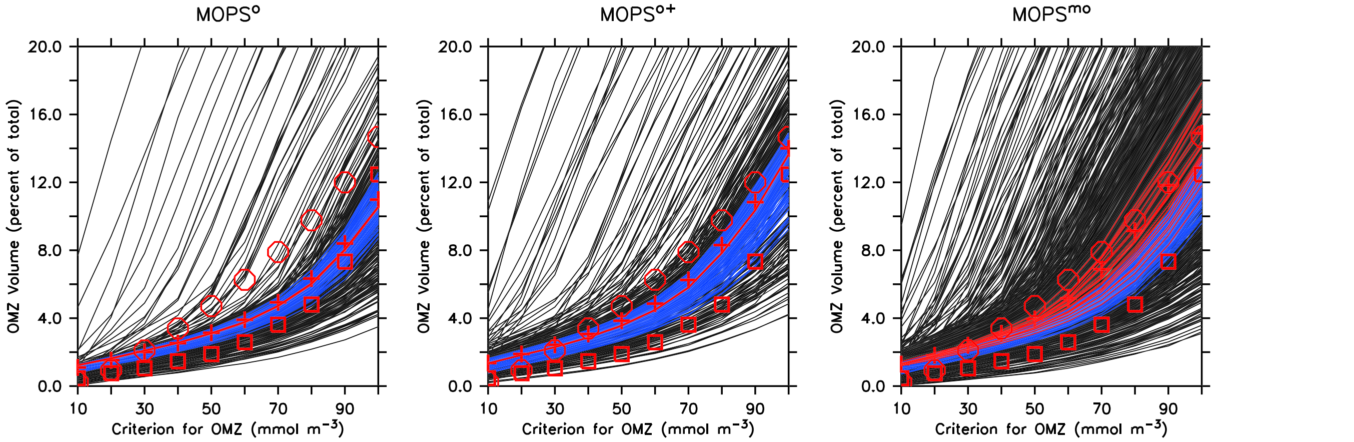

Figure 10. Total suboxic volume (as percent of total ocean volume), for model solutions of optimization trajectory, plotted against the criterion for OMZ definition. Panels from left to right: MOPS^o (left), MOPS^{o+} (middle), and MOPS^{mo} (right). Lines and symbols as of Figure 9. Open red circles: OMZ volume from observations (WOA05). MOPS = Model of Oceanic Pelagic Stoichiometry; OMZ = oxygen minimum zone.

physical, and statistical emulators have also been examined in order to reduce computational burden (also see Schartau et al., 2017, Chapter 8, and the references therein).

4.4. Different Criteria for OMZ Definition

By targeting one criterion for OMZ definition our optimization and metric never the less include a subjective element, because an oxygen content that is sufficient for one organism might not be so for another. Consider, for example, the vampire squid, *Vampyrotheutis infernalis*. These detritivores spend most of their life at depths where oxygen concentrations are very low (around 20 mmol/m^3) and where they can avoid their predators (Hoving & Robison, 2012). On the other hand, because of their large oxygen demand many top predators (e.g., commercially exploited fish such as tuna or swordfish) are limited by oxygen concentrations less than about 150 mmol/m^3 and reduce their vertical excursions when OMZs become shallow (Seibel, 2011). The effect of projected changes in OMZ extent on species distribution, global fisheries and the vulnerability of fish stocks (Stramma et al., 2012), and other animals can thus depend crucially on the criterion used for model calibration and OMZ volume assessment. Also, production and consumption of climate-relevant N_2O depends on the occurrence of regions with very low (down to a few millimoles per cubic meter, Farias et al., 2009) oxygen concentrations, which introduces another potential criterion for OMZ definition.

In our model calibrations we have applied a quite moderate yet for some fish detrimental criterion of $c = 50$ mmol/m^3 for the definition of OMZs. We here finally investigate whether model calibration against this criterion also improves the model's representation of OMZ volume defined via different criteria. To do this, we evaluated the model's fit to OMZ a posteriori for different criteria (c of equation (2)), over a range from 10–100 mmol/m^3. Compared to the reference run, there is already an improvement over the whole range of criteria when using only J_{RMSE} as the sole objective (Figure 9, left panel). Adding J_{OMZ} to the misfit in optimization MOPS^{o+} results in a further increase in model performance (Figure 9, middle panel). Most encouraging, the best solutions of optimization MOPS^{mo} show a considerable improvement particularly in the range between 80 and 100 mmol/m^3 when compared to the reference solution (Figure 9, right panel). Although the fit is not different to the best (neglecting J_{RMSE}) solutions of MOPS^{o} and MOPS^{o+}, the advantage of MOPS^{mo} is that these solutions do not have to sacrifice too much of J_{RMSE}; that is, they still include some information on the model's fit to nutrients. The beneficial effect of MO is even more obvious when looking at the simulated total volume of OMZs over the range of criteria from 10 to 100 mmol/m^3, as shown in Figure 10: Here, particularly MOPS^{mo} results in an improvement of the model close toward observations, for the range between about 70 and 80 mmol/m^3. Thus, a criterion such as $c = 50$ mmol/m^3 seems appropriate for model calibration toward observations even outside the range considered during optimization.

5. Conclusions

Optimization against nutrient and oxygen concentrations can help to improve the model fit to different, independent diagnostics and therefore help to improve the general model performance. However, different objectives (as formulated from the research question the model is applied to) might require very different model parameters for an optimal model performance. Combining different objectives into one single metric is not a trivial task, adds a subjective element, and may only result in compromise solutions. MO may help to set up models that are reliable over a range of different criteria corresponding to different organism habitats or biogeochemical processes. As biogeochemical models are used more and more in complex Earth system models with many users, and for many different scientific objectives and tasks, it is therefore useful to explore algorithms and methods that allow for a wide applicability of these models. MO and its exploration of the Pareto front can be one way to set up versatile model simulations, so that different users can, for different applications/objectives, use the (results from the) best model setup for their specific task.

Dealing with two objectives, we find that a moderate population size of 10 to 20 individuals suffices to obtain a collection of compromise solutions which is (in objective space) well distributed and contains satisfying extremes, as compared to single-objective model calibration results. The obtained Pareto approximate collection of parameter sets may ease model assessment by revealing more dependencies between parameter values and model skills. It would also be interesting to simultaneously consider more than two uncorrelated objectives, although the required population size (computational effort) is supposed to grow exponentially with the number of objectives, and visualizing/analyzing results will become more difficult beyond three objectives.

Our experiments suggest that model parameters, which have been fitted against concentrations of nutrients and oxygen, are not necessarily best suited to simulate the location and extent of OMZs. These regions are best represented with parameters that induce in a high fixed nitrogen turnover. However, these results may depend on the circulation, into which the biogeochemical model is embedded and may change if, for example, a resolution that represents tropical upwelling systems in a better way, is used. Therefore, a further, likely important modification would be the incorporation of more reliable physical processes, which resolve the very detailed dynamics in equatorial and coastal regions, which are of relevance for commercial fisheries and hot spots of climate gas emissions. Encouragingly, especially MO shows that once parameters have been fitted against OMZs defined by a single criterion, the model also represents OMZs better across a range of criteria, thereby rendering the model more applicable to different research questions.

Appendix A: MO

A1. Ranking by Level of Nondominance

Consider a set $S = \{s_1, \dots, s_{\lambda_{\mathrm{MO}}}\} \subseteq X \subseteq \mathbb{R}^n$ of candidate solutions. The ranking by level of nondominance is illustrated in Figure A1. The dots represent the elements of S, mapped to the objective space. The subset $\mathrm{ndom}(S)$ corresponds to the blue dots. All elements of $\mathrm{dom}(S)$ are to the right above of the blue line connecting the elements of $\mathrm{ndom}(S)$. We set $S_0 := \mathrm{ndom}(S)$. For all $s \in S_0$ we define the level of nondominance of s in S by $\mathrm{level}_S(s) := 0$. If $S_0 \neq S$ we continue with $S_1 := \mathrm{ndom}(S \setminus S_0)$ and define $\mathrm{level}_S(s) := 1$ for all $s \in S_1$. In Figure A1, S_1 corresponds to the cyan dots. All elements that do neither belong to S_0 nor to S_1 are to the right above of the cyan connection line of the elements in S_1. The procedure is repeated until all elements of S are assigned to a level of nondominance, setting $S_j := \mathrm{ndom}(S \setminus (S_0 \cup \dots \cup S_{j-1}))$ and $\mathrm{level}_S(s) := j$ for all $s \in S_j$ as long as $S_0 \cup \dots \cup S_{j-1} \neq S$. For our example in Figure A1 we have $S = S_0 \cup S_1 \cup S_2 \cup S_3$, defining four levels of nondominance 0,1,2,3.

Now, the primary (partial) rank of a candidate solution $s \in S$ is simply set to the total number of candidate solutions with a lower level of nondominance, which is $\mathrm{rank}^1_S(s) := 0$ for $s \in S_0$ and

$$\mathrm{rank}^1_S(s) := \sum_{i=0}^{\mathrm{level}_S(s)-1} |S_i|$$

for $s \in S \setminus S_0$. The assignment allows to derive a unique ranking by applying a secondary ranking criterion, $\mathrm{rank}^2_{\mathrm{S_j}}$, within each S_j and setting

$$\mathrm{rank}_S(s) := \mathrm{rank}^1_S(s) + \mathrm{rank}^2_{S_j}(s)$$

for $s \in S_j$. In the next section we describe the secondary ranking criterion, which is applied for the multiobjective CMA-ES described in section A3.

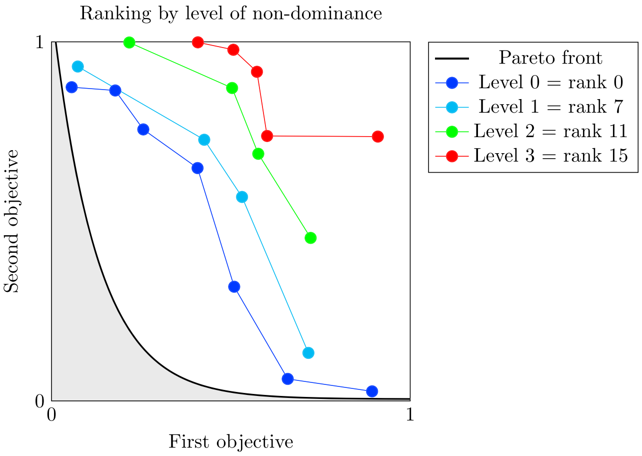

Figure A1. Ranking of the set of candidate solutions S of one generation with regard to two objectives. Dots represent the set S when mapped to the objective space. Each solution $s \in S$ is ranked with regard to its level of nondominance $\mathrm{level}_S(s)$. The different levels are indicated by different colors. All solutions on the same level of nondominance are assigned to the same (partial) rank, which is the total number of candidate solutions on lower levels of nondominance.

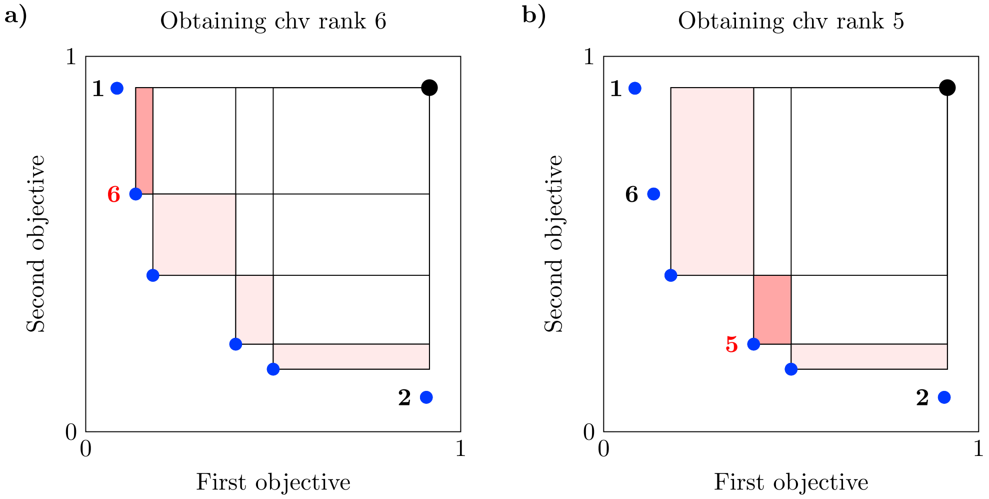

Figure A2. Ranking of a set of nondominated candidate solutions $S_j \subseteq S$ by the contributing hypervolume (chv) criterion with regard to two objectives. The blue dots represent the set S_j when mapped to the objective space. The black dot is a reference point x^{ref} of objective values such that all solutions in S_j are better with respect to both objectives. The first two ranks are assigned to the solutions that provide the best first objective value and the best second objective value, respectively. (a) The contributing hypervolumes of the remaining candidate solutions correspond to the red rectangles. Among these, the last rank is assigned to the solution with the smallest contributing hypervolume. (b) The procedure is repeated with the remaining unranked solutions to assign the second last rank, and so on.

A2. Ranking by Contributing Hypervolume

We now consider the subset $S_j \subseteq S$ for some level j of nondominance. We choose a reference point x^{ref} in the objective space $\mathbb{R}^k$, which is dominated by all $s \in S_j$; that is, $f_i(s) < x_i^{\text{ref}}$ for all $s \in S_j$ and all $i = 1, \dots, k$. Figure A2 illustrates the situation for $k = 2$ objectives. In panel (a), the blue dots represent the set S_j (in objective space) and the black dot is the reference point. For each $s \in S_j$ the *hypercube* between s and the reference point is the set

$$\text{hc}(s, x^{\text{ref}}) = \{x \in \mathbb{R}^k \mid f_i(s) \le x_i \le x_i^{\text{ref}} \quad \texttt{for all } i = 1, \dots, k\}.$$

The *hypervolume* (Zitzler & Thiele, 1998) of S_j with respect to x^{ref} can be defined as the volume of the union of all *hypercubes* between the elements of S_j and the reference point (Coello Coello et al., 2002); that is,

$$\text{vol}(S_j, x^{\text{ref}}) = \text{vol}\left(\bigcup_{s \in S_j} \text{hc}(s, x^{\text{ref}})\right).$$

In Figure A2, the hypercubes of interest are the rectangles between each blue dot and the black dot. The entire area that is covered by these rectangles corresponds to $\text{vol}(S_j, x^{\text{ref}})$. Finally, the *contributing hypervolume* of a single $s \in S_j$ (with regard to S_j) is

$$\text{chv}\,(s, S_j) = \text{vol}\,(S_j, x^{\text{ref}}) - \text{vol}\,(S_j \setminus \{s\}, x^{\text{ref}}).$$

In our example, the contributing hypervolume of a candidate solution corresponds to the highlighted nonoverlapping area of the associated rectangle.

The ranking that is based on contributing hypervolumes proceeds as follows: The best $\min(k, |S_j|)$ ranks are assigned to the *extreme* candidate solutions which (in S_j) provide the smallest value with respect to some of the objective functions $f_1, \dots, f_k$ (confer ranks 1 and 2 in Figure A2). In the remaining set of yet unranked candidate solutions, say, S_j^*, rank $|S_j|$ is assigned to a solution with minimum contributing hypervolume. Updating S_j^*, the latter step is repeated with decreasing ranks $|S_j| - 1, |S_j| - 2, \dots$, until all candidate solutions in S_j are ranked (until S_j^* is empty). Emmerich et al. (2005) proposed to use the contributing hypervolume as selection criterion for a multiobjective evolutionary optimization algorithm. They argue that in contrast to the crowding distance criterion, which is meant to distribute points uniformly along the Pareto frontier, the contributing hypervolume selection is meant to distribute points in a way that maximizes the covered hypervolume. For a typically knee-shaped Pareto frontier, the latter property implies a

Table A1
Operational Constants of the (1 + 1)-CMA-ES Algorithm

Step size control	Covariance matrix adaption
$q = \frac{1}{5.5}$	$c_c = \frac{2}{n+2}$
$c_p = \frac{q}{2+q}$	$c_1 = \frac{2}{n^2+6}$
$d = 1 + \frac{n}{2}$	$p_{\text{thresh}} = 0.44$

higher density of compromise solutions close to the knee of the frontier; that is, there will be more candidate solutions providing fair trade-offs. The contributing hypervolume criterion also proved well in theory (e.g., Berghammer et al., 2010). Igel et al. (2007) adopted the ranking by contributing hypervolume for their $\lambda_{\text{MO}} \times (1,1)$-CMA-ES. They observed superior behavior compared the crowding distance criterion using a couple of benchmark problems.

A3. The $\lambda_{\text{MO}} \times (1,1)$-CMA-ES

Based on our success with our parallelization (Kriest et al., 2017) of the most commonly used single-objective CMA-ES algorithm, the $(\mu/\mu_w, \lambda)$-CMA-ES (Hansen, 2016), we will apply a corresponding parallelized version for multiple objectives, the $\lambda_{\text{MO}} \times (1,1)$-CMA-ES by Igel et al. (2007). It maintains and iteratively updates a population of λ_{MO} normal distributions. It is based on another single-objective CMA-ES variant, the (1 + 1)-CMA-ES. Compared to the $(\mu/\mu_w, \lambda)$-CMA-ES, the (1 + 1)-CMA-ES provides some "simplifications" that are well suited to pass over to MO:

1. Only a single candidate solution x is sampled per iteration.
2. The sample replaces the distribution mean $\bar{x}$ directly if and only if $f(x) < f(\bar{x})$.
3. The covariance matrix is only updated by a so-called *rank-one update*, using an evolution path vector p_c, which cumulates the move of the mean of former iterations (the $(\mu/\mu_w, \lambda)$-CMA-ES additionally uses its μ good samples for the so-called *rank-μ update* of the covariance matrix)
4. Similar to the $(\mu/\mu_w, \lambda)$-CMA-ES, an additional over all scale s is used for step size control but must be adapted in a different way.

An outline of the (1 + 1)-CMA-ES is given below. We suppose the algorithm to operate on the unit cube $[0, 1]^n$ since, prior to evaluating the objective function f, any point x in $[0, 1]^n$ can be uniquely mapped to the actual parameter space by shifting and scaling x with regard to the boundaries of the biogeochemical parameters. Thus, distribution mean $\bar{x}$ and scaling factor s are initialized with $(0.5, \dots, 0.5)^T$ and 0.5, respectively. The operational constants of the main algorithm are defined in Table A1 and supposed to be accessible within calls of the subprocedures "updateS" and "updateC."

Algorithm 1 The (1 + 1)-CMA-ES

```
Set $q$, $c_p$, $d$, $c_c$, $c_1$ according to Table 3
Set $\bar{x} = (\frac{1}{2}, \dots, \frac{1}{2})^T$
Set $p_c = 0$, $C = I$, $s = 0.5$ and $\bar{p} = q$
while stopping criterion is not met do
    Sample $x \in \mathbb{R}^n$ from $\mathcal{N}(\bar{x}, s^2 C)$
    if $f(x) \leq f(\bar{x})$ then $p = 1$ else $p = 0$
    updateS$\big((\bar{x}, C, s, \bar{p}, p_c), p\big)$
    if $f(x) \leq f(\bar{x})$ then
        updateC$\big((\bar{x}, C, s, \bar{p}, p_c), \frac{\bar{x}-x}{s}\big)$
        $\bar{x} \leftarrow x$
    end if
end while
```

Algorithm 2 Procedure updateS$\left((\bar{x}, C, s, \bar{p}, p_c), p\right)$

```
$\bar{p} \leftarrow (1 - c_p)\bar{p} + c_p p$
$s \leftarrow s \cdot \exp\left(\frac{1}{d}\frac{q-\bar{p}}{1-\bar{p}}\right)$
```

Algorithm 3 Procedure updateC$\left((\bar{x}, C, s, \bar{p}, p_c), x_{\text{step}}\right)$

```
if $\bar{p} < p_{\text{thresh}}$ then
    $p_c \leftarrow (1 - c_c)p_c + \sqrt{c_c(2 - c_c)} \cdot x_{\text{step}}$
    $C \leftarrow (1 - c_1)C + c_1 \cdot p_c p_c^T$
else
    $p_c \leftarrow (1 - c_c)p_c$
    $C \leftarrow (1 - c_1)C + c_1 \cdot \left(p_c p_c^T + c_c(2 - c_c)C\right)$
end if
```

The (1 + 1)-CMA-ES maintains its normal distribution $\mathcal{N}(\bar{x}, s^2C)$ using five variables. Three of the variables are the shape-defining ones, namely, the mean vector $\bar{x}$, the covariance matrix C, and the factor s which scales the one-standard-deviation domain of the distribution, controlling the overall step size of the algorithm. Additionally, two auxiliary variables are used for the distribution update: an average sampling success rate $\bar{p}$ and the evolution path of cumulated shifts of the mean, p_c.

Practically, sampling $x \in \mathbb{R}^n$ from the normal distribution $\mathcal{N}(\bar{x}, s^2C)$ in line 5 is realized by drawing n independent samples of the univariate standard normal distribution $\mathcal{N}(0, 1)$ as components of a vector z and calculating

$$x = \bar{x} + sBDz,$$

where the matrices B and D are obtained from an eigendecomposition

$$C = BD^2B^T$$

of C; that is, the columns of B are orthogonal eigenvectors of C, and D is a diagonal matrix of the corresponding eigenvalues.

After sampling a single new solution x, updateS adjusts the average success rate $\bar{p}$ (the rate of samples that are better than the distribution mean) and uses $\bar{p}$ to update the scaling factor s of the distribution. Whether s increases or decreases depends on whether $\bar{p}$ is less or greater than the target success rate q, which is set in accordance to the (1/5)th rule (Rechenberg, 1973). A success rate higher than q is supposed to imply sufficient likelihood that an optimum is within the area of one standard deviation of the current distribution, while lower success rates are assumed to indicate that the optimum lies beyond that area. The constant c_p is a learning rate and controls how fast the success count of earlier samples fades out from $\bar{p}$. The adaption strength of s is attenuated by d.

If the new sample x is better than the mean $\bar{x}$ of the current distribution updateC is invoked to adapt the evolution path p_c and the covariance matrix C before x replaces the distribution mean. Usually, the first case of updateC ($\bar{p} < p_{\text{thresh}} < 0.5$) applies and p_c is adjusted using the scale-normalized step $x_{\text{step}} = \frac{\bar{x}-x}{s}$ and the covariance matrix is updated with the rank-one matrix $p_c p_c^T$. Again, the constants c_p and c_c are learning rates, which control how fast information of earlier iterations fade out from p_c and C, respectively. In the case $\bar{p} \geq p_{\text{thresh}}$ the evolution path shrinks instead of being updated with x_{step}. This approach is suggested by Igel et al. (2007) in order to prevent too fast expansion of the distribution into certain directions when the step size s is too small. The latter might happen in regions of the searchspace where f behaves linear. The shrinking of p_c is compensated in the update of C.

Finally, the $\lambda_{\text{MO}} \times (1, 1)$-CMA-ES hierarchically combines the level of nondominance criterion and the contributing hypervolume criterion described above in order to rank and select λ_{MO} out of $2 \times \lambda_{\text{MO}}$ normal

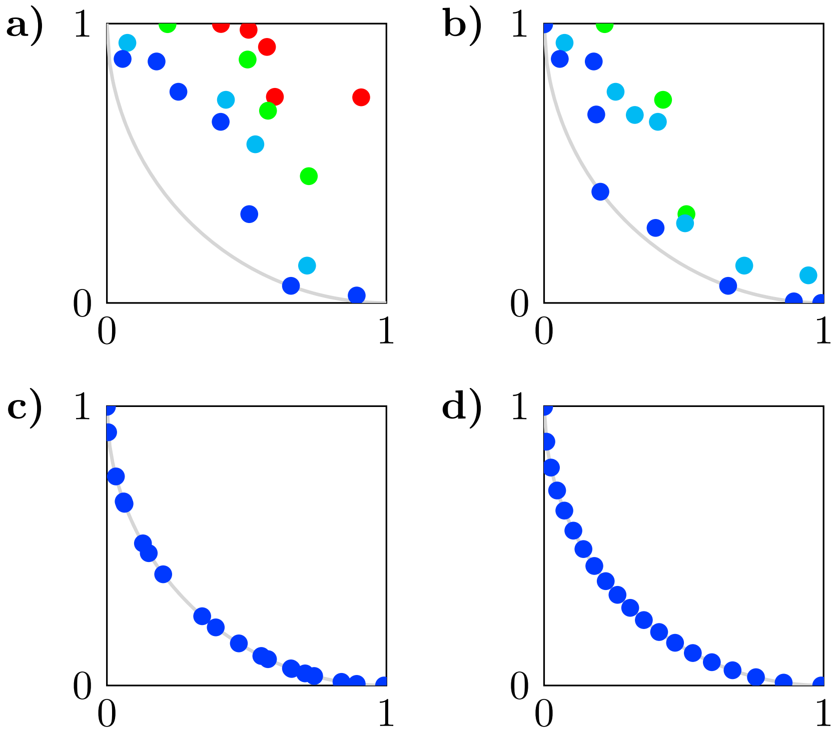

Figure A3. Convergence of the multiobjective CMA-ES for a test case with two objectives. Dots represent the population of candidate solutions in objective space (first objective versus second objective). In the first iterations, there might be more than one level of nondominance (indicated by the different colors in panels a and b). Later, when the set of candidate solutions has become a nondominated set (panels c and d), the selection process only depends on the ranking by contributing hypervolumes, which prevents "lump formations" and allows the algorithm to convergence to a good Pareto approximate set (panel d).

distributions as the next iterations' population. Each distribution itself is operated in a (1 + 1)-CMA-ES fashion. The algorithm outline is summarized below. For a distribution $a = (\bar{p}, \bar{x}, p_c, s, \mathrm{C})$ we write, for example, $a.s$ to refer to a single variable that is attributed with the distribution. Figure A3 illustrates how the interplay of the (1 + 1)-CMA-ES for single solutions and the suggested ranking procedure for the population of solutions enables convergence to well-distributed sets of compromises.

Algorithm 4 The $\lambda_{\mathrm{MO}} \times (1+1)$-CMA-ES

```
Set global constants $d$, $q$, $c_p$, $c_c$, $c_1$ according to Table 3
Initialize population $P = \{a_1, \dots, a_{\lambda_{\mathrm{MO}}}\}$ of normal distributions
for $i = 1, \dots, i_{\max}$ do
    Make a copy $a'_k = a_k$ of each distribution
    for $k = 1, \dots, \lambda_{\mathrm{MO}}$ do
        Sample $x \in \mathbb{R}^n$ from $\mathcal{N}(a_k.\bar{x}, a_k.C)$
        $a'_k.\bar{x} \leftarrow x$
        if $\bar{x} \prec a_k.\bar{x}$ then $p = 1$ else $p = 0$
        updateS$(a_k, p)$
        updateS$(a'_k, p)$
        updateC$\left(a'_k, \frac{a'_k.\bar{x} - \bar{x}}{a_k.s}\right)$
    end for
    $P \leftarrow \{a_1, \dots, a_{\lambda_{\mathrm{MO}}}\} \cup \{a'_1, \dots, a'_{\lambda_{\mathrm{MO}}}\}$
    Trim $P$ (such that $|P| = \lambda_{\mathrm{MO}}$) using ranking-based selection
end for
```

Acknowledgments
This work is a contribution to Kiel Marine Science (KMS), the DFG Cluster of Excellence "The Future Ocean," the DFG-supported project SFB754, and to BMBF joint project PalMod (FKZ 01LP1512A). 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 acknowledge financial support by Land Schleswig-Holstein within the funding program Open Access Publikationsfonds.

References

Battaglia, G., & Joos, F. (2018). Marine N_2O emissions from nitrification and denitrification constrained by modern observations and projected in multimillennial global warming simulations. *Global Biogeochemical Cycles*, *32*, 92–121. https://doi.org/10.1002/2017GB005671

Berghammer, R., Friedrich, T., & Neumann, F. (2010). Set-based multi-objective optimization, indicators, and deteriorative cycles, *Proceedings of the genetic and evolutionary computation conference* (pp. 495–502). Portland, Oregon, USA: July 2010 (GECCO 2010). https://doi.org/10.1145/1830483.1830574

Bopp, L., Resplandy, L., Orr, J. C., Doney, S. C., Dunne, J. P., Gehlen, M., et al. (2013). Multiple stressors of ocean ecosystems in the 21st century: Projections with CMIP5 models. *Biogeosciences*, *10*, 6225–6245. https://doi.org/10.5194/bg-10-6225-2013

Buesseler, K. O., Lamborg, C. H., Boyd, P. W., Lam, P. J., Trull, T. W., Bidigare, R. R., et al. (2007). Revisiting carbon flux through the ocean's twilight zone. *Science*, *316*, 567–570. https://doi.org/10.1126/science.1137959

Cabre, A., Marinov, I., Bernadello, R., & Bianchi, D. (2015). Oxygen minimum zones in the tropical Pacific across CMIP5 models: Mean state differences and climate change trends. *Biogeosciences*, *12*, 5429–5454. https://doi.org/10.5194/bg-12-5429-2015

Carr, M.-E., Friedrichs, M. A. M., Schmeltz, M., Aitac, M. N., Antoine, D., Arrigo, K. R., et al. (2006). A comparison of global estimates of marine primary production from ocean color. *Deep-Sea Research Part II*, *53*, 741–770. https://doi.org/10.1016/j.dsr2.2006.01.028

Ciais, P., Sabine, C., Bala, G., Bopp, L., Brovkin, V., Canadell, J., et al. (2013). Carbon and other biogeochemical cycles. In T. F. Stocker, D. Qin, et al. (Eds.), *Climate change 2013: The physical science basis. Contribution of working group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (pp. 465–570). Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press. https://doi.org/10.1017/CBO9781107415324.015

Cocco, V., Joos, F., Steinacher, M., Frölicher, T. L., Bopp, L., Dunne, J., et al. (2013). Oxygen and indicators of stress for marine life in multi-model global warming projections. *Biogeosciences*, *10*, 1849–1868. https://doi.org/10.5194/bg-10-1849-2013

Coello Coello, C. A., Van Veldhuizen, D. A., & Lamont, G. B. (2002). *Evolutionary algorithms for solving multi-objective problems*. New York, NY: Kluwer Academic/Plenum Publishers.

Deb, K. (2009). Multi-objective optimization using evolutionary algorithms. Reprint of the 2001 hardback ed.

Deb, K., Agarwal, S., Pratap, A., & Meyarivan, T. (2002). A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolultionary Computation*, *6*(2), 182–197. https://doi.org/10.1109/4235.996017

Doney, S., Lindsay, K., Caldeira, K., Campin, J., Drange, H., Dutay, J., et al. (2004). Evaluating global ocean carbon models: The importance of realistic physics. *Global Biogeochemical Cycles*, *18*, GB3017. https://doi.org/10.1029/2003GB002150

Dunne, J. P., Sarmiento, J. L., & Gnanadesikan, A. (2007). A synthesis of global particle export from the surface ocean and cycling through the ocean interior and on the seafloor. *Global Biogeochemical Cycles*, *21*, GB4006. https://doi.org/10.1029/2006GB002907

Dutay, J.-C, Bullister, J. L., Doney, S. C., Orr, J. C., Najjar, R., Caldeira, K., et al. (2002). Evaluation of ocean model ventilation with CFC-11: comparison of 13 global ocean models. *Ocean Modelling*, *4*, 89–120. https://doi.org/10.1016/S1463-5003(01)00013-0

Dutkiewicz, S., Follows, M. J., & Parekh, P. (2005). Interactions of the iron and phosphorous cycles: A three-dimensional model study. *Global Biogeochemical Cycles*, *19*, GB1021. https://doi.org/10.1029/2004GB002342

Efstratiadis, A., & Koutsoyiannis, D. (2010). One decade of multi-objective calibration approaches in hydrological modelling: A review. *Hydrological Sciences Journal*, *55*(1), 58–78. https://doi.org/10.1080/02626660903526292

Emmerich, M., Beume, N., & Naujoks, B. (2005). An EMO algorithm using the hypervolume measure as selection criterion, *Evolutionary multi-criterion optimization. EMO 2005. Lecture Notes in Computer Science* (pp. 62–76). Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-540-31880-4_5

Evans, G. T. (2003). Defining misfit between biogeochemical models and data sets. *Journal of Marine Systems*, *40–41*, 49–54. https://doi.org/10.1016/S0924-7963(03)00012-5

Farias, L., Castro-Gonzalez, M., Cornejo, M., Charpentier, J., Faundez, J., Boontanon, N., & Yoshida, N. (2009). Denitrification and nitrous oxide cycling within the upper oxycline of the eastern tropical South Pacific oxygen minimum zone. *Limnology and Oceanography*, *54*(1), 132–144. https://doi.org/10.4319/lo.2009.54.1.0132

Garcia, H. E., Locarnini, R. A., Boyer, T. P., & Antonov, J. I. (2006a). World Ocean Atlas 2005, Vol. 3: Dissolved oxygen, apparent oxygen utilization, and oxygen saturation. In S. Levitus (Ed.), *NOAA Atlas NESDIS 63* (342 pp.). Washington, DC: U.S. Government Printing Office.

Garcia, H. E., Locarnini, R. A., Boyer, T. P., & Antonov, J. I. (2006b). World Ocean Atlas 2005, Vol. 4: Nutrients (phosphate, nitrate, silicate). In S. Levitus (Ed.), *NOAA Atlas NESDIS 64* (396 pp.). Washington, DC: U.S. Government Printing Office

Graven, H. D., Gruber, N., Key, R., Khatiwala, S., & Gireaud, X. (2012). Changing controls on oceanic radiocarbon: New insights on shallow-to-deep ocean exchange and anthropogenic CO_2 uptake. *Journal of Geophysical Research*, *117*, C10005. https://doi.org/10.1029/2012JC008074

Hansen, N. (2006). The CMA evolution strategy: A comparing review. In J. A. Lozano, P. Larranaga, et al. (Eds.), *Towards a new evolutionary computation: Advances on estimation of distribution algorithms* (pp. 75–102). Heidelberg: Springer.

Hansen, N. (2016). The CMA evolution strategy: A tutorial. arXiv:1604.00772v1 [cs.LG].

Hansen, N., Auger, A., Ros, R., Finck, S., & Posík, P. (2010). Comparing results of 31 algorithms from the black-box optimization Benchmarking BBOB-2009, *Proceedings of the genetic and evolutionary computation conference, GECCO 2010*. Portland, Oregon, USA, july 7-11, 2010, companion material, pp. 1689–1696. ACM.

Hauschild, M., & Pelikan, M. (2011). An introduction and survey of estimation of distribution algorithms. Retrieved from http://medal-lab.org/files/2011004_rev1.pdf

Honjo, S., Manganini, S. J., Krishfield, R. A., & Francois, R. (2008). Particulate organic carbon fluxes to the ocean interior and factors controlling the biological pump: A synthesis of global sediment trap programs since 1983. *Progress in Oceanography*, *76*, 217–285. https://doi.org/10.1016/j.pocean.2007.11.003

Hoving, H. J. T, & Robison, B. H. (2012). Vampire squid: Detritivores in the oxygen minimum zone. In *Proceedings of the Royal Society B* (pp. 279). London. https://doi.org/10.1098/rspb.2012.1357

Igel, C., Hansen, N., & Roth, S. (2007). Covariance matrix adaptation for multi-objective optimization. *Evolutionary Computation*, *15*(1), 1–28.

Jeballa, M., Auger, A., & Hansen, N. (2011). Log-linear convergence and divergence of the scale-invariant (1+1)-ES in noisy environments. *Algorithmica*, *59*(3), 425–460. https://doi.org/10.1007/s00453-010-9403-3

Khatiwala, S. (2007). A computational framework for simulation of biogeochemical tracers in the ocean. *Global Biogeochemical Cycles*, *21*, GB3001. https://doi.org/10.1029/2007GB002923

Kriest, I. (2017). Calibration of a simple and a complex model of global marine biogeochemistry. *Biogeosciences*, *14*, 4965–4984.

Kriest, I., Khatiwala, S., & Oschlies, A. (2010). Towards an assessment of simple global marine biogeochemical models of different complexity. *Progress in Oceanography*, *86*, 337–360.

Kriest, I., & Oschlies, A. (2013). Swept under the carpet: Organic matter burial decreases global ocean biogeochemical model sensitivity to remineralization length scale. *Biogeosciences*, *10*, 8401–8422.

Kriest, I., & Oschlies, A. (2015). 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. https://doi.org/10.5194/gmd-8-2929-2015

Kriest, I., Sauerland, V., Khatiwala, S., Srivastav, A., & Oschlies, A. (2017). Calibrating a global three-dimensional biogeochemical ocean model (MOPS-1.0). *Geoscientific Model Development*, *10*(1), 127–154. https://doi.org/10.5194/gmd-10-127-2017

Langenbrunner, B., & Neelin, J. D. (2017a). Multiobjective constraints for climate model parameter choices: Pragmatic Pareto fronts in CESM1. *Journal of Advances in Modelling Earth Systems*, *9*, 2008–2026. https://doi.org/10.1002/2017MS000942

Langenbrunner, B., & Neelin, J. D. (2017b). Pareto-optimal estimates of California precipitation change. *Geophysical Research Letters*, *44*, 12,436–12,446. https://doi.org/10.1002/2017GL075226

Laufkötter, C., Vogt, M., & Gruber, N. (2013). Long-term trends in ocean plankton production and particle export between 1960–2006. *Biogeosciences*, *10*(11), 7373–7393. https://doi.org/10.5194/bg-10-7373-2013

Laufkötter, C., Vogt, M., Gruber, N., Aumont, O., Bopp, L., Doney, S. C., et al. (2016). Projected decreases in future marine export production: The role of the carbon flux through the upper ocean ecosystem. *Biogeosciences*, *13*(13), 4023–4047. https://doi.org/10.5194/bg-13-4023-2016

Lutz, M. J., Caldeira, K., Dunbar, R. B., & Behrenfeld, M. J. (2007). Seasonal rhythms of net primary production and particulate organic carbon flux to depth describe biological pump efficiency in the global ocean. *Journal of Geophysical Research*, *113*, C10011. https://doi.org/10.1029/2006JC003706

Marshall, J., Adcroft, A., Hill, C., Perelman, L., & Heisey, C. (1997). A finite-volume, incompressible Navier-Stokes model for studies of the ocean on parallel computers. *Journal of Geophysical Research*, *102*, 5733–5752. https://doi.org/10.1029/96JC02775

Martin, J. H., Knauer, G. A., Karl, D. M., & Broenkow, W. W. (1987). VERTEX: Carbon cycling in the Northeast Pacific. *Deep-Sea Research Part A*, *34*(2), 267–285. https://doi.org/10.1016/0198-0149(87)90086-0

Matear, R. J., & Holloway, G. (1995). Modeling the inorganic phosphorous cycle of the North Pacific using an adjoint data assimilation model to assess the role AOF dissolved organic phosphorous. *Global Biogeochemical Cycles*, *9*(1), 101–119. https://doi.org/10.1029/94GB03104

Matsumoto, K., Sarmiento, J. L., Key, R. M., Aumont, O., Bullister, J. L., Caldeira, K., et al. (2004). Evaluation of ocean carbon cycle models with data-based metrics. *Geophysical Research Letters*, *31*, L07303. https://doi.org/10.1029/2003GL018970

Moore, J. K., & Doney, S. C. (2007). Iron availability limits the ocean nitrogen inventory stabilizing feedbacks between marine denitrification and nitrogen fixation. *Global Biogeochemical Cycles*, *21*, GB2001. https://doi.org/10.1029/2006GB002762

Mostafaie, A., Forootan, E., Safari, A., & Schumacher, M. (2018). Comparing multi-objective optimization techniques to calibrate a conceptual hydrological model using in situ runoff and daily GRACE data. *Computational Geosciences*, *22*, 789–814. https://doi.org/10.1007/s10596-018-9726

Paulmier, A., Kriest, I., & Oschlies, A. (2009). Stoichiometries of remineralisation and denitrification in global biogeochemical ocean model. *Biogeosciences*, *6*, 923–935. https://doi.org/10.5194/bg-6-923-2009

Price, A. R., Myerscough, R. J., Voutchkov, I. I., Marsh, R., & Cox, S. J. (2009). Multi-objective optimization of GENIE Earth system models. *Philososphical Transactions of the Royal Society*, *367*(1898), 2623–2633. https://doi.org/10.1098/rsta.2009.0039

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

Orr, J. C., Najjar, R., Sabine, C. L., & Joos, F. (2000). Abiotic- howto. internal OCMIP report (*Tech. Rep. Nos. revision: 1.16*): Saclay, Gif-sur-Yvette, France: LSCE/CEA. Retrieved from ocmip5.ipsl.jussieu.fr/OCMIP/phase2/simulations/Abiotic/HOWTO-Abiotic.html (lastaccess:28November2013), 2000.

Rechenberg, I. (1973). *Evolutionsstrategie—Optimierung Technischer Systeme Nach Prinzipien Der Biologischen Evolution*. Stuttgart, Germany: Frommann-Holzboog.

Sauerland, V. (2018). calibrate2O: A parallel MO-CMAES implementation for the optimisation of computational demanding functions w.r.t. two objectives. https://doi.org/10.5281/zenodo.1432629

Sauerland, V., Loeptien, U., Leonhard, C., Oschlies, A., & Srivastav, A. (2018). Error assessment of biogeochemical models by lower bound methods (NOMMA-1.0). *Geoscientific Model Development*, *11*(3), 1181–1198. https://doi.org/10.5194/gmd-11-1181-2018

Schartau, M., Oschlies, A., & Willebrand, J. (2001). Parameter estimates of a zero-dimensional ecosystem model applying the adjoint method. *Deep Sea Research Part II: Topical Studies in Oceanography*, *48*(8–9), 1769–1800. https://doi.org/10.1016/S0967-0645(00)00161-2

Schartau, M., Wallhead, P., Hemmings, J., Löptien, U., Kriest, I., Krishna, S., et al. (2017). Reviews and syntheses: Parameter identification in marine planktonic ecosystem modelling. *Biogeosciences*, *14*, 1647–1701. https://doi.org/10.5194/bg-14-1647-2017

Segschneider, J., & Bendtsen, J. (2013). Temperature-dependent remineralization in a warming ocean increases surface pCO_2 through changes in marine ecosystem composition. *Global Biogeochemical Cycles*, *27*, 1214–1225. https://doi.org/10.1002/2013GB004684

Seibel, B. A. (2011). Critical oxygen levels and metabolic suppression in oceanic oxygen minimum zones. *Journal of Experimental Biology*, *214*, 326–336.

Stramma, L., Oschlies, A., & Schmidtko, S. (2012). Mismatch between observed and modeled trends in dissolved upper-ocean oxygen over the last 50 yr. *Biogeosciences*, *9*, 4045–4057. https://doi.org/10.5194/bg-9-4045-2012

Stramma, L., Prince, E. D., Schmidtko, S., Luo, J., Hoolihan, J. P., Visbeck, M., et al. (2012). Expansion of oxygen minimum zones may reduce available habitat for tropical pelagic fishes. *Nature Climate Change*, *2*(1), 33–37.

Urban, N. M., & Fricker, T. E. (2010). A comparison of Latin hypercube and grid ensemble designs for the multivariate emulation of an Earth system model. *Computers & Geosciences*, *36*, 746–755. https://doi.org/10.1016/j.cageo.2009.11.004

Van Mooy, B. A. S., Keil, R. G., & Devol, A. H. (2002). Impact of suboxia on sinking particulate organic carbon: Enhanced carbon flux and preferential degradation of amino acids via denitrificiation. *Geochimica et Cosmochimica Acta*, *66*, 457–465. https://doi.org/10.1016/S0016-7037(01)00787-6

Wallmann, K. (2010). Phosphorus imbalance in the global ocean? *Global Biogeochemical Cycles*, *24*, GB4030. https://doi.org/10.1029/2009GB003643

Ward, B. A., Friedrichs, M. A. M., Anderson, T. R., & Oschlies, A. (2010). Parameter optimisation techniques and the problem of underdetermination in marine biogeochemical models. *Journal of Marine Systems*, *81*, 34–43.

Yapo, O., Gupta, H. V., & Sorooshian, S. (1998). Multi-objective global optimization for hydrologic models. *Journal of Hydrology*, *204*(1–4), 83–97. https://doi.org/10.1016/S0022-1694(97)00107-8

Zitzler, E., & Thiele, L. (1998). Multiobjective optimization using evolutionary algorithms—A comparative case study. In A. E. Eiben, T. Bäck, M. Schoenauer, & H.-P. Schwefel (Eds.), *Proceedings of the 5th international conference on parallel problem solving from nature, Amsterdam, the Netherlands, September 27–30, 1998 (PPSN 1998)* (pp. 292–304), Lecture Notes in Computer Science. Berlin: Springer. https://doi.org/10.1007/BFb0056872