Reply to Referee 1

Once again, we thank referee 1 for the helpful and detailed comments. Below is our point-by-point reply to the reviewers suggestions (reviewers comments in blue italics).

General Comments:

The authors significantly improved the Introduction and the Methods part, although some of the newly added text are not entirely clear (see below, detailed comments). My concern about the applicability of the proposed approach is not resolved. The authors significantly revised the relevant section 3.2, improving its readability, but preserved the previously presented example, which is too synthetic in terms of its simplicity and data density, failing therefore to make a convincing case (see below, the specific comments 1-2). Finally, my experience with the software was not entirely convincing either (specific comment 3). I still can not support publication of this work in GMD in its current form. I recommend the authors either to make a convincing case of the applicability of their approach with a non-synthetic data set with a more realistic data density, and displaying a realistic level of complexity. If this would be really beyond the scope of the intended work, authors may want to consider a more theoretically oriented journal specializing on statistical optimization.

Specific Comments

1) Applicability of the approach in terms of complexity The authors argue in P.17,L.3 that application of their approach to more complex systems is 'beyond the scope of this contribution'. I do not agree. For demonstrating that the approach is not only of theoretical interest, but intended to be of practical relevance to the target audience of GMD, an application to a realistically complex data set is needed.

Thank you for insiting on that point. We now demonstrate our approach with global satellite observations of chlorophyll-a (Subsection 3.2.2). We feel that the global example might indeed help and convince other researchers to join our thoughts and to size on our approach for their own model applications.

2) Applicability of the approach in terms of data density As an issue that seem to have emerging anew, the authors warn that their method require dense observations. Again, in reality, such dense observations (for the systems the authors claim their method is suitable for) almost never exist. So the applicability of the approach to a realistic observation set (e.g., with monthly or bi-weekly observations) is needed, with a thorough elaboration of the specific problems, and proposed solutions.

We agree with the reviewer and apologise for being imprecise: Now that we added the new example treating noise in space instead of time (Subsection 3.2.2), the original statement can be generalised in that we need observations which are either dense in time or space. Thus one might either apply the presented method on long term observing stations (like e.g. BATS) or satellite data. We also stress this aspect in our conclusions (page 21, line 20-21), now.

3) Software and its documentation I downloaded and installed the code (NOMMA-1.0) and the prerequisite CPLEX software (IBM ILOG CPLEX Optimization Studio V12.7.0 for Linux), which went all fine by following the instructions in README files. The compiled 'regEx' binaries in both 'regression' and 'regressionCPX' folders seemed to run successfully, producing some .txt output. However, documentation of the software beyond installation seems to be entirely missing. In the absence of such documentation, my possibly incorrect understanding is that, the software is currently hardwired only to operate on a given data set, run a predefined set of tests with predefined parameters (such as the definition of number of extremes). If the code is provided for potential users, some documentation should describe how basic user options can be specified, or if the present code is provided only for exemplifying, an interpretation of examples should be provided.

We agree that a software manual must be provided. We added one to our GIT repository, now, decribing what the basic example is about and how the implemented methods can be used for own data and applications.

Detailed Comments:

P5, L7-8. I don't see the logical link of this sentence to the previous text suggested by 'Therefore'. – We removed 'Therefore' as a reason for our idea follows in the next sentence.

P5, L17: 'intuitively clear': not to me really. – We removed 'intuitively clear'.

Reply to Referee 2

We are grateful to referee 2 for his report. Below is our point-by-point reply to the reviewers suggestions (reviewers comments in blue italics).

In their manuscript the authors address the important issue of an effective error assessment in biogeochemical modeling. The reliability of model simulations and future projections strongly depend on the models ability to reproduce observational data as good as possible. Therefore it is necessary to apply effective concepts of model parameter calibration techniques. Notably, a technique is required which ensures to end up in the global minimum of the parameter space when minimizing the root mean square error (RMSE).

In the present study the authors have developed the new concept of a "non-parametric optimization" which allows to calculate a lower bound of the RMSE by introducing a non-parametric function which fits the assemblage of observational data at best. This concept of a surrogate formulation of the optimization problem was successfully tested on synthetic and observed data (chlorophyll) from the Bornholm Basin in the Baltic Sea.

This is a valuable study addressing a new method of model parameter calibration. However, although the guidance by referee 1 led to a considerable improvement of the manuscript the Methods chapter 2 it is still hard to read for non specialists. As proposed by referee 1 I would like to see a schematic graphic illustration of the whole procedure for better understanding. I also miss a short paragraph addressing the applicability of the method to fully 3D ocean general circulation models.

We agree and are also convinced now, that a global example helps to convince other researchers to size on our approach for their own applications. We demonstrate a global example in a new Subsection 3.2.2.

One minor technical remark: At page 4 line 19 the brackets in the equation are missing. – Changed correspondingly.

Error assessment of biogeochemical models by lower bound methods (NOMMA-1.0)

Volkmar Sauerland[1], Ulrike Löptien[2,3], Claudine Leonhard[1], Andreas Oschlies[2], and Anand Srivastav[1]

[1]Department of Mathematics, Christian-Albrechts-University of Kiel, Christian-Albrechts-Platz 4, 24118 Kiel, Germany
[2]GEOMAR Helmholtz Centre for Ocean Research Kiel, Düsternbrooker Weg 20, 24105 Kiel, Germany
[3]Institute of Geosciences, Christian-Albrechts-University of Kiel, Ludewig-Meyn-Strasse 10, 24118 Kiel, Germany

Correspondence to: Volkmar Sauerland (vsa@informatik.uni-kiel.de)

Abstract. Biogeochemical models, capturing the major feedbacks of the pelagic ecosystem of the world ocean, are today often embedded into Earth System models which are increasingly used for decision making regarding climate policies. These models contain poorly constrained parameters (e.g., maximum phytoplankton growth rate) which are typically adjusted until the model shows a reasonable behavior. Systematic approaches determine these parameters by minimizing the misfit between the model and observational data. In most common model approaches, however, the underlying functions mimicking the biogeochemical processes are non-linear and non-convex. Thus, systematic optimization algorithms are likely to get trapped in local minima and might lead to non-optimal results. To judge the quality of an obtained parameter estimate, we propose to determine a preferably large lower bound for the global optimum, that is relatively easy to obtain and that will help to assess the quality of an optimum, generated by an optimization algorithm. Due to the unavoidable noise component in all observations, such a lower bound is typically larger than zero. We suggest to derive such lower bounds based on typical properties of biogeochemical models (e.g., a limited number of extremes and a bounded time-derivative). We ~~evaluate this approach with synthetic observations and demonstrate a real-world example that consists of phytoplankton observations in~~ illustrate the applicability of the method based on two real word examples. The first example uses real world observations of the Baltic Sea ~~, proving that an exemplary model of NPZD type is quite well-calibrated.~~ in a box-model setup. The second example considers a 3 dimensional coupled ocean circulation model in combination with satellite chlorophyll-a.

1 Introduction

Earth system models are widely used to assess the consequences of climate change and explore climate engineering options (e.g., Brovkin et al., 2009; Keller et al., 2014; Mengis et al., 2015; Cao and Caldeira, 2008, 2010, and many more to follow). In order to capture the development of climate relevant greenhouse gases such as CO_2 and N_2O a pelagic biogeochemical component, embedded into a numerical ocean model, is essential.

In contrast to ocean physics, which is derived from first principles, current biogeochemical modules are based on empirical relationships. Thus, several studies compare models of different complexities (e.g., Friedrichs et al., 2006). Still there is no consensus yet which complexity is needed to capture the major processes and how exactly the model should be formulated (e.g., Anderson, 2005; Löptien, 2011). Today, various model formulations exist. Popular examples are the BLING model

with 4 prognostic variables only (Galbraith et al., 2010) versus the PICES model, containing 24 prognostic variables (Aumont et al., 2015). Another related major problem, beside the model complexity, is generally the multitude of poorly known model parameters which exert crucial control on the model behavior (e.g., Kriest et al., 2010). To assess and compare the quality of the different model formulations, it is crucial to chose these parameters such that the fit to observations is as good as possible. Due

5   to the large computational expenses of 3-dimensional coupled biogeochemical ocean models, it is common practice that few parameters are adjusted "by hand" until the model shows a "reasonable" agreement with some observations. More advanced approaches use automatized optimization techniques to adjust the model parameters. These techniques require an objective metric (e.g., Evans, 2003) that measures the model-data misfit and can be minimized automatically. Due to computational limitations, various studies estimate the model parameters at singular stations and adopt these for the full model (e.g. Kane et al.,

10   2011; Kaufman et al., 2017; Matear, 1995; Schartau and Oschlies, 2003). For this approach, it can be problematic to determine one parameter set for several sites (e.g., Kidston et al., 2011). Thus, other studies use fast approximations (Kennedy et al., 2006; Khatiwala, 2007) to optimize full 3-dimensional models. A drawback is, that these approaches are generally restricted to the estimation of few parameters only (e.g., Mattern et al., 2012; Kriest et al., 2017; Prieß et al., 2013a, b; Piwonski and Slawig, 2016; Rückelt et al., 2010). In addition, limited data availability (e.g., Lawson et al., 1996) and deficient processes

15   in the ocean circulation model (e.g., Dietze and Löptien, 2013) cumber the optimization process. In summary, the systematic optimization of a 3D coupled biogeochemical ocean model remains a difficult task and requires an ongoing advancement of the existing methods (Schartau et al., 2017).

Biogeochemical processes are non-linear, non-convex and complexly entangled. Therefore, as stressed by several studies, associated model-data misfit measures comprise an unknown number of local optima and admit no proof whether an obtained

20   parameter set is globally optimal or not (e.g., Faugeras et al., 2003; Hurtt and Armstrong, 1996). Many parameter optimization studies invoke deterministic methods that use gradient information about the objective function (the model-data misfit measure) to iteratively approach a locally optimal set of parameters in an efficient way, starting from some initial guess. Most of these studies calculate gradient information by the adjoint method (introduced for biogeochemical models by Lawson et al. (1995) since it is efficient if there are more parameters than model states) and use the gradient to determine a direction and an efficient

25   stepsize to change the parameters, often by applying a quasi-Newton method (e.g., Fennel et al., 2001; Friedrichs, 2001, 2002; Spitz et al., 1998; Tjiputra et al., 2007; Xiao and Friedrichs, 2014). Other attempts focus on stochastic search algorithms which rely on random decisions. Examples for stochastic search algorithms that have been applied to optimize parameters of biogeochemical models are simulated annealing (e.g., Hurtt and Armstrong, 1996, 1999; Matear, 1995; Kidston et al., 2011), genetic algorithms (e.g., Hemmings and Challenor, 2012; Kaufman et al., 2017; Schartau and Oschlies, 2003) and estimation

30   of distribution algorithms (Kriest et al., 2017). Vallino (2000) compares the performance of a couple of optimization algorithms of both types, tuning the parameters of an ecosystem model against mesocosm data. Stochastic search algorithms require more model simulations (computation time) than gradient-based methods to converge but are less likely to get trapped in a "first available" local optimum (cf. Vallino, 2000), which might possibly be far off the global optimum. On the other hand, several contributions which focus on gradient-based methods aim to increase confidence in the quality of an obtained parameter set

35   by repeating the optimization procedure many times (20–600), while using various random starting points (e.g., Garcia-Gorriz

et al., 2003; Hemmings et al., 2004; Schartau et al., 2001). Also this approach increases the number of required simulations considerably.

Still, it is crucial to find a global optimum to assess the quality of a certain model formulation. Lacking a proof about the global optimality of chosen parameters it is difficult to determine whether a model-data misfit is mainly caused by the parameter choice or attributed to other sources of uncertainty, like those concerning model equations or observational data (cf., e.g., Faugeras et al., 2003; Spitz et al., 1998; Schartau et al., 2001). Facing this situation, we have a strong interest to estimate the deviation of a model-data misfit for a given parameter set relative to the unknown global optimum. As the minimal accomplishable model-data misfit (i.e., the global optimum) is unknown, a good (i.e., preferably large) and easy to obtain lower bound on that value would help to judge the quality of a minimum obtained by an automated optimization algorithm. Provided that such a lower bound is close to the obtained model-data misfit, a continuation of the parameter optimization process would not be necessary. In the present study, we introduce an approach to determine such lower bounds. We suggest to consider a surrogate formulation that is easier to solve and determine the global optimum based on this "relaxed" problem. Our approach is based on certain properties of typical biogeochemical models which are likewise fulfilled by non-parametric functions. We propose to search for the best fit to the observations among these functions – which is a much easier and faster optimization problem than minimizing the model-data misfit based on the full biogeochemical model. Optimizing these non-parametric functions provides the desired bounds on the lowest possible misfit of the actual model, since the properties we choose to constrain the generalized optimization problems are satisfied by each solution of the original problem.

The following section focuses on some typical properties of biogeochemical models which lead to the intended relaxed problems. The choice of model properties is also owed to the fact that efficient tailored algorithms to solve the associated problem relaxations are easily available. In Section 3, we examine the proposed method with regard to both characteristics of observational data: their noise-level and coverage. For this purpose, we generate synthetic observations by adding random Gaussian noise to samples of a parameterized exemplary model trajectory. In the next step, we compare the results with our lower bound approaches, i.e., with the global optima of the corresponding easier optimization problems. We systematically examine the relation between both values in dependence of sparseness of the observational data and noise level. We further consider ~~a~~ two real-world applications. The first application is a based on a box model and considers a common NPZD-type biogeochemical model in combination with ~~real-world~~ phytoplankton observations in the Baltic Sea. Our second example is based on global satellite observations of chlorophyll-a and a coupled biogeochemical ocean general circulation model.

2   Methods

Comparing model output to observational data requires a criterion to measure the misfit between both data sets. To apply an automated optimization algorithm, such a measure needs to be reduced to a single real number. We give commonly used measures in the following Subsection. In Subsection 2.2, we introduce a mathematical notation for the optimization problem based on the given measure for the model-data misfit. Additionally, we provide a mathematical formulation for the (simplified) non-parametric approach. We then give specifications of the non-parametric data-fit problem based on frequency limits on the

parameterized models (Subsection 2.3 and Subsection 2.4), bounds on their derivatives (Subsection 2.5), and the combination of both (Subsection 2.6). These non-parametric relaxations will be used to calculate lower misfit bounds as outlined above.

2.1 Model-data misfit

A quality assessment of biogeochemical models usually compares available observational data $\boldsymbol{o} = (o_1, \ldots, o_N)$ with corre-

5 sponding model output (model predictions) $\boldsymbol{p} = (p_1, \ldots, p_N)$.

For the sake of simplicity, we will consider scalar data in the following, assuming that both $\boldsymbol{o}$ and $\boldsymbol{p}$ are uni-variate ~~time series. Clearly, with~~ . Actually, comprehensive global ocean models and observational data sets ~~, both time series consist of vectors, comprising~~ both comprise multiple quantities of interest on spatial grids. The presented lower bound methods can be ~~transferred to~~ utilized for that multi-variate case by applying them ~~to~~ chunk-wise and for each quantity and ~~each grid box and~~

10 summing up the obtained results, optionally using weights for the single terms.

Objective judgment about the differences between observational data and model output requires an associated measure f_{err} that assigns a real number to the model-data misfit. Such an objective model-data misfit measure f_{err} further has the advantage that it allows to apply mathematical optimization algorithms to parametric models, where otherwise only manual parameter tuning can be done until the model output shows a "reasonable" behavior.

15 There are several possible measures for the model-data misfit that have been used to evaluate biogeochemical models (see, e.g., Evans, 2003; Gregg et al., 2009; Stow et al., 2009). Common measures are the mean absolute error (MAE)

$$f_{\mathrm{mae}}(\boldsymbol{p}, \boldsymbol{o}) = \frac{1}{N} \sum_{i=1}^{N} |p_i - o_i|$$

and the root mean squared error (RMSE)

$$f_{\mathrm{rmse}}(\boldsymbol{p}, \boldsymbol{o}) = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (p_i - o_i)^2}.$$

20 It is sufficient to consider the following expression (sum of squared errors)

$$\sum_{i=1}^{N} (p_i - o_i)^2$$

instead of RMSE as this transformation does not change the ranking on considered model outputs p. We will exemplarily work with RMSE which is the most commonly used misfit measure for biogeochemical models. However, our approach and the corresponding algorithms are transferable to other misfit measures like MAE.

25 2.2 The optimization problem

As mentioned above, we consider scalar observations $\boldsymbol{o} = (o_1, \ldots, o_N)$ taken at times $t_1 < t_2 < \cdots < t_N$. Moreover, we introduce a scalar parametric model function $\varphi : S \times \mathbb{R} \to \mathbb{R}$, where the set $S \subseteq \mathbb{R}^n$ is the domain of the free parameters of the model. For a given parameter vector $\boldsymbol{s} \in S$, the model prediction $\boldsymbol{p}(\boldsymbol{s}) = (p(\boldsymbol{s})_1, \ldots, p(\boldsymbol{s})_N)$ at times $t_1 < t_2 < \cdots < t_N$

is given by $p(\boldsymbol{s})_i = \varphi(\boldsymbol{s}; t_i)$. So, in order to determine optimal model parameters, we want to minimize the model-data misfit measure

$$\min \quad \sum_{i=1}^{N} (\varphi(\boldsymbol{s}; t_i) - o_i)^2, \tag{1}$$
$$\text{s.t.} \quad \boldsymbol{s} \in S,$$

that is, we want to determine the minimal sum of squared errors over all possible parameter values. As discussed in the introduction, for global biogeochemical ocean models a full scan of the parameter space is hampered by computationally expensive models that would have to be evaluated several times for differing parameter sets during the optimization. Moreover, we usually neither know if selected parameters $\boldsymbol{s} \in S$ correspond to a global optimum of the associated data-fit problem (1) nor how good (or bad) these parameters are in relation to a global optimum of (1). ~~Therefore, the~~ Our idea is to find a value as large as possible which we know to be smaller than the minimum of the optimization problem (1), i.e., a lower bound. Then, if the minimum obtained by the optimization is close to this value, we may terminate the procedure. In mathematical terms we seek a number $\alpha \in \mathbb{R}_{>0}$ which is as large as possible while satisfying

$$\alpha \leq \sum_{i=1}^{N} (\varphi(\boldsymbol{s}; t_i) - o_i)^2 \quad \text{for all } \boldsymbol{s} \in S. \tag{2}$$

Now, if α satisfies Inequality (2) and it holds for some model parameters $\boldsymbol{s} \in S$ that the corresponding model-data misfit is close to α, then $\boldsymbol{s}$ is a good parameter set with respect to the observational data (as well as α is a good lower bound on the unknown optimal model-data misfit).

In order to find such a lower bound α, our approach is to replace the parametric optimization problem (1) by a formulation that can be solved more easily. Therefore, we specify a number of properties of the original model that hold for all parameter sets $\boldsymbol{s} \in S$ and which we require the alternative formulation to fulfill. ~~It is intuitively clear that the~~ The global optimal value of such a relaxed problem is a lower bound α on the best possible model-data misfit of the original model. If the relaxed problem is convex, in contrast to the original optimization task, its global optimum can be calculated efficiently (cf., e.g., Boyd and Vandenberghe, 2004). We also refer to the information box below. Mathematically, our relaxations are modifications of the original optimization task (1) in the sense that the parametric model function φ is replaced by a non-parametric function Φ from a class $\mathcal{F}$ of all functions that satisfy the considered property. In particular, $\mathcal{F}$ contains $\varphi(\boldsymbol{s}; \cdot)$ for all $\boldsymbol{s}$ from the parameter domain S of the actual model. The associated non-parametric optimization problem on the "extended search space" reads

$$\min \quad \sum_{i=1}^{N} (\Phi(t_i) - o_i)^2, \tag{3}$$
$$\text{s.t.} \quad \Phi \in \mathcal{F}.$$

The model-data misfit of a global optimum of the relaxed problem (3) satisfies Inequality (2), meaning that it is a lower bound on the model-data misfit for all allowed parameters $\boldsymbol{s}$ of the original problem (1). We refer to Section 3 for thoughts on how the lower bound is employed in applications to judge the quality of the optimization outcome. In short, the main idea of the lower bound method can be summarized as

- pick some properties that the model comprises for all parameters $s \in S$

- solve the optimization problem detached from the parametric model. Precisely, we minimize the sum of squared errors over all functions $\Phi \in \mathcal{F}$ that fulfill the selected properties.

- The procedure yields a lower bound for the original optimization problem as the set of possible solutions is larger for the relaxed problem and contains the original model output.

In the following sections, we give examples on the properties of the model that we choose.

Terms and background information

Given a function $f : \mathbb{R}^n \to \mathbb{R}$ and a subset X of $\mathbb{R}^n$, a general mathematical optimization problem is

(MP) minimize $f(\boldsymbol{x})$,

  subject to $\boldsymbol{x} \in X$.

An example is the parameter optimization problem (1).

Convex optimization problem. If f is a convex function and X is a convex set, then (MP) is a called a convex optimization problem (CP). All relaxed problem formulations considered below are CPs. An important property of a CP is, that every local optimum of f over X is already a global optimum (cf., e.g., Boyd and Vandenberghe, 2004).

Quadratic program. A special CP is a convex quadratic program (QP). A QP has a convex quadratic objective function f and its function domain X is described in terms of some k linear constraints, i.e.,

$$X = \left\{ \boldsymbol{x} \in \mathbb{R}^n \,|\, g_i(\boldsymbol{x}) \geq 0 \text{ for } i = 1, \ldots, k \right\},$$

where g_i, $i = 1, \ldots, k$, are linear functions. The surrogate model formulations (4)–(7) are QPs. Tools to calculate global optima of arbitrary QPs exist, but for most of our surrogates we can apply tailored algorithms which are more efficient.

2.3 Bounds for monotonic models

We start with an example that is not directly related to biogeochemical models but which serves as a basis for the approaches in Subsection 2.4 and Subsection 2.6, respectively. The task here is to fit ~~a time series of~~ observations by a monotonically

10 increasing data set. Measuring the model-data misfit by its sum squared error, the associated non-parametric optimization problem can, e.g., be stated as a convex quadratic program as follows

$$\min \ \sum_{i=1}^{N} (p_i - o_i)^2,$$

s.t. $\boldsymbol{p} \in \mathbb{R}^N,$

  $p_i \leq p_{i+1}$ for $i \in \{1, \ldots, N-1\}.$

(4)
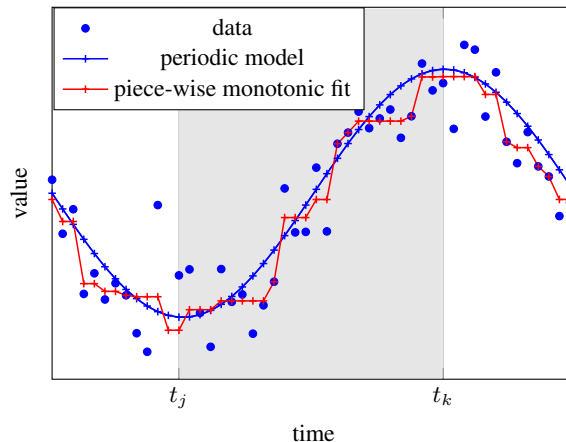
Figure 1. Synthetic data (blue dots) and corresponding output (blue crosses) of a periodic model function (blue curve). As the model frequencies are low, both the model function and its samples take only two local extremes. The segments before, after, and between the extremes (times t_j and t_k) are monotonically decreasing/increasing. The respective monotonic fits to the data (drawn in red) are therefore "better" than the model output.

This yields a vector $p \in \mathbb{R}^N$ with monotonically increasing entries, where p_i is the data point that corresponds to time t_i. These entries are selected such that the sum of the squared deviations from the observations is minimized. Note, that (4) corresponds to the general non-parametric optimization problem (3) if $\mathcal{F}$ is the class of all monotonically increasing functions. If we want to work with monotonically decreasing functions instead, we just need to replace "$\leq$" with "$\geq$" in the monotonicity constraints or we can apply (4) to $-o$ instead of o and negate the ~~resulting time series~~result.

The optimization problem (4) can be solved efficiently. The pool adjacent violators (PAV) algorithm (Barlow et al., 1972) solves it with linear effort, i.e. in less than $c \cdot N$ computer operations for some constant c and all N. Another possibility is to use a general optimization tool for convex quadratic programs like, e.g., CPLEX or Matlab quadprog. Clearly, solutions of (4) provide a lower bound on the optimal model-data misfit for every parametric model that is monotonically increasing.

2.4 Bounds for periodic models

When simulating periodic systems, the model might (intentionally or un-intentionally) not resolve all frequencies that occur in the corresponding observational data. Models that resolve low frequencies with respect to data frequency (e.g., NPZD models that aim to capture the main characteristics of an annual cycle) take a correspondingly limited number of extreme values within a given time interval, e.g., a seasonal cycle. This situation is sketched in Figure 1. The fact that each segment between two subsequent extreme values is monotonically increasing/decreasing, allows to apply the methods introduced in Subsection 2.3. A corresponding series $p_1, \ldots, p_N$ of discrete samples has (at most) the same number of local extremes as the model. For illustration, suppose that the series has exactly two extreme values p_j and p_k with $j < k \in [N]$ as sketched in the example in Figure 1. These must be one minimum and one maximum. Assume that the time points j and k are known in advance and

the minimum appears at position j. Then, an optimal data-fit is a solution of a convex quadratic program similar to (4)

$$\min \quad \sum_{i=1}^{N} (p_i - o_i)^2,$$

$$\text{s.t.} \quad \boldsymbol{p} \in \mathbb{R}^N,$$

$$p_i \geq p_{i+1} \quad \text{for } i \in \{1, \ldots, j-1\},$$

$$p_i \leq p_{i+1} \quad \text{for } i \in \{j, \ldots, k-1\}, \tag{5}$$

$$p_i \geq p_{i+1} \quad \text{for } i \in \{k, \ldots, N-1\},$$

$$\left[p_N \geq p_1 \right],$$

where the optional last constraint appears, if the considered interval represents a full cycle of a periodic model. This yields a vector $p \in \mathbb{R}^N$ with entries that decrease up to entry j, then start to increase and which fall after entry k. At the same time this
5   vector minimizes the deviation from the observational data. The negated solution of (5) applied to $-\boldsymbol{o}$ instead of $\boldsymbol{o}$ is an optimal data-fit to observations $\boldsymbol{o}$ that has a maximum at position j and a minimum at position k. Now, if the positions j and k of the extremes are unknown, repeating the optimizations with $\boldsymbol{o}$ and $-\boldsymbol{o}$ for every $j < k \in [N]$, the best of all results is an optimal data-fit subject to the property that there are (at most) two local extremes in the ~~time~~ series. Similar to the case of two extremes, we can consider more than two, say m, extremes. Dealing with all possible combinations of the positions of m extremes would
10   imply a computational effort of $c_1 \cdot m \cdot N^2$ operations (c_1 constant, N arbitrary), but using a tailored algorithm (Demetriou and Powell, 1991) we can calculate a best piece-wise monotonic fit in only $c_2 \cdot m \cdot N^2$ computer operations.

2.5   Bounds for models with bounded derivatives

The change rates of biogeochemical processes like growth and decay have natural limits. In the presence of noise, observational data is very likely to exhibit higher variations than a model that is devoted to comparatively slow interactions. In other words,
15   noise (or unresolved periodic processes with high frequencies and high amplitudes) cannot be well approximated by models that mimic processes of lower variation, i.e., models with small changes in a given time step. These processes are characterized by a small absolute derivative. If we are able to postulate general bounds on the derivatives of a parametric model function φ with respect to time, we can try to utilize this property in order to calculate lower bounds on the optimal misfit of φ.

General bounds on the first time-derivative (steepness) of φ are given as real numbers $D_{\min} < D_{\max}$ such that $D_{\min} \leq$
20   $\frac{\partial \varphi}{\partial t}(\boldsymbol{s}, t) \leq D_{\max}$ holds for all allowed parameter sets $\boldsymbol{s}$ and time points t. Using the function space $\mathcal{F} = \{\Phi : \mathbb{R} \to \mathbb{R} \mid D_{\min} \leq \Phi' \leq D_{\max}\}$ in (3), we obtain a relaxation of the parametric problem (1) that can be expressed as the convex quadratic program

$$\min \quad \sum_{i=1}^{N} (p_i - o_i)^2,$$

$$\text{s.t.} \quad \boldsymbol{p} \in \mathbb{R}^N, \tag{6}$$

$$p_i + (t_{i+1} - t_i) D_{\min} \leq p_{i+1} \quad \text{for } i \in [N-1],$$

$$p_i + (t_{i+1} - t_i) D_{\max} \geq p_{i+1} \quad \text{for } i \in [N-1].$$

A solution of this problem yields a lower model-data misfit bound for all parameter sets $\boldsymbol{s}$ such that $\varphi(\boldsymbol{s}, \cdot)$ satisfies the steepness bounds, $D_{\min} \leq \frac{\partial \varphi}{\partial t}(\boldsymbol{s}, t) \leq D_{\max}$. Here, we approximated the derivative $\Phi'(t)$ by finite differences which yields

5   $D_{\min} \leq \frac{\Phi(t_{i+1}) - \Phi(t_i)}{t_{i+1} - t_i} \leq D_{\max}$.

It is also possible to add linear constraints to the QP which consider bounds on higher-order derivatives of φ in terms of higher-order finite differences. For example, the property $D_{2,\min} \leq \frac{\partial^2 \varphi}{\partial t^2}(\boldsymbol{s}, t) \leq D_{2,\max}$, $s \in \boldsymbol{s}$, $t \in [t_1, t_N]$, can be accounted for with second-order differences by, e.g., posing the (compactly written) constraints

$$D_{2,\min} \leq \frac{p_{i+2} - 2p_{i+1} + p_i}{(t_{i+2} - t_i)^2} \leq D_{2,\max} \quad \text{for } i \in [N-2].$$

10   The knowledge of tight bounds on derivatives of increasing order allows to obtain increasingly tight lower bounds on the model-data misfit. However, since bounds on higher-order derivatives are more difficult to derive in practice, we restrict our studies to steepness bounds.

2.6   Bounds for models with combined properties

Clearly, we can combine model properties into a joint QP, e.g., if the model has both, two local extremes within a window of
15   interest and bounded steepness. We can apply the combination of (5) and (6) and obtain the joint QP

$$\min \quad \sum_{i=1}^{N} (p_i - o_i)^2,$$

$$\text{s.t.} \quad \boldsymbol{p} \in \mathbb{R}^N,$$

$$p_i \geq p_{i+1} \geq p_i + (t_{i+1} - t_i) D_{\min}$$
$$\quad \text{for } i \in \{1, \ldots, j-1\} \cup \{k, \ldots, N-1\}, \tag{7}$$

$$p_i \leq p_{i+1} \leq p_i + (t_{i+1} - t_i) D_{\max}$$
$$\quad \text{for } i \in \{j, \ldots, k-1\},$$

$$\left[p_N \geq p_1 \geq p_N + (T + t_1 - t_N) D_{\min} \right].$$

Here, again, $j < k$ are the indices of the unique minimum and the unique maximum, respectively, $D_{\min} < 0$ and $D_{\max} > 0$ are the universal lower and upper bounds on the model's first derivative, and T is the optional period of the model.

Similar to the approach in Subsection 2.4, the optimal solution of (7) applied to $\boldsymbol{o}$ and $-\boldsymbol{o}$ for all $j < k \in [N]$ will provide
20   the lower bound on the model-data misfit of the parametric model. As an alternative to a QP solver, we can use an extension

of the PAV algorithm that additionally considers steepness bounds with monotonic regression (LPAV, Yeganova and Wilbur, 2009) in order to solve (7).

3 Experiments

3.1 Method evaluation

5   We first aim to examine the extend to which the minimum model-data misfit of a parameterized model can deviate from the corresponding minimum misfit of a proposed non-parametric relaxation. Clearly, the difference between both misfits also depends on the characteristics of the observational data, that is, noise level and data density. We therefore derive statistics about that dependency using synthetic observations.

3.1.1 Test statistics

10   We generate the synthetic observations by adding white noise to N discrete samples $\Phi(t_i)$ of a model function $\Phi : \mathbb{R} \longrightarrow \mathbb{R}$, varying both the noise level and the number of samples. Our noise levels will be relative to the range

$$r := \max_{i \in [N]} \Phi(t_i) - \min_{i \in [N]} \Phi(t_i)$$

of the model output. As a simple parametric test function we use a cubic polynomial $\Phi(t) = \varphi(\boldsymbol{s}, t) = \sum_{i=0}^{3} s_i t^i$. We simulate one-year time series of observational data by considering the interval $[0, 365]$ and taking N equidistant samples, $t_i = \frac{i}{N} \cdot 365$,

15   for a polynomial with fix coefficients s^* (a fixed parametrization), and $N \in \{12, 25, 50, 100, 200, 300\}$. We add zero-mean white noise $\mathcal{N}(0, \sigma)$ to the time series values using one of 6 different noise levels with standard deviations $\sigma = \sigma^* \cdot r$, $\sigma^* \in \{0.1, 0.2, 0.3, 0.5, 0.7, 1.0\}$. Figure 2 shows the exemplary cubic polynomial

$$\Phi(t) = 2 + 0.035 \cdot t - 0.0003 \cdot t^2 + 5.592 \cdot 10^{-7} \cdot t^3$$

and $N = 300$ synthetic observations o_i obtained by adding white noise with standard deviation $\sigma = 0.2 \cdot r$ to the corresponding

20   function values, i.e., $o_i = \Phi(t_i) + \mathcal{N}(0, 0.2 \cdot r)$. The figure further shows the minimum root mean squared error data-fit by a function that has at most two local extremes as introduced in Subsection 2.4. The related root mean squared error between the synthetic data and this piece-wise monotonic fit is 0.445. We know that this error cannot be larger than the corresponding error between the fix polynomial Φ and the data since any cubic polynomial takes at most 2 extremes, too. Indeed, the latter error is 0.501, which is the root mean square of the white noise we added. By solving a convex optimization problem we can

25   efficiently identify the coefficients $s^* = (s_0^*, s_1^*, s_2^*, s_3^*)$ of a polynomial $\Phi^* = \sum_{i=0}^{3} s_i^* t^i$ that provides the best data-fit of all cubic polynomials. Unsurprisingly, the re-optimized polynomial Φ^* differs only slightly from the original one and yields a root mean squared error of 0.497.

For our statistics about the proposed error assessment methods we are interested in the ratio

$$q := \frac{f_{\text{rmse}}(\boldsymbol{p}^{\text{rel}}, \boldsymbol{o})}{f_{\text{rmse}}(\boldsymbol{p}^{\text{par}}, \boldsymbol{o})}$$
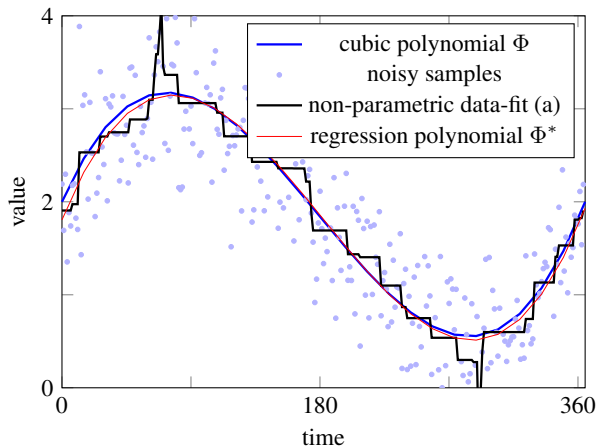
Figure 2. A cubic polynomial, synthetic observational data generated by adding white noise to 300 equidistant samples of the polynomial, and a minimum root mean squared error data-fit with regard to the property that no more than 2 extremes are taken.
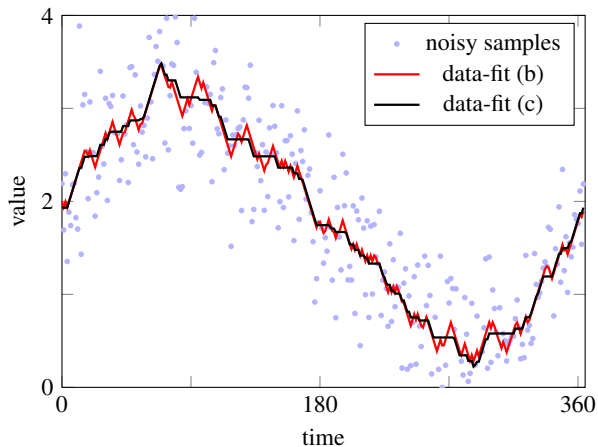
Figure 3. The synthetic observational data of Figure 2 and minimum root mean squared error data-fits with regard to a steepness bound (data-fit (b)) as well as regarding both properties, bounded steepness and the existence of at most 2 extremes (data-fit (c)).

between the lower error bound given by the optimal output of a non-parametric model relaxation p^{rel} and the corresponding data-fit with the original parametric model p^{par}. In the above example this ratio is $q_a = \frac{0.445}{0.501} = 0.888 \sim 89\%$. We repeat the calculation of a lower error bound and the corresponding error ratio with two other relaxations assuming only a bounded model steepness (cf. Subsection 2.5) and a combination of both properties, bounded steepness and the existence of at most 2

5 (local) extremes (cf. Subsection 2.6), respectively. The results are depicted in Figure 3. Here, for both relaxations we assume a maximum model steepness of 0.05 which is approximately 28% more than the maximum steepness of the original polynomial in the interval $[0, 365]$. The resulting root mean squared errors of the property-based optimal data-fits are 0.442 if only the

Table 1. Ratios (times 100) between the misfit of the parametric model (cubic polynomial) to synthetic observations (the model output plus white noise) and the misfit of the corresponding non-parametric regression model. We state the ratios for different noise levels σ and numbers of samples N. The values in each cell are the range of the ratios over 100 trials followed by their average and standard deviation. Non-parametric regression was done by (a) only assuming that at most 2 local extremes exist, (b) only assuming a steepness bound of 0.05, (c) assuming both properties.

property	σ			range mean st.dev.			
		$N = 300$	$N = 200$	$N = 100$	$N = 50$	$N = 25$	$N = 12$
(a)	0.1	82–88 85 1.2	74–85 81 2.2	63–79 71 3.3	36–69 56 6.6	9–58 36 10.2	0–51 12 13.8
	0.2	86–92 89 1.1	80–90 86 1.9	70–85 78 3.0	50–79 66 5.9	22–70 50 9.6	0–65 27 15.7
	0.3	88–94 91 1.0	83–92 88 1.8	74–87 81 2.8	56–82 71 5.6	24–74 56 9.4	1–70 36 15.2
	0.5	90–95 93 0.9	86–94 90 1.6	77–90 84 2.6	60–84 75 5.2	29–80 62 9.4	7–69 44 14.7
	0.7	91–96 94 0.9	88–95 91 1.5	79–92 86 2.5	62–86 77 5.2	31–81 64 9.2	13–71 48 14.4
	1.0	91–97 95 0.9	89–96 92 1.5	80–93 87 2.5	63–87 78 5.1	34–82 66 8.9	19–75 50 14.0
(b)	0.1	77–84 81 1.4	69–80 75 2.1	51–70 61 3.5	22–52 41 6.4	0–48 16 10.7	0–15 0 1.5
	0.2	85–91 88 1.2	79–88 84 1.7	67–81 75 2.9	45–69 60 5.6	12–63 38 10.3	0–42 7 10.2
	0.3	88–93 91 1.0	83–91 88 1.5	74–86 81 2.6	54–77 69 5.0	27–69 51 9.1	0–53 19 13.7
	0.5	91–95 93 0.9	87–94 91 1.3	81–91 86 2.2	63–84 77 4.4	44–77 64 7.7	0–66 37 14.1
	0.7	92–96 95 0.8	89–95 93 1.2	84–93 89 2.0	67–88 82 4.0	52–82 70 6.7	10–72 47 12.9
	1.0	94–97 96 0.7	91–96 94 1.1	87–95 91 1.8	71–91 85 3.7	59–86 76 5.8	25–77 57 11.5
(c)	0.1	85–92 89 1.1	80–88 85 1.7	70–83 76 2.8	44–73 61 5.6	15–58 39 9.5	0–51 12 13.8
	0.2	89–95 93 1.0	87–93 90 1.4	79–89 84 2.3	59–82 74 4.8	35–72 57 8.2	0–65 31 14.8
	0.3	91–96 94 0.9	89–95 92 1.2	82–92 88 2.1	66–86 79 4.4	47–78 66 7.2	1–70 42 13.5
	0.5	93–97 96 0.7	92–96 94 1.1	86–95 91 1.8	71–90 84 3.9	54–84 74 6.3	8–78 55 12.0
	0.7	94–98 97 0.7	93–97 95 0.9	88–96 92 1.7	73–93 87 3.6	59–87 79 5.7	18–83 62 11.1
	1.0	95–99 97 0.6	94–98 96 0.8	90–97 94 1.5	76–94 89 3.3	65–90 83 5.1	29–87 68 10.2

steepness bound is assumed (data-fit (b)) and 0.464 if both properties are assumed (data-fit (c)). The corresponding error ratios are $q_b = 0.883$ and $q_c = 0.927$, respectively.

To derive robust statistics, we repeat the experiment 100 times using different zero-mean white noise with the same standard deviation σ. Now, we do the same for all 6×6 combinations of N and σ, i.e., we apply the 3 model relaxations (a), (b), and (c) with regard to each of the 36 data property assumptions to 100 data sets of corresponding synthetic observations. The results are shown in Table 1.

The approach to calculate lower bounds on the model-data misfit by using property-based model relaxations stems from the intuition that the overall shape of the optimized parametric model and that of the non-parametric relaxation should be similar,

if the relaxation describes the main properties of the original model well. The amount of similarity is reflected by the ratios stated in Table 1. Values that are close to 100 percent provide evidence that the parametric model is suitably shaped with regard to the corresponding general model property assumptions. Here, by construction of the synthetic data, we already know that the original polynomials are "correctly shaped". Therefore, the numbers in the table actually reflect the tightness of the property-based relaxations and serve as an orientation under which circumstances the lower bound approach can succeed.

We observe that the data must be rather dense in order to reach good error ratios, especially with low levels of noise. This dependence is plausible because small numbers of observations as well as low levels of noise cause small difference quotients $\frac{o_{i+1}-o_i}{t_{i+1}-t_i}$ of the observations. However, the explicit steepness bounds (property (a)) or implicit steepness bounds (property (b)) which we use for the model relaxation must be considerably smaller than the difference quotients in order to provide a lower bound that is close to the model-data misfit of the optimized parametric model.

For example, consider a target ratio of 85% that is desired to be reached for all 100 sets of random observations, i.e., the left (worst case) number in a cell of Table 1 should be greater than 85. For up to $N = 50$ observations none of our experiments reaches the 85% in the worst case. For $N = 100$ it is reached with property (b) and noiselevel 1.0 and with property (c) and noise levels 0.5, 0.7, 1.0 (multiplied with the range of the underlying true process). Regarding the lowest applied noise level of 0.1 and property (a), the 85% ratio is never reached in the worst case but only in the average case and only with $N = 300$ observations.

3.1.2 A countercheck

Having evidence that the lower bounds on the model-data misfit become tight with sufficiently dense observations, we want to countercheck if an optimized parametric model that slightly differs from the actual process behind the observational data has a significantly worse model-data misfit in comparison with its non-parametric relaxation. This time, we generate 300 synthetic observations by disturbing the sum of two sine waves

$$\Phi(t) = \sin(t) + 0.3 \cdot \sin(2t)$$

and start with a noise level of 10% relative to the range of the function values ($\sigma = 0.1 \cdot r$). As the data might be mistaken as noisy measurements of a single sine process at first glance, we use a general sine model to fit the observations. From the data, we estimate that both the frequency and the amplitude of the sine are at most 1.2. This implies a maximum steepness of 1.44 and that the sine model takes no more than 2 extremes in $[2, \pi]$, that is, according to the above notation, we use a type (c) model relaxation. Optimization yields a solution with a root mean squared model-data misfit of 0.275 and the corresponding property-based lower misfit bound 0.2. The data and both model outputs are shown in Figure 4. With regard to the data density, the ratio $q_c = 0.727$ of both values is not completely convincing and, indeed, one can recognize a "failure in the model shape". Now, we suppose the more precise process

$$\varphi(s,t) = s_1 + s_2 \cdot \sin(s_3(t - s_4)) + s_5 \cdot \sin(s_6(t - s_7)), \tag{8}$$

resolving a second sine wave of higher frequency. We further suppose to know general bounds on both amplitudes $|s_2| \leq 1.2$, $|s_5| \leq 0.35$ and on both frequencies $|s_3| \leq 1.2$, $|s_6| \leq 2.2$. This implies that the steepness of φ is bounded by $1.2^2 + 0.35 \cdot 2.2 =$
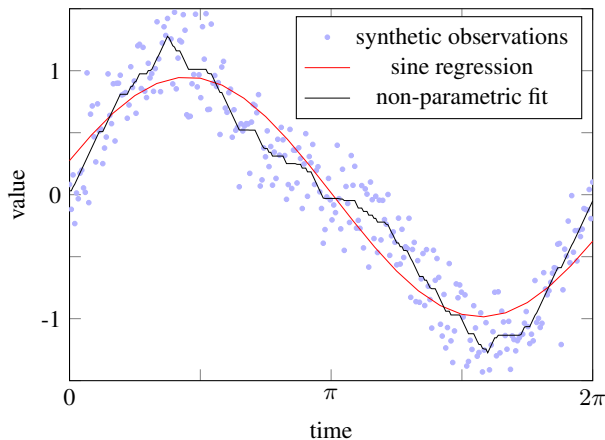
Figure 4. Synthetic data obtained by adding noise to the function $\Phi(t) = \sin(t) + 0.3 \cdot \sin(2t)$ and the optimized data-fit by a clean sine wave (parametric model) and its property-based non-parametric relaxation (steepness ≤ 1.44, at most 2 extremes), respectively.
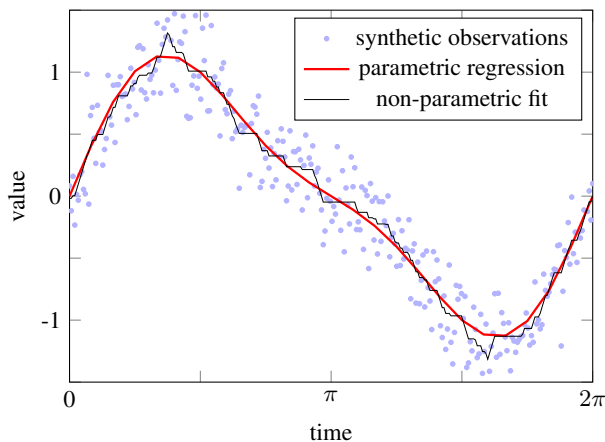
Figure 5. Synthetic observations as in Figure 4 but optimizing the "correct" parametric model and its property-based non-parametric relaxation (steepness ≤ 2.21, at most 4 extremes), respectively.

2.21. From the data, we can expect that the new model with optimized parameters does only take 2 extremes in the interval $[0, 2\pi]$, too. However, for the given bounds on the frequency and arbitrary parameters the model can take up to 4 extremes. Consequently, in addition to the steepness bound on the model, at most 4 extremes must be assumed to calculate the lower error bound on the best possible model-data misfit. Applying both assumptions, (i.e. using model property (c) from above) the exact optimum value of the model relaxation is ~ 0.217 while the optimized new parametric model comes down to ~ 0.193 providing a clearly better model-data misfit ratio $q_c = 0.891$ than the pure sine model. The optimized parametric model curve is shown in Figure 5.

Table 2. Ratios (times 100) between the misfit of the parametric model to synthetic observations (the model output plus white noise) and the misfit of the corresponding non-parametric regression model. The ratios are given for different noise levels σ and numbers of samples N. The 4 entries in each cell are the mean ratio of 100 trials, the standard deviation for the pure sine model, and the corresponding values for the "true" model that uses a sum of two sine waves. Non-parametric regression was done by (a) only assuming that at most 2 (4) local extremes exist, (b) only assuming a model specific steepness bound (see text), (c) assuming both properties.

property	σ	mean (sine) $N=500$	st.dev. (sine) $N=300$	mean ("truth") $N=200$	st.dev. ("truth") $N=100$
	0.05	84 1.1 50 1.8	79 1.5 47 2.0	73 2.5 43 2.6	60 3.6 36 3.3
(a)	0.1	89 0.9 74 2.0	85 1.2 70 2.2	81 2.2 67 3.1	71 3.2 60 4.2
	0.2	92 0.8 88 1.5	89 1.1 85 1.8	86 2.0 82 2.6	78 3.0 76 3.9
	0.05	81 1.1 52 1.8	73 1.7 49 2.0	65 2.5 46 2.4	46 4.1 39 3.3
(b)	0.1	88 0.9 77 1.9	83 1.4 74 2.2	78 2.0 71 2.5	65 3.5 65 3.6
	0.2	92 0.7 90 1.3	89 1.1 89 1.6	86 1.6 87 1.8	77 2.9 82 2.6
	0.05	88 0.9 55 1.8	84 1.4 52 2.0	78 2.1 50 2.5	65 3.2 44 3.1
(c)	0.1	93 0.7 79 1.9	89 1.1 77 2.2	86 1.7 75 2.6	77 2.6 70 3.5
	0.2	95 0.6 92 1.3	93 0.9 90 1.6	91 1.4 89 1.8	85 2.2 86 2.6

We repeat the experiment with noise levels of $\sigma = 5\%$ and $\sigma = 20\%$ for different numbers of equidistant observations $N \in \{500, 300, 200, 100\}$ and for all three property-based model-relaxation types (a), (b) and (c) used in Subsubsection 3.1.1. Again, we generate 100 different random sets of observations for each combination of σ and N. The results are depicted in Table 2. The experiments help to identify conditions under which we may distinguish the "truth" from "distortions of the truth". Sufficient conditions are given if the misfit ratio for the true parametric model, say q_1, is not too small, e.g., $q_1 \geq 0.5$ but the ratio for a moderate distortion of the true parametric model, say q_2, is essentially smaller. Depending on how close q_1 is to 1, we may concretize the predicate "essentially smaller" to hold if either of the fractions $\frac{q_1}{q_2}$ and $\frac{1-q_2}{1-q_1}$ is convincingly less than 1, say $\frac{q_1}{q_2} \leq 0.75$ or $\frac{1-q_2}{1-q_1} \leq 0.5$. We find that a rather low noise level is necessary to satisfy these conditions. As already observed in Subsection 3.1, high noise levels σ provide rather tight lower bounds on the minimum attainable model-data misfit of the "correct model type" if sufficiently many observations are available. Unfortunately, the corresponding lower bounds for a less accurate model become similarly close in this case. For properties (b) and (c) and fewer observations, they can even exceed the lower misfit bounds for the "correct model" since we apply different uniform steepness bounds.

3.2 Application on real-world observations

We now consider a two real-world example examples with the aim to fit chlorophyll observationschlorophyll-a observations.
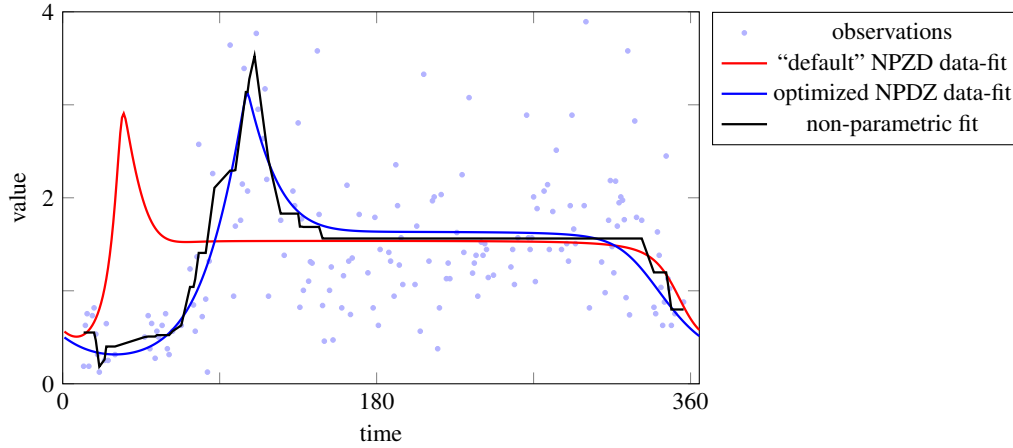
Figure 6. Bornholm seasonal adjusted observation time series of phytoplankton, and data-fits by the considered NPZD-model using the parameters which were adjusted for global model configuration (red) and optimized parameters for the local model version (blue). The (black) reference plot is the minimum error data-fit with regard to the properties that no more than 2 extremes are taken and the steepness is at most 0.14.

3.2.1 Baltic Sea observations

Our first example considers observations from the Bornholm Basin in the Baltic Sea at 55.15 N, 15.59 E, dubbed station BY5. The data were provided by the Swedish Oceanographical Data Center (SHARK) at the Swedish Meteorological and Hydrological Institute (SMHI). BY5 was repeatedly sampled during 1962–2009. As there are long data gaps, we merge all data into a climatological seasonal cycle. To derive phytoplankton (in nitrate units) from ~~chlorophyll~~chlorophyll-a, we use a constant chlorophyll-a to nitrate ratio of 1.59 Chla/mol N. The considered seasonal adjusted time-series comprises 175 observations of phytoplankton.

We fit a box model of NPZD type to the data. It is based on a model of Oschlies and Garcon (1999). The original version was set up and tuned for the global ocean, but we consider a simplified version which is described in detail by Löptien and Dietze (2015). Its model equations are given in the Appendix and its free parameters and their assumed limits can be found in Table A1. As there is no temperature dependence in this model version an average temperature of 10 degree Celsius is assumed for the growth period. Further, the "assimilation efficiency of herbivores" parameter is omitted (implicitly set to 1). Figure 6 shows the Baltic Sea phytoplankton data set and simulations of the NPZD model simulations (red curve) using the parameter values taken from Oschlies and Garcon (1999). As this model-fit appears to be poor, we optimize the parameters with regard to the Baltic Sea observations (these optimized parameter values are also depicted in Table A1). The result is a more adequate model output (blue curve) lowering the associated RMSE model-data misfit from 0.896 to 0.717.

Now, we assess our result by lower bounds. Following the procedure outlined at the end of Subsection 2.2, we need to identify properties of the NPZD model that are satisfied for every credible (allowed) parameter choice and lead to an easily solvable

surrogate problem. The ideal is to give a mathematical proof of such properties. However, it is justified to postulate a model property which is based on sound biological experience, even if this property is not satisfied for all feasible combinations of the parameter values. In this context it is important to note, that the relatively simple model structure of our NPZD-type model with fixed (non-temperature dependent) rates does not suffice to describe the seasonal cycle after the spring bloom (Fennel and Neumann, 2004, page 35ff). Generally, model versions which fit the spring bloom satisfactory do not capture the observed chlorophyll increase in autumn. We thus assume only two extremes to determine a compatible lower bound. A practical approach for more complex systems ~~(which are beyond the scope of this contribution)~~ will be to iteratively increase the number of extremes of the non-parametric model relaxation until the obtained lower bound hardly increases anymore (this approach will also require quite dense observational data). Using the algorithm of Yeganova and Wilbur (2009), we find that the best attainable RMSE misfit between a time series with 2 extremes and our data is $\sigma_a = 0.557$, a first lower error bound for the applied NPZD model. The corresponding error ratio between this bound and the error of the optimized model is $q_a = 0.777$. In order to tighten our lower error bound, we additionally postulate a model steepness limit of 0.14, which we justify by the fact that the optimized model curve has two extremes, a plausible position of its maximum, and a maximal steepness of ~ 0.1. The associated best possible data-fit, which we can calculate using the "piecewise monotonic regression" algorithm (Yeganova and Wilbur, 2009) in combination with the LPAV algorithm (Demetriou and Powell, 1991) instead of the classical PAV algorithm (black curve in Figure 6) has an RMSE of $\sigma_c = 0.66$ which yields a quite high ratio of $q_c = \frac{\sigma_c}{0.717} = 0.921$. Thus, it is confirmed that the main portion of the model-data misfit of the optimized NPZD model is not caused by a sub-optimal choice of the parameter set but by other sources of uncertainty. For the sake of completeness we calculate the best data-fit with regard to a limited model steepness of 0.14 solely (disregarding its number of extremes), using CPLEX to solve the corresponding formulation in terms of a quadratic program (6). In this case, the RMSE is $\sigma_b = 0.619$ and the corresponding error ratio is $q_b = 0.864$.

Some indication of an even smaller gap between the attained model-data misfit and the globally optimal misfit of the NPZD model is given by the following additional step. The RMSE error of the calibrated NPZD model is the empirical standard deviation σ between model simulations and observations. The lower bounds with regard to the general model properties (a), (b) and (c), i.e., $\sigma_a = 0.557$, $\sigma_b = 0.619$ and $\sigma_c = 0.66$, are approximately 20%, 22% and 23.5% of the range of the model output, respectively. Experiments with random noise might help to further assess the quality of our parametric solution. Similar to the statistics in Subsubsection 3.1.1, Table 1, we generate 100 sets of synthetic observations for each of the 3 standard deviations σ_a, σ_b and σ_c by simply adding white noise to the model output and calculate the average error of the corresponding optimal non-parametric data-fit. The obtained average ratios are $q_{a,\mathrm{emp}} = 0.719$, $q_{b,\mathrm{emp}} = 0.891$, and $q_{c,\mathrm{emp}} = 0.919$ which is encouragingly close to the respective ratios for the true observations. We have to note, however, that the assumed normal distribution property is not actually satisfied by the errors between phytoplankton observations and the optimized NPZD model.

3.2.2 Global satellite observations

In a second example, we consider global observations of maritime chlorophyll-a. We use annual mean chlorophyll-a observations in units [mg/m^3], which were derived from SeaWiFS data from http://seadas.gsfc.nasa.gov using 8-day composites binned on
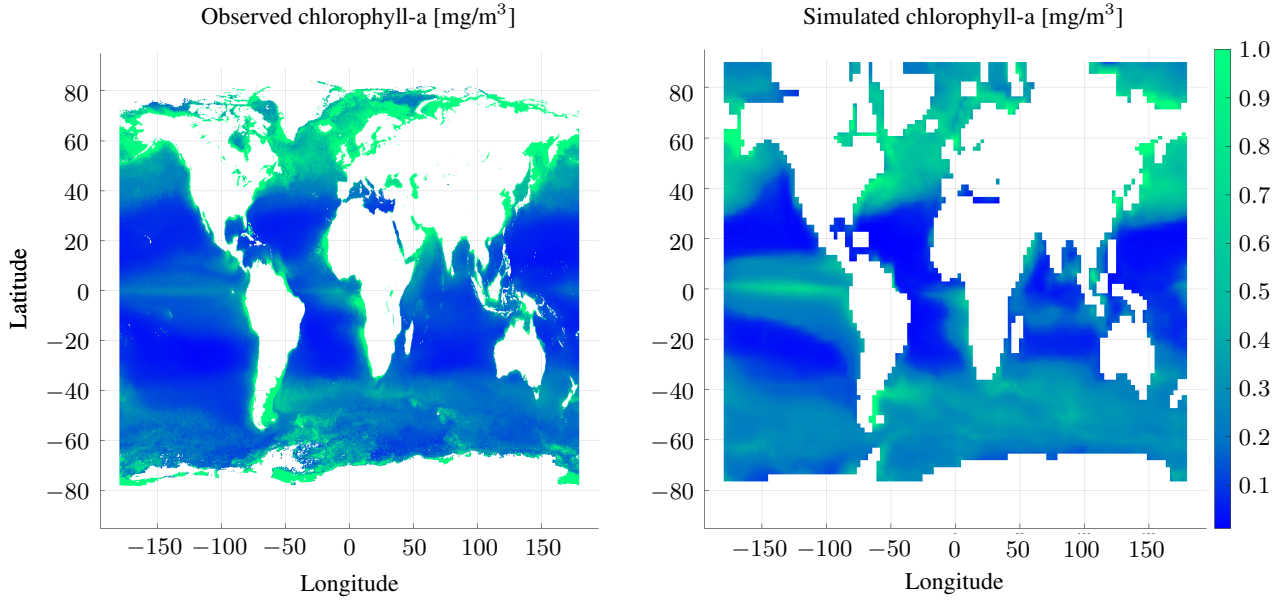
Figure 7. Observed and simulated annual mean chlorophyll-a mapped onto a 1/12° spatial grid. Left panel: SeaWiFS level 3 mapped 8-day composites from http://seadas.gsfc.nasa.gov, 1998–2005 annual means of binned climatology. Right panel: Corresponding spatially interpolated model simulations.

a 1/12° spatial grid (4320×2160 boxes). Note that this means might be seasonally biased in regions with sparse data. As we consider annual mean values, we apply our method in space instead of time. We do not consider coastal areas, since these can not be well represented in the coarse resolution model and are also likely to contain a considerable degree of observational uncertainty. We thus outmask all grid boxes with chlorophyll-a concentration above 1 mg/m^3. We compare the observed chlorphyll-a concentrations to simulated values. The simulation is based on the CM2Mc configuration, described by Galbraith et al. (2010). Spinup procedure and boundary conditions follow Dietze et al. (2017) (cf. Tab. 1 their FMCD configuration). The resolution of the model comprises 120×80 boxes (3° for longitudes and 2°–3° for latitudes) and is coarse, compared to the spatial resolution of the observational data. The annual mean model simulations are interpolated onto the observation grid in order to compute the corresponding RMSE for the model-data misfit. Figure 7 visualizes the observed chlorophyll-a and the corresponding model simulations. The observational data is quite rugged for larger regions of the ocean while the simulations are comparably smooth everywhere, due to the resolution of the model. Therefore, we can expect positive lower bounds on the model-data misfit.

The RMSE model-data misfit is 0.138 mg/m^3. Dealing with 2-dimensional data, our one-dimensional lower bound methods can be applied chunk-wise. We decide to traverse each longitude of our spatial grid in chunks of 200 consecutive boxes (where the last chunk for each longitude consists of its ≤ 200 remaining boxes). It provides us a lower bound on the sum squared errors between observations and simulations with regard to each chunk, say $\alpha_{i,k}$ for the k-th of n_i chunks of the i-th longitude, $i = 1, \ldots, 4320$. A lower bound on the (unweighted) RMSE model-data misfit is then given by

$$\alpha = \sqrt{\frac{1}{N_{\text{obs}}} \sum_{i=1}^{4320} \sum_{k=1}^{n_i} \alpha_{i,k}}\,,$$

where N_{obs} is the total number of considered observation values. Note, that the proposed method works equally well on weighted RMSEs. As a general model property for our bound approach we use a slightly higher maximum chlorophyll-a variation per distance than the maximum variation of our model simulation results (similar to the Baltic Sea example, we multiply the maximum simulated variation with 1.4 for each chunk). The result is a lower bound of 0.049 mg/m^3 which is 35% of the misfit we achieved with our model simulations. Thus, the lower bound is in the same order of magnitude, but still considerable lower than the actual model-data misfit. One might conclude that there is room for model-improvement when it comes to chlorophyll-a. Still one needs to keep in mind that the model was presumably never optimized to simulate chlorophyll-a as good as possible and focused on many other factors as well.

We repeated our experiment restricted to the Southern Ocean (below $60°$ S). Here, the RMSE model-data misfit is 0.170 mg/m^3 and the calculated lower bound is 0.108 mg/m^3 (63%).

4   Discussion

Our aim is to complement research on the calibration of biogeochemical models by calculating lower bounds on their best attainable model-data misfit. We utilize two general model properties for our purpose; a limited number of extremes and a bounded model steepness. We also consider the combination of both properties. The reason to consider such non-parametric model properties is that they yield efficiently solvable (relaxed) optimization problems whereas optimizing the original parametric model is computationally demanding.

4.1   Applicability

In our experiments (Section 3.1.1), the solitary assumption of a bounded model steepness leads to tight model relaxations (tight lower error bounds), if enough observational data is available and the steepness bound is chosen to be close to the maximum steepness of a calibrated model output. The task to derive a maximal bound for the steepness of a respective model output can be difficult in practice and relies on (1) model equations and (2) observational data. A rigorous mathematical model analysis, e.g., considering single model parameters like the maximum growth of phytoplankton, provides maximal limits which are valid for the entire parameter domain. However — resting on observation-based experience with the modeled processes — it might be justified to assume a smaller, empirical steepness bound, irrespective of that bound being valid for all permitted parameters.

In ~~Subsection ??~~Subsubsection 3.2.1, we assumed a steepness bound that is $\sim 40\%$ larger than the maximum steepness of the NPZD model with optimized parameters. In future, we aim to target iterative procedures to derive tight universal (likely time variate) model steepness bounds, e.g., using some kind of branch-and-bound approach.

Our second constraint, a limited number of extremes, is generally relatively easy to determine for common, rather smooth biogeochemical models. An applicable number of extremes can be determined if regression with more extremes does hardly reduce the misfit anymore. But here one should also keep the model structure in mind. Simple models can be limited in reproducing specific shapes of the seasonal cycle. Based on the model structure, we assumed only two extremes for our NPZD real-world example in ~~Subsection ??~~Subsubsection 3.2.1. Note, however, that assuming 4 extremes yields better fits in this case: the RMSE decreases from 0.619 to 0.559 without bounding the steepness (from 0.66 to 0.62 with steepness bound). Note that the "low number of extremes" condition indirectly implies a bounded (average) model steepness, too. In our Baltic Sea experiments, the assumption about the number of extremes resulted in better bounds than the sole assumption of a bounded steepness. In our experiments with global ocean data, we observed opposite results.

Unsurprisingly, the combination of tight steepness bounds with a limited number of extremes yields even better lower bounds on the minimum attainable model-data misfit than both properties separately. ~~However~~Finally, all our model relaxations require a rather large number of observations (per chunk) in order to yield convincingly tight bounds (cf. Table 1).

4.2 Generalizations

Our contribution considers the root mean squared error (RMSE) as an objective measure of the model-data misfit because it eases the task to formulate certain model properties in terms of convex optimization problems and to resort to corresponding tailor-made efficient algorithms. However, the suggested model properties also allow to deduce efficiently solvable optimization problems if other misfit measures are used. For example, the sum absolute error can be dealt with in terms of linear programs (LPs) by including auxiliary variables and auxiliary linear constraints to express absolute values. Also, the efficient methods of Demetriou (1995) and Yeganova and Wilbur (2009) (we provide RMSE implementations as article supplements) can be realized with other misfit measures than RMSE.

Concerning the number of local extremes, our proof-of-concept experiments are restricted to a maximum of 2 (4) extremes, according to the properties of the respective parametric models. However, solutions can even be calculated efficiently if the model output is assumed to take a large maximum number of extremes (Demetriou and Powell, 1991; Demetriou, 1995). As mentioned above, a suitable approach to work with that property is to increase the maximum assumed number of extremes until the corresponding lower bound on the minimum attainable model-data misfit hardly increases anymore, indicating that further extremes contribute to fit noise rather than processes of interest.

4.3 Cautionary notes

Contrary to the fact that a small gap between the misfit of some property-based model relaxation and the misfit of the optimized original model proves that further parameter calibration is not required, a large gap between both misfits does not necessarily mean that the calibration of the chosen model is bad nor that the model is a wrong representation of the processes of interest.

Our experiments indicate that a large gap only then tends to serve as an inadequacy proof of a model (calibration) if enough observations are available. Otherwise, the chosen property-based relaxations might fit observations too well.

On the other hand, a small gap between the optimal misfit of a property-based non-parametric relaxation and the misfit of the original parametric model can even be reached with an inappropriate parametric model structure if there is too much noise in the data. The experiments in Section 3.1.2 are setup to estimate conditions that allow to distinguish the "truth" from a "moderate distortion of the truth". With regard to the experimental results in Table 2, a rather low noise level is necessary to satisfy these conditions.

5   Conclusions

We presented a concept to prove that a parametric model is well calibrated, i.e., that changes of its free parameters cannot lead to a much better model-data misfit anymore. The intention is motivated by the fact that calibrating global biogeochemical ocean models is important but computationally expensive.

Generally, the aim is to determine an optimal parameter set such that a predefined metric of the model-data misfit is minimal. To keep the number of required expensive model simulations as small as possible, we suggest to calculate "tight" lower bounds on the lowest achievable model-data misfit. Our idea is to utilize properties of the original model that are satisfied for all permitted parameters and lead to easily solvable optimization problems. Here, we focus on two such model properties to derive our lower bounds on the model-data misfit; a maximum time-derivative and a maximum number of extremes per time unit.

Indeed, our experiments show that the achieved bounds can come quite close to the optimized misfit of the original model if many observations are available. However, a problem with global observational data (e.g., World Ocean Atlas data) is that it is often sparse in time. For example, if we examine annual cycles of periodic processes with monthly observations, our lower bound approach will only succeed if we overlay (seasonally adjust) measurement data of several years in order to reach the required data-coverage. Long-term time series from observing platforms like BATS (Steinberg et al., 2001) provide enough data on the temporal dimension but are limited in space and are only available for certain sites. However, we can also apply our method with data that is dense in space. A suitable global application of our method to biogeochemical models ~~might be~~ is related with dense satellite observations of chlorophyll-a (Volpe et al., 2007; Dogliotti et al., 2009). Subsubsection 3.2.2 illustrates, how our methods can be applied in order to cope with such data.

Assuming the error between model output and observations to be Gaussian distributed noise, an obtained lower bound on the root mean squared error is also a lower bound on the empirical standard derivation σ of the noise. We suggest the following rule-of-thumb procedure, which is illustrated for a real world example in ~~Subsection ??~~ Subsubsection 3.2.1:

1. Optimize the model parameters w.r.t. the corresponding model-data misfit

2. Calculate lower error-bounds on the model-data misfit by using appropriate assumptions about the model properties

3. Be satisfied if the ratio q between 1 and 2 is close to 1 or

4. Consider the lower bound from 2 to be the standard deviation σ of the noise in the observations and check whether q corresponds to the empirical ratio q_{emp} that is obtained by adding random noise of level σ to the output of the optimized parametric model and fitting the obtained synthetic observations with the non-parametric relaxation.

Code availability. Implementations of the applied methods are available on GitHub (Sauerland, 2017). We provide two packages of C++ sources:

- `regressionCPX` includes QP formulations and requires the CPLEX solver.

- `regression` is a subset that does not require CPLEX but only uses QP free and tailored regression algorithms: PAV (Barlow et al., 1972), LPAV (Demetriou and Powell, 1991), PMR (Yeganova and Wilbur, 2009), and a combination of LPAV and PMR, PMRS.

For compilation, usage, and further notes, we refer to the README files contained in both packages.

Appendix A: NPZD model parameters and equations

We explicitly state the free parameters and equations of the NPZD type box model that has been studied in Löptien and Dietze (2015) and is used for our real world example in ~~Subsection ??~~Subsubsection 3.2.1. The prognostic variables are nitrate (N), phytoplankton (P), zooplankton (Z) and detritus (D) and scaled to units of $\mathrm{mmol\,N\,m}^2$. The temporal change of the prognostic variables depends on 10 free parameters outlined in Table A1 and is determined by the following equations

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{N} = -\mu_{\max} \cdot g_l \cdot g_N \cdot \mathrm{P} + m_{\mathrm{PN}} \cdot \mathrm{P} + m_{\mathrm{ZN}} \cdot \mathrm{Z} + m_{\mathrm{DN}} \cdot \mathrm{D},$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{P} = \mu_{\max} \cdot g_l \cdot g_N \cdot \mathrm{P} - m_{\mathrm{PN}} \cdot \mathrm{P} - G(\mathrm{P}) \cdot \mathrm{Z} - m_{\mathrm{PD}} \cdot \mathrm{P},$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{Z} = G(\mathrm{P}) \cdot \mathrm{Z} - m_{\mathrm{ZN}} \cdot \mathrm{Z} - m_{\mathrm{ZD}} \cdot \mathrm{Z}^2,$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{D} = m_{\mathrm{ZD}} \cdot \mathrm{Z}^2 + m_{\mathrm{PD}} \cdot \mathrm{P} - m_{\mathrm{DN}} \cdot \mathrm{D}.$$

Here, the hyperbolic MM equations $g_{\mathrm{I}} = \frac{\mathrm{PAR}}{\mathrm{PAR} + H_{\mathrm{PAR}}}$ and $g_{\mathrm{N}} = \frac{\mathrm{N}}{\mathrm{N} + H_{\mathrm{N}}}$ describe the limiting effect of light and nitrate concentration on the nitrate uptake of phytoplankton and

$$G(\mathrm{P}) = \frac{g_{\max} \cdot \mathrm{P}^2}{\mathrm{P}^2 + H_{\mathrm{Z}}}$$

is a "Holling III-type" term. The maximum growth rate of phytoplankton $\mu_{\max}$ and the maximum grazing rate of zooplankton $g_{\max}$ are obtained by substitutions

$$\mu_{\text{new}} := \mu_{\max} - m_{\mathrm{PN}} - m_{\mathrm{PD}},$$

$$g_{\text{new}} := g_{\max} - m_{\mathrm{ZN}},$$

in order to enforce net phytoplankton growth and net zooplankton grazing by the positive lower limits on μ_{new} and g_{new}, respectively.

Table A1. Parameters of the considered NPZD model with their physical units, allowed ranges, and optimized values.

Parameter	Symbol	Unit	Range	Optimized value
Net max. phytoplankton growth rate	μ_{new}	day^{-1}	0.1–0.9	0.1
Half-sat. const for light	H_{PAR}	$\text{W}\,\text{m}^{-2}$	5.0–40.0	24.7832
Half-sat. const for nutrient uptake	H_{N}	$\text{mmol}\,\text{N}\,\text{m}^{-3}$	0.05–1.2	0.05
Max. grazing/prey-capture rate	H_{Z}	$\text{mmol}\,\text{N}\,\text{m}^{-6}$	0.2–1.1	0.2
Net max. grazing rate	g_{new}	day^{-1}	0.01–1.2	1.2
Phytoplankton loss to N	m_{PN}	day^{-1}	0.01–0.6	0.01
Zooplankton loss to N	m_{ZN}	day^{-1}	0.01–0.65	0.01
Remineralization rate of Det.	m_{DN}	day^{-1}	0.02–0.15	0.02
Zooplankton loss to Det.	m_{ZD}	$\text{day}^{-1}(\text{mmol}\,\text{N}\,\text{m}^{-3})^{-1}$	0.01–0.9	0.507
Phytoplankton loss to Det.	m_{PD}	day^{-1}	0.01–0.9	0.0191

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. This work is a contribution to the project "Reduced Complexity Models" (supported by the Helmholtz Association), the DFG-supported project SFB754 and ~~to the research platforms of the DFG cluster of excellence~~ the DFG-cluster of excellence "The Future Ocean". The authors received very helpful comments from two anonymous reviewers! We are grateful to Heiner Dietze for technical assistance.

References

Anderson, T.: Plankton functional type modelling: running before we can walk?, Journal of Plankton Research, 27, 1073–1081, doi:10.1093/plankt/fbi076, 2005.

Aumont, O., Ethé, C., Tagliabue, A., Bopp, L., and Gehlen, M.: PISCES-v2: an ocean biogeochemical model for carbon and ecosystem studies, Geoscientific Model Development, 8, 2465–2513, doi:10.5194/gmd-8-2465-2015, 2015.

Barlow, R. E., Bartholomew, D. J., Bremner, J. M., and Brunk, H. D.: Statistical Inference under Order Restrictions. Theory and Application of Isotonic Regression, Wiley Series in Probability and Mathematical Statistics, John Wiley & Sons, London, doi:10.1111/j.1467-9574.1973.tb00228.x, 1972.

Boyd, S. and Vandenberghe, L.: Convex optimization, Cambridge University Press, 2004.

Brovkin, V., Petoukhov, V., Claussen, M., Bauer, E., Archer, D., and Jaeger, C.: Geoengineering climate by stratospheric sulfur injections: Earth system vulnerability to technological failure, Climatic Change, 92, 243–259, doi:10.1007/s10584-008-9490-1, 2009.

Cao, L. and Caldeira, K.: Atmospheric CO2 stabilization and ocean acidification, Geophysical Research Letters, 3, doi:10.1029/2008GL035072, 2008.

Cao, L. and Caldeira, K.: Atmospheric carbon dioxide removal: long-term consequences and commitment, Environmental Research Letters, 5, doi:10.1088/1748-9326/5/2/024011, 2010.

Demetriou, I. C.: Discrete piecewise monotonic approximation by a strictly convex distance function., Mathematics of Computation, 64, 157–180, doi:10.2307/2153327, 1995.

Demetriou, I. C. and Powell, M. J. D.: Least squares smoothing of univariate data to achieve piecewise monotonicity., IMA Journal of Numerical Analysis, 11, 411–432, doi:10.1093/imanum/11.3.411, 1991.

Dietze, H. and Löptien, U.: Revisiting nutrient trapping in global coupled biogeochemical ocean circulation models, Global Biogeochemical Cycles, 27, 265–284, doi:10.1002/gbc.20029, 2013.

Dietze, H., Getzlaff, J., and Löptien, U.: Simulating natural carbon sequestration in the Southern Ocean: on uncertainties associated with eddy parameterizations and iron deposition, Biogeosciences, 14, 1561–1576, doi:10.5194/bg-14-1561-2017, 2017.

Dogliotti, A. I., Schloss, I. R., Almandoz, G. O., and Gagliardini, D. A.: Evaluation of SeaWiFS and MODIS chlorophyll-a products in the Argentinean Patagonian Continental Shelf (38°S-55°S), International Journal of Remote Sensing, 30, 251–273, doi:10.1080/01431160802311133, 2009.

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

Faugeras, B., Lévi, M., Mémery, L., Verron, J., Blum, J., and Charpentier, I.: Can biogeochemical fluxes be recovered from nitrate and chlorophyll data? A case study assimilating data in the Northwestern Mediterranean Sea at the JGOFS-DYFAMED station, Journal of Marine Systems, 40–41, 99–125, doi:10.1016/S0924-7963(03)00015-0, 2003.

Fennel, K., Losch, M., Schröter, J., and Wenzel, M.: Testing a marine ecosystem model: sensitivity analysis and parameter optimization, Journal of Marine Systems, 28, 45–63, doi:10.1016/S0924-7963(00)00083-X, 2001.

Fennel, W. and Neumann, T.: Introduction to the Modelling of Marine Ecosystems, Elsevier Science, https://www.123library.org/ebook/isbn/9780080534978/, 2004.

Friedrichs, M. A. M.: A data assimilative marine ecosystem model of the central equatorial Pacific: Numerical twin experiments, Journal of Marine Research, 59, 859–894, doi:10.1357/00222400160497544, 2001.

Friedrichs, M. A. M.: Assimilation of JGOFS EqPac and SeaWiFS data into a marine ecosystem model of the Central Equatorial Pacific Ocean, Deep Sea Research Part II: Topical Studies in Oceanography, 49, 289–319, doi:10.1016/S0967-0645(01)00104-7, 2002.

Friedrichs, M. A. M., Hood, R. R., and Wiggert, J. D.: Ecosystem model complexity versus physical forcing: Quantification of their relative impact with assimilated Arabian Sea data, Deep Sea Research Part II: Topical Studies in Oceanography, 53, 576–600, doi:10.1016/j.dsr2.2006.01.026, 2006.

Galbraith, E. D., Gnanadesikan, A., Dunne, J. P., and Hiscock, M. R.: Regional impacts of iron-light colimitation in a global biogeochemical model, Biogeosciences, 7, 1043–1064, doi:10.5194/bg-7-1043-2010, 2010.

Garcia-Gorriz, E., Hoepffner, N., and Ouberdous, M.: Assimilation of SeaWiFS data in a coupled physical–biological model of the Adriatic Sea, Journal of Marine Systems, 40–41, 233–252, doi:10.1016/S0924-7963(03)00020-4, 2003.

Gregg, W. W., Friedrichs, M. A., Robinson, A. R., Rose, K. A., Schlitzer, R., Thompson, K. R., and Doney, S. C.: Skill assessment in ocean biological data assimilation, Journal of Marine Systems, 76, 16–33, doi:10.1016/j.jmarsys.2008.05.006, 2009.

Hemmings, J. C., Srokosz, M. A., Challenor, P., and Fasham, M. J.: Split-domain calibration of an ecosystem model using satellite ocean colour data, Journal of Marine Systems, 50, 141–179, doi:10.1016/j.jmarsys.2004.02.003, 2004.

Hemmings, J. C. P. and Challenor, P. G.: Addressing the impact of environmental uncertainty in plankton model calibration with a dedicated software system: the Marine Model Optimization Testbed (MarMOT 1.1 alpha), Geoscientific Model Development, 5, 471–498, doi:10.5194/gmd-5-471-2012, 2012.

Hurtt, G. C. and Armstrong, R. A.: A pelagic ecosystem model calibrated with BATS data, Deep Sea Research Part II: Topical Studies in Oceanography, 43, 553–583, doi:10.1016/0967-0645(96)00007-0, 1996.

Hurtt, G. C. and Armstrong, R. A.: A pelagic ecosystem model calibrated with BATS and OWSI data, Deep Sea Research Part I: Oceanographic Research, 46, 27–61, doi:10.1016/S0967-0637(98)00055-7, 1999.

Kane, A., Moulin, C., Thiria, S., Bopp, L., Berranda, M., Tagliabue, A., Crépon, M., Aumont, O., and Badran, F.: Improving the parameters of a global ocean biogeochemical model via variational assimilation of in situ data at five time series stations, Journal of Geophysical Research – Oceans, 116, 1–14, doi:10.1029/2009JC006005, 2011.

Kaufman, D. E., Friedrichs, M. A. M., Hemmings, J. C. P., and Jr., W. O. S.: Assimilating bio-optical glider data during a phytoplankton bloom in the southern Ross Sea, Biogeosciences Discussions, doi:10.5194/bg-2017-258, in review, 2017.

Keller, D. P., Feng, E. Y., and Oschlies, A.: Potential climate engineering effectiveness and side effects during a high carbon dioxide-emission scenario, Nature Communications, 5, doi:10.1038/ncomms4304, 2014.

Kennedy, M. C., Anderson, C. W., Conti, S., and O'Hagan, A.: Case studies in Gaussian process modelling of computer codes, Reliability Engineering and System Safety, 91, 1301–1309, doi:10.1016/j.ress.2005.11.028, 2006.

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

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

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

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

Lawson, L. M., Spitz, Y. H., Hofmann, E. E., and Long, R. B.: A data assimilation technique to a predator-prey model, Bulletin of Mathematical Biology, 57, 593–617, doi:10.1016/S0092-8240(05)80759-1, 1995.

Lawson, L. M., Hofmann, E. E., and Spitz, Y. H.: Time series sampling and data assimilation in a simple marine ecosystem model, Deep Sea Research Part II: Topical Studies in Oceanography, 43, 625–651, doi:10.1016/0967-0645(95)00096-8, 1996.

5 Löptien, U.: Steady states and sensitivities of commonly used pelagic ecosystem model components, Ecological Modelling, 222, 1376–1386, doi:10.1016/j.ecolmodel.2011.02.005, 2011.

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

Matear, R. J.: Parameter optimization and analysis of ecosystem models using simulated annealing: A case study at station P, Journal of
10 Marine Research, 53, 571–607, doi:10.1357/0022240953213098, 1995.

Mattern, J. P., Fennel, K., and Dowd, M.: Estimating time-dependent parameters for a biological ocean model using an emulator approach, Journal of Marine Systems, 96–97, 32–47, doi:10.1016/j.jmarsys.2012.01.015, 2012.

Mengis, N., Keller, D. P., Eby, M., and Oschlies, A.: Uncertainty in the response of transpiration to CO2 and implications for climate change, Environmental Research Letters, 10, doi:10.1088/1748-9326/10/9/094001, 2015.

15 Oschlies, A. and Garcon, V.: An eddy permitting coupled physical-biological model of the North Atlantic. Part I: Sensitivity to advection numerics and mixed layer physics., Global Biogeochemical Cycles, 13, 135–160, doi:10.1029/98GB02811, 1999.

Piwonski, J. and Slawig, T.: Metos3D: the Marine Ecosystem Toolkit for Optimization and Simulation in 3-D – Part 1: Simulation Package v0.3.2, Geoscientific Model Development, 9, 3729–3750, doi:10.5194/gmd-9-3729-2016, 2016.

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

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

Rückelt, J., Sauerland, V., Slawig, T., Srivastav, A., Ward, B., and Patvardhan, C.: Parameter Optimization and Uncertainty Analysis in a Model of Oceanic CO_2 Uptake Using a Hybrid Algorithm and Algorithmic Differentiation, Nonlinear Analysis B: Real World Applica-
25 tions, 11, 3993–4009, doi:10.1016/j.nonrwa.2010.03.006, 2010.

Sauerland, V.: Non-parametric optimization methods for model assessment (NOMMA-1.0), doi:10.5281/zenodo.1162769, 2017.

Schartau, M. and Oschlies, A.: Simultaneous data-based optimization of a 1D-ecosystem model at three locations in the North Atlantic: Part I – Method and parameter estimates, Journal of Marine Research, 61, 765–793, doi:10.1357/002224003322981147, 2003.

Schartau, M., Oschlies, A., and Willebrand, J.: Parameter estimates of a zero-dimensional ecosystem model applying the adjoint method,
30 Deep Sea Research Part II: Topical Studies in Oceanography, 48, 1796–1800, doi:10.1016/S0967-0645(00)00161-2, 2001.

Schartau, M., Wallhead, P., Hemmings, J., Löptien, U., Kriest, I., Krishna, S., Ward, B. A., Slawig, T., and Oschlies, A.: Reviews and syntheses: Parameter identification in marine planktonic ecosystem modelling, Biogeosciences, 14, 1647–1701, doi:10.5194/bg-14-1647-2017, 2017.

Spitz, Y., Moisan, J. R., Abbott, M. R., and Richman, J. G.: Data assimilation and pelagic ecosystem model: parameterizations using time
35 series observations, Journal of Marine Systems, 16, 51–68, doi:10.1016/S0924-7963(97)00099-7, 1998.

Steinberg, D. K., Carlson, C. A., Bates, N. R., Johnson, R. J., Michaels, A. F., and Knap, A. H.: Overview of the US JGOFS Bermuda Atlantic Time-series Study (BATS): a decade-scale look at ocean biology and biogeochemistry, Deep Sea Research Part II: Topical Studies in Oceanography, 48, 1405–1447, doi:10.1016/S0967-0645(00)00148-X, 2001.

Stow, C. A., Jolliff, J., McGillicuddy, D. J., Doney, S. C., Allen, J. I., Friedrichs, M. A. M., Rose, K. A., and Wallhead, P.: Skill assessment for coupled biological/physical models of marine systems, Journal of Marine Systems, 76, 4–15, doi:10.1016/j.jmarsys.2008.03.011, 2009.

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

Vallino, J. J.: Improving marine ecosystem models: Use of data assimilation and mesocosm experiments, Journal of Marine Research, 58, 117–164, doi:10.1357/002224000321511223, 2000.

Volpe, G., Santoleri, R., Vellucci, V., d'Alcalá, M. R., Marullo, S., and D'Ortenzio, F.: The colour of the Mediterranean Sea: Global versus regional bio-optical algorithms evaluation and implication for satellite chlorophyll estimates, Remote Sensing of Environment, 107, 625–638, doi:10.1016/j.rse.2006.10.017, 2007.

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

Yeganova, L. and Wilbur, W. J.: Isotonic regression under Lipschitz constraint., Journal of Optimization Theory and Applications, 141, 429–443, doi:10.1007/s10957-008-9477-0, 2009.